

# An Acoustic-Phonetic and Articulatory Study of Speech-Speaker Dichotomy

Parham Mokhtari



A thesis submitted for the degree of  
Doctor of Philosophy  
of the University of New South Wales

School of Computer Science  
University College, University of New South Wales  
Australian Defence Force Academy  
Canberra, Australia

June 1998



## Statement

I hereby declare that this submission is my own work and to the best of my knowledge it contains no material previously published or written by another person, nor material which to a substantial extent has been accepted for the award of any other degree or diploma at UNSW or any other educational institution, except where due acknowledgement is made in the thesis. Any contribution made to the research by colleagues, with whom I have worked at UNSW or elsewhere, during my candidature, is fully acknowledged.

I also declare that the intellectual content of this thesis is the product of my own work, except to the extent that assistance from others in the project's design and conception or in style, presentation and linguistic expression is acknowledged.

---

(Parham Mokhtari)

(date)



*“It seemed to me a superlative thing — to know the explanation of everything, why it comes to be, why it perishes, why it is.*

So spoke Socrates nearly 2500 years ago...”

Robyn Williams,  
in Foreword of “More Big Questions,  
Paul Davies in conversation with Phillip Adams” (Piper Films, 1998)



Dedicated

to my mother,

*Pari*

to my father,

*Parviz*

and to my sister,

*Tara*



## Acknowledgements

During the course of my research endeavours which have ultimately led to the completion of this doctoral dissertation, I have had the extremely good fortune of being associated with a number of people whose support I herein wish to acknowledge.

I extend my gratitude first and foremost to Dr Frantz Clermont, whom I have been proud to consider my research supervisor, professional colleague, and personal friend. Ever since our intellectually explosive Australian summer of 1993/94 when the seeds of the present research work were sown, he has consummately transcended his supervisory role, and provided intense and almost brotherly guidance at every stage of the research process. Our uncountable and immeasurably inspiring, free-spirited discussions, have enriched my experience as a doctoral candidate, and fuelled my own passion for the Science in human speech. I am boundlessly grateful to Frantz for having imparted his scholarly outlook, his visionary thinking, and his exacting attitude in scientific research. May he, my scientific mentor, find frequent satisfaction in this permanent record of our often challenging but always exciting and inspired journey.

I amicably thank Dr Michael (Spike) Barlow, for his selfless friendship, both on and off the tennis court; and for his numerous and insightful comments and suggestions both informally and at the Speech Group meetings held at the School. My sincere thanks also extend to Dr Santha Sampath, for her moral support and encouragement during the entire duration of my candidature, and for her involvement and suggestions at the Speech Group meetings at which my research work was the topic of discussion.

I gratefully thank the following people, both past and present at the School: Dr Michael Wagner for having taken me on board as a doctoral candidate early in 1993, for his initial suggestions of pursuing research in the area of speaker characterisation, and for his many hours of involvement and guidance during approximately the first year of my candidature; Dr Hiroaki Oasa for his early involvement, interest, and suggestions; Cmdr. Clive Cooper for his amiable and inspiring, but all too brief presence, especially at those three-way meetings with Frantz at the Gorman House café during 1994; Mr

Andy Quaine for his moral support and guidance during a particularly traumatic period in 1995; Dr Xin Yao for sharing some of his expertise in numerical optimisation methods; Mrs Willma Nelowkin, Mr Phillip Clark, Mr Jeff Howard, Mr Wen Ung, Mr Joe Holloway, Mr Aaron Mihe, Mr Crispin Russell and Mr Colin Stevens for their technical support in computer-related matters; Mrs Eileen Trot, Mrs Pam Gianakakis and Ms Alison McMaster for their invaluable, secretarial assistance; and Dr Charles Newton, Head of the School, for his invariable support.

I thank Ms Peta Kennedy and Ms Diane Cook of the Postgraduate Office of the Academy, for their kind assistance in matters relating to University administration; and the University College itself, for granting me the financial support of a Postgraduate Research Scholarship.

My special thanks are extended to Ms Marija Tabain and Ms Catherine Watson, whose kind invitation and encouragement to present a seminar at the Speech, Hearing and Language Research Centre at Macquarie University in Sydney (in July 1997) did significantly help to crystallise my thoughts on some of the work relevant to Chapter 5.

I warmly thank my dear friends Mr Bruce Taloni and Mr Brendan Cooper; the latter for his candid friendship and encouragement even at a distance, and the former whose frequent enquiries as to my state of health and that of the “baby” thesis did help me to retain an overall sense of balance. In that regard, I also extend my heartfelt thanks to Mr Ehsan Keyhani, Mrs Mehrassa Keyhani, Dr Hashem Etminan, Mrs Mina Etminan, Dr Fereidoon Ghasemi, Mrs Mina Ghasemi, and to Mr Shahrokh Vaziri, whose invariable encouragements, warm conviviality, and refined culture are deeply impressed in my mind.

Finally, I wish to thank my dear family, to whom this work is dedicated. Their unflinching and unwavering love and multi-faceted support through times of crisis and joy alike, have been the mainstay of my passage not only through this period of doctoral research, but more inclusively, throughout my years. The strength I carry with me beyond these formative years, is firmly rooted in their singular belief in my ability to succeed. I hope that in this volume they will find at least a small part of the fruits of their labours.

## Abstract

The acoustic speech signal naturally contains information pertaining to both the linguistic message and the identity of the speaker. Separation of these two, primary sources of variability has been of considerable concern to speech scientists for many years, yet we still lack a model of the speech communication process that can account for a wide variety of speakers even for a restricted set of vowels of a given dialect of spoken English.

Basic research was therefore undertaken to gain a better understanding of the separate influences and of the interactions between phonetic and speaker-specific attributes of vowel sounds. First, a new methodology for performing computer vowel recognition was developed, which has the ability to reveal the contrastive influence of vowel-speaker interactions across different regions of the frequency spectrum. The results obtained from applying this methodology have: (i) yielded significant insights into the long-standing problem of phonetic-speaker dichotomy; and (ii) prompted a search for an even more fundamental explanation in terms of the physical properties of the speech production mechanism.

To this end, a new functional representation of vocal-tract shapes was derived, which depends directly on resonance parameters while retaining the uniqueness properties of the Linear-Prediction model of speech production. This hybrid modelling approach was used together with a new, articulatory method of speaker normalisation, to quantify speaker differences in vocal-tract shapes, and thus to define physical correlates of the phonetic-speaker interactions which were earlier shown to adversely affect vowel recognition accuracy in certain frequency bands.

In sum, this research work embraces two major domains of computer speech science and technology — namely, speech acoustics and speech production. Beyond advancing knowledge in speaker characterisation, it ultimately has implications in the still unresolved problem of speech or speaker recognition by computer.



# Contents

<b>Acknowledgements</b> . . . . .	vii
<b>Abstract</b> . . . . .	ix
<b>Contents</b> . . . . .	xi
<b>1 General Introduction</b> . . . . .	1
1.1 Research Context: The Speech-Speaker Dichotomy . . . . .	1
1.1.1 On the Sources of <i>Speech</i> Variability . . . . .	2
1.1.2 On the Sources of <i>Speaker</i> Variability . . . . .	3
1.2 Research Scope . . . . .	5
1.3 Thesis Outline . . . . .	6
<b>2 Speech-Speaker Dichotomy: An Interpretive Review of the Literature</b> . . . . .	9
2.1 Introduction . . . . .	9
2.1.1 An Historical Account of the Dichotomy . . . . .	9
2.1.2 Structural Profile of Literature . . . . .	12
2.2 Acoustic-Auditory Evidence and Influential Role . . . . .	13
2.2.1 Human Auditory-Perceptual Evidence of Spectral Dichotomy . . . . .	13
2.2.2 Overcoming the Dichotomy by Auditorily-motivated Spectral Emphasis . . . . .	16
2.2.2.1 Frequency-scale Transformations . . . . .	16
2.2.2.2 Importance of Relative Spectral Measures . . . . .	19
2.2.3 Emerging Perspective . . . . .	21
2.3 Acoustic-Phonetic Evidence and Approaches . . . . .	24
2.3.1 Spectral Manifestations and Descriptive Parameters . . . . .	25
2.3.1.1 Phonetic Quality of Spoken Vowels . . . . .	25
2.3.1.2 Speaker Individuality. . . . .	27
2.3.2 Consequences in Vowel/Speaker Recognition by Machine . . . . .	30
2.3.2.1 Speaker Recognition . . . . .	30
2.3.2.2 Vowel Recognition . . . . .	34

2.3.3 Approaches to Overcoming the Dichotomy . . . . .	38
2.3.3.1 Towards Robustness in Parameterisation. . . . .	39
2.3.3.2 Speaker Normalisation . . . . .	42
2.4 Acoustic-Articulatory Evidence and Approaches . . . . .	46
2.4.1 The Speaker Factor: Sources of Variability and Consequences . . . . .	47
2.4.1.1 Theoretical Framework . . . . .	47
2.4.1.2 Most Common Empirical Observations . . . . .	48
2.4.2 Persistent Approaches to Speaker Normalisation . . . . .	51
2.4.3 Speaker Normalisation Beyond Vocal-Tract Length . . . . .	55
2.4.3.1 Under-exploited Evidence . . . . .	55
2.4.3.2 Articulatory Parameterisation and the Inverse Problem . . . . .	62
2.5 Concluding Discussion . . . . .	69
<b>3 Speech Materials and Acoustic Parameterisation . . . . .</b>	<b>73</b>
3.1 Introduction . . . . .	73
3.2 Speech Materials and Speaker Sets . . . . .	73
3.3 Acoustic Data used to Unfold the Dichotomy . . . . .	81
3.3.1 Vocalic Steady-State Detection . . . . .	81
3.3.2 Formant Estimation . . . . .	86
3.3.2.1 Goodness Criteria. . . . .	87
3.3.2.2 Formant-tracking Method and Evaluation Aids . . . . .	87
3.3.2.3 IFV-related Problems and Corrections . . . . .	89
3.3.2.4 LP-related Problems and Corrections . . . . .	92
3.3.3 Cepstral Analysis . . . . .	97
3.4 Acoustic Data used to Validate the Dichotomy . . . . .	98
3.4.1 Formant Estimation . . . . .	99
3.4.1.1 Amplitude-Section Approach . . . . .	99
3.4.1.2 Spectrogram Approach . . . . .	100
3.4.2 Simplified Cepstrum . . . . .	101
3.5 Summarising Perspective . . . . .	104
<b>4 Acoustic-Phonetic Analysis of Speech-Speaker Dichotomy . . . . .</b>	<b>107</b>
4.1 Introduction . . . . .	107
4.2 Methodology for Unfolding the Dichotomy. . . . .	109
4.2.1 Approach . . . . .	110

4.2.2 Spectral Representations . . . . .	112
4.2.3 Classification Methods and Distance Formulations . . . . .	114
4.2.3.1 Classifier based on Hyper-Planar Decision Boundaries . . . . .	116
4.2.3.2 Classifier based on Hyper-Quadratic Decision Boundaries . . . . .	121
4.3 The Dichotomy Unfolded. . . . .	135
4.3.1 Behaviour of Classification Accuracy Using Linear Classifier . . . . .	135
4.3.2 Behaviour of Classification Accuracy Using Quadratic Classifier . . . . .	137
4.4 The Dichotomy Explained . . . . .	146
4.4.1 Spectral Regions of Vowel-Speaker Dichotomy . . . . .	146
4.4.2 Acoustic-Phonetic Explanation of the Dichotomy . . . . .	151
4.4.3 Dependence on Spectral Representation . . . . .	155
4.4.4 Dependence on Speaker Homogeneity. . . . .	162
4.4.5 Dependence on Idiolectal Differences . . . . .	174
4.5 Concluding Summary . . . . .	183
<b>5 Vocal-Tract Shape Parameterisation and Estimation . . . . .</b>	<b>187</b>
5.1 Introduction . . . . .	187
5.2 Rationale . . . . .	189
5.2.1 Resonance-Based Parameterisation of Area-Functions . . . . .	189
5.2.2 Inherent Uniqueness in LP-Derived Area-Functions . . . . .	195
5.3 Parameters of Unique LP-derived Area-Functions . . . . .	201
5.3.1 Dependence of LP-derived VT-Shapes on Formant Frequencies . . . . .	203
5.3.1.1 Partial Theoretical Proof . . . . .	203
5.3.1.2 Empirical Validation . . . . .	207
5.3.2 Dependence of LP-derived VT-Shapes on Formant Bandwidths . . . . .	212
5.3.2.1 Theoretical Motivation . . . . .	212
5.3.2.2 Empirical Justifications . . . . .	214
5.4 Method of Area-Function Parameterisation and Estimation . . . . .	221
5.4.1 Description . . . . .	221
5.4.2 Evaluation of Proposed Area-Function Parameterisation . . . . .	227
5.4.2.1 Inter-parameter correlations. . . . .	227
5.4.2.2 Representation of Directly Measured Area-Functions . . . . .	229
5.5 Evaluation of Estimation Method . . . . .	234
5.5.1 Inter-repetition Alignment of Area-Functions . . . . .	236

5.5.2 Re-estimation of Directly Measured Area-Functions . . . . .	239
5.5.2.1 Model-matched Conditions . . . . .	239
5.5.2.2 More Realistic Conditions . . . . .	246
5.5.2.3 Model-based Formant Correction . . . . .	249
5.5.3 Estimation of Area-Functions from Measured Formants . . . . .	253
5.5.3.1 Formant-induced Variability of LP Area-Functions . . . . .	254
5.5.3.2 Closed-glottis Correction of Formant Bandwidths . . . . .	260
5.6 Concluding Summary . . . . .	264
<b>6 An Articulatory Explanation of Speech-Speaker Dichotomy . . . . .</b>	<b>269</b>
6.1 Introduction . . . . .	269
6.2 Articulatory Approach. . . . .	270
6.3 Methodology for Articulatory Speaker Normalisation . . . . .	273
6.3.1 Speaker Normalisation of VT Fixed Structure . . . . .	273
6.3.2 Speaker Normalisation of Articulatory Setting . . . . .	276
6.3.3 Speaker Normalisation of Vowel-Specific Articulatory Strategy . . . . .	281
6.4 Articulatory Explanation of the Dichotomy . . . . .	283
6.4.1 Contributions of VT Structure to the Dichotomy . . . . .	284
6.4.2 Contributions of Articulatory Setting to the Dichotomy . . . . .	293
6.4.3 Contributions of Articulatory Strategy to the Dichotomy. . . . .	303
6.5 The Dichotomy Undone . . . . .	321
6.5.1 Articulatory Consequences . . . . .	323
6.5.2 Acoustic-Phonetic Consequences . . . . .	328
6.6 Concluding Summary . . . . .	333
<b>7 Concluding Discussion . . . . .</b>	<b>337</b>
7.1 Acoustic-Phonetic Perspective . . . . .	337
7.1.1 Experimental Methods . . . . .	337
7.1.1.1 <i>A Posteriori</i> Selection of Frequency Sub-bands. . . . .	337
7.1.1.2 Recognition Framework for Diagnosing the Dichotomy . . . . .	338
7.1.2 Spectral Manifestations of the Dichotomy and Implications . . . . .	340
7.2 Acoustic-Articulatory Perspective . . . . .	343
7.2.1 Experimental Methods . . . . .	343
7.2.1.1 Vocal-tract Shape Parameterisation . . . . .	343
7.2.1.2 Articulatory Speaker Normalisation . . . . .	346

7.2.2 Articulatory Explanation of the Dichotomy . . . . .	347
<b>A Formant-tracking Analysis Conditions and Sequence Charts . . . . .</b>	<b>353</b>
<b>B Mathematical Derivation of the SM Model . . . . .</b>	<b>363</b>
<b>C A Lossy Vocal-Tract Acoustic Model . . . . .</b>	<b>369</b>
<b>D The LP Method of Inversion “Stepped Down” . . . . .</b>	<b>375</b>
<b>E Representation of Directly Measured Area-Functions . . . . .</b>	<b>383</b>
<b>F Re-estimation of Directly Measured Area-Functions . . . . .</b>	<b>391</b>
<b>G Closed-glottis Correction of Formant Bandwidths . . . . .</b>	<b>399</b>
<b>Bibliography . . . . .</b>	<b>403</b>



# Chapter 1

## General Introduction

### 1.1 Research Context: The Speech-Speaker Dichotomy

If there is a single objective which epitomizes over half a century of international research efforts concerned with the processes of human speech communication, it is to find order and regularity in the *variability* which exists therein. Owing to their intrinsic dimensionality and complexity, the sources of variability which pervade the three domains of speech production, acoustics, and perception, still pose a challenge to speech scientists and technologists alike. Indeed, whether it is to contribute toward the ultimate goal of a unified theory of speech communication, or toward the application of such knowledge in spoken language systems or in aids for the handicapped, variability is still a significant and an eminent research problem.

Human speech communication naturally implies auditory perception of an acoustic speech signal, which itself is produced by the vocal apparatus of a speaker. The acoustic speech signal therefore carries not only the linguistic message which the speaker wishes to convey, but also a host of para-linguistic information about the speaker's identity, gender, emotions, state of health, and other characteristics which may be perceived by the listener. The kernel of the variability problem is therefore made up of two main components: the *speech* or the linguistic message itself, and the idiosyncrasies of the *speaker* who produces it. Furthermore, the influences of these two components are so intricately intertwined in the acoustic speech signal, that a complete definition of their separate manifestations continues to be elusive. The inherent difficulties of this long-standing problem are better appreciated, and an investigative approach more easily formulated, by considering the potential sources of variability which underlie the two components of what we henceforth will refer to as the *speech-speaker dichotomy* or,

interchangeably, as the *dichotomy* in short.

### **1.1.1 On the Sources of *Speech* Variability**

Sources of variability which are normally attributed to the speech component of the dichotomy are broadly categorised in terms of either *prosodic* or *segmental* properties of the acoustic speech stream. The former include signal energy (or stress), syllable duration (or rhythm), and fundamental frequency (or intonation) which characterise utterances across syllables, words, and sentences; they embody generally poly-segmental aspects of speech. By contrast, segmental properties are known to more directly embody the phonetic or the linguistic aspects of the spoken message, and are described in terms of the individual phonemes, which themselves are generally believed to be the building blocks (or phonetic units) of spoken syllables, words, and phrases.

The segmental units of the speech stream are, in turn, broadly categorised as either *consonants* or *vowels*, depending on the degree to which the airstream through the vocal-tract is obstructed during the production of those speech sounds. Consonants, which include stops, fricatives, and approximants, involve a relatively greater degree of obstruction than vocalic sounds; they can therefore give rise to non-periodic, noise-like, or sudden bursts of acoustic energy, caused either by air turbulence at a very narrow constriction, or by a release following complete closure in the vocal-tract. By contrast, spoken vowels are produced with relatively more open vocal-tract configurations and, more importantly, their articulatory, acoustic, and perceptual properties are relatively better apprehended. In particular, the acoustic properties of spoken vowels are firmly rooted in the spectral domain, where they have been studied mainly in terms of the compact, articulatorily- and perceptually-relevant description afforded by the first few formants (or acoustic resonances of the vocal-tract).

The importance of vowels, and our relatively superior understanding of their properties, can be explained partly by their high rate of occurrence in spoken language (Denes (1963) has estimated that the vowel phonemes occur over 30% of the time in spoken English). In addition, they exhibit relatively stable spectral characteristics and consistently high levels of acoustic energy, and are therefore regarded as “islands of reliability” (Lea et al., 1975) in the acoustic speech stream, particularly in context of

stressed syllables. Even so, the segmental properties of spoken vowels are known to be influenced by potential sources of variability which include speaking rate, linguistic stress, and coarticulation with preceding and following phonemes. Traditionally, these types of variability are controlled by embedding the vowel (generically represented by the symbol V) in the quasi-neutral phonetic context afforded by the formation of /hVd/ words; consistency in stress and speaking rate is then achieved when those monosyllabic words are read aloud (in citation form) from lists, in which the order of the words are randomised to reduce inter-repetition bias.

### **1.1.2 On the Sources of *Speaker* Variability**

It is often taken for granted that any of the types of variability discussed above could potentially arise from, and could therefore be used to characterise, either *inter-* or *intra-speaker* variations. Whilst the former refers to the speaker differences which are implied in the speech-speaker dichotomy, the latter refers to the repeatability of speech sounds by a single speaker on different occasions; variability of this kind might be induced in the short-term by varying emotions, in the medium-term by different states of health, and in the longer-term by the consequences of ageing. In this regard, Nolan (1983, p.59) aptly puts forward the view that each speaker's vocal apparatus determines "merely the range within which variation in a particular parameter is constrained to take place", with the assertion that there is "no acoustic feature which escapes the plasticity of the vocal tract."

Leaving aside the possibilities of intended disguise or mimicry, or indeed of extrinsically-induced variations in speech production beyond what might be regarded as a speaker's *normal* manner of speaking, it is generally accepted that a given speaker (particularly, a phonetically naive speaker) will not be able to exactly reproduce a given set of speech sounds, even in an emotionally neutral, contemporaneous setting, and even after accounting for measurement variability. In this regard, however, Broad (1972) has noted that the magnitude of the resulting, so-called inter-repetition variation (of measured formant frequencies) is of the same order as the human auditory-perceptual tolerance to perturbations in those acoustic parameters of vocalic speech sounds. Indeed, it is generally assumed that under controlled conditions (i.e., in the

absence of the types of systematic variability discussed above, or attempts at disguise or mimicry), the variability of acoustic parameters measured in any given vowel of a single speaker is considerably less than the variability measured in the same vowel of two or more speakers; consequently, the phonetic equivalence (Broad, 1976) of the measured parameter values is more easily defined in the single-speaker case.

Perhaps the most well-known consequence of the greater magnitude of inter- than intra-speaker variability, is the often dramatic deterioration in the performance of an automatic speech recogniser, when used, without proper training, by more than one speaker. The extent of that deterioration — a presumed consequence of the dichotomy problem — depends partly on the nature and the degree of the differences between the speakers involved. In order to cope with this problem, state-of-the-art automatic speech recognition (ASR) systems often use statistically-based algorithms for speaker adaptation of either the training data or the model parameters (e.g. Young, 1996). However, at best (as we shall see in Chapter 2), such algorithms account only incompletely for the basic sources of inter-speaker variability; their effectiveness is therefore ultimately curtailed by under-exploitation of our limited but expanding knowledge of the fundamental components of speaker variability.

In this vein, it may be of some advantage to note the obvious fact that the acoustic speech signal originates from the vocal apparatus of the speaker. Differences between the acoustic realisations of phonetically-equivalent speech sounds uttered by different speakers can therefore be related to the size, shape, and idiosyncratic usage of their articulatory organs. The two commonly accepted, broad categories which describe inter-speaker differences at an articulatory level, are known as *organic* and *learned*, respectively (Garvin and Ladefoged, 1963). Whilst the former refers to anatomical differences in the size and shape of the speakers' articulatory organs such as the tongue, teeth, laryngeal and other vocal-tract fixed structures, the latter refers to the behavioural differences in their individual usage of the moveable articulators during speech production.

The most substantial types of organic differences are undoubtedly those which are found between the vocal-tract anatomical structures of men, women, and children. Analogously, the most substantial types of learned differences are those which

distinguish between the dialects or idiolects of a given language. Whilst both of these types of differences serve to differentiate between *groups* of speakers, within each group (for example, adult male speakers of a particular dialect of English) there exist *individual* or *intrinsic*, inter-speaker differences which, although generally smaller in magnitude than group differences, can still give rise to substantial acoustic variability. Indeed, as Nordström (1977, p.81) has pointed out in his interpretation of a wide range of vowel formant data, “one does not need to go beyond comparisons within the same speaker category (male speakers of the same dialect) to find immense variability.”

An important consideration in this regard is the degree of *homogeneity* (or conversely, *heterogeneity*) of a given speaker group. Speakers who possess similar, vocal-tract anatomical and behavioural characteristics, might be expected to exhibit a fair degree of homogeneity in their measured acoustic data, and are therefore less likely to be problematic from an ASR point of view. By contrast, a more heterogeneous group of speakers might be expected to exhibit larger amounts of acoustic, inter-speaker variability, which may arise from a combination of organic and learned, articulatory differences; interestingly, whilst the acoustic manifestations of these differences might detrimentally affect the performance of an ASR system, they are generally of distinct advantage in the converse task of automatic speaker recognition.

## **1.2 Research Scope**

It is well-known, and our introductory exposition of the problem has made quite clear, that the speech-speaker dichotomy is the kernel of a vast and intricate network of variability in the processes of human speech communication. Whilst the two primary components of variability manifest in the acoustic speech signal can be attributed, respectively, to the linguistic message (the speech) and the communicator of that message (the speaker), each component implies a multitude of potential sources of variability, whose intertwined influences have rendered a complete decoupling of *speech* and *speaker* formidable. Clearly, one cannot hope to undertake in a single study, an investigation of the speech-speaker problem in its entirety. We must therefore narrow the scope of our research, by imposing limits on the potential “determinants of acoustic parameter values” (Broad, 1982).

As noted earlier, our understanding of the vowel sounds of English, accumulated over decades of research, perhaps exceeds that of any other class of speech sounds. Indeed, our greater knowledge of the properties of spoken vowels is reflected not only in the acoustic domain where their spectral properties are best known in terms of the formant parameters, but also in the articulatory domain, where a number of models have been proposed which account for the physiologically and acoustically relevant properties of vowel articulation. The research scope of this thesis will therefore be limited to the segmental properties of the spoken (non-nasalised) vowels of English. Our emphasis on the importance of relating the acoustic manifestations of the speech-speaker dichotomy to its physical causes, will become apparent in later chapters where we shall adopt a vocal-tract modelling approach to provide such physical explanations for the vowel sounds considered.

As for the speaker component of the dichotomy, we shall restrict our attention to differences amongst adult, male speakers with no known speech impairments. The potential sources of inter-speaker variability will include intrinsic and idiolectal differences, which may be related to a combination of vocal-tract anatomical and behavioural idiosyncrasies. Furthermore, by eliciting natural productions of vowel sounds in a contemporaneous, laboratory environment, we forsake the wide range of intra-speaker variability which has been attributed to emotions, states of health, ageing, and either intended disguise or socially-induced variations. As a result, the full range of intra-speaker variability in vowel production is kept to a minimum, and a relatively invariant *norm* is established for each vowel of each speaker, from which may proceed a systematic investigation of purely *phonetic* and *inter-speaker* influences.

### **1.3 Thesis Outline**

As outlined below, the results of our acoustic-phonetic and articulatory investigations of the speech-speaker dichotomy are presented in two core chapters (4 and 6). Those chapters are preceded by an interpretive review of the relevant literature (Chapter 2) and a complete description of our speech data (Chapter 3); they are connected by our detailed treatment of the problem of vocal-tract shape parameterisation and estimation (Chapter 5); and they are followed by a concluding discussion (Chapter 7).

In Chapter 2 we address the question of how we are justified in pursuing a problem which has long been central to speech science and technology. Past attempts to find regularity in the intertwined influences of speech and speaker variability, whether qualitatively suggested or quantitatively shown, are recalled and placed in context. Previous methods which have been proposed to overcome the influences of speech-speaker interactions are then reviewed, together with an overall perspective on the literature concerning the speech-speaker dichotomy. In short, we will argue in Chapter 2 that despite our relatively more comprehensive knowledge of both the acoustic and articulatory properties of spoken vowels, the dichotomy problem manifested in those speech sounds still warrants further research.

In Chapter 3 we present and justify the range of speech materials and methods of acoustic parameterisation used in our investigation of the speech-speaker dichotomy. In particular, we describe three different datasets of vowels, recorded in the time-honoured /hVd/ context by adult, male speakers of Australian and American English. Differences in the degree of phonetic and speaker complexity of those datasets, afford the opportunity of examining the dependence of the dichotomy both on the degree of speaker homogeneity, and on the presence or absence of idiolectal speaker differences. We also recount the methods of acoustic parameterisation used to carefully extract the formants and the linear prediction (LP) cepstra at the steady-state of each vocalic nucleus.

Those sets of parameters are then used in Chapter 4 to investigate the vowel-speaker dichotomy in the acoustic-phonetic domain. In particular, the spectral manifestations of vowel-speaker interactions are unfolded by way of vowel recognition experiments, designed specifically to lay bare the influences of phonetic and speaker variabilities along the spectral continuum. An indispensable tool in our methodology for unfolding the dichotomy, is a new cepstral distance measure (Clermont and Mokhtari, 1994) which facilitates the selection of frequency-bands in machine classification of vowels. The interpretively superior formant parameters then provide an acoustic-phonetic explanation of the observed spectral manifestations of the dichotomy. We conclude our acoustic-phonetic investigations by addressing the questions raised earlier, which concern speaker homogeneity and the influence of idiolectal speaker differences.

In search of a more complete elucidation of the problem of speech-speaker dichotomy, we then extend our investigations into the domain of speech production. However, a methodological framework for investigating the physical correlates of the speech-speaker dichotomy would be impractical if it were to rely on direct articulatory measurements, which are in general difficult to acquire. In Chapter 5 we therefore reconsider the well-known *inverse problem* of mapping the geometry of the human vocal-tract from the acoustic speech signal. In particular, we derive a new functional representation of vocal-tract shapes, which depends directly on resonance parameters (formants) while retaining the uniqueness properties of the well-known linear prediction (LP) method of inversion.

The vocal-tract shapes thus obtained from the formant data, which themselves are measured in Chapter 3 and used in Chapter 4 to unfold and explain the acoustic-phonetic manifestations of the dichotomy, then afford a physical explanation of the dichotomy in Chapter 6. Towards that end, we describe a new method of articulatory speaker normalisation, which facilitates a tripartite decomposition of the articulatory sources of inter-speaker variability. Our investigative loop is then closed via resynthesis of acoustic parameters from the speaker-normalised vocal-tract shapes, thus allowing a direct interpretation of the articulatory sources of speaker variability in terms of our acoustic-phonetic methodology of cepstrum-based vowel recognition and formant-based explanation of the dichotomy.

In sum, Chapters 4, 5, and 6 collectively provide more complete insights into, and thus advance our understanding of the speech-speaker problem in the two domains of speech acoustics and speech production. In Chapter 7 we then draw conclusions regarding the various contributions which arise from the research work described in this thesis.

## Chapter 2

### Speech-Speaker Dichotomy: An Interpretive Review of the Literature

#### 2.1 Introduction

The speech-speaker dichotomy is not a new problem. Its long history is first given due consideration in Section 2.1.1, where we will recall recurrent references to the opposing and complementary roles attributed to the lower and the higher spectral regions of spoken vowels in particular. This historical profile will also highlight that previous contributions towards elucidating the intertwined influences of phonetic and speaker-specific sources of variability in the segmental properties of spoken vowels, span the acoustic as well as the articulatory and auditory domains of speech communication. Consequently, in Section 2.1.2, we will propose to re-examine the dichotomy problem by attempting a tripartite review of the literature, from which will emerge our contention that previous works do not provide a cohesive perspective on the problem and that further elucidation is warranted.

##### 2.1.1 An Historical Account of the Dichotomy

Scattered amongst the vast assemblage of previous contributions, are a number of isolated statements of a largely qualitative nature, which nevertheless betray an underlying consensus regarding the likely acoustic correlates of the dichotomy.

One of the earliest of such quotes, dates back to a decade before the invention of the sound spectrograph in 1946. Using very laborious techniques to estimate the resonance frequencies of sung vowels, Lewis (1936, p.97) very cautiously offered the following conjecture:

“It may be that the *typical* quality of a vowel is determined mainly by the two resonators of lowest frequency, with individual voice differences resulting from the action of the

other resonators...”

Following relatively more comprehensive spectral analyses of the vowels “Ah” and “O” sung by six trained, male singers, Lewis and Tuthill (1940, p.156) were then able to conclude

“...that individual uniqueness in voice is significantly related to the operation of one or more high frequency resonators which are relatively invariable. ... The two low frequency resonators are undoubtedly crucial in the determination of vowel character, while the high frequency resonators probably contribute only to less basic tonal qualities.”

More substantial evidence to support this view, had to await the invention of the sound spectrograph (Koenig et al., 1946). Indeed, this experimental tool paved the way for less laborious, more large-scale acoustic-phonetic studies, which firmly established the importance of the formant frequencies (or resonance frequencies of the vocal tract)<sup>1</sup> in acoustic-level descriptions of spoken vowels. Spectrographic analyses were soon complemented by auditory-perceptual experiments, using the so-called “pattern playback” machine which synthesised speech sounds from hand-painted spectrograms. Equipped with considerable expertise in both spectrographic analyses and perceptual experiments, Delattre (1951, p.872) was able to assert the following:

“No doubt that formant 3 is much less responsible than formants 1 and 2 for the linguistic color of vowels. Formant 3 is mainly to be considered as one of the many higher resonances... . Being the lowest of these, it has the most perceptible effect. ... as a whole these high resonances above formant 2 have very little effect on color, they mostly add intelligibility without changing the color and are probably responsible for voice quality.”

However, perhaps the most direct interpretation of the formants remains firmly rooted in the articulatory domain. For example, transcending the inherent complexity of the relations between formants and articulatory configurations assumed during the production of voiced speech sounds, Peterson (1959, p.151) noted that

“...the lower formant frequencies depend more upon gross cavity sizes and constrictions than upon exact shapes. The adjustment of the higher formant frequencies in general depends progressively more upon specific cavity sizes and shapes.”

More specifically, in his classic work on the acoustic theory of speech production,

---

<sup>1</sup> Hereafter denoted in symbolic form as  $F_1$ ,  $F_2$ ,  $F_3$ , ... in ascending order of frequency.

Fant (1960, p.48) begins his analytical treatment of the formant-based constraints on the spectra of voiced speech sounds thus:

“The relative importance of the separate formants of voiced sounds decreases with increasing order above *F2*. *F1* and *F2* are the main determinants of vowel quality. *F3* and *F4* contribute significantly to the phonetic quality of front vowels, but in back vowels they are of minor importance only. *F3* and *F4* ... also provide certain information on personal voice characteristics.”

Further evidence of the relative importance of the formants is given by Bernard (1967, p.103) who, in the course of his painstaking, spectrographic measurements of the first three formant frequencies of vocalic speech sounds recorded by 170 speakers of Australian English, was compelled to disregard the third formant in determining the so-called onset, steady-state or target, and offglide of those monophthongal and diphthongal sounds:

“It seems clear that the first and second formants together make a much greater contribution to the specification of phonetic colour than the third formant and accordingly the third formant was not considered when deciding at which points to make measurements.”

Nor has this common view failed to persist in more recent times. For example, Saito and Nakata (1985, p.35) offer the following general observations, confirmed by formant measurements of vowels recorded by speakers of Japanese:

“Characteristics of a vowel depend mainly on formants in the low frequency range, that is, the first and second formants. ... The higher-order formant frequencies show a relatively small range of variation between vowels that depends mostly on the speaker.”

More specifically in the context of synthesis of voiced sounds using vocal-tract analogues, Linggard (1985, p.15) notes that

“...in the frequency range of interest, upto 5kHz, there are only four or five formants. In general terms, two of these are necessary to specify vowel quality, a third is required to establish speaker identity, and the fourth/fifth may be added to give natural voice quality.”

A similar point of view is offered by Ladefoged (1993, p.211):

“...the position of the fourth and higher formants in most vowels is indicative of a

speaker’s voice quality rather than the linguistic aspects of the sounds.”

All the statements quoted above are in general agreement; together, they provide a common conception — accumulated over 50 years of speech research and spanning all three domains of speech production, acoustics, and perception — of the spectral correlates of phonetic quality, as distinct from para-linguistic attributes which have been alternately referred to as “voice quality”, “tonal quality”, “personal characteristics”, “speaker identity”, or “individual uniqueness” in spoken vowels. Indeed, the common view on the acoustic correlate of the vowel-speaker dichotomy is that of a bipartite separation of vowel and speaker influences along the frequency scale — phonetic variability is widely regarded to be manifest in the low spectral regions which contain the first two formants, while speaker-specific influences are believed to be manifest mainly in the higher spectral regions which contain the third and higher formants. Moreover, these spectral correlates are of importance both perceptually and articulatorily.

### **2.1.2 Structural Profile of Literature**

As stated earlier, the literature relevant to the dichotomy problem does encompass the three major domains of auditory-perception, acoustics, and production of speech. Not unexpectedly, these three bodies of literature contribute unequally to elucidating the problem; in particular, it will transpire that the acoustic literature plays the more substantial role. Indeed, it is the acoustic manifestations of the dichotomy to which the perception and articulatory literature constantly return. Hence, our interpretive review of the *acoustic-auditory*, the *acoustic-phonetic*, and the *acoustic-articulatory* literature, in search of evidence and insights into the speech-speaker dichotomy.

Nor is this particular order of presentation purely incidental. As we shall see in Section 2.2, the perception literature not only has contributed to our understanding of the dichotomy, but it has also exerted an influential role on the spectral representations commonly used in acoustic analyses and in machine recognition of speech and speaker. By contrast, the more substantial, acoustic literature reviewed in Section 2.3, provides relatively more explicit evidence of the spectral manifestations of phonetic quality, speaker individuality, and the consequences of their interactions in recognition by

machine. The latter has provided the impetus for *speaker normalisation* as an approach to overcome the dichotomy — but has this approach yielded any fundamental insights regarding the dichotomy problem itself? In Section 2.4 we then review the articulatory literature, which provides perhaps the most sparse evidence of the speech-speaker dichotomy. Although the articulatory domain is likely to hold the greatest potential for yielding fundamental insights into the nature of the dichotomy, exploitation of that domain has admittedly been thwarted by methodological obstacles. In particular, nonuniqueness in acoustic-to-articulatory mapping, and the limitations in methods of articulatory parameterisation, have together impeded progress towards more complete descriptions of speaker differences in articulatory terms, which might then be expected to yield more informed approaches to overcome the dichotomy by speaker normalisation. Finally, in Section 2.5 we provide a concluding perspective on the various bodies of literature reviewed, which will sum up the weaknesses and relative states of incompleteness of previous works concerned with the speech-speaker dichotomy, and thus point towards our own contributions presented in the following chapters.

## **2.2 Acoustic-Auditory Evidence and Influential Role**

The literature on human auditory-perception of speech provides quantitative evidence which does substantiate the statements quoted in Section 2.1. Indeed, we will first recall evidence to support the role of the higher spectral regions in providing perceptual cues of speaker identity, and of the low spectral regions in providing the primary cues to perceived phonetic quality of spoken vowels. Equally importantly, we will also attempt to reconstruct the influential role which the perception literature has played in shaping the popular approaches to parameterisation of speech in the acoustic domain.

### **2.2.1 Human Auditory-Perceptual Evidence of Spectral Dichotomy**

The perceptual studies which have shed light on the speaker-related potency of the higher spectral regions, have invariably adopted a speaker identification approach whereby a panel of listeners reasonably familiar with the voices of the speakers recorded, are asked to identify the speaker when presented with auditory stimuli

comprising the (acoustically manipulated) recorded vowel sounds. One of the earliest of such studies is that of Peters (1954, cited in Hecker, 1971, p.40), who manipulated recorded speech sounds by passing them through single octave-band filters; the accuracy of listeners' identification of the speaker was found to be highest for the octave-band spanning the range from 1.2 to 2.4 kHz.

Similar results were reported by Compton (1963) and Dukiewicz (1970) using filtered versions of the vowel /i/ recorded, respectively, by 9 speakers of American English and 5 male speakers of Polish. For example, Compton used various low-pass and high-pass filters with cut-off frequencies of 255Hz, 510Hz, and 1020Hz; listeners' accuracy in speaker identification (found to be significantly greater than chance even for stimuli of duration 25msec) was adversely affected only under the low-pass conditions, clearly indicating perceptual preference for spectral information above 1020Hz. Along the same lines, Dukiewicz used single band-pass filters spanning the range from 128Hz to 8192Hz in single-octave, double-octave, and triple-octave pass-bands; and found listeners' accuracy in speaker identification of the stimuli (recorded with monotonous pitch, and with a duration of approximately 1 sec) to be the highest in the single-octave band from 2048Hz to 4096Hz, in the double-octave band from 2048Hz to 8192Hz, and in the triple-octave band from 1024Hz to 8192Hz. Although the results for two other vowels /a/ and /u/ reported by Dukiewicz were less consistent, significant improvements in accuracy were obtained when the single-octave range 4096Hz to 8192Hz was extended downwards to 2048Hz.

More direct and perhaps more conclusive results were reported recently by Kitamura and Akagi (1994, 1996), who were able to claim that "voice quality can be controlled" simply by replacing the higher frequency band of one speaker with that of another speaker. Their auditory stimuli consisted of 5 vowels recorded by speakers of Japanese; and then resynthesised (with normalised pitch and energy contours) using an FFT cepstrum-based analysis-resynthesis method, which consisted of joining one speaker's low frequency range with another speaker's higher frequency range for the same vowel. Listeners' accuracy in speaker identification was found to be a monotonic function of the cross-over frequency defining the boundary between the low and the higher spectral ranges. In their earlier study, Kitamura and Akagi found the listeners'

preference for one speaker over the other to occur at about 2340Hz; in their later study (in which there is no indication of any difference in methodology compared with the earlier study) they found the cross-over frequency to occur at 1740Hz.

Consistent with the implied speaker-specificity of the higher spectral regions, Kuwabara and Takagi (1991) found human perception of speaker individuality to be most sensitive to percentage changes in  $F_3$  alone. Their result was obtained using two male, Japanese speakers' recordings of the all-sonorant nonsense word /aoiue/, the spectral characteristics of which were modified using an all-pole analysis-resynthesis method whereby the frequency and bandwidth of spectral poles corresponding to individual formant peaks could be directly, and independently manipulated.

Although the studies reviewed thus far have been more directly concerned with the importance of spectral information in providing perceptual cues to speaker identity, one could almost infer from the more recent, Japanese studies in particular, that the auditory percept of vowel identity is more strongly linked with the low spectral regions. For example, Kitamura and Akagi's (1994) method of spectral swapping between speakers did no harm to their listener's identification of the phonetic identity of the vowel stimuli.

Much of the earlier literature on human auditory-perception of speech does also provide a solid foundation for the perceptual salience of the low spectral regions which encompass the first two formants, as the primary carriers of phonetic information in spoken vowel sounds. For example, perceptual experiments carried out by Delattre et al. (1952) using the famous "pattern playback" machine, showed that two formants (or spectral prominences) are sufficient for the auditory perception of 16 cardinal vowels. In particular, they found that a single formant is sufficient for the perception of back vowels, while two formants are required for categorical perception of front vowels. The perceptual role of  $F_3$  was then relegated to that of supporting the spectral prominence formed by the high  $F_2$  in front vowels.

Evidence to support the phonetic importance of the low spectral regions is also provided in the classic studies of Pols et al. (1969) and Singh and Woods (1971). In those studies, which were primarily concerned with the perceptual dimensionality of the Dutch and American English vowel spaces, respectively, the steady-state portions of

recorded vowels of a single speaker were presented (in randomised triads and pairs, respectively) to listeners who were asked to rate the degree of dissimilarity in those speech sounds. Multidimensional scaling analyses of the perceptual similarity matrices then revealed that the two principal dimensions of their listeners' auditory vowel space were highly correlated with the first two formant frequencies (Pols et al., 1969), and also with the articulatorily-motivated features "height" and "advancement" (Singh and Woods, 1971).

## **2.2.2 Overcoming the Dichotomy by Auditorily-motivated Spectral Emphasis**

### **2.2.2.1 Frequency-scale Transformations**

The auditory-perceptual findings reviewed in the previous section provide quantitative evidence (i) that at least on a per-speaker basis, the perceptual structure of spoken vowel sounds depends primarily on the spectral information in the low-frequency range which encompasses the first two formants, (ii) and that the perception of speaker individuality relies heavily on the spectral information carried in the higher-frequency ranges which encompass the third and higher formants. It therefore follows that one method of overcoming the dichotomy is to suppress the speaker-specific information whilst retaining the phonetic information, by re-scaling the frequency axis such that the higher frequency range is de-emphasised (by contraction along the axis).

In this vein, it is of interest to note that the human auditory system already appears to have such a mechanism. Indeed, it is well-known (Békésy, 1960) that acoustic-mechanical stimulation of the basilar membrane (in the coiled structure of the cochlea, situated in the inner ear) maps approximately equal length-intervals onto a non-linear frequency scale which is approximately logarithmic for frequencies above about 500Hz, thus effectively de-emphasising the higher spectral ranges by a reduction in frequency resolution. An empirically-determined approximation of the non-linear frequency scale of the auditory system was given in tabular form by Zwicker (1961), and later in analytical form by Zwicker and Terhardt (1980); indeed, according to the so-called *critical band* (or Bark) scale, the bandwidth of each critical band within which spectral integration occurs at the auditory level, increases with the centre-frequency of the band.

Perhaps the most influential role of perceptually-motivated frequency-scale transformations as exemplified in critical band processing of the acoustic speech signal, is in machine recognition of speech. For example, Dautrich et al. (1983) obtained higher accuracies in an isolated word recognition task, using band-pass filters spaced along a critical band (frequency) scale which was uniform below, and highly nonlinear above 1500Hz. Similarly, Davis and Mermelstein (1980) obtained superior performance in monosyllabic word recognition by machine, using cepstral parameters derived on the comparable *mel*-frequency scale. In particular, Davis and Mermelstein (1980, p.364) concluded that the *mel*-frequency cepstral coefficients (MFCC's) "allow better suppression of insignificant spectral variation in the higher frequency bands." In light of the evidence reviewed earlier regarding the contrastive roles of the lower and the higher spectral regions, it is not at all surprising that MFCC's should perform so well in automatic *speech* recognition (ASR), and that they have consequently been adopted as the acoustic parameters of choice in most state-of-the-art ASR systems (e.g. Young, 1996).

Nor has the well-known linear-prediction (LP) method of speech analysis (Markel and Gray, 1976) been spared from auditorily-motivated methods of spectral emphasis. For example, Strube (1980) derived a computationally efficient and flexible method of transforming LP parameters such that the all-pole spectrum is represented on a warped (or a non-linear) frequency scale. He then showed that an LP-based method of analysis-resynthesis on the critical band (or Bark) scale yielded superior intelligibility of German monosyllabic words. Consistent with the findings of Delattre et al. (1952) reviewed earlier (in Section 2.2.1), Strube found the perceptual superiority of the frequency-warped method to be particularly prominent for a 5th-order model (which can yield only up to two spectral peaks), intelligibility scores for which were comparable to a normal (linear frequency scale) LP model of twice that order.

Perhaps the most celebrated importation of auditory processing methods in the LP model, is the *perceptual linear predictive* (PLP) analysis of speech (Hermansky, 1990). Prior to autoregressive modelling, the short-term power spectrum is warped onto a Bark scale along the frequency axis, submitted to spectral reduction by critical band processing, and its amplitude is equalised in order to account for the non-equal

perceptual sensitivity of loudness at different frequencies. A relatively low-order LP analysis is then afforded by the considerable reduction in spectral resolution; thus, across a spectral range which normally would encompass the first four formants, a 5th-order PLP model yields a gross spectral shape which contains either one peak (for most back vowels) or only two peaks (for mid- to front vowels). Similarly to our interpretation of the results of Strube (1980), the phonetic relevance of the one- or two-peak spectral representation is consistent with the perceptual findings of Delattre et al. (1952) reviewed earlier.

In addition, however, it is Hermansky's interpretation of the results of speaker-dependent and speaker-independent, phoneme and word recognition accuracy obtained using PLP (and LP) models of increasing order, which is of particular concern in our search for insights regarding the speech-speaker dichotomy. Not only did the PLP model generally outperform the conventional LP model in those automatic speech recognition tasks, it also exhibited two different types of behaviour of machine recognition accuracy as a function of the model order: in speaker-dependent experiments, accuracy was found to increase nearly asymptotically with the order of the model; by contrast, in inter-speaker experiments the highest accuracy was obtained for a 5th-order PLP model.

Consequently, Hermansky (1990, p.1743) was able to conjecture "that the linguistically relevant speaker-independent cues lie in the gross shape of the auditory spectrum" which is captured "by the one or two spectral peaks of the 5th-order PLP model", and that the "finer details of the auditory spectrum, modeled by additional poles of the higher-order PLP models, carry more speaker-dependent information." In sum, Hermansky states that the PLP model

*"...allows for the effective suppression of the speaker-dependent information by choosing the particular model order."*

Thus, the phonetic component of the acoustic speech signal is apparently emphasised by the perceptually-based frequency-scale transformation and the subsequent reduction of spectral information to only two peaks in the 5th-order PLP spectrum. Furthermore, Hermansky has shown that the second spectral peak of the 5th-order PLP analysis is

consistent with the so-called *effective second formant*  $F_2'$  which is of considerable significance in the *perception* of two-formant vowel sounds (Fant and Risberg, 1962).

### 2.2.2.2 Importance of Relative Spectral Measures

Whilst the studies reviewed thus far in this section do lend support to the contrasting roles of the low and the higher spectral regions of spoken vowel sounds, there is long-standing evidence to suggest that a mere re-scaling of the frequency axis is generally insufficient to account for the gross spectral differences manifest between groups of speakers such as adult males, adult females, and children. In this vein, Chiba and Kajiyama (1958, pp.193-194) attempted to explain the human auditory-perceptual ability to recognise the phonetic identity of vowels spoken by a wide variety of speakers spanning gender and age groups, by the “space pattern theory”:

*“A vowel is characterized by its relative formants, provided the centres of the formants are situated within certain frequency regions fixed for a given vowel...”*

More specifically in terms of the physical mechanisms of the auditory system, Potter and Steinberg (1950, p.812) suggested that

*“...within limits, a certain spatial pattern of stimulation on the basilar membrane may be identified as a given sound regardless of position along the membrane.”*

Ladefoged and Broadbent (1957) arrived at the same general conclusion by observing that listeners’ preference for the phonetic identity of a vowel embedded in a phrase, is influenced by the absolute values of the formants in the parts of the phrase preceding the test vowel. Thus, they put forward the suggestion that whilst the phonetic information is encoded in the relative positions of the formants, speaker-specific information may be encoded in their absolute values.

Indeed, it is well known (e.g., Potter and Steinberg, 1950; Peterson and Barney, 1952) that although the formants of a given vowel (presumed phonetically equivalent) spoken by an adult male, an adult female, and a child, do not have the same *absolute* positions along the frequency axis (no matter what the scale), they do appear to have similar *relative* locations. In particular, Potter and Steinberg’s (1950) measurements of the formant frequencies of 10 vowels spoken by 25 speakers of American English

including children, adult females and males, revealed a tendency for the formant distributions of some vowels to elongate along lines passing approximately through the origin. Potter and Steinberg (1950, p.816) then suggested that “different speakers tend to hold ratios of the formant frequencies constant for a given sound.” To illustrate this point, they showed that the ratios of adjacent formant frequencies in mels ( $M_2/M_1$  and  $M_3/M_2$ ) “appear to be fairly constant over the range of speakers” for each vowel, and that the ratio  $M_3/M_2$  in particular “is significantly different for the two groups, front and back vowels.”

A relatively more recent, but similar approach to overcoming gender and age-group differences in vowel formant frequencies, is Syrdal and Gopal’s (1986) proposal of a so-called “inherent or speaker-independent” normalisation method whereby the first three formant frequencies (and the fundamental frequency  $F_0$ ) are first transformed onto a Bark scale, then three new dimensions are defined by taking the Bark-scale differences  $F_1 - F_0$ ,  $F_2 - F_1$ , and  $F_3 - F_2$ . Applying this transformation to the well-known set of formant frequencies of 10 vowels recorded by 33 adult male, 26 adult female, and 15 child speakers of American English (Peterson and Barney, 1952), Syrdal and Gopal (1986, p.1095) found that it “greatly reduced between-speaker variability”. Similarly to Potter and Steinberg (1950), they also found that the Bark-difference  $F_3 - F_2$  dimension “clearly distinguished” between the front and back vowels of the 76 speakers (as did Clermont and Broad (1995), using the steady-state vowel formant frequencies of 4 adult, male speakers of Australian English).

By contrast with the preceding, formant-based approaches, Hermansky (1990) has demonstrated that the additional spectral processing involved in obtaining a 5th-order PLP spectrum, is able to reduce the significant differences in the underlying formant structure of the vowels spoken by an adult male and a child, and thus yield two-peaked spectra which reflect essentially the phonetic or the linguistic content of the spoken sounds. However, whilst Hermansky’s (1990) illustration of 5th-order PLP-based spectrograms of an utterance spoken by an adult male and a child speaker shows the PLP’s effectiveness in suppressing speaker-group differences and enhancing the linguistic message, other types of auditorily-based whole-spectrum representations have not been as successful in overcoming the dichotomy without the help of a subsequent

normalisation step. For example, Bladon et al. (1983) have shown that a reasonable first-order approximation to the normalisation of the quasi-auditory spectra of vowels spoken by adult males and females, is to shift the spectra of female speakers downward in frequency by 1 Bark, thus accounting for the generally higher-valued vowel formant distributions of female compared with male speakers.

In a similar vein, Nearey's (1978) so-called Constant Log Interval Hypothesis (CLIH) suggests that a given speaker's vowel formant data can be normalised with respect to some reference data simply by an additive constant on  $\log(F_1)$  and  $\log(F_2)$ . Nearey's results of experiments similar to those of Ladefoged and Broadbent (1957) using synthetic vowel stimuli were found to support that hypothesis, indicating that a shift in the perceived identity of a test vowel could be altered by the absolute formant values of a single, carrier vowel. The CLIH was also tested on the 76-speaker vowel formant data of Peterson and Barney (1952), and was found to greatly reduce the differences between the three groups of speakers. However, in light of the perceptual evidence reviewed earlier regarding the speaker-specificity of the higher formants, Nearey (1978) might have gained further valuable insights by extending his experimental methodology to include formants higher than the second. Clearly, a constant additive term applied to formant frequencies in the logarithmic domain, is equivalent to a constant multiplicative factor applied to the formants on a linear (Hz) scale (indeed, Nearey's Constant Ratio Hypothesis, or CRH); and the greatest effect of such a normalisation method is therefore on the higher, rather than the lower formants.

### **2.2.3 Emerging Perspective**

Our foregoing review was driven primarily by the contributions of the auditory-perception literature in elucidating the problem of speech-speaker dichotomy. Indeed, there is sufficient evidence to suggest that in human perception of spoken vowel sounds, phonetic and speaker-specific information is not evenly distributed throughout the entire spectral range. Rather, they appear to be perceptually manifest in terms of a bipartite separation of the frequency axis into a predominantly phonetic, low spectral range which encompasses the first two formants or spectral prominences, and a predominantly speaker-specific, higher spectral range which encompasses the third and

higher formants.

In addition, it emerges that in attempting to overcome the speech-speaker dichotomy by perceptually-motivated spectral emphases which are intended to suppress the speaker component while enhancing the phonetic component, the perception literature has played an influential role in shaping the commonly adopted methods of acoustic parameterisation of speech. For example, the critical band concept (Zwicker, 1961) which stems from pioneering research on the acoustic-mechanical properties of the auditory system (Békésy, 1960), has been applied with some success in machine recognition of speech; the mel-frequency cepstrum enjoys a popularity today which is clearly founded in those early, auditory-perceptual findings. Indeed, the auditorily-motivated, non-linear transformations of the frequency scale, are entirely consistent with the perceptual evidence which appears to place greater phonetic emphasis on the low spectral regions of spoken vowel sounds.

Nevertheless, the relatively large spectral differences in the spoken vowels of adult male, adult female, and child speakers, have been found to require rather more elaborate methods of spectral emphasis. For example, relative formant measures (Potter and Steinberg, 1950; Syrdal and Gopal, 1986), formant scaling (Nearey, 1978), and shifting of a quasi-auditory spectral representation along the Bark scale (Bladon et al., 1983), are just three of the perceptually-motivated methods which have been proposed to overcome those types of speaker-group differences. On the other hand, the one- or two-peaked, 5th-order PLP spectrum is purported to already suppress such speaker-group differences, and thereby yield a predominantly phonetic spectral representation (Hermansky, 1990).

Quite clearly, spectral representation is an important issue which permeates throughout the auditory-perception literature. In particular, each of the wide range of studies reviewed thus far, can be categorised on the basis of their preference for either a *continuous* (whole-spectrum) or a *discrete* (formant) spectral representation. On the one hand, the *formants* which have proved to be indispensable in acoustic-phonetics, are more directly interpreted as the resonance frequencies of the vocal-tract during speech production. On the other hand, there are conflicting points of view as to the perceptual saliency of those parameters, and a *whole-spectrum* representation is often

regarded as more akin to the human auditory representation of speech sounds.

Indeed, there has been much debate as to the appropriateness of formant and whole-spectrum representations of speech, and the debate is far from being entirely settled. One of the central issues in the continuing debate, is the ability of either the formant or a whole-spectrum representation to adequately account for the human auditory-perceptual judgement of phonetic distances between vowel sounds. For example, Bladon (1982) argues that a distance measure based on a quasi-auditory whole-spectrum representation is better able to predict the nearly equal, perceptual phonetic distances between the pairs of front vowels /i-/e/ and /i-/æ/, whilst a formant-based distance would clearly yield a much smaller distance for the former compared with the latter pair of vowels.

Bladon (1982) also argues in favour of a whole-spectrum approach, on grounds of the general indeterminacy of formants. In support of this view, Klatt (1982) notes that whilst formant-based approaches to speech recognition by machine can be quite error-prone due to gross errors in formant-tracking, such gross errors do not usually occur in human auditory-perception of speech. Consequently, Klatt (1982, p.1280) hypothesises that maybe “humans do not detect and label formant peaks when making phonetic judgements.”

Further support for this hypothesis is found, for example, in the classic study of Chistovich and Lublinskaya (1979), which suggests that the auditory system performs spectral integration of formant peaks when the centre-frequencies of those peaks are closer than about 3.5 Bark (or critical bands). Thus, the  $F_2$  and the  $F_3$  of certain front vowels appear to be auditorily represented as a single, merged spectral prominence. Both the second spectral peak of Hermansky’s (1990) 5th-order PLP model, and Delattre et al.’s (1952) observations regarding their listeners’ preference for somewhat higher second formants in two-formant vowel stimuli, are consistent both with the auditory integration theory and with the concept of an effective second formant  $F_2'$ .

The above considerations collectively raise doubts as to the perceptual compatibility of acoustic-phonetic distance measures based simply on raw formant frequencies. However, Klatt (1982) has shown that the sensitivity of perceptual phonetic quality judgements to acoustic changes in vowel sounds is largest for

percentage shifts in the formant frequencies (the first two formant frequencies in particular), compared with changes in other spectral-shape parameters such as formant bandwidths, relative formant amplitudes, and overall spectral tilt. As argued both by Klatt (1982, 1986) and by Bladon and Lindblom (1981), distance measures based on quasi-auditory, whole-spectra should therefore benefit from enhancement of the formant peaks, and suppression of other, phonetically less relevant spectral variations. Indeed, the literature does suggest that while the formant frequencies play a central role in the auditory-perceptual characterisation of spoken vowel sounds, their perceptual relevance is conditioned by the combined and interactive influences of the formant peaks in the whole-spectrum representation of those sounds.

We have begun our search for evidence of the speech-speaker dichotomy in this section by considering the acoustic-auditory literature. On the weight of the evidence found in previous works, it is clear that the perceptual characterisation of the dichotomy is enlightening and important, to the point of having influenced modern approaches to acoustic parameterisation and spectral representation of speech. However, it is also understood that perceptual evidence alone would provide only a partial understanding of the dichotomy in the acoustic-phonetic domain, to which we now turn for more substantial, and perhaps more direct insights.

## **2.3 Acoustic-Phonetic Evidence and Approaches**

In our survey of the acoustic-phonetic literature, we will first search (in Section 2.3.1) for evidence of the spectral manifestations of the speech-speaker dichotomy, whilst also attending to the range of acoustic parameters used to describe those manifestations. We will then examine (in Section 2.3.2) the consequences of the dichotomy in the two, converse tasks of automatic speaker and speech recognition, from which we expect to gain further insights into the spectral manifestations and parameterisation issues regarding the dichotomy. The latter task in particular will be shown (in Section 2.3.3) to have prompted various approaches to overcome the dichotomy, either by way of more robust spectral parameterisation, or by more explicit methods of speaker normalisation or adaptation in speech recognition by machine.

### 2.3.1 Spectral Manifestations and Descriptive Parameters

In this section we seek to gain a better understanding of the spectral manifestations of the speech-speaker dichotomy in spoken vowel sounds, by reviewing first the basic literature which has established the standard view on their *phonetic quality*, and then the more dispersed but substantial literature concerned with the manifestations of *speaker individuality* in those sounds.

#### 2.3.1.1 Phonetic Quality of Spoken Vowels

Although basic concepts in acoustic-phonetics had been known from many years earlier, the invention of the sound spectrograph (Koenig et al., 1946) made it possible to perform acoustic analyses of speech more easily and more extensively than ever before. This development spawned a number of pioneering studies concerned with acoustic descriptions of spoken language; and by far the most widely studied sounds have since been the non-nasalised vowels with which we are here concerned.

From the very first series of spectrographic observations, the regularity in the patterns of dark bands which mark the spectral regions of high acoustic energy across the duration of spoken vowels, words, and phrases, confirmed that the phonetic content of those utterances could be deciphered by what is now regarded as the art and the science of spectrogram reading. In this context, Kopp and Green (1946, p.87) noted that “similarities in the oral use of English are more numerous than differences, despite variations in vocal quality and sectional pronunciation differences.” Denoting the dark bands or formants as “bars”, Kopp and Green (1946, p.82) made the following general observations regarding an acoustic-phonetic description of the vowel space afforded by spectrograms of one speaker’s recordings of several diphthongs:

“Bar 1 is limited in movement to a small area at the bottom of the pattern, and bar 3 is both limited in the area of its movement and is frequently weak or absent from the pattern. Because bar 2 is changed in position and shape more than either of the other bars, and because of the consistent manner in which the second bars of all sounds are connected, the position of bar 2 is considered to be the principal zone or the hub for all sounds.”

Thus was established the importance of the second formant frequency  $F_2$  in describing

the principal acoustic dimensions of vocalic speech sounds.

More extensive spectrographic measurements of 10 steady-state vowels recorded by an American English phonetician, then led Potter and Steinberg (1950, p.814) to state that the “first and second formant positions show the greatest movement from vowel to vowel.” Indeed, the vowel formant distribution of that speaker plotted on the  $F_1F_2$  plane was clearly found to be “such as to identify each vowel distinctly.” Whilst even more extensive measurements of vowel formant frequencies have shown that there exists significant overlap amongst vowel classes when data of a large or heterogeneous group of speakers are pooled (e.g., Peterson and Barney, 1952; Bernard, 1967, cf. Clermont, 1996; Hillenbrand et al., 1995), these and many other acoustic-phonetic studies have since established the standard view that the  $F_1F_2$  plane is the most descriptive, two-dimensional representation of the phonetic quality of spoken vowel sounds.

It follows therefore, that the bulk of phonetic information in those speech sounds should be contained in the low spectral regions which encompass the first two formants. In this vein, Pols et al. (1973) have shown that the maximally vowel-discriminating plane found by rotation of the first two dimensions of a principal components analysis of the  $\frac{1}{3}$ -octave band spectra (spanning the frequency range from 100Hz to 10kHz) of 12 Dutch vowels recorded by 50 adult male speakers, is highly correlated with the plane defined by  $\log(F_1)$  and  $\log(F_2)$  measured from the same data. This classic result attests both to the phonetic significance of the first two formant frequencies of spoken vowels, and to that of the low spectral regions wherein they reside.

By contrast with the major role of  $F_1$  and  $F_2$  in describing the phonetic quality of spoken vowels, the higher formants (and by extension, the higher spectral regions) would appear to play only a minor or subsidiary role. For example, the third and fourth formants are often regarded as complementary to the high  $F_2$  in shaping and emphasising the spectral prominence of high front vowels; retroflexion, as in the American English vowel / $\text{ɝ}$ /, is known to substantially lower the  $F_3$ . On the other hand, these higher resonances and the spectral regions in which they are contained, appear to manifest a great deal of speaker-specific information, as our following review of the acoustic-phonetic literature on speaker individuality will bear out.

### 2.3.1.2 Speaker Individuality

Indeed, the speaker-specificity of the higher formants has been observed in numerous studies, and in different research contexts. For example, in a spectrographic study of 7 vocalic sounds including diphthongs and diphthongised monophthongs recorded by speakers of two main varieties (or idiolects) of Australian English, Burgess (1969) observed several, idiolect-specific trends in the behaviour of  $F_3$  in those speech sounds, thus implying that the third formant might provide a reliable cue for the speaker's idiolect. In a different spectrographic study of spoken texts recorded by 5 speakers and by two well-known imitators who attempted to imitate the voices of those 5 speakers, Endres et al. (1971, p.1845) observed that the formant structure in the vowels /i:/ and /a/ of the original speakers and of the imitators “do not agree, especially in the higher-frequency bands” above about 2kHz.

More quantitative, analyses of variance of the formant frequencies of 12 Dutch vowels recorded by 50 adult, male speakers, clearly showed (Pols et al., 1973, Table II, p.1095) the *speaker* variance in  $\log(F_3)$  to be much higher than that in either  $\log(F_1)$  or  $\log(F_2)$  (23% of the total variance in the data, compared with 5% and 2%, respectively); by contrast, the *vowel* variance in the same data was observed to behave conversely (43%, compared with 86% and 94%, respectively). Similarly, analyses of variance of the first four formant frequencies measured in 5 vowels contained in 4 words, recorded several times over a duration of 5 years by 9 adult, male speakers of Japanese, led Saito and Itakura (1983, p.150) to observe that the variances of the “speaker factor” in “the higher formant frequencies are higher than those of the lower ones.”

Evidence of the speaker-specificity of the higher spectral regions also appears in studies adopting a whole-spectrum approach to parameterisation of the acoustic speech signal. For example, histograms of the so-called “average absolute difference” between 22 speakers' long-term spectral correlation matrices obtained using a 35-band filter-bank spanning the range from 270Hz to 10kHz, was found by Li and Hughes (1974, p.835) to have a higher distribution when the frequencies below 2200Hz were de-emphasised with a +6 dB/oct filter; thus, they concluded that “low-frequency

deemphasis helps to retain more intertalker differences.” In a study of the long-term averaged spectra of a phonetically-balanced list of 25 spondee words recorded by 20 male, 18 female, and 2 child speakers of Finnish, Kiukaanniemi et al. (1982, p.23) found, for each group of speakers separately, “surprisingly large individual variations in the distribution of speech power especially at frequencies 1000-3000Hz.”

Using more contemporary methods of speech signal analysis to obtain the long-term average, cepstrally-smoothed spectra (by truncation of the FFT cepstrum to 11 terms) of 2 words recorded by 9 adult, male speakers of Japanese, Furui and Akagi (1985, p.6) conclude that

“...the individuality of the spectral envelope derived from the cepstrum is almost completely represented by the difference of spectral levels between 2.5 and 3.5 kHz.”

Similar conclusions emerge from Kitamura and Akagi’s (1994, Figure 3, p.1185) profiles of the vowel and speaker components of spectral variance, computed for 5 vowels recorded by 10 male speakers of Japanese. Representing each vowel spectrum with 60 FFT cepstral coefficients (and using an auditorily-motivated frequency-scale), they confirmed their perceptual findings (discussed earlier, in Section 2.2.1) that whilst “vowel characteristics” are concentrated mainly in the lower spectral region from about 603Hz to 2212Hz, “speaker individualities” are to be found mainly above 2212Hz.

Whilst the studies reviewed thus far have provided evidence concerning the spectral correlates of speaker individuality in spoken utterances which contain vocalic sounds, the question arises whether those spectral correlates are related to, or might be predicted by, the greater speaker-specificity of certain phonetic (or vowel) subspaces. In this regard, Stevens et al. (1968) provide an indication of the importance of the phonetic feature “front-back” for vowels. In particular, identification of speakers by visual inspection of voice spectrograms was found to be more reliable for utterances containing stressed front vowels, than for those containing stressed back vowels. Consistent with the emerging view on the spectral correlates of speaker individuality, Stevens et al. (1968, p.1602) then suggest that “the quality of a talker’s voice may be judged, both visually and aurally, on the basis of high-frequency components, say in the range 2000-3000Hz.”

The speaker-related potency of front vowels is also borne out in the formant modelling study of Sato et al. (1982). Applying the bi-planar model of Broad and Wakita (1977) to the first three formant frequencies of 5 vowels extracted from sentences recorded by each of 4 adult, male speakers of Japanese, they obtained larger, inter-speaker differences in the coefficients of the planar equations for the front vowels than for the back vowels. On the other hand, van den Heuvel and Rietveld (1992) obtained a higher “ratio of speaker specificity” (RSS) for the vowel /a/ compared with the vowels /i/ and /u/, in 24 /CVCə/ nonsense words recorded by 5 male and 5 female speakers of Dutch. However, their whole-spectrum representation (LP cepstrum) and computation of spectral distance using dynamic time warping (DTW), did not yield insights regarding the contributions of different spectral regions to the RSS.

On the basis of the accruing evidence regarding both the spectral correlates and the phonetic subspaces of greater *speaker* potency, the acoustic-phonetic literature on speaker individuality seems to implicate the higher spectral regions which encompass not only  $F_3$  and the higher resonances, but also the  $F_2$  of front vowels. By contrast, the more commonly accepted view on the acoustic correlates of *phonetic* information in spoken vowels, as reviewed earlier, clearly points to the importance of the low spectral regions which encompass the entire range of both  $F_1$  and  $F_2$ . Furthermore, these implications are also consistent with the auditory-perceptual findings reviewed earlier (in Section 2.2), which imply that whilst as many as two formants are necessary for the perception of vowel sounds, the higher spectral range which appears to carry perceptually important information on speaker individuality, may also include the  $F_2$  of front vowels.

Together, these conclusions then raise the question of whether the phonetic and the speaker-specific influences shown to be manifest *predominantly* in the low and in the higher spectral regions, respectively, might not overlap and impinge on each other to yield the phenomenon of *vowel-speaker interactions*. In this vein, it is worth noting that the acoustic-phonetic studies reviewed in this section have adopted mainly direct methods of observation such as analyses of variance, in order to quantify separately either the phonetic or the speaker component of spectral variability. By contrast, we might expect the effects of vowel-speaker interactions to be more explicitly manifest,

and therefore better described, in terms of their consequences in vowel and speaker recognition by machine.

### **2.3.2 Consequences in Vowel/Speaker Recognition by Machine**

The evidence reviewed in the previous section concerning the spectral manifestations and acoustic-phonetic correlates of vowel quality and of speaker individuality, to a large extent foreshadow the consequences of the speech-speaker dichotomy in automatic vowel and speaker recognition. Indeed, those spectral regions or spectral parameters which have been directly observed to manifest the bulk of either speaker individuality or phonetic quality, can be expected to also yield the highest accuracies in automatic speaker and vowel recognition, respectively. In addition, these converse tasks offer the potential to observe the spectral manifestations of the vowel-speaker dichotomy as follows: in speaker recognition by machine emerges the question of the most appropriate vowel subspaces; in automatic vowel recognition, vowel-speaker interactions may be induced by performing recognition on an inter-speaker (or a speaker-independent) basis. In the following two sections, therefore, further insights into the phenomenon of speech-speaker dichotomy are sought by reviewing the relevant literature on automatic speaker and vowel recognition, respectively.

#### **2.3.2.1 Speaker Recognition**

One of the earliest studies to consider the relative importance of formants in speaker recognition by machine, appears to be that of Carré (1971). Speaker identification was attempted using pairwise combinations of the first three formant frequencies extracted from 2 repetitions of a given, dynamic vocalic utterance (described as containing three syllables within a duration of about 500msec extracted for formant analysis) recorded by 10 male speakers of French. Formant trajectories on the  $F_2F_3$  plane were found to be more effective in discriminating between the speakers, than those on the  $F_1F_2$  plane.

The effectiveness of different formants of different vowels in speaker identification by machine has since received attention in several studies. For example, Sambur (1975) found that the values of the second, third, and fourth formant frequencies measured in the steady-state of certain vowels contained in sentences recorded over a period of 3.5

years by 21 adult, male speakers of American English, were amongst the most important acoustic parameters for automatic identification of those speakers. For the front vowels /i/, /ɪ/ and /æ/, the best combination of high inter-speaker variability and low intra-speaker variability was found in  $F_2$  and  $F_4$ , whilst for the back vowel /u/, the most effective parameter was found to be  $F_3$ . The first four formant frequencies of 11 vowels recorded in /hVd/ context by 1 female and 9 male, adult speakers of British English were assessed for their speaker discriminating power by Paliwal (1984a), who found the best performance in both  $F_2$  and  $F_3$  of the mid- to front vowels /ɜ/ and /ɪ/. More recently, Mella (1994) found that on average, the performance of  $F_3$  in automatic speaker identification is superior to that of  $F_1$  and  $F_2$ . Speaker identification performed using the first three formant frequencies in 4 repetitions of 7 vowels in /pVr/ and /bVr/ context in sentences recorded by 10 male speakers of French (of the Lorraine region), showed that for  $F_2$ , the front vowels /i/, /e/, /ɛ/, and the central vowel /œ/ yielded the highest accuracies, whilst for  $F_3$ , the back vowel /u/ performed the best. Similarly, Cooper and Clermont (1994) showed that the  $F_2$  of front vowels and the  $F_3$  of back vowels yielded the highest accuracies in speaker identification, using 5 repetitions of /CVd/ monosyllables recorded by 4 adult, male speakers of Australian English. These authors also reported that speaker-identification experiments performed on Peterson and Barney's (1952) formant data of 33 adult, male speakers of American English, yielded the highest accuracy (96%) using only  $F_2$  and  $F_3$  of 8 of the 10 vowels, having excluded the two back vowels /ɑ/ and /ʊ/.

Evidently, the contribution of formants in the upper spectral range (including  $F_2$  of front vowels) appears to be greater than that of the formants in the low spectral range. However, formant extraction is admittedly a non-trivial task; not surprisingly, therefore, the majority of studies in automatic speaker recognition have opted for the computationally less expensive and more robust, whole-spectrum approach to acoustic parameterisation. Unfortunately, this approach is often taken to implicitly condone the use of the entire available spectral range (i.e., up to half the sampling frequency of the speech signal), without further regard for the relative effectiveness of different spectral regions, as shown in the numerous studies reviewed thus far in this chapter. Nevertheless, previous works which have adopted a whole-spectrum approach in

speaker recognition, do offer insights regarding the effectiveness of different phonemes.

For example, Kashyap (1976) used a rather limited set of phonemes (four front vowels, one high back vowel, and two nasal consonants) extracted from a list of words recorded by 4 speakers, and obtained the highest inter-speaker distances (aimed at being used more effectively in automatic speaker recognition) for the front vowels /ɪ/ and /ɛ/. A more comprehensive study of the relative importance of 24 different phonemes (Höfker, 1976, cited by Jesorsky, 1978, p.117), found that amongst the vowels, the front vowels /i/, /e/, /ɛ/, and /ɪ/ yielded the highest accuracies in the identification of 12 speakers. Using 12th-order LP cepstra and FFT mel-frequency cepstra measured in 6 sentences recorded (at a sampling frequency of 8kHz) over the telephone by 125 speakers, Eatock and Mason (1994) conducted an even more comprehensive study of the effectiveness of 75 different phonemes in a speaker verification task based on vector quantisation (VQ), and obtained a rank-ordering of those phonemes, the top four of which were the nasal /ŋ/, the low vowel /ɑ/, and the two high front vowels /i/ and /ɪ/. By comparison, Liou and Mammone (1995) used 12th-order mel-frequency cepstral coefficients, the signal energy, and the first and second-order differences of those parameters (across time) measured for 30 speakers, and found the statistical F-ratio of speaker discriminability yielded by a neural tree network (NTN), to be the highest for the front vowel /e/, the retroflex /ɤ/, and the diphthong /aɪ/ which does involve a series of front-vowel configurations (Clermont, 1991).

In contrast with those studies which have adopted a whole-spectrum approach without attempting to seek further insights regarding the spectral ranges of importance, van den Heuvel et al. (1993) performed speaker discriminant analysis using the output of each of 19 frequency sub-bands spanning the frequency range up to 8kHz on a Bark scale. Their speech material consisted of the vocalic steady-state of 24 isolated /CVCə/ nonsense words recorded by 15 adult, male speakers of Dutch. Most importantly, despite their use of a filter-bank, their interpretations of the recognition results were offered in terms of the speakers' formant frequency ranges. In particular, they found the best speaker discrimination in the frequency bands around  $F_3$  of /a/, and around  $F_2$  and  $F_3$  of /ɪ/. Interestingly, these results qualify the earlier finding of van den Heuval

and Rietveld (1992) regarding the higher ratio of speaker specificity (RSS) obtained for the vowel /a/ of their speakers of Dutch (as reviewed in Section 2.3.1.2). Although van den Heuvel and Rietveld did not venture in their earlier study to spectrally decompose their cepstrum-based, and thus whole-spectrum representation, it is indeed likely that the higher spectral regions containing  $F_3$  made the largest contribution to the high RSS obtained for that vowel.

By contrast, the following two studies were able to confirm the relative importance of the higher spectral regions in speaker recognition by machine using a cepstrum-based approach, albeit at the cost of cepstral re-analysis in each frequency band of interest. First, Hayakawa and Itakura (1994) performed automatic speaker recognition using cepstral coefficients obtained by *selective* LP analysis (Makhoul, 1975b) within various frequency sub-bands (in the range up to 16kHz) of 5 repetitions of 5 words, recorded every 3 months over a duration of 1 year by 5 male speakers of Japanese. Whilst they could not therefore guarantee an equivalent accuracy of spectral representation across the different frequency bands, they were able to conclude that the band spanning 4–10 kHz carries as much speaker-specific information as the low-frequency range extending up to 4kHz. A methodologically more consistent method of re-analysis was adopted recently by Lin et al. (1996), who obtained 20th-order FFT cepstral coefficients in two, *equal*-interval frequency ranges spanning 0–5 kHz and 3–8 kHz, respectively. Text-independent, VQ-based speaker identification using sentences recorded by 24 male and 14 female speakers then yielded a lower error-rate for the higher spectral range (1.1%) than for the lower spectral range (5.8%).

Prompted perhaps mainly by a desire to gain higher levels of accuracy in automatic speaker recognition, recent studies clearly reflect an increasing interest in the relative importance of different spectral regions and phonetic subspaces. Further examples include some recent attempts at separate analyses and weighted recombination of FFT-based frequency sub-bands (e.g., Besacier and Bonastre, 1997; Auckenthaler and Mason, 1997), and at least one attempt to incorporate an empirically-determined phonetic weighting in speaker verification (e.g. Liou and Mammone, 1995). However, the speech-speaker dichotomy does have serious implications in speaker-independent, speech recognition which need to be elucidated as well. To this end, therefore, a careful

review of studies of vowel recognition in particular, is attempted in the next section.

### **2.3.2.2 Vowel Recognition**

Whilst in automatic speaker recognition there is the question of determining the most speaker-specific acoustic-phonetic subspaces, the converse problem of speaker-independent vowel recognition by machine offers the potential of observing the manifestations of *vowel-speaker interactions*, albeit indirectly via recognition accuracy. For example, it is well known that in general, a speech (or vowel) recognition system will perform worse on an inter-speaker, than on an intra-speaker basis (e.g., Rabiner and Juang, 1993, p.285). However, it is quite clear from our earlier review of the literature on phonetic quality, speaker individuality, and automatic speaker recognition, that vowel recognition accuracy will depend not only on the presence of inter-speaker variability, but also on the spectral range or the spectral parameters selected.

Indeed, Paliwal (1984b, p.104) obtained results which confirm the importance of the low spectral range in vowel recognition by machine. In particular, he found spectral pre-emphasis to be detrimental to vowel recognition on an intra-speaker basis, and attributed the deterioration in recognition accuracy to the “undue weight on high frequency components” which do not carry reliable information pertaining to vowel identity.

Using an LP-based whole-spectrum representation, Ainsworth and Foster (1985) obtained results which more directly confirm the implications of vowel-speaker interactions in recognition. They performed vowel classification experiments both on an intra- and an inter-speaker basis, using 10 repetitions of 11 vowels recorded in /hVd/ context by 4 male and 4 female speakers of English. Each vocalic steady-state was first represented by a 128-point spectrum up to 5kHz (obtained by Fourier transformation of the 10th-order LP autoregressive coefficients). Vowel classification was then performed using first the entire spectral range, then only the spectral points up to about 3200Hz. Whilst intra-speaker accuracy suffered a 3% drop (to 93%) as a result of the reduction in spectral range, inter-speaker vowel classification accuracy was found to *increase* by 3% (to 43%). These results not only confirm the relative phonetic unimportance of the spectral range between 3200Hz and 5000Hz, but also demonstrate

its speaker-related, *detrimental* influence in vowel recognition.

As that particular spectral range includes perhaps only the high end of the third formant distribution of most adult female speakers (e.g. Peterson and Barney, 1952), Ainsworth and Foster's (1985) result is more likely indicative of the phonetic unimportance and speaker-specificity of mainly the fourth and fifth formant frequency ranges. However, Arslan and Hansen's (1997) results do bear on the frequency ranges of the lower formants. They performed speaker-independent, isolated word recognition using a hidden Markov model (HMM), in each of 16 frequency bands spaced uniformly across the spectral range up to 4kHz, and with a dataset of 48 adult male speakers, including equal numbers of native speakers of American English and speakers of Turkish-, Mandarin-, and German-accented English. Speech recognition accuracy was found to be highest in the frequency pass-bands below about 1800Hz; by contrast, accuracy was found to be much lower in the pass-bands above about 1800Hz. On the other hand, the accuracy of classification of the speakers' accents was found to be highest in the so-called mid-range frequencies spanning 1500 – 2500 Hz.

In addition to their whole-spectrum approach using a filter-bank analysis, Arslan and Hansen (1997) also performed speech recognition and accent classification using each of the first four formant frequencies separately. In speaker-independent speech recognition,  $F_2$  followed by  $F_1$  were found to yield the highest scores; by contrast,  $F_2$  and  $F_3$  yielded the best performance in accent classification. Although accent classification involves characteristics of groups of speakers rather than individuals, these results are nevertheless consistent with those reviewed earlier, which together substantiate the contrastive roles of the low and the higher spectral regions in regard to the speech-speaker dichotomy.

In contrast to Arslan and Hansen's (1997) use of each formant separately in recognition, most previous works have used the formants *in crescendo* (i.e. first only  $F_1$ , then  $F_1$  and  $F_2$  together, then the first three formants, and so on). In light of the evidence reviewed earlier, this approach implicitly acknowledges the phonetic importance of the first two formants; in addition, it may afford the potential of observing the influences of vowel-speaker interactions on recognition accuracy, as first the phonetic, and then the presumably more speaker-specific formants are recruited in

the recognition process. Indeed, somewhat analogously to the whole-spectrum results of Ainsworth and Foster (1985) reviewed above, one might expect vowel classification accuracy to be detrimentally affected upon recruitment of  $F_3$  and higher formants.

Table 2.1 shows a summary of the results reported in eight different studies. Summarised in each column from left to right, following the authors and the year of publication, are the vowel and speaker materials used, the type of decision boundaries used in the classifier, and the frequency scale of the formants. In the last two columns are listed the inter-speaker vowel classification accuracies obtained using only the first two, then the first three formants.

Consistent with our initial expectations, recruitment of  $F_3$  is reported to affect inter-speaker vowel classification accuracy *detrimentally* (by about 3.5% on average), but in only three of the studies listed (Assmann et al., 1982; Chung et al., 1988; and Kumar, 1996). By contrast, the remaining five studies report an *improvement* in accuracy (by about 8.3% on average). Apparently therefore, the effectiveness of the third formant in manifesting vowel-speaker interactions, is not consistently observable in this type of recognition experiment. Nor can this discrepancy in the role of the third formant be explained entirely in terms of methodological differences — both rises and falls in accuracy are reported using both linear and logarithmic frequency scales, both linear and quadratic classifiers, and with both smaller and larger populations of speakers of different languages or dialects.

It is perhaps not surprising that this apparent discrepancy should not have been noted in previous studies on vowel recognition, which have been primarily concerned with accuracy and not with explaining the speech-speaker dichotomy problem. It would appear, however, that spectral representation can play a determining role in revealing manifestations of the dichotomy in recognition. On the one hand, the inconsistency in the contribution of the third formant noted above, suggests that formant-based spectral representations in vowel recognition may undermine the possibility of gaining insights into the dichotomy. On the other hand, quite a number of studies reviewed earlier suggest that a whole-spectrum representation is likely to more consistently exhibit the detrimental consequences of speaker-induced variability in the higher spectral regions, on inter-speaker vowel classification accuracy.

Previous Study	Vowels	Speakers	Classifier	Frequency scale	Vowel Recognition Accuracy (%)		
					$F_1F_2$	$F_1F_2F_3$	
Welch & Wimpres (1961)	10 American English	33 male & 26 female (Peterson & Barney, 1952) <sup>†</sup>	Quadratic	Hz	87	91	
Pols et al. (1973)	12 Dutch	50 male	Quadratic	log(Hz)	71.3	75.3	
Assmann et al. (1982)	10 Canadian English	5 male & 5 female	Linear (Mahalanobis)	ln(Hz)	<b>80</b>	<b>77</b>	
Chung et al. (1988)	8 Korean	20 male	Quadratic	Hz	<b>92.4</b>	<b>92.2</b>	
Hillenbrand & Gayvert (1993)	10 American English	33 male, 26 female, & 15 child (Peterson & Barney, 1952)	Quadratic	Hz	74.9	83.6	
				log(Hz)	75.2	83.6	
				Bark	76.1	83.6	
Zahorian & Jagharghi (1993)	11 American English	10 male, 10 female, & 10 child	Quadratic	Hz	60.5	73.0	
Hillenbrand et al. (1995)	12 American English	45 male, 48 female, & 46 child	Quadratic	Hz	68.2	81.0	
Kumar (1996)	11 Australian English	36 male (of 3 idiolectal groups) <sup>‡</sup>	Linear (Mahalanobis)	Hz	<b>82.8</b>	<b>72.0</b>	
				Quadratic	Hz	<b>84.4</b>	<b>83.6</b>
				Quadratic	Hz	<b>68.2</b>	<b>64.0</b>
				Quadratic	Hz	<b>80.6</b>	<b>76.4</b>
				Quadratic	Hz	<b>72.7</b>	<b>71.5</b>

Table 2.1: Summary of speech materials, methods, and inter-speaker (or speaker-independent) automatic vowel classification accuracies obtained in 8 previous acoustic-phonetic studies, using the *first two*, then the *first three* formant frequencies. Accuracies in bold-font indicate *detrimental* effect of the third formant. [<sup>†</sup>Only those vowels which were correctly identified unanimously by a panel of 26 listeners; <sup>‡</sup>Unpublished results of Kumar’s (1996) Master’s thesis.]

In an attempt to reconcile these two behaviours, it seems important to reconsider the whole-spectrum versus formants issue raised earlier in Section 2.2.3, from the point of view of recognition. First, by invoking basic principles of pattern recognition, it can be argued that if the ratio of the amount of training data to the dimensionality of the feature set is sufficiently high, and provided also that the probabilistic assumptions of the classifier are reasonably met in the given data, then an additional, independent parameter may either improve, or at worst leave unaffected the classification accuracy

yielded by a nearly optimal classifier (Duda and Hart, 1973, p.67), even if that extra parameter ( $F_3$ ) were to exhibit potentially confounding, inter-speaker variations. Indeed, this argument finds support in the very low feature-set dimensionality of only 2 or 3 formants, and in the often reasonable assumption of a normal distribution of formant data on a per-vowel basis (as required by a quadratic classifier).

It is also clear that in a recognition framework, the formant parameters are treated as independent, mutually orthogonal dimensions. By contrast, a whole-spectrum representation used in recognition naturally embodies formant-peak interactions which occur both on an intra-frame (i.e., merging of proximate formant peaks) and on an inter-frame basis (e.g., inter-vowel or inter-speaker overlap of different formant peaks). In addition, both the perceptual and acoustic-phonetic literature reviewed earlier implicate the speaker-specificity of not only the third and higher formants, but also the  $F_2$  of front vowels, which cannot simply be excluded from a vowel recognition experiment using the first two formants. It is therefore reasonable to speculate that the dichotomy should be more clearly manifest in vowel recognition using a whole-spectrum approach, rather than the discrete, formant-based representation which would appear not only to discard important formant-peak interactions, but also to limit the degree to which vowel and speaker influences might be decoupled.

### **2.3.3 Approaches to Overcoming the Dichotomy**

Having reviewed the relevant literature on the spectral manifestations of the speech-speaker dichotomy and its consequences in vowel and speaker recognition by machine, we now direct our attention to acoustic-phonetic approaches intended to overcome the dichotomy. In contrast to the frequency-scale transformations and relative spectral measures which (as reviewed in Section 2.2.2) are motivated by properties of human auditory-perception, the approaches to be described in the following two Sections are driven mainly by the criterion of higher accuracies in speaker-independent speech recognition by machine, and do not have a perceptual basis. In Section 2.3.3.1, we consider one of the more important trends in gaining robustness of spectral parameters, hence higher accuracies, in recognition. In Section 2.3.3.2 we then consider some of the more explicit methods of speaker normalisation and adaptation. A recurrent question is

whether any of these approaches have yielded, irrespective of their success in dealing with or normalising speaker differences, any new insights into the dichotomy problem.

### 2.3.3.1 Towards Robustness in Parameterisation

We deduced from the literature reviewed in the previous section, that the formants may not be the best spectral parameters for elucidating the speech-speaker dichotomy in a vowel recognition framework. However, this does not imply that the formants are not useful parameters in vowel recognition; on the contrary, their potential in that context (and more generally in speech recognition) has often been advocated (e.g., Broad, 1972; Zahorian and Jagharghi, 1993; Holmes et al., 1997). Unfortunately, the formants do suffer the serious limitation of being difficult to measure robustly — a limitation which has prevented a wider use of those parameters in speech recognition systems. It is therefore a fortunate coincidence that the more robustly measured, whole-spectrum parameters such as the cepstrum, do hold the potential of elucidating the dichotomy problem via recognition.

Independently of its potential ability to yield insights into the dichotomy problem, the cepstrum has indeed enjoyed tremendous popularity and relative success both in speech recognition (e.g., Davis and Mermelstein, 1980; Paliwal and Rao, 1982; Young, 1996) and in speaker recognition (e.g., Luck, 1969; Atal, 1974; Furui, 1981; Mammone et al., 1996). The demand for higher accuracies, however, has inevitably led to continued efforts aimed at improving the robustness of the cepstrum in either of those tasks. As noted in Section 2.2, the perception literature has been particularly influential in this regard, motivating the transformation of the cepstrum onto the phonetically more robust, mel-frequency scale. An equally important method for improving the robustness of the cepstrum in recognition, is that of *liftering* (or weighting of the cepstrum in the quefrequency domain), the objective of which is aptly summarised by Rabiner and Juang (1993, p.169):

“... to control the noninformation-bearing cepstral variabilities for *reliable* discrimination of sounds.”

In line with Rabiner and Juang’s prescription, are the perceptually-motivated arguments put forward by proponents of the whole-spectrum approach (e.g., Klatt,

1982, 1986; Bladon and Lindblom, 1981, as discussed in Section 2.2.3), who advocate spectral enhancement of the formant peaks for improved vowel recognition in particular. Klatt's (1982) proposal of a spectral slope measure based on an auditory filter-bank, was indeed an important first step towards enhancing formant peaks in spectral distance computation. However, a major advance in that regard had already been made by Yegnanarayana (1978), who first showed that the negative derivative of the LP phase spectrum (NDPS, also known as the group delay spectrum) has several desirable properties, including not only the flattening of the phonetically unimportant spectral slope and the enhancement of formant peaks (see also Fuchi and Ohta, 1978), but also the additive (rather than multiplicative) influence of formant peaks in that spectral representation. Most importantly, Yegnanarayana showed that the NDPS is equivalent to the so-called quefrequency-weighted cepstrum (QCEP, also known as the root power sum), and that its phonetically relevant properties are therefore embodied both in the QCEP and, by extension, in the cepstral distance measure based on the QCEP (Yegnanarayana and Reddy, 1979).

The QCEP has since been shown to perform better than the unweighted LP cepstrum in speaker-dependent vowel recognition by machine (Paliwal, 1982), and in speaker-dependent isolated word recognition in the presence of noise (Hanson and Wakita, 1987). More specifically in the context of the speech-speaker dichotomy, Tohkura (1986) has shown that a cepstral lifter based on the reciprocal of the empirically-determined variance in each cepstral coefficient, is not only very similar to the QCEP, but also performs much better than the unweighted cepstrum in *speaker-independent* isolated word recognition. Indeed, Tohkura (1986, p.761) concluded that the relative de-weighting of the low-order cepstral coefficients "tended to equalize the performance of the recognizer across different talkers."

However, the best performance was reported by Tohkura (1986) with the number of cepstral coefficients approximately equal to the LP analysis order; the higher-order coefficients were found to be detrimental to recognition accuracy. Indeed, the role of the higher cepstral coefficients had earlier been questioned by Davis and Mermelstein (1980, p.364) who found, using unweighted, mel-frequency cepstra in monosyllabic word recognition, that the first six coefficients (based on a sampling frequency of

10kHz) already yielded high accuracies, and that the “importance of the higher cepstrum coefficients appears to depend on the speaker.” The importance of spectral smoothing (by cepstral truncation) in order to reduce the sensitivity of distance measures used in speech recognition, is now well-established, as indicated for example by Shikano and Itakura (1992, p.422): “... broadening of the peaks is indispensable, since just a slight deviation in formant frequencies between two patterns would result in a large distance value.”

A compromise between the desirable, high-frequency emphasis of the QCEP, and the efficacy of spectral smoothing by cepstral truncation, was proposed by Itakura and Umezaki (1987) in the form of a Gaussian lifter, and by Juang et al. (1987) in the form of a raised-sinusoidal, band-pass lifter. The former was evaluated in speaker-dependent, isolated word recognition, and found to yield the highest recognition accuracies with an equivalent frequency resolution (up to 5kHz) of about 300Hz. The latter, band-pass lifter was evaluated in both vowel and isolated digit recognition on an *inter-speaker* basis, and was found to yield the highest accuracies compared with several competing, cepstrum-based distance measures. Juang et al. (1987, p.949) then attributed the success of the band-pass liftering approach, to the effective suppression of the following two sources of variability: that of higher frequency terms which determine the more detailed, higher-resolution components of the spectral shape, are “inherent artifacts of the analysis procedure”; and that of low-frequency terms which affect mainly the gross spectral shape and tilt, are “primarily due to variations in transmission, speaker characteristics, and vocal efforts ...”.

Admittedly, cepstral liftering achieves a *wholesale* modification of the entire shape of the spectrum by a combination of enhancement and smoothing of spectral peaks, the effect of which is difficult if not impossible to predict in specific frequency sub-bands. In view of the evidence accumulated in the auditory-perception and acoustic-phonetic literature reviewed earlier in this chapter, attempts to control the so-called “noninformation-bearing variabilities” in the spectrum by cepstral liftering are therefore, at best, *indirect* approaches to dealing with the speech-speaker dichotomy. Indeed, whilst the importance of cepstral liftering is itself acknowledgment of the dichotomy problem, that method does not allow direct control of any particular frequency sub-

band. Nor are current cepstral distances capable of yielding similarity measures in selectable frequency sub-bands. Further research, which more directly addresses the problem of speech-speaker dichotomy by allowing the flexibility of selecting spectral ranges in either cepstrum derivation or cepstral distance computation, is clearly warranted!

### 2.3.3.2 Speaker Normalisation

In contrast to the indirect approach of cepstral liftering to overcome the dichotomy, the harmful influences of inter-speaker variability in speech recognition have been more explicitly acknowledged, and more directly dealt with, by so-called methods of *speaker normalisation*. As with the dual approach to spectral parameterisation evident in the various bodies of literature reviewed thus far, methods of speaker normalisation may also be categorised as adopting either a formant or a whole-spectrum approach. A relevant question, then, is to what extent either of these direct approaches have acknowledged or further elucidated the acoustic-phonetic manifestations of the speech-speaker dichotomy.

Whilst the following three methods of speaker normalisation are theoretically applicable to any number of formants, they were originally intended for only the first two. Gerstman (1968) linearly mapped the maximum range of each speaker's  $F_1$  and  $F_2$  separately onto a normalised scale from 0 to 999. Using the classic dataset of Peterson and Barney (1952), Gerstman (1968, p.79) reported “no significant  $F_1$  or  $F_2$  differences” in the per-vowel mean formant values computed for each of the three groups of speakers (adult male, adult female, and child) after normalisation. Vowel classification performed by Gerstman using the normalised formant values, their sums, and their differences (with  $F_3$  used solely to identify the retroflex vowel), then yielded misclassifications (approximately 12%) for only those vowel tokens which had also failed of unanimous correct identification by a panel of 26 listeners. The first two formants were also the focus of Lobanov's (1971) normalisation of the overall mean and standard-deviation of each speaker's vowel formant distribution. Applied to the steady-state vowel formants of 5 adult, male speakers of Russian, his method yielded a higher index of adjacent-vowel separability, compared either with Gerstman's range-

normalisation, or with a linear scaling to normalise the speakers' maximum formant values. More recently, Di Benedetto and Liénard (1992) proposed an “isomorphic transformation” of each speaker's  $F_1F_2$  vowel formant distribution, onto a normalised plane where each vowel token retains its relative distances from three reference vowels (chosen as the three point vowels /i/, /a/, and /u/). Applied to the Peterson and Barney (1952) dataset of all 76 speakers' formants, the method was found to yield degrees of pairwise vowel discrimination comparable to those yielded by a modified version of Gerstman's method whereby only the three point vowels are used in the range-normalisation. However, by definition, this method untenably assumes that those reference vowels are phonetically, precisely equivalent across speakers. An equally drastic assumption of all three methods reviewed above, has led to their blatant disregard of the speaker-related influences of formants higher than the second, which might potentially have an important role in speaker normalisation.

By contrast, Assmann et al. (1982) applied Nearey's (1978) CLIH method (see Section 2.2.2.2) to the first *three* formant frequencies of 10 vowels recorded by 5 male and 5 female speakers of Canadian English. As expected, they obtained higher vowel classification accuracies with the speaker-normalised, than with the original formants. However, in contrast to the *drop* in accuracy obtained originally upon recruiting the third formant (from 80% to 77%, as listed earlier in Table 2.1), the speaker-normalised data yielded a small *rise* in accuracy from 91% to 93%. Whilst these results indirectly confirm the speaker-related potency of the third formant in particular, apparently contradictory results were obtained by Pols et al. (1973), who had earlier applied a similar method of speaker normalisation (i.e., translation of each speaker's, logarithmic vowel formant data to a common, overall mean value) to the first three formants of 12 vowels recorded by 50 male speakers of Dutch. Indeed, further attesting to the inconsistency of formant-based vowel recognition results, their speaker-normalised data yielded a relatively *smaller* rise than did their original data, upon recruiting the third formant (from 78.3% to 80.5%, in contrast with the larger rise from 71.3% to 75.3% as listed earlier in Table 2.1).

The apparently inconsistent behaviour of formants in vowel recognition as typified in those studies, and the general lack of insights regarding the contribution of acoustic-

phonetic subspaces or of higher formants, raises the question of whether a whole-spectrum approach to speaker normalisation is more likely, or indeed has previously been used, to provide such insights into the speech-speaker dichotomy. One of the simplest approaches to speaker normalisation using whole-spectrum parameters, is that of long-term mean subtraction (which has also been used with the aim of removing frequency-dependent recording or transmission-channel distortions in automatic speaker recognition, as proposed for example by Atal (1974) using the LP cepstrum). A similar method was indeed used by Plomp et al. (1967) in order to remove the vowel-averaged speaker component in each of their  $\frac{1}{3}$ -octave frequency bands spanning 100Hz to 10kHz; later, the same method applied by Klein et al. (1970), Pols et al. (1973), and van Nierop et al. (1973) to the principal-components space of those filter-bank spectra, was found to be quite effective in speaker normalisation of Dutch vowel acoustic data. However, it would appear that further insights into the spectral regions or phonetic subspaces most affected by mean spectrum (or cepstrum) normalisation, have not been forthcoming in the literature.

In contrast to the above method which assumes that the speaker component of the acoustic speech signal is primarily contained in the overall mean spectral parameters, *frequency warping* (Matsumoto and Wakita, 1978) is a more general and potentially a more powerful method of speaker normalisation. By analogy with the popular method of dynamic time warping (DTW, see Sakoe and Chiba, 1978) used in ASR, dynamic frequency warping (DFW) is intended to optimally align two given spectra, with constraints on the warping path applied in a dynamic programming framework.

Paliwal and Ainsworth (1985) showed specific examples of the effectiveness of DFW in matching the LP-derived spectrum of a given vowel of one speaker, with that of another speaker. However, they also found that the same method could equally well match the spectra of two *different*, but neighbouring vowels. Not surprisingly therefore, the use of DFW in inter-speaker vowel classification using 4 male and 4 female speakers' data, yielded *worse* results on average.

A similar degradation in inter-speaker vowel classification performance was reported by Matsumoto and Wakita (1986) using vowel data recorded by 12 male and 12 female speakers of American English. However, DFW was found to *improve*

classification accuracy when the overall spectral slope was concurrently speaker-normalised. Indeed, Sejnoha and Mermelstein (1983) had also found that DFW applied to mel-frequency filter-bank vowel spectra of 7 male and 6 female speakers, improved both within- and across-gender, vowel classification accuracy, when applied together with a speaker-dependent compensation of overall spectral tilt. These results suggest that DFW should perhaps yield improved performance in vowel classification using the NDPS representation (as embodied in the QCEP), which inherently suppresses the influence of the overall spectral slope.

Regarding the type of DFW best suited to improving vowel classification accuracy, Sejnoha and Mermelstein (1983) found the “best frequency warp paths ... to be nonlinear and strongly dependent on the speech segment.” By contrast, Matsumoto and Wakita (1986) found that the largest contribution to improved accuracy was yielded by a *linear* frequency warp (where the frequency locations of spectral peaks or formants are effectively scaled by a common factor). However, acoustic-phonetic insights into the dichotomy were offered only by Matsumoto and Wakita, who reported that the (linear) frequency warp improved inter-speaker classification of mainly the mid- to high front vowels.

Insights of a similar kind were offered by Suomi (1984), who also adopted a whole-spectrum approach to speaker normalisation. Superimposing the gender-dependent mean, critical-band spectra (in perceptually-motivated frequency and loudness units of Bark and phons, respectively) of 8 vowels recorded by 8 male and 8 female speakers of Finnish, Suomi (1984, p.204) first observed that “the differences appear phoneme-specific, particularly in the upper part of the spectrum.” However, Suomi (1984, p.205) then obtained (cf. Bladon et al., 1983, as reviewed earlier in Section 2.2.2.2) “a close match ... in the upper bands by shifting the male spectra one Bark upwards and three phons downwards relative to the female spectra.” Following normalisation, the accuracy of classifying the female speakers’ vowels based on the male speakers’ training data was found to improve from 70.7 % to 88.1%. Prior to speaker normalisation, those gender differences had been found to detrimentally affect the classification accuracy of mainly the mid- to front vowels.

In view of these results, and indeed of the much earlier studies of Potter and

Steinberg (1950) and Peterson and Barney (1952), it is not at all surprising that Umesh et al. (1996) recently found that the optimum, non-linear frequency warp which would allow gender-normalisation of the warped spectra by only a linear translation, was in fact very similar to the perceptually-motivated, mel-frequency scale. Thus, in addition to its inherent emphasis of the low spectral range, the popular mel-frequency cepstrum is expected to hold the potential of facilitating normalisation of inter-gender differences in spoken vowels. However, it seems clear that spectral warping approaches are aimed at alleviating the effects of and not at explaining the speech-speaker dichotomy.

Nor is one further equipped to unfold the dichotomy problem, by turning to methods of speaker adaptation such as those advocated by Furui (1989; 1991; 1992). While they embody the appealing idea of incrementally improving recognition performance, and have therefore the potential of securing technological advances, it is fair to say that the problem is thus acknowledged but not explained. No more elucidating are those common approaches adopted in ASR, which simply rely on vast amounts of data in the hope of capturing sufficient variability amongst speakers.

## **2.4 Acoustic-Articulatory Evidence and Approaches**

We have reviewed in Sections 2.2 and 2.3, respectively, the relevant parts of the auditory-perceptual and acoustic-phonetic bodies of literature which bear on the problem of speech-speaker dichotomy. As intimated earlier (both in Section 2.1 and in Chapter 1), an even more fundamental perspective on the dichotomy problem is expected if the intertwined sources of variability are examined in the speech production domain. In that context, we first consider (in Section 2.4.1) previous attempts to theoretically formalise, and to empirically describe the articulatory sources of inter-speaker variability and their acoustic-phonetic consequences. We then review (in Section 2.4.2) the most persistent approaches to overcoming the dichotomy by articulatorily-motivated methods of speaker normalisation. We shall then argue (in Section 2.4.3) that if more complete articulatory insights are to be gained, then there is a need to advance beyond those persistent approaches, and to more inclusively account for the articulatory sources of speaker variability.

## 2.4.1 The Speaker Factor: Sources of Variability and Consequences

In contrast to the physical correlates of the *phonetic* quality of spoken vowels, which have long been the subject of investigation both directly by way of articulatory measurements and indirectly in terms of articulatory modelling which has helped to establish the basic relations between speech production and acoustics (e.g., Chiba and Kajiyama, 1958; Fant, 1960; Flanagan, 1972), the literature on the physical correlates of *speaker differences* in spoken vowels is considerably more sparse. Nevertheless, in the following two sections, we first review some prominent attempts to formalise the potential articulatory sources of speaker variability in a theoretical or a descriptive framework, and then we re-examine some of the most common, directly measured and acoustically inferred observations of speaker differences in the production of spoken vowels.

### 2.4.1.1 Theoretical Framework

As briefly mentioned in our introductory chapter, Garvin and Ladefoged (1963) qualitatively categorised the physical sources of inter-speaker variability in terms of *organic* and *learned* differences. Whilst organic differences refer to the size and shape of speakers' fixed vocal-tract anatomies, learned differences imply idiosyncracies in both the static and dynamic aspects of articulation during speech production. However, this bipartite categorisation is not a mutually exclusive one — as pointed out for example by Stevens (1971), the so-called learned characteristics may partly depend on the speaker's vocal-tract anatomical structure and the constraints which it imposes on both the static postures and the dynamic gestures of articulators.

Nolan (1983), who first points out that the literature on speaker recognition in particular had for some twenty years ventured no further than Garvin and Ladefoged's bipartite categorisation of speaker differences (an observation which is almost equally valid to this day), then himself proceeds to “reject” that framework on grounds of its oversimplification of the speaker variability problem. Nolan's criticism of the organic versus learned framework is founded partly on the non-trivial influence of so-called organic attributes, on both learned (speaker-related) and linguistic (or phonetic) aspects

of speech production; and partly also on the complexity of the so-called learned attributes, which themselves are conditioned by such factors as communicative intent. Ultimately, Nolan advances the notion that a speaker's fixed vocal-tract anatomy merely sets limits on the range of possible articulatory implementations available in the acoustic transmission of a given linguistic message; furthermore, that there should be made a more explicit distinction between the long-term qualities and the short-term or realisational rules, in describing the so-called learned speaker differences.

In fact, Laver (1980) had earlier expounded a detailed framework for the phonetic description of voice quality (including both laryngeal and supralaryngeal structures and articulators), which Nolan (1983) then used in his own investigations of the acoustic correlates of long-term segmental properties. Although Laver's descriptive framework is initially motivated by voice quality distinctions at an auditory-perceptual level, it is defined entirely in terms of the so-called *articulatory settings*, such as "raised-larynx voice", "labiodentalised voice", or "palatalised voice" (to name only three). Settings are thus described in relation to an overall "neutral" articulatory configuration and, whilst each setting is liable to exert different amounts of influence (both at an articulatory and an acoustic level) on different types of speech segments, it does afford an articulatory description of a speaker's long-term voice quality.

Unfortunately, it would appear that neither Laver's (1980) descriptive framework, nor Nolan's (1983) cautions against the inadequacies of the organic-learned framework, have had much direct influence in the literature concerned with the speech-speaker dichotomy. As the overall anatomical size of a speaker is perhaps the most clearly visible and easily accessible of individual characteristics, it can almost be expected that gross differences of the *organic* type should have dominated previous approaches to describing speaker variability in speech production. Indeed, this is borne out in the next section, in which we review the most common, empirical observations of the articulatory sources and acoustic consequences of speaker differences in the static, segmental properties of spoken vowels.

#### **2.4.1.2 Most Common Empirical Observations**

As noted above, perhaps the most obvious difference between speakers is in the overall

size of their vocal-tract. For example, it is well known that on average, the vocal-tract of an adult male is larger than that of an adult female, which in turn is larger than that of a child. The most acoustically significant, physical measurement relating to vocal-tract size, is the total length of the vocal-tract airway from the glottis to the lips (e.g. Fant, 1960). In this vein, if one assumes that different speakers adopt very similar articulatory configurations for a given vowel, and that the only source of variability is in the total length of their vocal-tract, then acoustic theory predicts that the corresponding formant frequencies will be scaled inversely to the ratio of the speakers' vocal-tract lengths. Indeed, the acoustic manifestations of vocal-tract length-related, *organic* differences between speaker *groups* (Garvin and Ladefoged, 1963), are the progressively higher, average vowel formant frequencies of men, women, and children, as portrayed for example in the classic study of Peterson and Barney (1952). In this vein, it has been shown that the formant frequencies of (presumed) phonetically equivalent vowels spoken by a wide range of speakers, can be modelled to a first approximation by lines of constant formant ratio (e.g. Peterson, 1952; 1961).

Nevertheless, the accuracy of that basic, linear relation in measured acoustic data, has generally been observed to be greater in regard to the so-called *individual* (Garvin and Ladefoged, 1963), organic speaker differences (e.g., differences between speakers of the same gender). For example, Stevens and House (1963) were able to show a reasonable agreement between the percentage differences in 3 adult, male speakers' mean vocal-tract lengths measured from mid-sagittal X-ray images, and the percentage differences in those speakers' first and second formant frequencies averaged over 8 vowels. Later, Stevens (1971, p.216) obtained a reasonably linear relation (as predicted by theory) between those same, three speakers' average vocal-tract lengths, and their mean values of  $F_3$ , arguing that

“... since the third formant does not change markedly from vowel to vowel, ... it provides a more precise indication of average vocal-tract length than does the first or second formant.”

Implicit in that statement is the greater speaker-specificity of the third formant, by virtue of its more direct relation to the vocal-tract length of the speaker.

By contrast, the relation between the vocal-tract lengths (and of the formants) of

speakers of different gender and age groups, has been shown to be more complicated. Despite some recent evidence to the contrary (e.g. Yang and Kasuya, 1995, 1996), mid-sagittal X-ray images have revealed that the lengths of the *oral* and *pharyngeal* parts of the vocal-tract (approximately the distance from the lips to the back wall of the pharynx, and from the soft palate to the glottis, respectively) are, in general, scaled differently from speaker to speaker, particularly across gender and age groups. For example, Chiba and Kajiyama (1958) found that, although the total length of the vocal-tract of an 8 year-old girl was 70% of the total vocal-tract length of a 26 year-old male, the young girl's oral and pharyngeal cavity lengths were 77% and 64% of the corresponding cavity lengths of the adult male. Similar, physiological findings were reported by Goldstein (1979) in her modelling study of the childhood and adolescent growth curves of vocal-tract anatomical distances. Furthermore, the acoustic consequences of inter-gender differences of that type were studied in detail by Fant (1966, 1975a), who used formant frequency measurements from speakers of eight different languages to show the importance of so-called *non-uniform* formant scaling, and to explain the non-uniformity in terms of female speakers' relatively smaller ratio of pharyngeal to oral cavity lengths, compared to adult males. Both the acoustic and articulatory evidence thus support Chiba and Kajiyama's (1958, p.193) general statement, that

“... the length of the mouth cavity does not vary so much with different individuals as that of the pharynx cavity.”

Not incompatible with that statement, is the long-held view that the speaker-specific dimensions of the so-called “larynx tube” which extends above the glottis, may provide a phonetically-independent, acoustic cue which is characteristic of the speaker. If such a structure, typically about 2 – 3 cm in length, were assumed to be acoustically uncoupled from the rest of the pharyngeal cavity (as is approximately the case when the relatively narrow larynx tube opens to a much wider pharynx), then it may indeed be responsible for the relatively fixed, higher formant observed (e.g. Ladefoged, 1993) in the vicinity of 3kHz in some speakers' spectrograms of vocalic speech sounds. This hypothesis was put forward more than sixty years ago by Bartholomew (1934), and has

since been revisited in describing potential articulatory sources of inter-speaker variability (Stevens, 1971), in providing reliable cues to speaker identification by spectrograms (Stevens et al., 1968), and in explaining the phenomenon of the “singer’s formant” (Sundberg, 1974, 1995). From an acoustic point of view, the phonetic independence and the potential speaker-specificity of the so-called larynx resonance is consistent with its typical frequency location being higher than the normal range of the first two formants. However, robust identification of such a fixed, higher formant, which admittedly has been more clearly observed in sung (rather than spoken) vowels, can generally be expected to be problematic.

The emphasis on speaker differences in vocal-tract fixed anatomy has recently been reaffirmed, from both an acoustic and an articulatory point of view. Owren and Bachorowski (1997) reported an accuracy of 99% in automatic gender identification using only the fundamental frequency and a formant-based estimate of the vocal-tract length, in the first (mid-front) vowel of the utterance “Test n Test” recorded by 50 male and 75 female speakers. On the other hand, direct, physical measurements were used by Honda et al. (1997) to study the effects of the size and shape of speakers’ orofacial structure on their production of spoken vowels. Whilst the “lower facial height or the size of the mandibular symphysis” was found to be inversely correlated with the first formant frequency, the speaker-dependent “spread of the second formant” was found to be correlated with the “oral cavity aspect ratio and mandibular symphysis shape”.

In sum, the most common empirical observations of the physical sources of inter-speaker variability and their acoustic manifestations in the static, segmental properties of spoken vowels, clearly emphasise the importance of vocal-tract fixed anatomical structures. A relevant question then arises whether previous, articulatorily-motivated approaches to overcoming the speech-speaker dichotomy by speaker normalisation, also bear the influence of this persistent emphasis on speaker differences in vocal-tract fixed structure. Our review of the relevant literature in the following section will shed some light on this issue.

#### **2.4.2 Persistent Approaches to Speaker Normalisation**

Analogously to the acoustic-auditory and acoustic-phonetic approaches to overcoming

the speech-speaker dichotomy (as reviewed in Sections 2.2.2 and 2.3.3, respectively), the speaker normalisation problem has also been approached from an *acoustic-articulatory* point of view. Similarly to those approaches reviewed earlier, a primary objective of importing acoustic-articulatory principles in speaker normalisation has been to overcome the detrimental influences of inter-speaker variability, and thereby gain higher accuracies in speaker-independent, automatic speech recognition. Previous methods which have aimed to account for articulatory sources of speaker differences in speech recognition, have either explicitly estimated certain articulatory parameters from acoustic measurements, or only implicitly accounted for the acoustic consequences of speaker-induced variations in those articulatory parameters.

One of the earliest of the explicit approaches is the study by Wakita (1977), who used the linear-prediction (LP) model to first estimate the instantaneous vocal-tract length (VTL) for any given frame of vocalic speech data (the method will be described in Chapter 5). Both the frequencies and bandwidths of the LP poles corresponding to the formants of the measured frame of speech, were then scaled by the ratio of the estimated VTL and a fixed, reference length of 17cm (a typical length of an adult male's vocal-tract). Whilst normalisation of both the speaker *and* phonetic variation in VTL led to an overall shrinkage of the formant space of the 9 vowels of the 14 male and 12 female speakers of American English used in that study, the VTL-related speaker normalisation of the first three formant frequencies yielded a nearly consistent increase in the ratio of inter- to intra-vowel variance between pairs of neighbouring vowels, and led to an improvement from 78.9 % to 84.4 % in inter-speaker vowel classification accuracy using a Bayes' classifier.

One of the advantages of Wakita's (1977) method of speaker normalisation, particularly in the context of recognition, is that it makes no *a priori* assumption about the phonetic identity of the analysis frame. However, if the phonetic identity can be assumed to be known, a more physically meaningful approach (Ishizaki, 1978a; Broad and Wakita, 1978) is to first estimate the VTL of all the vowels of each speaker, then normalise each speaker's *average* VTL to a fixed reference value. This approach retains each speaker's inter-vowel distribution of vocal-tract length, and thereby avoids the potentially harmful (and unrealistic) shrinkage of the entire vowel formant space as

noted by Wakita (1977). It is also interesting to note that this method of speaker normalisation is similar to Pals et al.'s (1973) translation of speakers' logarithmic formant frequencies to a common mean value (as reviewed earlier in Section 2.3.3.2), if that translation were constrained to equal intervals along each dimension.

Indeed, uniform speaker normalisation of mean vocal-tract length entails a linear expansion or contraction of each speaker's vocal-tract *area-functions*<sup>2</sup> by a single, speaker-dependent scaling factor. Broad and Wakita (1978, p.85) computed the scaling factor for each of their 26 speakers, using all 9 vowels available in a reference dataset, and noting only the phonetic equivalence of the "canonical allophones" of those vowel phonemes to the corresponding vowels in each speaker's data. On the other hand, Ishizaki (1978a, p.67) recognised the potential speaker-specificity of "lip protrusion and the up-and-down motion of the larynx" in certain rounded vowels, and therefore used only the three unrounded vowels /i/, /e/, and /a/ to obtain a more realistic, and perhaps a more purely organic or anatomically-based estimate of the differences between 8 adult, male speakers of Japanese.

In contrast to the explicit estimation of VTL prior to speaker normalisation, a number of studies have shown that improvements in speaker-independent speech recognition accuracy can be gained by allowing certain, vocal-tract length-related transformations of the acoustic (usually whole-spectrum) parameters. For example, Golibersuch (1983) used per-speaker histograms of the first three formant frequencies (measured from about 40sec of speech recorded by each speaker), to determine the optimum *linear* frequency warp between pairs of speakers' filter-bank spectra (spanning centre-frequencies from 350Hz to 3400Hz, on a logarithmic scale). In a speaker-independent, isolated word recognition task using each of 4 male speakers' data in turn for training, and 10 other male and 6 female speakers for testing, the linear frequency warp was found to improve recognition accuracy from about 32% to 42%, on average; as the training speakers were all male, the largest improvements in accuracy were obtained for the female speakers. In this vein, Matsumoto and Wakita (1986) also reported improvements in inter-gender vowel recognition accuracy (cf. Section 2.3.3.2)

---

<sup>2</sup> The profile of cross-sectional area as a function of the distance along the vocal-tract airway from the glottis to the terminating plane at the lips.

using linear frequency warping. However, they avoided the error-prone task of formant estimation, by using a suitably constrained dynamic programming framework to determine the optimum, linear frequency warp path; in particular, their frequency warping function was constrained to allow variations in vocal-tract length between 13cm and 21cm.

More recent approaches to articulatorily-motivated speaker normalisation, have attempted to obtain higher accuracies in speaker-independent speech recognition by allowing for VTL-related transformations of the cepstrum parameters. For example, Claes et al. (1997) derived a transformation of the mel-frequency cepstrum which linearly warps the original frequency axis, based on speaker-group differences in the median of the third formant frequency. Thus implicitly accounting for VTL-related differences between adult and child speakers, they obtained a reduction in the word error rate yielded by a continuous-density HMM-based, continuous word recogniser, from 4.3% to 1.9%.

On the general assumption (similar to that of Matsumoto and Wakita, 1986) that the formant parameters themselves cannot be reliably estimated, Lin and Che (1995), and very recently Lee and Rose (1998), proposed to maximise speech recognition accuracy by performing recognition on several, differently computed cepstrum vectors, based on a finite number of degrees of linear expansion or compression of FFT-based spectra. Lin and Che (1995) reported a monotonic improvement in isolated word recognition accuracy using 18 mel-frequency cepstral coefficients in a discrete-density HMM recogniser trained on the data of 20 male speakers, when the effective half-sampling frequency of the test spectra of 10 female speakers was increased in three steps, from the original 6.0 kHz up to 7.8 kHz (thus compressing those spectra, and effectively lengthening the female speakers' vocal-tract lengths by a factor of 1.3). They also obtained a monotonic improvement in recognition accuracy when the recogniser was trained on 20 female speakers' data and tested on 10 male speakers' data, as the effective half-sampling frequency of the latter spectra was decreased in two steps, from 7.2 kHz down to 6.0 kHz (thus expanding those spectra, and effectively shortening the male speakers' vocal-tract lengths by a factor of about 0.83). In this latter case, Lin and Che (1995, p.203) pertinently noted that "use of less spectral

information can lead to a higher word recognition accuracy...”, even — one should note — with a powerful HMM recogniser trained on presumably sufficient amounts of data.

In this vein, Lee and Rose (1998) also reported improved performance in an HMM-based, speaker-independent (mixed gender) speech recognition task, using a continuous-digits dataset recorded over the telephone. Mel-frequency cepstra of 12th order were computed from spectra which were originally represented by 22 mel-spaced filter-bands up to about 4kHz, but subsequently allowed to be compressed or expanded along the frequency axis, with 13, equally-spaced linear warping factors between 0.88 and 1.12. The speaker-dependent warping factors were optimised both in training and testing the recogniser, and the average factor for the male and female speaker groups was found to be 1.00 and 0.94, respectively.

### **2.4.3 Speaker Normalisation Beyond Vocal-Tract Length**

Clearly, the literature on articulatorily-motivated speaker normalisation in speech recognition, is dominated by either direct or indirect approaches to account for speaker differences in vocal-tract length. This is perhaps not surprising, given that the so-called *organic* differences between speaker *groups* such as adult males, adult females, and children, are the most significant and the most readily interpretable, both physiologically and acoustically. However, the overriding emphasis on vocal-tract length has impeded progress towards a better understanding of the so-called *learned* differences between speakers which, as we shall see in the next section, is far from insignificant.

#### **2.4.3.1 Under-exploited Evidence**

Indeed, the existing theoretical frameworks for describing speaker differences in vowel production (as reviewed in Section 2.4.1.1) already claim that a minimal description of those differences should include both vocal-tract anatomy and articulatory behaviour. Amongst the studies which provide empirical evidence to support this view, are those which are concerned with elucidating the relative significance of uniform and non-uniform scaling of vocal-tract length. For example, in advocating non-uniform scaling between adult male and female speakers, Fant (1975 a, p.13) claims that it can *halve* the speaker-group variance remaining in the first three formant frequencies after uniform

length scaling; thus, “it can be expected that non-uniform normalization brings out dialectal differences between speakers more clearly than uniform scaling.” Similarly, Nordström (1977, p.91) concludes from his vocal-tract modelling study of male-female-child differences, that “*anatomical differences ... only explain part of the formant differences*”, and that it is “probable that the vocal tract form varies between men and women/children.” The speaker differences which Fant grossly categorises as “dialectal”, and which Nordström refers to as “vocal tract form”, are indeed the so-called learned differences in articulatory behaviour, the acoustic consequences of which can be at least as significant as those induced by anatomical or organic differences.

In this vein, some of the acoustic implications of learned speaker differences were noted by Saito and Itakura (1983, p.151), whose contribution to elucidating speaker individuality in vowel formant data was reviewed earlier in Section 2.3.1.2. In particular, they noted that

“...the normalization of frequency spectral differences by the use of the ratios between the formant frequencies ..., or corrections based on the vocal tract length ..., will be insufficient, because variations in the ratio of formant frequencies are not always in a radial direction, but in a circular one...”

Similarly, in his bi-planar modelling of the first three formant frequencies of 30 vowels recorded by each of 3 male and 3 female, adult (and phonetically-trained) speakers of American English, Broad (1981, p.1428) concluded thus:

“The significant inter-speaker variations in the orientations of the planes demonstrates a statistically detectable departure from uniform scaling.”

Applying the same, bi-planar model to the formants of 5 vowels recorded by 4 adult, male speakers of Japanese, Sato et al. (1982) made similar observations, and arrived at a very similar conclusion (as one might have predicted, given that differences in vocal-tract length are perhaps of secondary importance between speakers of the same gender and age group). Analogously in terms of a whole-spectrum representation, we have already reviewed (in Section 2.3.3.2) those studies in which *non-linear* (rather than linear) frequency warping was found to more completely normalise the spectral differences between speakers (e.g., Sejnoha and Mermelstein, 1983; Paliwal and Ainsworth, 1985; Matsumoto and Wakita, 1986).

All of this evidence, which arises from acoustic and vocal-tract modelling studies, clearly points *beyond* vocal-tract length (or more generally, vocal-tract anatomical size) as an articulatory source of speaker differences. In this vein, one wonders whether direct articulatory measurements have at all been used to shed light on the *learned* types of speaker differences which lie beyond vocal-tract length. Admittedly, this question itself presupposes a methodology for comparing articulatory data of different speakers (who in general have different vocal-tract anatomical sizes and shapes), such as to bring out those vocal-tract *shape*-related differences which might be attributed to articulatory behaviour rather than anatomy.

One particularly insightful approach to inter-speaker comparison of vocal-tract shapes was devised by Liljencrants (1971), who computed inter-speaker correlations between corresponding pairs of Fourier coefficients used to model two adult, male speakers' tongue shape profiles during vowel production. The profiles themselves were obtained from mid-sagittal X-ray images, by sampling at intervals of about 5 mm along the length of the vocal-tract airway, the distance to the tongue contour with respect to a *speaker-dependent* coordinate system. Although only the "DC term" and the first two Fourier components were extracted by Liljencrants, he found that the speakers' tongue shapes were highly correlated in their first Fourier components ( $r = 0.895$  and  $0.957$  for the cosine and sine terms, respectively), and far less correlated in their second Fourier components ( $r = 0.365$  and  $0.637$ , respectively). Thus, Liljencrants (1971, p.16) concluded that whilst the "gross shapes are similar for the two subjects uttering similar vowels", there appears to be "a more subject dependent fine structure." Apart from Mermelstein's (1967, p.1292) qualitative assertions regarding the phonetic importance of the lower formants and therefore of the low-frequency components of vocal-tract shapes, this is perhaps the first and only, direct evidence to support Peterson's (1959) general statement quoted earlier (in Section 2.1.1), which itself also implicates the speaker-specificity of the higher formants.

Empirical observations of speaker differences in directly measured vocal-tract shapes (whether mid-sagittal traces or area-functions obtained therefrom), are indeed scarce in the literature. Another notable exception is Harshman et al.'s (1977) three-way factor analysis (so-called PARAFAC) of tongue shapes obtained from mid-sagittal

tracings of 10 vowels produced by 5 speakers of American English. A distinct advantage of the PARAFAC analysis adopted in that study, is its explicit modelling of both vowel and speaker-related statistical variations along each tongue contour; furthermore, as opposed to principal components analysis, the factors yielded by PARAFAC are not constrained to be orthogonal. Regarding the principal dimensions of vowel variation, two factors were found which were then described, respectively, as “front raising” and “back raising” of the tongue (Ladefoged et al., 1977, p.21); the vowel distribution on the corresponding factor plane was found to be very similar to the traditional vowel chart (albeit slightly rotated), with opposing and extreme vowel configurations of /o/ and /i/, and of /ɑ/ and /u/, respectively, along each dimension. Although the PARAFAC analysis also yielded factors describing the speaker variation in the tongue-shape data, Harshman et al. (1977) did not offer new insights from a speaker-related point of view, other than to point out a correlation between the first speaker factor and the speakers’ oral cavity lengths, and between the second speaker factor and the ratio of the speakers’ pharynx to oral cavity lengths. Thus, although data on detailed tongue shapes were available, those authors did not venture beyond considerations of gross, anatomical speaker differences, which indeed appear to have been the dominant influence in the literature.

In line with the notion of a more speaker-specific, pharyngeal cavity length, Hermansky and Broad’s (1989) hypothesis represents a significant refinement of the structurally-motivated partitioning of the vocal-tract into oral and pharyngeal sections. In particular, they advanced that the *phonetic* and the *speaker-specific* attributes of vocal-tract articulatory configurations can be associated, respectively, with the *front* and the *back* cavities which are divided by the main place of lingual constriction. Direct evidence to support this hypothesis was shown in tracings of mid-sagittal X-ray images of vocal-tract vowel configurations of an adult male and a child, which revealed that although the child’s vocal-tract was much smaller and had a proportionally shorter pharynx, the size and shape of the child’s vocal-tract “front cavity” (anterior to the main place of lingual constriction) was comparable to that of the adult male speaker, both for a front and a back vowel. The acoustic consequence of this hypothesis was demonstrated by showing that the second spectral prominence of the 5th-order PLP

model (Hermansky, 1990) is highly correlated with the resonance of the front cavity (assumed to be decoupled from the back cavity) of a vocal-tract model. Hence, the phonetic salience of the two-peaked PLP spectrum (as discussed earlier in Section 2.2.2) was re-interpreted articulatorily, in terms of its ability to capture the effective second formant  $F_2'$ , or the phonetically important *front cavity resonance*. However, it remains to be tested whether their hypothesis of speaker differences being manifest mainly in the shape and size of the vocal-tract cavity behind the main place of lingual constriction, is equally tenable for speakers having similar vocal-tract anatomical sizes and proportions (e.g., for speakers of the same gender and age group).

Two further exceptions to the dominant influence of vocal-tract structure, include the magnetic resonance imaging (MRI) studies of Baer et al. (1991) and of Narayanan et al. (1997), who superimposed the MRI-measured, vocal-tract area-functions for each of 4 point vowels (/i/, /æ/, /ɑ/ and /u/) sustained by two male speakers, and for each of 3 point vowels (/i/, /a/ and /u/) sustained by two male and two female speakers, respectively. Regarding speaker differences, Baer et al. (1991, p.812) noted the largest amount of variability in the vowel /æ/, for which one speaker's area-function was "generally wider" and showed "more peaks and dips" than that of the other speaker; the lip opening of one speaker was also systematically found to be larger than that of the other speaker. However, whilst one speaker's entire area-function for /u/ was found to be "shifted to the left" with respect to that of the other speaker, this difference was cautiously attributed to a systematic measurement error in the position of the pharyngeal starting point, at which all area-functions were aligned. By contrast, Narayanan et al. (1997) aligned their MRI-measured area-functions of two male and two female speakers at the lips, thereby allowing the position of the glottal end to vary on each graph, according to the inter-speaker differences in vocal-tract length for each vowel. Speaker differences were then qualitatively noted in the anterior region (or front cavity) for the vowels /u/ and /a/, and in the back cavity for the vowels /u/ and /i/. These observations led Narayanan et al. (1997, p.1008) to suggest that speaker differences are most pronounced in vocal-tract regions of large cross-sectional area (i.e., in vocal-tract cavities, rather than at places of constriction), since in those regions, "individual anatomical differences are allowed to show prominently."

However, perhaps the most detailed, articulatory study of speaker differences, was recently conducted by Högberg (1995), who compared the vocal-tract area-functions of 10 non-nasalised vowels obtained from X-ray images of one adult male and one adult female speaker of French. A particularly noteworthy, methodological advance relative to previous studies, which have clearly stumbled on the problem of how best to align and thereby quantitatively compare area-functions or vocal-tract shapes of different speakers, was Högberg's assumption that at least on a per-vowel basis, the purely anatomical differences between the two speakers could be normalised by separately scaling the lengths of the oral and pharyngeal parts of the vocal-tract. Although Högberg did acknowledge the potential contribution of lip protrusion and larynx height to speaker differences in vocal-tract length (similarly to Ishizaki, 1978a, as reviewed earlier in Section 2.4.1.2), the two speakers were found to exhibit similar degrees of lip rounding and larynx raising from vowel to vowel. Thus having scaled down the lengths of the male speaker's area-functions (by an overall average of 17%), Högberg (1995, p.45) found the largest amount of speaker differences in "articulation strategies", for the vowels "of the most neutral character", namely /ɛ/, /œ/, and /ø/. In addition, he found that the female speaker "used more extreme articulations" for the vowels /i/ and /a/. Averaged over all 10 vowels, the male-to-female ratios of vocal-tract areas were found to be largest in the lower pharynx, and at the lips. Also somewhat consistent with Hermansky and Broad's (1989) hypothesis reviewed earlier, Högberg (1995, p.41) found the largest inter-vowel variation of those ratios in the oral cavity, thus "implying that the differences in the mouth are quite vowel dependent."

Whilst the studies reviewed above do provide some direct articulatory evidence which underscores the importance of speaker differences beyond vocal-tract length, they also attest to the difficulties of quantitatively comparing articulatory data across speakers in a meaningful way. For example, three different methods of accounting for organic differences prior to direct comparisons of vocal-tract shapes across speakers, are the anatomically-based, speaker-dependent, mid-sagittal reference grid used by Harshman et al. (1977); the uniform vocal-tract length normalisation effectively carried out by Liljencrants (1971); and the inter-gender, non-uniform length normalisation used by Högberg (1995). On the other hand, the organic versus learned distinction has been

completely ignored in those studies where the original vocal-tract area-functions of different speakers are simply aligned either at the glottis (Baer et al., 1991) or at the lips (Narayanan et al., 1997). Indeed, despite the potential guidance of existing theoretical frameworks (as reviewed in Section 2.4.1.1), there does not appear to be a consistent or a generally accepted method of separately accounting for the so-called learned, articulatory sources of speaker variability. It is therefore conceivable that the exploitation of such direct articulatory evidence as reviewed above, has been partly inhibited by the very absence of an accepted methodology for adequately describing and accounting for the various sources of speaker differences in vowel production.

In this vein, the valiant efforts of Payan and Perrier (1993) and of Perrier et al. (1995) to incorporate acoustic-articulatory knowledge in speaker normalisation of vowel formant data, are to be applauded. Their proposed method is founded on the acoustic-articulatory principle that in certain vowels for which the vocal-tract cavities can be assumed to be acoustically, nearly uncoupled, there exist strong formant-cavity affiliations. Using Fant's (1960) well-known, model-based formant nomograms, Perrier and his colleagues first infer the formant-cavity affiliations for a given speaker's recordings of the point vowels /i/, /a/, /u/, and various vocalic transitions such as /iy/, /iu/, and /ua/, during which sudden changes in formant-cavity affiliations are known to occur. They then obtain so-called *normalisation coefficients* which map, on a per-vowel basis, the effective lengths of the speaker's vocal-tract cavities (determined by the measured frequencies of the formants which are affiliated with those cavities) to those of a standardised, mid-sagittal articulatory model (Maeda, 1979; 1988). Having thus computed the normalisation coefficients for the point vowels, those of intermediate vowels are found by interpolation, and the entire, vowel formant space of the speaker can then be mapped onto that of the articulatory model. Thus, rather than attempt to separate the organic and the learned components of vowel production, Payan and Perrier (1993, p.418) acknowledge that

“...the consequences of anatomical and gestural differences on the size of back and front cavities could be variable, depending on the location of the articulation in the vocal-tract.”

In view of the acoustic-articulatory knowledge involved in their method of speaker

normalisation, it is surprising that Perrier et al. (1995) then found it to perform only slightly better on  $F_2$ , and actually worse on  $F_1$  of measured vowel data, compared with the mainly heuristic, acoustic-phonetic methods reviewed earlier in Section 2.3.3.2 (i.e., Gerstman, 1968; Lobanov, 1971; Di Benedetto and Liénard, 1992). On the other hand, in view of the significant amount of literature reviewed in this chapter which points to the speaker-specificity of the third and higher formants, it is also surprising that Perrier et al. (1995) did not evaluate their method of speaker normalisation on  $F_3$ , which nevertheless played a crucial role in determining formant-cavity affiliations. Furthermore, whilst their method of normalisation is significantly more informed than previous approaches, the large amount of expert knowledge (and often subjective judgement) which is required in order to infer formant-cavity affiliations from measured formants alone, may perhaps be overly prohibitive in applying the method to more extensive datasets.

As aptly stated by Payan and Perrier (1993, p.417), a method of speaker normalisation “is efficient if it can account for the causes of variability.” However, the simplicity of this statement belies the lack of cohesion and incompleteness of previous approaches to overcome the speech-speaker dichotomy by speaker normalisation, whether perceptually motivated, statistically-based, or vocal-tract length-related. In this vein, it would seem that a more inclusive approach to accounting for the physical sources of speaker variability, must involve both vocal-tract *length*- and *shape*-related articulatory parameters, and must be guided by a sufficiently complete, theoretical framework for describing speaker differences in the speech production domain.

#### **2.4.3.2 Articulatory Parameterisation and the Inverse Problem**

If a more complete description of the vocal-tract shape is to be used either to overcome the speech-speaker dichotomy in speech recognition, or more generally to define physical correlates of speaker differences in an acoustic dataset of spoken language, then direct articulatory measurements, whether by non-intrusive methods such as X-ray, MRI, or ultra-sound imaging, or by more intrusive methods such as electro-magnetic mid-sagittal articulometry, are clearly impractical; nor are those methods immune to various sources of measurement error. For example, the mapping from two-dimensional

(usually mid-sagittal) images to a more complete, three-dimensional representation (usually the area-function) is still an active research problem; the three-dimensional reconstruction afforded by MRI still requires the speaker to sustain the speech sound for many seconds; magnetic tracking of small pellets attached to the tongue or the lips can only provide localised information on specific articulators.

A more generally applicable, and certainly a more practical approach, would be to estimate articulatory parameters from the acoustics of speech. However, acoustic-to-articulatory mapping (or the so-called speech *inverse problem*) is itself an unresolved and controversial area of research (e.g. Schroeter and Sondhi, 1994) — even for the subset of vocalic sounds, the acoustic and articulatory properties of which are relatively better understood. Indeed, the most significant problem in estimating articulatory parameters from acoustic parameters, is that in general, the inverse mapping is *non-unique* (as epitomized for example by the ventriloquist, who is able to articulate speech sounds without the normal movements of the lips and the jaw).

Some limited conclusions on the extent of the non-uniqueness problem can be drawn from studies of human articulatory activity during speech production. A number of such studies (e.g., Lindblom et al., 1979; Gay et al., 1981) have shown that when speakers are physically inhibited from assuming their normal articulatory configuration for a given vowel (e.g., by inhibiting either lip or jaw movement), they are able to compensate for the acoustic perturbation, by recruiting the help of other articulators (e.g. the tongue). In particular, Gay et al. (1981) inferred from X-ray images, that when speakers perform such *articulatory compensation* in unnaturally perturbed conditions, they attempt to attain a vocal-tract *area-function* which is appropriate for the given vowel (especially at the acoustically more sensitive place of maximum constriction).

The phenomenon of articulatory compensation has also been observed in normal (or uninhibited) speech. For example, Hughes and Abbs (1976) found that in repetitions of the vowels /i/, /æ/ and /ε/ in /hVbVb/ context, a speaker achieves a nearly invariant degree of vertical lip opening near the centre of each acoustic vowel nucleus, but with highly varying, vertical positions of the lower lip, the jaw, and the upper lip. Those three articulators were thereby shown to exhibit “motor equivalence”, insofar as their individual “target” positions are subordinate to the more acoustically important, vocal-

tract “goal” of an appropriate lip aperture for the given vowel. Such “trading relations” between articulators were also observed by Maeda (1991), who first used his own articulatory model (Maeda, 1979) to determine the components of the mid-sagittal profile pertaining to the positions of the tongue-dorsum and the jaw, at the “target” of the vowels /i/ and /a/ recorded in various consonantal contexts by two female speakers of French. He then found, for each of those vowels, a positive correlation between tongue fronting (or backing) and jaw opening (or closing); and determined that by subtracting the articulatory variation due to this correlation, it was possible to reduce the variability in those articulatory parameters to a degree which is comparable to that of the first two formants measured in the corresponding acoustic data.

Such findings of compensatory articulation support the view that a more acoustically-relevant representation is afforded by the entire vocal-tract *area-function*, than by a model of the mid-sagittal profile which includes the positions of individual articulators. In addition to the potential exacerbation of the non-uniqueness problem, such mid-sagittal articulatory models also suffer the plight of *speaker-dependence*. Indeed, as rightly stated by Hogden et al. (1996, p.1821):

“...many of the ... vocal tract models used to study the inverse mapping problem are essentially models of a single speaker.”

Nowhere is this more obvious than in mid-sagittal models (e.g., Mermelstein, 1973; Coker, 1976; Sorokin, 1992) which necessarily include both moveable articulators (e.g., the tongue and the lips) and the vocal-tract hard-structure (e.g., the outline of the palate and the rear pharyngeal wall). In this vein, there have been some recent attempts to normalise certain anatomical dimensions of mid-sagittal articulatory models to those of a particular speaker, prior to using that speaker’s acoustic data in the inverse mapping (e.g., McGowan, 1997; Mathieu and Laprie, 1997). Whilst this procedure may be effective in decoupling the so-called organic and learned components of vocal-tract shapes, it still relies partly on the availability of a mid-sagittal X-ray or MR image of each speaker, without which the possibilities of articulatory compensation (in this case, between moveable articulators and the vocal-tract hard-structure) might potentially undermine the uniqueness of the normalisation itself. On the other hand, explicit

modelling of the vocal-tract area-function can provide a *speaker-independent* framework for acoustic-to-articulatory mapping, while still being reasonably amenable to the decoupling of anatomical and behavioural characteristics.

Nevertheless, some studies of human articulatory behaviour imply that even the area-function is not immune to the non-uniqueness problem. For example, Perkell et al. (1993) found that some speakers exhibit a (negative) correlation between tongue-body raising and lip rounding in the vowel /u/ recorded in a wide range of phonetic contexts. Owing to the relative independence of those two articulators, and their spatially distal influence on the shape of the vocal-tract airway, variations in the measured positions of the tongue and lip pellets were assumed to imply changes in the vocal-tract area-function; however, the degree of acoustic compensation was determined using only  $F_2$ . By contrast, Savariaux et al. (1995) offered a more complete set of formant measurements in their acoustic and articulatory study of 11 speakers' lingual compensations to an inhibiting lip tube, in the production of the vowel /u/. Six of the speakers were found to partially compensate for the unnaturally wide lip-opening, and one speaker was found to compensate nearly completely, by retracting the tongue body towards a more velo-pharyngeal (than velo-palatal) constriction. Consistent with the phonetic importance of the low formant space (as reviewed earlier in this chapter), Savariaux et al. (1995) used only  $F_1$  and  $F_2$  as a measure of their speakers' success in acoustic compensation; by contrast, their speakers were generally not as successful in compensating for the effect of the inhibiting lip tube on  $F_3$ .

The human-based studies of articulatory behaviour reviewed above, together substantiate the *reality* of the non-uniqueness problem. However, model-based experiments afford more precise control on the variability of articulatory and acoustic parameters, and potentially more direct insights into the fundamental nature of the non-uniqueness. In this vein, one of the most celebrated studies of the non-uniqueness problem is that of Atal et al. (1978), who used one of the most "realistic" vocal-tract acoustic models then available (i.e., one which includes acoustic energy losses arising from viscosity, heat conduction, radiation at the lips, yielding vocal-tract walls, and a mean glottal resistance to airflow). Their well-known illustrations of so-called "fibers" in articulatory parameter space, have since been repeatedly cited as evidence that

essentially the same set of acoustic parameters can be generated using a realistic vocal-tract model, with a continuously-varying combination of articulatory parameter-values. However, in describing vocal-tract area-functions either in terms of their individual section-areas, or even in terms of a more economical parameterisation adapted from Stevens and House's (1955) three-parameter model, Atal et al. (1978) used a greater number of articulatory than acoustic parameters; consequently, as they had earlier shown mathematically, non-uniqueness is guaranteed by the inevitable creation of a "null-space". For example, Atal et al. (1978, p.1547) note that "the mouth opening for vowel /i/ can vary considerably without affecting the formant frequencies", and they indeed give evidence that the first three formant frequencies (and also the first two formant bandwidths) are relatively invariant for the wide range of mouth openings considered (approximately  $0.3\text{cm}^2$  to  $12\text{cm}^2$ ). However, their evidence also clearly shows that for the same range of articulatory parameter variation, the fourth formant frequency varies from 3973Hz down to 3672Hz, and the third formant bandwidth varies from 102Hz to a completely unrealistic 2439Hz (cf. Atal et al., 1978, Figure 13, p.1548); furthermore, it was found that such "differences in the formant bandwidths and the higher formant frequencies were perceptible to expert listeners." These observations underscore the importance not only of constraining model-generated vocal-tract shapes within realistic ranges of variation (as might appear to be more directly possible using a mid-sagittal articulatory model), but also of ensuring that the non-uniqueness problem is not unnecessarily exacerbated by inappropriate articulatory parameterisation which could lead to *under-specification* of the acoustic-to-articulatory mapping.

However, in addition to the issue of articulatory parameterisation, the literature does also highlight the importance of the type of acoustic vocal-tract model. Indeed, it has been repeatedly shown that, regardless of the articulatory parameterisation, the non-uniqueness problem is almost inevitable when using a fully lossy, so-called "realistic" vocal-tract acoustic model. For example, contrary to the methods adopted by Atal et al. (1978), Charpentier (1984) intentionally *over-specified* the inverse mapping by using a greater number of acoustic parameters (the first five formant frequencies and the first three formant amplitudes) than articulatory parameters (the six parameters of Flanagan et al.'s (1980) area-function model). Yet, Charpentier's (1984) detailed analyses, using

a vocal-tract acoustic model which included all standard losses except at the glottis, clearly showed the rampant non-linearities and potential ambiguities of the inverse mapping. Approaching the problem from a more basic angle, Fant (1980) considered a single-resonance, single-section acoustic tube, and showed that whilst the length of the tube could be determined by the frequency of the single resonance (using the well-known quarter-wavelength formula), its cross-sectional area was a *non-unique* function of the resonance bandwidth, owing to the complementary effects of lip radiation loss and internal surface losses included in the model.

Partly as a result of the potential multiplicity of positions of individual articulators in yielding a nearly invariant area-function (as implied by certain types of articulatory compensation), and perhaps more fundamentally owing to the intrinsic non-uniqueness of so-called realistic articulatory and vocal-tract acoustic models, stringent constraints have had to be applied in acoustic-to-articulatory mapping (e.g., McGowan, 1994; Gupta and Schroeter, 1993; Sorokin, 1992; Schroeter and Sondhi, 1992; Parthasarathy and Coker, 1992). Indeed, in order to combat the theoretical and practical inevitability of non-uniqueness, previous methods have relied on pre-determined codebooks based on vector quantisation of articulatory space (Larar et al., 1988); static and dynamic constraints on articulatory parameters imposed in a dynamic programming framework (Schroeter and Sondhi, 1989); pitch-synchronous analysis to increase the reliability of measured acoustic parameters, and to jointly estimate the parameters of a glottal model (Gupta and Schroeter, 1993); some acoustic-phonetic knowledge in order to ensure a reasonable, initial articulatory configuration, thereby avoiding local optima in an analysis-by-synthesis framework (Lin and Fant, 1989); and linear interpolation between pre-determined (logarithmic) area-functions of three extreme articulatory configurations of a given speaker (Butler and Wakita, 1982).

Clearly, both human-based and model-based studies imply that non-uniqueness is *inherent* to the “realistic” vocal-tract<sup>3</sup>. However, non-uniqueness at the acoustic-to-area-function level has also been shown to exist when the vocal-tract is treated as a

---

<sup>3</sup> Indeed, a further complicating factor is the potential non-uniqueness believed to be caused by compensatory relations between the supralaryngeal vocal-tract and the glottis, especially those which might affect the formant bandwidths (e.g. Schroeter and Sondhi, 1994).

completely *lossless* acoustic tube. For example, in their pioneering investigations of the acoustic properties of the so-called three-parameter model (which specifies the place and the degree of lingual constriction, and the degree of lip opening), Stevens and House (1955) noted the existence of two separate articulatory configurations yielding the same set of first three formant frequencies for the vowel /u/. Not surprisingly, those two configurations are similar to the human-based, articulatory compensation strategies observed more recently by Perkell et al. (1993) and Savariaux et al. (1995).

However, perhaps the most general formulation of the non-uniqueness properties of a completely lossless vocal-tract model, was derived by Schroeder (1967) and Mermelstein (1967). Briefly, those authors found that if the logarithmic area-function is modelled in terms of the coefficients of the *cosine*-series, then for reasonably small perturbations about a uniform area-function, a change in each *odd*-indexed coefficient will induce a proportional change in a unique formant frequency; by contrast, perturbation of any of the *even*-indexed coefficients will leave the formants unaffected. This fundamental result provides the first, direct link between acoustic and vocal-tract shape parameters. Equally importantly, it provides a theoretically-grounded explanation of *the non-uniqueness which is inherent to any, completely lossless model of the vocal-tract* — in particular, *symmetric* perturbations of an area-function (where the axis of symmetry is at the mid-length of the area-function) are *acoustically inconsequential*; furthermore, there exists an infinite range of such, acoustically-inconsequential vocal-tract shape-perturbations, at every possible value of the entire vocal-tract *length*.

To summarise, the literature clearly indicates that acoustic-to-articulatory mapping with either a fully lossy (i.e., a “realistic”) or a completely lossless vocal-tract acoustic model, is prone to non-uniqueness; moreover, this inherent non-uniqueness is prone to exacerbation by inappropriate choice of articulatory parameters. Assuming that the entire shape of the vocal-tract (as embodied in the area-function) can be parameterised in an acoustically relevant way, the question then arises whether there exists any particular combination of vocal-tract losses which might render the non-uniqueness problem somewhat less problematic. In fact, it is well known (Atal, 1970; Wakita, 1973) that for the linear-prediction (LP) vocal-tract model which has only a single

source of energy dissipation (either at the glottis or at the lips), uniqueness<sup>4</sup> is theoretically guaranteed!

If, as implied at the start of this section, a more complete description of the length and shape of the vocal-tract can be expected to yield more informed methods of overcoming the speech-speaker dichotomy by speaker normalisation, then perhaps the best approach would be to adopt a minimal and orthogonal parameterisation of the vocal-tract area-function, and entrust to the uniqueness properties of the LP model the task of acoustic-to-articulatory mapping. However, whilst the LP-based method of area-function estimation does resolve the non-uniqueness problem, it remains problematic from several other points of view. Indeed, since its inception nearly three decades ago, it has suffered considerable neglect, primarily on grounds of its presumed incapability to yield anything but *pseudo*-area-functions which are generally regarded as *unreliable* and too far removed from “real” vocal-tract shapes (e.g., Sondhi, 1979). Nevertheless, we may yet remain hopeful as perhaps Broad and Shoup (1975, p.254) were, when they made the following statements in regard to the work of Atal (1970) and Wakita (1973), which

“...suggests that the combination of formant frequency and formant bandwidth information can give fairly reasonable estimates of vocal tract shapes. These results are very important for understanding the behavior of the acoustic speech wave in terms of underlying articulatory processes, although this possibility has not yet been exploited very extensively.”

Dare we hope to find the LP model useful in advancing our understanding of the speech-speaker dichotomy from a speech production point of view?

## 2.5 Concluding Perspective

Our interpretive review of the auditory-perception, acoustic-phonetic, and articulatory literature on the long-standing problem of speech-speaker dichotomy has, as far as we are aware, for the first time assembled together the numerous but often disparate works which provide evidence of both the separate and intertwined influences of phonetic and speaker-specific attributes of spoken vowels. Indeed, despite this large and growing

---

<sup>4</sup> Strictly, uniqueness of the discrete-sectioned, relative area-function of known length.

body of evidence which is held together perhaps by the frequent return to the acoustic-domain descriptions of vowel and speaker variability, there is a distinct lack of cohesion and an immaturity which is reflected both in the diversity of previous approaches, and in their still incomplete explanation of the speech-speaker dichotomy.

In the acoustic-auditory literature there is ample evidence to suggest that human perception of vowel identity and of speaker identity is inclined towards favouring, respectively, the information contained in the low and in the higher parts of the frequency spectrum. In that context, the perception literature has played an influential role in shaping current approaches to spectral parameterisation; particularly the auditorily-motivated methods of frequency scale-transformation and relative spectral measures, which are partially successful in suppressing the speaker-specific and enhancing the phonetic information in spoken vowel sounds. In addition, the perception literature has clearly fuelled the ongoing debate of formant versus whole-spectrum representation, and thereby almost inadvertently served as a catalyst for the widespread appropriation of whole-spectrum parameters such as the mel-frequency cepstrum, in machine analysis and recognition of speech.

However, the most substantial body of evidence concerning the vowel-speaker duality of the spectral continuum, is found in the acoustic-phonetic literature. Indeed, the well-known, classic studies which have firmly established the phonetic importance of the low spectral range encompassing the first two formants of spoken vowel sounds, are complemented by numerous, but relatively unacknowledged works which provide direct evidence of the speaker individuality contained in the higher spectral regions which include and extend beyond  $F_2$  of front vowels. Indirect, but equally compelling evidence is provided in studies of speaker recognition by machine which, consistently with the studies on human auditory-perception, report higher accuracies when using the higher spectral regions or the higher formants of mid- to front vowels in particular.

While studies on speaker recognition and speaker individuality have thus afforded insights into the acoustic-phonetic subspaces of relatively greater speaker potency, the vowel recognition paradigm appears to hold the potential of more completely revealing the spectral manifestations of *vowel-speaker interactions*. However, the literature on automatic vowel recognition suggests that the detrimental influences of inter-speaker

variability on recognition accuracy may not be consistently observed using the formants. Indeed, apart from the sheer popularity and demonstrated superiority of the cepstrum in state-of-the-art ASR, the whole-spectrum representation itself would appear to embody more completely (and with greater auditory-perceptual relevance) the types of formant-peak interactions which define the very essence of vowel-speaker interactions.

Nevertheless, previous attempts to overcome the speech-speaker dichotomy by suppressing speaker-related influences, have failed to exploit the full potential of the whole-spectrum representation which is commonly used in ASR. For example, the popular method of cepstral liftering does not afford direct control of frequency sub-bands; nor have the more direct methods of speaker normalisation, whether based on the whole-spectrum or the formants, addressed the fundamental issue of the apparent phonetic-speaker duality of the low and the higher spectral ranges. Clearly, a new approach to cepstral distance computation is called for which overcomes the limitation of operating over the entire available spectral range.

Even so, it emerges from the acoustic-articulatory literature that a more informed, and therefore a more effective method of overcoming the speech-speaker dichotomy by speaker normalisation, is potentially afforded by transcending the acoustics of speech, and accounting for the physical sources of speaker differences. However, whilst reasonably effective, articulatorily-motivated methods of speaker normalisation have previously been proposed, they almost invariably account for only the gross differences in vocal-tract length. Although the overall size of the vocal-tract is admittedly the most obvious and the most easily estimated, physical source of variation between adult male, adult female, and child speakers, it is also well-known that such anatomical differences can only partly explain the wide range of inter-speaker variations in the acoustics of speech; even more so for speakers of the same gender and age group, who may possess vocal-tracts of comparable size, and who may therefore exhibit mainly learned (whether inter- or intra-idiolectal), rather than organic differences.

Indeed, importation of acoustic-articulatory principles in speaker normalisation, is still in its infancy. On the one hand, a method of speaker normalisation which relies on direct articulatory measurements would in general be too prohibitive and impractical. On the other hand, the inherent non-uniqueness of acoustic-to-articulatory mapping is a

major stumbling block which continues to divert research efforts away from the potential use of estimated articulatory parameters. However, even if the non-uniqueness problem were alleviated by using the inherently unique, LP vocal-tract acoustic model, there still remains the non-trivial problem of adopting a suitable parameterisation of the vocal-tract area-function, such that the uniqueness properties of the LP model are retained, and more importantly, that the estimated vocal-tract shapes may be quantitatively compared in a physically meaningful way across different speakers. In that context, adherence to a sufficiently complete, theoretical framework for describing speaker differences in articulatory terms, is highly desirable.

Quite clearly, our interpretive review of the literature has shown that whilst an acoustic-phonetic approach to elucidating the speech-speaker dichotomy is necessary, it is not sufficient. Our acoustic-phonetic investigations in Chapter 4 will therefore be complemented by an acoustic-articulatory investigation of the dichotomy in Chapter 6. However, the latter presupposes an appropriate method of vocal-tract shape estimation and parameterisation, which we do consider in Chapter 5. In order to stage the scene, the phonetic and the speaker complexity of the spoken vowel data used in this thesis, and the acoustic parameterisation of those data, are described in the following chapter.

## Chapter 3

### Speech Materials and Acoustic Parameterisation

#### 3.1 Introduction

In this chapter we describe the speech materials and the methods of acoustic parameterisation which have yielded the data used in the remainder of this thesis. As foreshadowed in our analysis of the literature (in Chapter 2), the very definition of our proposed study of speech-speaker dichotomy raises the issue of both the speech and speaker components of the implied datasets, which is given careful consideration in Section 3.2. Our review of the literature (in Chapter 2) also highlighted the distinction between the so-called discrete representation of the acoustics of speech in terms of the formant parameters, and the so-called continuous or whole-spectrum representation afforded by the cepstrum. In view of the numerous, salient features of both types of spectral representation, our study of speech-speaker dichotomy will rely heavily on both formants and cepstra. Sections 3.3 and 3.4 are therefore devoted to describing the methods of acoustic parameterisation of the datasets used to investigate the speech-speaker dichotomy. A summarising perspective is offered in Section 3.5, which points to the subsequent chapters of this thesis by drawing attention to the usage therein of the acoustic speech data described in this chapter.

#### 3.2 Speech Materials and Speaker Sets

An acoustic-phonetic study of speech-speaker dichotomy would ideally require an experimental set of spoken language data which embody the two key attributes of *phonetic complexity* and *speaker heterogeneity*. The former refers to a sufficient coverage of the acoustic-phonetic space of each speaker, while the latter implies sufficient differences between speakers, as to render speech-speaker interactions

measurable by experimental acoustic-phonetic methods.

The *phonetic* aspect of the present study was established in our analysis of the literature (in Chapter 2), from which it emerges that even for the relatively better understood acoustic-phonetic properties of vowel sounds, there still exist fundamental gaps in our knowledge of the speaker-specific aspects of those properties. It is well-known, however, that the acoustic-phonetic properties of vowels vary not only with the speaker, but also according to intra-speaker variability, coarticulation with neighbouring speech segments, and prosodically-induced segmental variations such as those arising from the presence or absence of linguistic stress. The monosyllabic /hVd/ context has traditionally been used for studying vowel sounds, firstly because it offers the formation of common words in English, while at the same time nearly nullifying coarticulatory influences from the preceding consonant, since the articulatory configuration adopted during the /h/ normally approximates that of the following vowel. In addition, the reading aloud of /hVd/ words in isolated-citation form tends to encourage well-articulated vocalic configurations or gestures; and randomisation of the words in each list helps to maintain consistency in prosodic variables such as intonation and linguistic stress.

The *speaker* aspect of the present study was also established in our review of the literature (in Chapter 2), where attention was drawn to the distinct lack of emphasis in previous works, in regard to explaining the types of differences that exist between speakers of a given language and of the same gender, be they idiolectal differences or within-dialect, so-called *intrinsic* differences. In this vein, it is by far more problematic to secure sufficient speaker heterogeneity, than it is to secure phonetic complexity, especially when one is limited to sampling only the adult male population of speakers of a particular variety of spoken English (e.g., Australian English, or American English). On the one hand, most of the current technology in automatic speech recognition (ASR) is based on a data-driven approach whereby large amounts of training speech data are obtained from as large as possible a population of speakers of the target language, in order to help increase the robustness of ASR systems against speaker variation, at least on a gender-specific basis. On the other hand, the larger the population of speakers, the greater are the chances that differences between speakers

may be obscured or blurred into a more homogeneous continuum of variations. Heterogeneity therefore does not necessarily imply great numbers of speakers. Indeed, acoustic data of just a few speakers may suffice to clearly observe certain manifestations of vowel-speaker interactions, provided only that those speakers are sufficiently different from one another, either in a physiological or an articulatory-behavioural sense.

With these issues in mind, we now proceed to describe the three sets of speech materials on which our acoustic-phonetic and articulatory study of vowel-speaker dichotomy is based. The flow-diagram in Figure 3.1 summarises the nature of the speech materials and speaker populations which comprise each of our three datasets, and gives an overview of the pre-processing steps taken (and described in the succeeding sections of this chapter) to acquire the formant and cepstrum parameters, which together form the main body of acoustic data used in the remainder of this thesis.

The first dataset (Clermont, 1991), used primarily to unfold the phenomenon of vowel-speaker dichotomy, comprises eleven non-nasalised vowels in /hVd/ context, recorded five times contemporaneously by each of four adult male, native speakers<sup>1</sup> of Australian English. The dataset, which is a subset of a larger corpus of /CVd/ monosyllables recorded in citation form by the said four speakers, contains the following vowels in syllable-stressed position: V = /i/ (as in *heed*), /ɪ/ (as in *hid*), /ɛ/ (as in *head*), /æ/ (as in *had*), /ɑ/ (as in *hard*), /ɒ/ (as in *hod*), /ɔ/ (as in *hoard*), /ʊ/ (as in *hood*), /u:/ (as in *who'd*), /ʌ/ (as in *hud*), and /ɜ/ (as in *herd*). Five repetitions of the monosyllabic words in randomised order, were presented to each speaker on a computer screen, at a comfortable rate of about 3.7 seconds per word. The recordings were made in an acoustically-treated booth, with a light-weight electret microphone (flat frequency response from 50 Hz to 20 kHz) “clipped on the speaker’s shirt, roughly 10 cm below the neck” (Clermont, 1991), and the speech waveforms were quantised to 12 bits and sampled at  $F_s = 10$  kHz. A semi-automatic, acoustic prosodic method was then used to break up the recorded speech stream into each individual (monosyllabic) word, with leading and trailing intervals of silence discarded. Although the above

---

<sup>1</sup> Henceforth referred to as speakers A, B, C, and D.

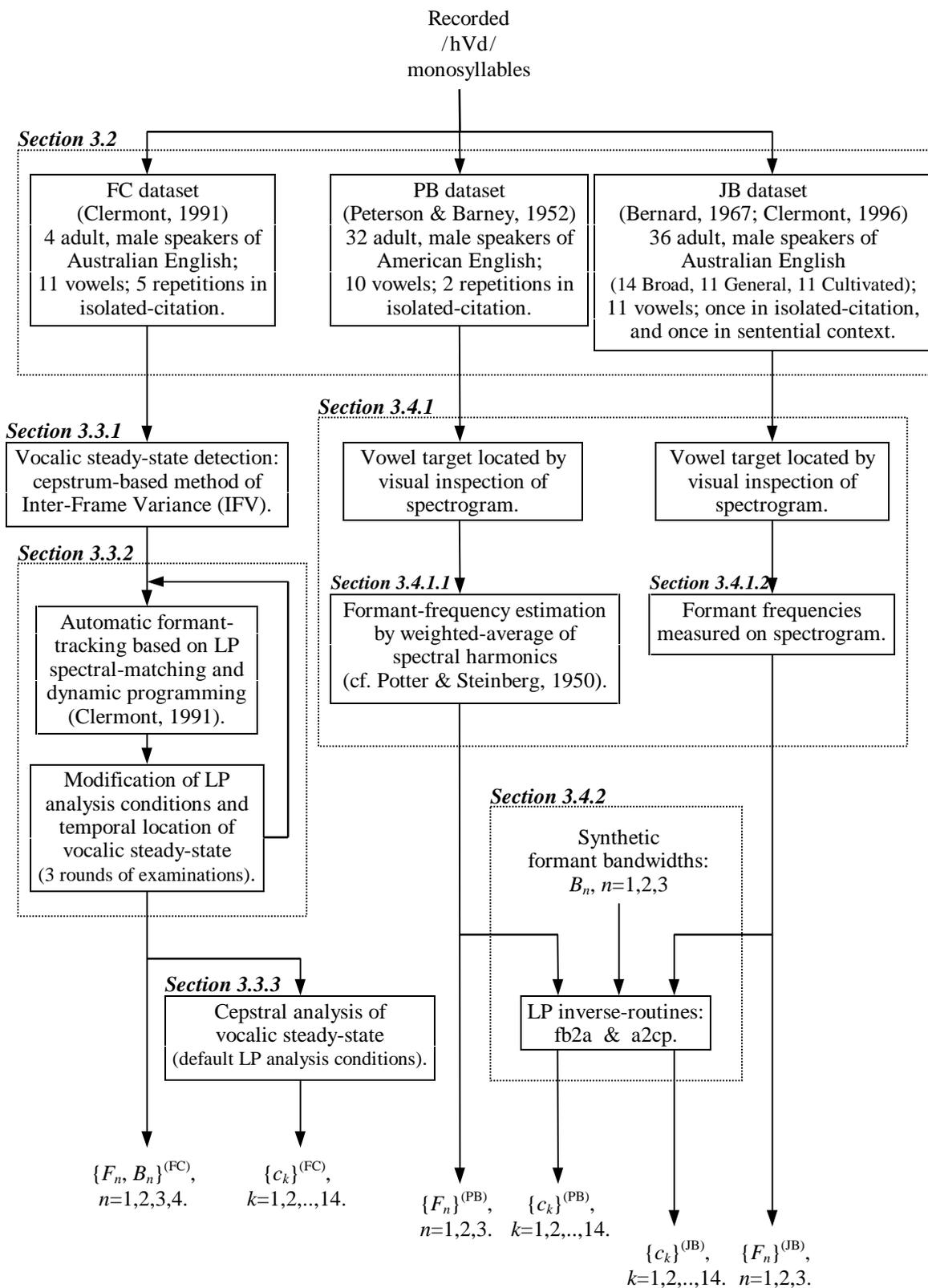


Figure 3.1: *Speech data map of the thesis.* Flow-chart summarising the three sets of speech materials, and the methods of analysis used to obtain the formant and cepstrum parameters, which comprise the acoustic data of this study. To the top-left of each module is enumerated the section of this chapter in which the respective bodies of data or methods of processing are discussed.

description suffices for our present purpose, further details regarding the speech materials and the methods of data collection and recording can be found in Clermont (1991, pp.75-108).

Although one could argue that the FC dataset (as hereafter referred to) detailed above leaves much to be desired by way of sampling a sizeable population of adult male speakers of Australian English, it nevertheless embodies a number of salient properties which render it suitable as a good starting point for investigating the question of vowel-speaker dichotomy. Firstly, it comprises a well-balanced and representative sampling of the acoustic vowel-space of each speaker, with the seven most steady-state, consecutive frames taken (as described later, in Section 3.3) from the vocalic nucleus of each of the five repetitions of each monosyllable. Admittedly, longer-term intra-speaker variability is not captured in the present dataset, but nor has this type of variability ever been so well-defined in the literature that its absence might be taken to undermine our intended explanations of acoustic-phonetic and articulatory differences between the four speakers. Secondly, the relatively small number of speakers does lend itself to an exploratory framework whereby the intricacies of vowel-speaker interactions might be accounted for and explained within reasonable bounds of computational complexity. Thirdly, the four speakers of the FC dataset have been determined by informal auditory-perceptual judgements (Clermont, 1991), to span a range of idiolectal variations which encompass the so-called General and Cultivated varieties of Australian English, thereby embodying presumably sufficient amounts of speaker differences to induce the vowel-speaker dichotomy. As stated earlier, the four-speaker FC dataset will therefore be used to unfold the dichotomy, first in the acoustic-phonetic (in Chapter 4) and then in the articulatory (in Chapter 6) domain.

With a view to extending our initial investigations to a dataset which contains a larger number of speakers and in addition, offers an advantage in terms of formant measurements having already been made, we therefore consider next the seminal study of Peterson and Barney (1952), whose published vowel formant (averaged) data of 33 adult male, 28 adult female, and 15 child speakers are well-known, and have been widely used in the literature. The speech materials of this dataset comprise ten monosyllabic /hVd/ words, read twice in randomised order by each of the said 76

speakers of American English. The ten vowels placed in syllable-stressed position are as follows: V = /i/ (as in *heed*), /ɪ/ (as in *hid*), /ɛ/ (as in *head*), /æ/ (as in *had*), /ɑ/ (as in *hod*), /ɔ/ (as in *hawed*), /ʊ/ (as in *hood*), /u/ (as in *who'd*), /ʌ/ (as in *hud*), and /ɚ/ (as in *heard*). The first three formant frequencies were originally measured (as described later, in Section 3.4.1) at a temporal location which was judged within each vowel nucleus to be the most steady-state.

Although only the per-vowel *averaged* formant frequencies for each of the three groups of speakers were originally published (Peterson and Barney, 1952, Table II, p.183), the full body of vowel formant data was recently restored and made publicly available by Watrous (1991), based on careful comparisons of three existing versions of the data. As the electronic location of the data given by Watrous (1991) no longer appeared to hold the data-file at the time that we wished to acquire it (January 1994), we obtained an electronic copy of the UNIX compressed tar file via anonymous ftp from “thumper.bellcore.com” in directory “pub”, where it had been deposited by Spiegel (“comp.speech” Internet newsgroup, 1994).

Only a subset of Peterson and Barney’s data will be used in the present study, and comprises the first three formant frequencies of the steady-state portion of each of two (contemporaneous) repetitions of the ten vocalic nuclei in /hVd/ context, spoken by 32 of the adult male speakers of American English. The data for male speaker number 2 were not used, because the first two formants ( $F_1$  and  $F_2$ ) of his vowel /ɔ/ (in both repetitions) are listed as having identical values, presumably owing to inseparable or merged formant peaks in the measured harmonic spectra. Evidence in support of this observation can be found in Figure 8 of Peterson and Barney (1952, p.182), in which the second repetition of all 76 speakers’ formants are plotted on the  $F_1F_2$  plane. That figure clearly shows a number of data points for the vowels /ɔ/ and /ɑ/, which lie along the  $F_1 = F_2$  line. The restored version of the formant data suggests that the /ɔ/ shown at  $F_1 = F_2 = 560$  Hz belongs to the second repetition of adult male speaker number 2. Whilst from a perceptual point of view the merged formant peaks may be completely acceptable (indeed, annotations which accompany the restored data suggest that both repetitions of that speaker’s /hɔd/ were unanimously identified correctly by a panel of 26 listeners), specification of a vowel with identical first and second formant

frequencies would clearly present difficulties in an acoustic-phonetic, and particularly in an articulatory investigation; hence, the formant data of speaker number 2 were disregarded.

It is not immediately clear as to the degree of (acoustic-phonetic or articulatory) heterogeneity of the group of 32 male speakers of the PB dataset (as hereafter referred to). Indeed, as discussed earlier, the larger the number of speakers drawn from the same variety of a given language, the greater are the chances of introducing homogeneity in the group as a whole. Unfortunately, we remain today with only an indication of the diversity of the PB speakers. Referring to all 76 speakers, Peterson and Barney (1952, p.177) state that only two of them “were born outside the United States”, while “a few others spoke a foreign language before learning English”. It is unknown whether any of those are amongst the subset of 32 male speakers. However, it is also stated that, in comparison with the adult female and child speakers, “the male speakers represented a much broader regional sampling of the United States; the majority of them spoke General American.” Certainly, the PB dataset is well-known and often-used in the literature; it does comprise formant measurements which are otherwise difficult to acquire accurately and consistently from a large group of speakers; and it would appear to embody a healthy mix of regional American-English dialects. We shall therefore use the subset of 32 adult male speakers primarily to validate the vowel-speaker dichotomy, in the acoustic-phonetic domain (in Chapter 4).

The presence of either dialectal (PB dataset) or idiolectal (FC dataset) variations in each of the two groups of speakers reviewed thus far, would presumably override the relatively more subtle differences which exist between speakers of the same dialect of American English, or of the same idiolect of Australian English, respectively. Vowel-speaker interactions arising from these latter, so-called *intrinsic* types of speaker differences might only be clearly observed by using a dataset in which the dialectal tendency of each speaker is known, and can therefore be taken into account. Indeed, Bernard’s (1967) pioneering study of Australian English provides exactly that, with each speaker having been categorised (by auditory impressions rendered by Bernard during an informal interview) as belonging to one of the three idiolects or varieties known as Broad, General, and Cultivated. Furthermore, the JB dataset (as hereafter

referred to) offers the convenience of formant frequency measurements, analogously to the PB dataset.

Whilst the original study involved a total of 170 adult male, native speakers of Australian English, subsequent publications by Bernard (1970, 1989) have only dealt with per-vowel *averages* of the formant data for each of the three idiolectal groups. Clermont's (1996) recent resurrection of Bernard's original formant measurements has revealed, however, that a *complete* subset of the first three formant frequencies measured in the steady-state of each vocalic nucleus (the same eleven vowels that were listed earlier in connection with the FC dataset) of /hVd/ monosyllabic words recorded once in isolated-citation and once in sentential context, could be obtained only by retaining 36 of the speakers. Fortunately, this subset includes a fairly even distribution of speakers of each idiolect, with 14 Broad, 11 General, and 11 Cultivated speakers. Moreover, Clermont (1996) has shown that the differences between the original dataset and the retained subset in terms of per-vowel formant averages computed for each of the three idiolects, are "quite tolerable", which indicates that the 36-speaker subset may indeed be considered a representative sampling of the original 170 speakers, both as a whole and in terms of each idiolectal group.

In contrast with the PB dataset where the dialectal distribution of the speakers is known only qualitatively at best, the idiolectal grouping of the 36-speaker JB dataset is known precisely. The JB dataset will therefore be used in Chapter 4 to further validate the phenomenon of vowel-speaker dichotomy, particularly from the point of view of elucidating vowel-speaker interactions induced by either *idiolectal* or *intrinsic* speaker differences.

The remainder of this chapter is devoted to a complete description of the methods of acoustic parameterisation used to construct the three bodies of acoustic data, as outlined in Figure 3.1. Section 3.3 describes the acoustic parameterisation of the more phonetically-emphasised (FC) dataset, which will be used to *unfold* the phenomenon of vowel-speaker dichotomy. Section 3.4 then describes the acoustic parameterisation of the two formant-based (PB and JB) datasets, which will permit *validation* of the dichotomy using not only a greater number of speakers, but also exploiting the contrast between the effects of idiolectal and intrinsic speaker differences clearly defined.

### 3.3 Acoustic Data used to Unfold the Dichotomy

As described in the preceding section, the speech materials used to first unfold the acoustic-phonetic manifestations of vowel-speaker interactions, were selected on the basis of embodying sufficient phonetic complexity and speaker heterogeneity, whilst ensuring that the involved and exacting task of acoustic-phonetic parameterisation of the entire dataset could still be carried out within a reasonable period of time. As shown earlier in Figure 3.1, the FC dataset of eleven non-nasals in /hVd/ context, recorded five times by four adult male, native speakers of Australian English, is first subjected to an automatic procedure (described in Section 3.3.1) to arrive at an initial estimate of the temporal location of the most steady-state portion of each vocalic nucleus. A semi-supervised approach (described in Section 3.3.2) is then adopted to maintain consistency in estimating the first four formants of the vocalic steady-states, using an automatic formant-tracker (Clermont, 1991) based on Linear Prediction (LP) spectral-matching and dynamic programming. Finally, the well-known LP cepstrum parameters are extracted (as described in Section 3.3.3) from the same, steady-state frames of vocalic speech data, in preparation for the acoustic-phonetic investigations to be described in Chapter 4.

#### 3.3.1 Vocalic Steady-State Detection

An implicit assumption in our proposed study of vowel-speaker interactions, as noted earlier in Section 3.2, is the selection of appropriate speech materials such that influences external to *vowel* and *speaker* variations are controlled or minimised. Thus, for example, the time-honoured /hVd/ monosyllabic context was selected to increase the chances of consistently stressed acoustic realisations of well-articulated vocalic gestures. Closely associated with the notion of well-articulated vowel production, is the concept of the so-called *vowel target* which the vocal-tract articulators presumably succeed in attaining some time after the initial consonant /h/ but before the subsequent articulatory gesture which leads into the final consonant /d/. The acoustic manifestation of that target articulatory configuration is the so-called *vocalic steady-state* during which there is, by definition, relatively little spectral change. Indeed, it is in the steady-

state portion of each vowel nucleus that we aim to make our measurements of formant and cepstrum parameters, which will then be taken to represent that speaker's acoustic target for the given vowel.

Towards facilitating our identification of the steady-state portion of each vowel nucleus (of which there are 5 repetitions of 11 vowels from 4 speakers, i.e., a total of 220 vowel targets to identify), we note the power and simplicity of the procedure used by Broad and Wakita (1977) to locate five consecutive, steady-state frames in smoothed formant trajectories which they obtained from acoustic recordings of sustained vowels. Their operational definition of steady-state was simply "minimum interframe formant frequency variation", which they found by first computing the inter-frame variance in the formant frequencies of each combination of five consecutive frames across the entire utterance, and then locating the temporal location of the global minimum.

However, if the formants themselves have not yet been estimated, the question then arises whether the criterion of minimum inter-frame *formant frequency* variation cannot be replaced by a more general criterion of minimum inter-frame *spectral* variation. In this vein, we note that the Euclidean distance between a pair of LP cepstra is equivalent to the root-mean-square (rms) distance between the corresponding pair of cepstrally-smoothed, LP log-magnitude spectra (Gray and Markel, 1976). We can therefore still embrace the spirit of Broad and Wakita's (1977) steady-state detection algorithm, while adopting a cepstrum-based criterion of minimum spectral variation, which uses a far more robust acoustic parameterisation than the formants. Furthermore, whilst the LP cepstrum can be used indiscriminately to represent the short-time LP spectrum of both sonorant and obstruent segments of the speech stream, the spectrally most steady-state region of an /hVd/ utterance can be expected to lie within the vocalic nucleus. Consequently, the inter-frame cepstral variance can be computed in all groups of any fixed number of consecutive frames throughout the entire duration of each /hVd/ utterance, without prior segmentation of the vocalic nuclei.

In addition to a whole-spectrum approach thus advocated mainly for practical reasons, we also need a spectral representation, and just as importantly a cepstral distance measure, which tends to emphasise the formant peaks and to de-emphasise the

spectral valleys in between those peaks. To this end, we note the “formant-sensitivity” property (Clermont, 1991, p.114) of the Negative Derivative of the LP Phase Spectrum (NDPS), as captured in the so-called NDPS cepstral distance measure proposed by Yegnanarayana and Reddy (1979) and shown as follows:

$$d_{\text{QCEP}}^2(\mathbf{c}, \mathbf{c}') = \frac{1}{\pi} \int_0^\pi \left[ \left( -\frac{d\phi}{d\theta} \right) - \left( -\frac{d\phi'}{d\theta} \right) \right]^2 d\theta \approx \frac{1}{2} \sum_{k=1}^{NCC} [k(c_k - c'_k)]^2, \quad (3.1)$$

where  $\phi$  and  $\phi'$  are the LP phase spectra of two frames of speech which are to be compared,  $c_k$  and  $c'_k$  are the  $k^{\text{th}}$  components of the corresponding pair of cepstral vectors ( $\mathbf{c}$  and  $\mathbf{c}'$ , respectively),  $\theta$  is the normalised frequency variable, and  $NCC$  is the number of cepstral coefficients (Gray and Markel (1976) have shown that the approximate equality in Equation 3.1 tends to an exact equality as  $NCC \rightarrow \infty$ , but that a value equal to the LP order of analysis  $M$  yields a sufficiently high correlation between the two distances evaluated over a large number of speech frames). Also known as the root-power-sum (RPS) or the quefrequency-weighted cepstrum (QCEP), the equivalence of the  $kc_k$  sequence to the LP-NDPS (also known as the *group delay spectrum*), and the remarkable “formant-enhancement” properties (Clermont, 1991, p.111) of the latter, were first shown by Yegnanarayana (1978) and by Fuchi and Ohta (1978). The so-called NDPS or quefrequency-weighted cepstral distance measure in Equation 3.1 inherits those properties, and is therefore equivalent to a spectral distance measure with enhanced sensitivity to variations in the formant frequencies. In addition, the cepstral distance measure on the right-hand side of Equation 3.1 is computationally far more economical than the exact NDPS distance measure (shown in the middle of that Equation).

Our algorithmic procedure for estimating the temporal location of the vocalic steady-state in each /hVd/ utterance, is therefore described by the following steps:

1. Using the autocorrelation method of LP analysis (Makhoul, 1975a; Markel and Gray, 1976), compute  $M = 14^{\text{th}}$ -order LP cepstrum in each frame of length 25.6 msec (speech waveform Hamming-windowed, and pre-emphasised using the digital filter  $1 - 0.98z^{-1}$ ), with a frame-advance of 5.0 msec, yielding a total of  $N$  frames across the entire /hVd/ utterance.

2. Initialise frame-group counter to  $n = 1$ .
3. Compute Inter-Frame Variance (IFV) in the group of  $NSSF = 7$  potential steady-state frames starting at frame  $n$ , as follows:

$$\text{IFV}(n) = \frac{1}{NSSF} \sum_{ssf=1}^{NSSF} d_{\text{QCEP}}^2(\mathbf{c}(n + ssf - 1), \bar{\mathbf{c}}(n)), \quad (3.2)$$

where  $\mathbf{c}(n)$  is the cepstral vector at the  $n^{\text{th}}$  frame,  $d_{\text{QCEP}}^2(\cdot)$  is the NDPS cepstral distance defined earlier in Equation 3.1, and:

$$\bar{\mathbf{c}}(n) = \frac{1}{NSSF} \sum_{ssf=1}^{NSSF} \mathbf{c}(n + ssf - 1). \quad (3.3)$$

4. Increment the frame-group counter  $n$  by 1; go to step 3 while  $n \leq N - NSSF + 1$ .
5. Determine the index  $n$  of the global minimum in the IFV profile; store the entire IFV profile for future reference.

A fine example of a correctly located vocalic steady-state is shown in Figure 3.2(a), which displays the acoustic speech waveform in the middle graph, the spectral poles obtained by LP analysis of each frame in the lower graph, and, time-aligned above those plots, the computed IFV profile across the entire duration of the first repetition of /hud/ recorded by speaker C. Although the IFV profile appears to be quite noisy, especially during the apparently noise-ridden parts of the waveform before the aspiration of the /h/ and also well after the /d/ release, there is clearly a global minimum which identifies the start of the seven-frame vocalic steady-state to occur shortly after the onset of voicing (after just one or two glottal periods, as shown by the left vertical marker superimposed on the waveform). As shown in the spectrogram-like display of the LP poles, the spectral variation within the automatically identified steady-state is indeed quite small compared with the rest of the utterance, and the formant structure which is apparent in that region provides visual confirmation of a correctly identified vowel target, just before the rapid transition towards the /d/ consonant.

Of all 220 /hVd/ utterances thus analysed, the temporal location of the global minimum in the IFV profile was found to lie outside the vowel nucleus (as judged by visual observation of the speech waveform and the time-aligned display of LP poles) for only four utterances of one speaker (A). The vowel target was incorrectly detected

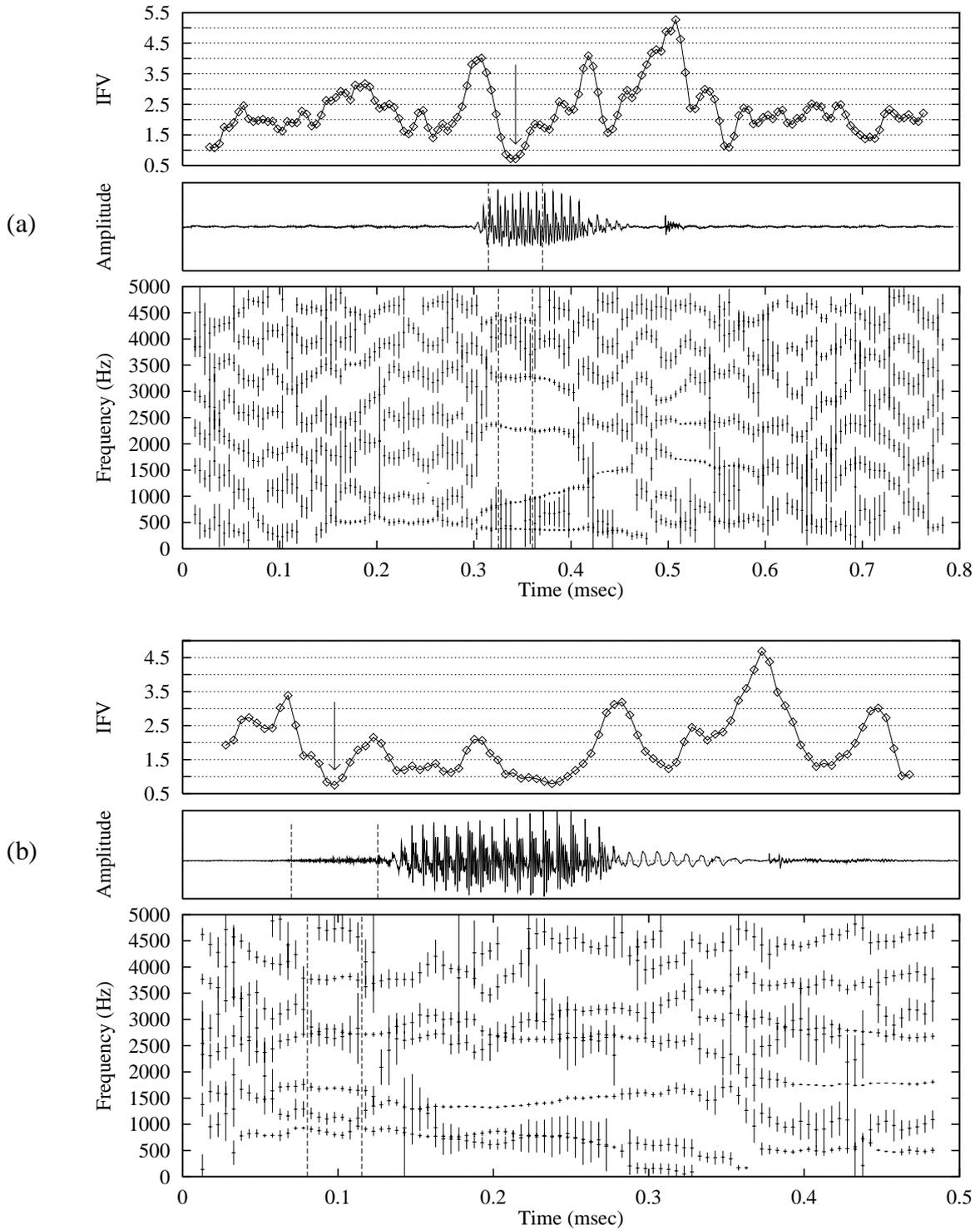


Figure 3.2: *Panel (a)*: a good choice of vocalic steady-state location, in the first repetition of /hʊd/ recorded by speaker C. *Panel (b)*: an erroneous choice of vocalic steady-state location, in the fifth repetition of /hʌd/ recorded by speaker A. *Time-aligned in each panel, from top to bottom*: Inter-Frame Variance (IFV) profile obtained by the algorithm described in Section 3.3.1 (arrow points to the global minimum); speech waveform, with vertical dashed lines demarcating the entire, automatically-selected steady-state; spectrogram-like display of LP poles (length of each vertical stroke proportional to the pole bandwidth) obtained with default analysis conditions (see IFV algorithm in Section 3.3.1), with vertical dashed lines demarcating the seven frames automatically selected as the steady-state.

once before the onset of voicing (during the /h/ aspiration), once after the /d/ release, and twice during the /d/ closure. In the first case (the fifth repetition of /hʌd/ recorded by speaker A), the spectrally most steady-state sequence of seven consecutive frames was found by the algorithm to lie before the onset of voicing, during the /h/ aspiration. As shown in Figure 3.2(b), there is indeed less inter-frame variation of the LP poles in the selected frames than in the vowel nucleus itself, where the first formant appears to split into two closely-spaced poles, and where the third and fourth formants are not at all clearly represented. Nevertheless, the second-ranked minimum in the IFV profile does fall inside the vowel nucleus, and might therefore serve as a more likely candidate for the vocalic steady-state in this utterance. Similar observations were made in the fifth repetition of /hʊd/ (steady-state detected just after the /d/ release) and in the third and fourth repetitions of /hʊd/ (steady-state detected during the /d/ closure) recorded by speaker A. In all four cases, the potential steady-state location was manually corrected by selecting the next candidate minimum in the IFV curve lying within the vocalic nucleus.

Although the completely unsupervised algorithm described above did appear to identify very likely, candidate steady-state locations for the great majority of the /hVd/ utterances recorded by our four speakers, we emphasise that it can only be regarded as providing an objective guide as to the vicinity of the steady-state within each vowel nucleus. As it was expected that manual intervention would be necessary to ensure correct identification of each vowel target, we made no further attempt to improve on the reliability and precision of the very simple, steady-state detection algorithm. Indeed, a more precise, final decision on the location of the vocalic steady-state awaits the formant estimation stage (described in the following section), where we also take steps to ensure that the first four formants are measured with as much care and consistency as possible, across the seven consecutive frames which best characterise the vowel target.

### **3.3.2 Formant Estimation**

As will become plainly evident in the remaining chapters of this thesis, both the acoustic-phonetic and articulatory interpretations of vowel-speaker dichotomy rely crucially on the formant parameters. One of our main objectives in acoustic

parameterisation of the dataset destined to unfold the dichotomy, was therefore to secure the best possible formants in the steady-state region of each of the 220 vowel nuclei recorded by our four speakers. As discussed in the preceding section, the completely unsupervised algorithm described therein was used to obtain an objective, initial estimate of the temporal location of the seven most steady-state, consecutive frames in each /hVd/ utterance. Only four of those estimates were found to lie (incorrectly) outside the vowel nucleus, thus requiring manual intervention to select a local minimum in the IFV profile within the vowel nucleus. Notwithstanding the apparent success of our automatic algorithm, more extensive manual intervention was unavoidably required in order to overcome both IFV- and LP-related problems in obtaining the best formants, as described in Sections 3.3.2.3 and 3.3.2.4, respectively.

### **3.3.2.1 Goodness Criteria**

It is first necessary to define “best possible formants” in the context of this study. Our definition is based on the requirement of formants which best characterise the *vowel target* of each /hVd/ monosyllable recorded. That is, we require the trajectories of the first *four* formant frequencies ( $F_1$ ,  $F_2$ ,  $F_3$ , and  $F_4$ ) to be as well-defined as possible throughout the selected, seven consecutive frames. However, “well-defined” is here taken to imply not only *smooth* and *nearly-horizontal* formant frequency trajectories, but also an avoidance of excessively wide formant bandwidths. In addition, whilst the utmost care is taken to preserve the natural occurrence of intra-speaker variability across the five repetitions of each utterance, our acoustic-phonetic knowledge is also brought to bear on the task of ensuring *inter-repetition consistency* in the per-vowel formant estimates of each speaker.

### **3.3.2.2 Formant-tracking Method and Evaluation Aids**

Having thus defined our goals in regard to formant measurements, we next describe the methods used to attain them. First, the task was carried out on the data of one speaker at a time, by two, acoustic-phonetically-trained examiners (the present author and his doctoral supervisor Frantz Clermont). The formant-tracker itself (Clermont, 1991) is an entirely automatic, analysis-by-synthesis method based on spectral matching and

dynamic programming. Briefly, it selects the first four formants of each frame from amongst all combinations of the poles yielded by LP analysis of that frame, then applies a formant-sensitive spectral distance measure (the NDPS cepstral distance measure described in the preceding section) to construct a distance matrix which spans the selected frames of data and finally yields, by the method of dynamic-programming, an optimum path which defines the formant trajectories across those frames. Given the temporal constraints imposed by the formant-tracker, each steady-state was therefore padded with a preceding and a following group of seven frames (a total of 21 frames in all) in order to secure greater continuity in the formant-tracks of the seven central frames.

Formant trajectories yielded by the tracker, were examined by superimposing them onto the corresponding, spectrogram-like display of LP poles (examples of which were shown earlier in Figure 3.2), together with a pair of vertical lines demarcating the seven central, potentially steady-state frames. Also displayed on the same screen were time-aligned plots of the corresponding speech waveform and of the energy profile across the entire utterance, both of which proved useful in visually confirming the location of the vowel nucleus. Our per-speaker examinations proceeded one vowel at a time, often displaying the aforementioned graphical plots for two or three repetitions side by side, thus increasing the chances of retaining in the examiners' minds, a semblance of cross-repetition consistency in assessing the formant-tracks and the relative temporal locations of the selected steady-states. Also quite useful in this regard, is the so-called "formant sequence chart" (cf. Potter and Steinberg, 1950), where all four formant frequencies are plotted (with frequency on the ordinate) in sequence from left to right, across the seven frames, five repetitions, and eleven vowels of each speaker. We found the formant sequence chart useful not only for confirming the vowel-to-vowel relations of the formant frequencies (whereby gross errors in formant labelling can be identified), but, more importantly, for examining their variability across the seven steady-state frames and especially across the five repetitions of each vowel. Particularly variable formant trajectories, or those which appeared to be markedly out of line in comparison with the neighbouring repetitions of the same vowel, could thus be rapidly identified and marked for a more careful re-examination of the pole-display and superimposed

formant trajectories of each individual utterance.

The vocalic steady-state location and the measured formant-tracks were visually examined as described above, and corrections administered in three consecutive passes. Each pass was preceded by automatic estimation of the formant trajectories across 21 frames centred about those seven frames which were selected at the previous round of examinations as the steady-state of each /hVd/ utterance recorded by that speaker. The formant-tracker was first applied to the steady-state of each vowel nucleus as determined by the global minimum in the corresponding IFV curve (with the four corrections as described Section 3.3.1). The first pass of examinations then prompted either a more suitable choice of steady-state location, a more appropriate set of LP analysis conditions, or both. Once a new set of instructions were formulated concerning the 21 frames and the LP analysis conditions for each utterance, the formant-tracker was then re-run to produce a new set of formant estimates. These were examined using the tools and displays described earlier, a fresh set of instructions were prepared for the formant-tracker based on the latest corrections to steady-state location and analysis conditions, and the entire process was repeated a third and final time, to yield the steady-state locations and formant trajectories ultimately to be used in the remainder of this study.

### **3.3.2.3 IFV-related Problems and Corrections**

The difficulties encountered at each examination of estimated formant data, can be broadly categorised as occurring either as a result of inter-repetition inconsistency, or simply as a result of one or more of the first four formants not being sufficiently well-defined by the LP poles. Included in the former category are those vowel nuclei which were found to carry acoustic-phonetic cues of diphthongisation (typically /i/ and /u:/ which, as also noted by Bernard (1967, 1989), tend to acquire an “onglide” for some speakers of Australian English, thus rendering the nominal monophthongs perceptually closer to /əi/ and /əu:/, respectively). If spectral variation persists throughout a diphthongised vowel nucleus, then admittedly that part of the nucleus determined by the IFV algorithm (described in Section 3.3.1) to have the least spectral change, will not always coincide with the segment which might be selected by a trained phonetician as

the steady-state or vowel target of the nominal monophthong. A case in point is the fourth repetition of /hid/ recorded by speaker B. As shown in Figure 3.3(a), the global minimum in the IFV profile occurs soon after the onset of voicing. However, the profile in the vicinity of the global minimum is quite shallow, and is followed towards the end of the vowel nucleus by comparably small inter-frame variances which are only local minima. The corresponding display of LP poles shows that the automatically selected seven frames (demarcated in Figure 3.3(a) by the left-most pair of vertical, dashed lines) lie in what Bernard (1967, 1989) might refer to as the “onglide”, where in this case the second formant is still rising towards its target for the nominal monophthong. Hence, in order to more accurately identify the intended vowel target, and thus also to retain a greater degree of inter-repetition consistency, it was necessary to manually correct the temporal location of the vocalic steady-state by choosing a more suitable location (as indicated in Figure 3.3(a) by the right-most pair of vertical, solid lines), displaced by a total of 27 frames (or 135 msec) towards the end of the vowel nucleus, just prior to the so-called offglide into the final /d/ consonant. On the other hand, Figure 3.3(b) shows a correctly located vocalic steady-state in the second repetition of /hʌ:d/ recorded by speaker D. Despite the nearly constant rise of the second formant throughout the vowel nucleus, all four formant frequencies measured in the automatically-selected steady-state region demarcated by the vertical dashed lines, are consistent across repetitions for this speaker.

By far the majority of problematic cases requiring manual intervention, however, were due to ill-defined formants, even when the temporal location of the automatically-detected steady-state would otherwise have been acceptable. After the third and final round of examinations, the location of the steady-state had been manually shifted, owing either to an incorrectly identified vowel target, or to ill-defined formants at what might have passed as a suitable vowel target, in 159 (or 72 %) of the 220 utterances. Although this figure might seem to undermine the IFV method of vocalic steady-state detection, the power of that simple method in yielding a good initial estimate of the vowel target is better appreciated by noting that the average shift in steady-state location was only 26 msec, which is less than half of the maximum length of each seven-frame window (55.6 msec).

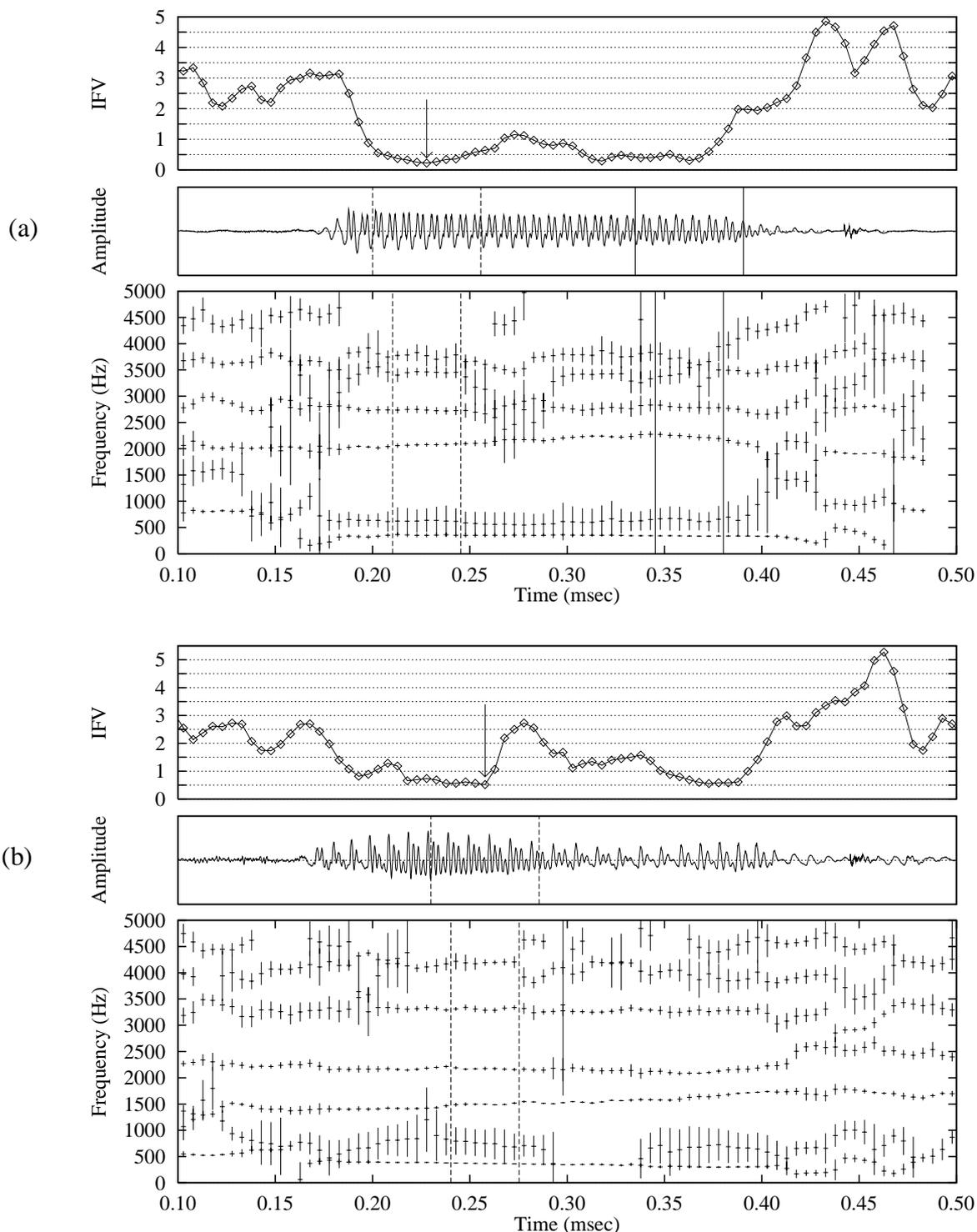


Figure 3.3: *Panel (a)*: steady-state automatically selected in the vocalic “onglide”, in the fourth repetition of /hid/ recorded by speaker B. *Panel (b)*: a good choice of vocalic steady-state location, for the second repetition of /hɪ:d/ recorded by speaker D. *Time-aligned in each panel, from top to bottom*: Inter-Frame Variance (IFV) profile obtained by the algorithm described in Section 3.3.1 (arrow points to the global minimum); speech waveform, with vertical dashed lines demarcating the entire, automatically-selected steady-state; spectrogram-like display of LP poles (length of each vertical stroke proportional to the pole bandwidth) obtained with default analysis conditions (see IFV algorithm in Section 3.3.1), with vertical dashed lines demarcating the seven frames automatically selected as the steady-state. In panel (a), the manually corrected steady-state is demarcated by vertical solid lines.

### 3.3.2.4 LP-related Problems and Corrections

As stated earlier, manual intervention entailed not only selecting a more suitable steady-state position within the vowel nucleus, but also modifying the LP analysis conditions used by the automatic formant-tracker. The default analysis conditions were the same as those used earlier to generate the IFV profiles, i.e., LP order of analysis  $M = 14$ , pre-emphasis coefficient  $PR = 0.98$ , analysis frame-length  $FL = 25.6$  msec, a Hamming-window of the same length applied to each frame prior to analysis, and the frame-advance  $FA = 5.0$  msec. However, for the sake of obtaining the “best possible formants” as defined earlier, it was found necessary either to increase or to decrease the order of analysis  $M$ , and often to select a higher, a lower, or a frame-by-frame adaptive (Gray and Markel, 1974) pre-emphasis  $PR$ .

Not surprisingly, the rarely-documented fourth formant proved particularly troublesome in this regard, almost independently of vowel or speaker. An example in passing was shown previously in Figure 3.3(a) where, in order to secure a more well-defined fourth formant trajectory in the manually corrected steady-state region, it was found necessary to reduce the LP order of analysis to  $M = 12$  and to increase the pre-emphasis to  $PR = 0.99$ . As a result of ill-defined formants alone, the LP analysis conditions required modification in 165 (or 75 %) of the 220 vowel nuclei.

The beneficial effect of *increasing* the LP order of analysis is illustrated in Figure 3.4, which shows the IFV profile, the speech waveform, and two displays of LP poles, for the fourth repetition of /hɛd/ recorded by speaker D. The pole-display shown directly below the waveform was obtained with the default LP analysis conditions, and the vertical lines indicate the steady-state region which for this utterance was correctly identified by the global minimum in the IFV profile. However, at the default value  $M = 14$  there appear spurious poles of wide bandwidth which tend to overlap with the second and the third formants, and, more importantly, the fourth formant (at approximately 3500 Hz) is appreciably unsteady and ill-defined. By contrast, an increase in the LP order of analysis to  $M = 18$ , as shown in the pole-display at the bottom of Figure 3.4, provides a far better definition of the fourth formant (as seen by the more nearly horizontal trajectory and the smaller bandwidths) and, presumably by

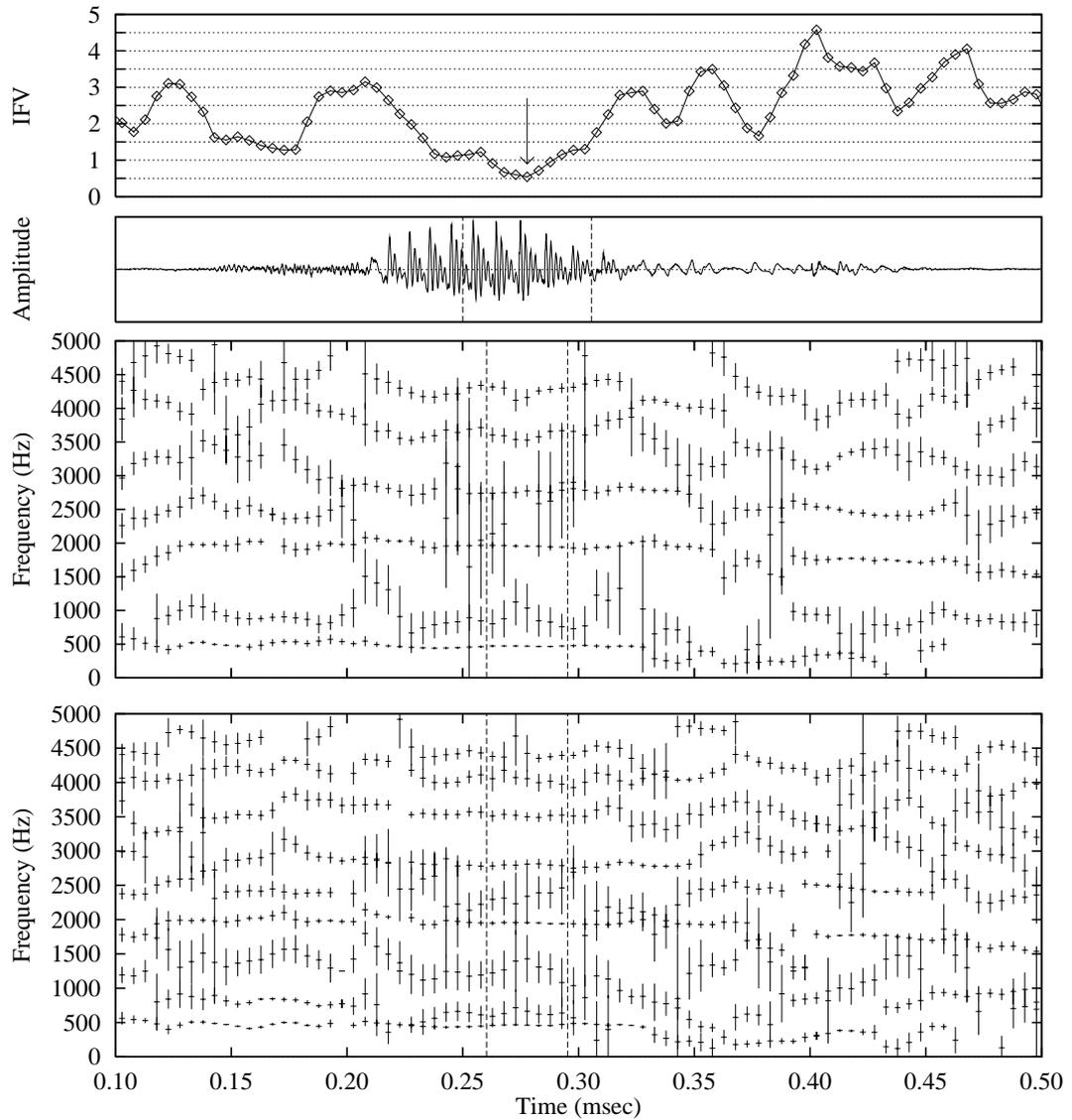


Figure 3.4: An example of the beneficial effect of *increasing* the LP order of analysis to obtain more well-defined formants ( $F_4$  in particular), in the fourth repetition of /hɛd/ recorded by speaker D. *Top two graphs*: Inter-Frame Variance (IFV) profile obtained by the algorithm described in Section 3.3.1 (arrow points to the global minimum); speech waveform, with vertical dashed lines demarcating the entire, automatically-selected steady-state. *Bottom two graphs*: spectrogram-like display of LP poles (length of each vertical stroke proportional to the pole bandwidth) obtained with default analysis conditions (*above*), and with the LP order of analysis increased to  $M=18$  (*below*), (vertical dashed lines demarcate the automatically selected, seven steady-state frames).

virtue of the two additional, conjugate pole pairs available in the analysis, yields spurious poles which are somewhat less erratic in behaviour and better separated from the true formants.

An illustration of the beneficial effect of *reducing* the LP order of analysis is provided by the fifth repetition of /hʌd/ recorded by speaker C. As shown in Figure 3.5, the IFV algorithm selected a very likely steady-state location (in fact, it was manually advanced by just one frame in order to secure slightly less variable formant trajectories). However, it is not immediately apparent from the LP analysis performed at  $M = 14$  (pole-display shown directly below the speech waveform), whether or not the second formant is represented by either of the two, closely-spaced and broader-bandwidth poles at approximately 1 kHz. By contrast, a reduction in the order of analysis to  $M = 13$  (pole-display shown at the bottom of Figure 3.5) yields a more acceptable, narrower-bandwidth second-formant trajectory which lies in between the two sets of poles obtained previously, and which, more importantly, is more consistent than either of the two previous sets of poles, with the steady-state second formant as measured in the first four repetitions of the back vowel in /hʌd/ recorded by the same speaker.

In general, an ill-defined formant was manifested on the pole-display either as a formant which appeared to be “split” into two poles of broader bandwidth (as just described and illustrated in Figure 3.5), a formant which simply disappeared for one or more consecutive frames, or a formant trajectory which appeared to be relatively variable or “noisy” (mainly in its frequency-location from frame to frame, but also in its bandwidth). As described above, an increase in the LP order of analysis  $M$  often helped to better define certain formants with the aid of additional spectral poles, whilst a decrease in  $M$  often resolved the ambiguity when a single formant was erroneously represented by more than one spectral pole. In addition, an increase or a decrease in the default pre-emphasis  $PR = 0.98$  (or in some cases, an adaptive pre-emphasis) often helped to better define formants by virtue of the greater or lesser degree, respectively, of overall spectral tilt which, as indeed was found, tended to have more influence on the poles in the higher frequency-bands.

However, when a formant-trajectory still showed signs of being variable or “noisy” despite changes in LP analysis conditions as described above, then one of two

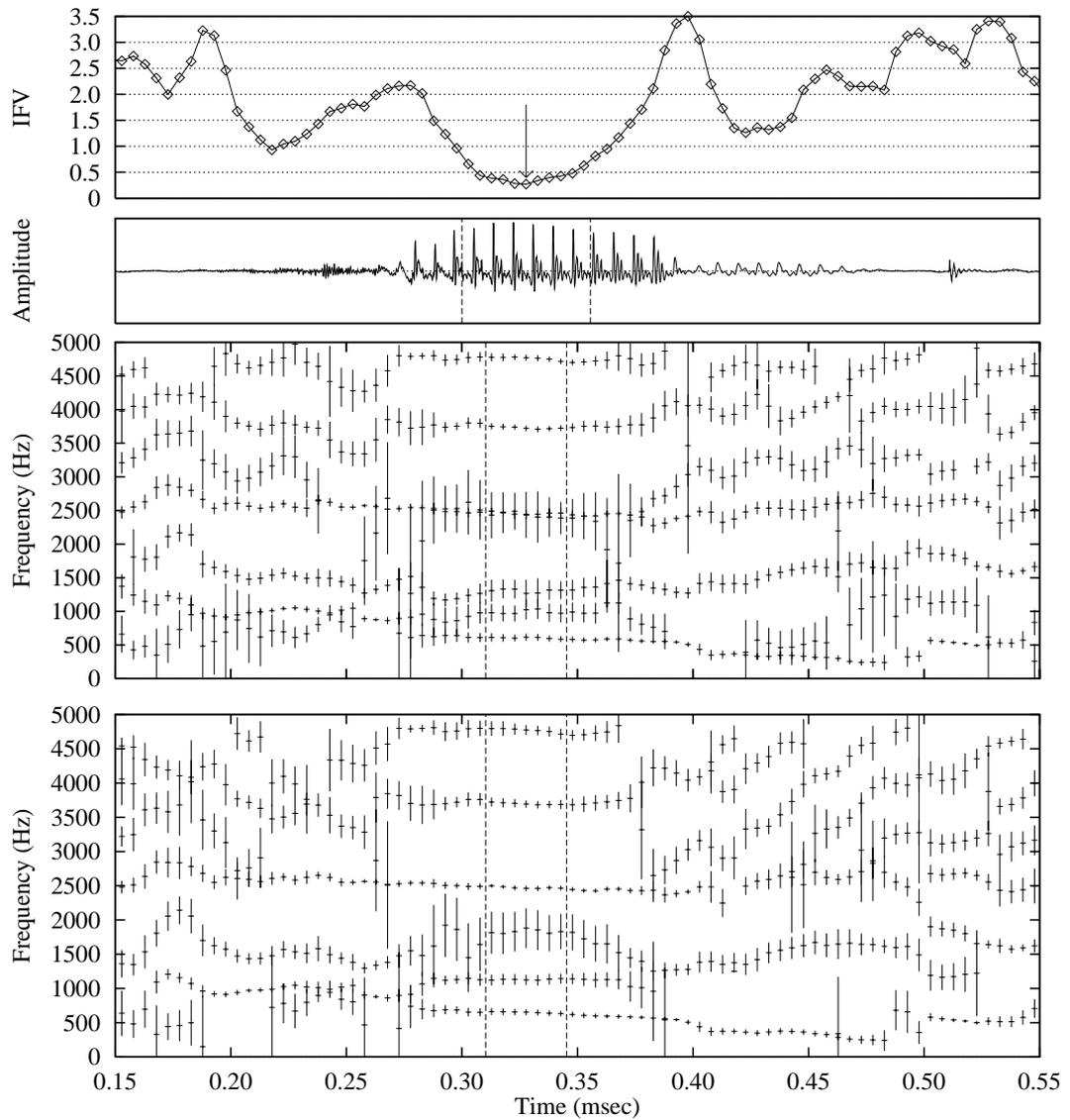


Figure 3.5: An example of the beneficial effect of *decreasing* the LP order of analysis to obtain more well-defined formants ( $F_2$  in particular), in the fifth repetition of /hʌd/ recorded by speaker C. *Top two graphs*: Inter-Frame Variance (IFV) profile obtained by the algorithm described in Section 3.3.1 (arrow points to the global minimum); speech waveform, with vertical dashed lines demarcating the entire, automatically-selected steady-state. *Bottom two graphs*: spectrogram-like display of LP poles (length of each vertical stroke proportional to the pole bandwidth) obtained with default analysis conditions (*above*), and with the LP order of analysis reduced to  $M=13$  (*below*), (vertical dashed lines demarcate the automatically selected, seven steady-state frames).

remaining options were chosen to remedy the situation. First, the so-called (DP-2) smoothing function of the formant-tracker (Clermont, 1991) was put into effect, whereby the frequencies of each of the first three, tracked formants would be perturbed in steps of  $\pm 30$  Hz, and the quefrency-weighted cepstral distance measure (Equation 3.1) used to construct a much larger, dynamic-programming distance matrix from which smoother formant trajectories might be selected, owing to the greater number of possibilities and also to a more stringent continuity constraint imposed on those trajectories. This option of the formant-tracker was indeed used a total of 7 times on  $F_1$ , 16 times on  $F_2$ , and 25 times on  $F_3$ , confirming our expectations that the higher formants might tend to be more prone to measurement errors induced by frame-to-frame variability in the frequency-locations of LP poles.

Failing the above procedure, our last resort was to decrease the frame-advance in order to allow all seven, consecutive frames to fit within the longest span of acceptable, steady-state formant trajectories that could be found in the vowel nucleus. Remarkably, this option was selected for only one of our four speakers (speaker A), whose fourth-formant trajectories proved particularly elusive. Only three of that speaker's vowels (those in /hɔd/, /hʊ:d/, and /hɜd/) were found not to require a reduction in the frame-advance in any of the five repetitions. On the other hand, all five repetitions of his two front vowels in /hɛd/ and /hæd/, four of the five repetitions of his two front vowels in /hid/ and /hɪd/, and four repetitions of his back vowel in /hʊd/, were found to require a reduced frame-advance, often in addition to changes in LP analysis order and pre-emphasis. However, in order to retain a reasonable lower limit on the length of each seven-frame steady-state, the frame-advance was never decreased below  $FA = 1.0$  msec (thus retaining a minimum steady-state duration of 31.6 msec).

Whilst the main criterion for assessing the goodness of measured formants traditionally refers to their *frequencies* (as exemplified here by the use of the NDPS cepstral distance measure in the formant-tracker), the formant *bandwidths* have also played an indirect role in that assessment, by visual examination of the LP pole displays in order to avoid excessively large formant bandwidths. A single exception to the rule occurred for the first repetition of /hʊ:d/ recorded by speaker A, where the formant-tracker yielded a fourth formant which was found to be inconsistent (by nearly 500 Hz)

with the fourth formant frequencies obtained for the other repetitions of the same vowel recorded by the same speaker. Examination of the LP pole display confirmed that the tracked, “spurious” pole lying in between the third and fourth formants, was indeed of a slightly narrower bandwidth than the more consistent, true fourth formant. Interestingly, the tracker was induced into tracking what we considered (as a result of cross-repetition comparisons) the true fourth formant, only after changing the NDPS-based cepstral distance measure to an unweighted one, which is equivalent to the rms logarithmic-magnitude distance between cepstrally-smoothed LP spectra. Thus, the reliability of bandwidth measurements cannot be expected to rival that of the formant frequencies, and their use in either an acoustic-phonetic or an articulatory investigation of vowel-speaker dichotomy will require further, careful examinations which we duly consider in subsequent chapters.

In sum, the methods described in this section were used to ensure the best possible conditions for estimation of the first four formants, thus securing a reliable, acoustic-phonetic description of the eleven nominal monophthongs recorded by our four speakers of Australian English. The formant sequence charts in Appendix A provide an overview of the formant frequency data thus obtained for each speaker after all three rounds of examinations. Above all, they confirm the inter-repetition consistency in each formant frequency measured in the vocalic steady-state of our four speakers’ recordings of /hVd/ monosyllables. Also included in Appendix A are tables which describe our final choice of LP analysis conditions used to obtain the formant measurements for each utterance.

### **3.3.3 Cepstral Analysis**

The methods of speech data analysis described in Sections 3.3.1 and 3.3.2 were used, respectively, to obtain an objective estimate of the temporal location of the vocalic steady-state in each of the 220 /hVd/ utterances recorded, and to adjust those estimates on the basis of acquiring the best possible trajectories of the first four formants which characterise each vowel target. Whilst the formants do provide a discrete and readily interpretable acoustic-phonetic parameterisation of vocalic speech data, the cepstrum offers a whole-spectrum representation which, as argued in Chapter 2, is likely to yield

key insights regarding manifestations of vowel-speaker interactions along the spectral continuum. To this end, the cepstrum was obtained by the autocorrelation method of LP analysis (Makhoul, 1975a; Markel and Gray, 1976) of the 220 utterances, in each of the seven consecutive frames which comprise the vocalic steady-state located by the methods described in the previous section.

As explained in the preceding section, the LP analysis conditions used by the automatic formant-tracker were determined manually for each utterance, such that the best set of formants could be selected from amongst the LP poles. Thus, the pre-emphasis coefficient  $PR$  and the order of analysis  $M$  were manually varied when necessary, in order to ensure that the first four formants in the vocalic steady-state were both well-defined (as observed on the spectrogram-like display of LP poles) and reasonably consistent across repetitions (as observed on the formant sequence chart which was prepared prior to each round of examinations). However, it is not usual practice, either in speech or speaker recognition systems or in the research laboratory, to obtain the cepstrum with analysis conditions which vary from one utterance to another for the sake of measuring the best formants. In order to maintain consistency in spectral representation across vowels, speakers, and repetitions of utterances, the default LP analysis conditions ( $M = 14$ ,  $PR = 0.98$ ) were therefore used to obtain 14 cepstral coefficients  $\{c_1, c_2, \dots, c_{14}\}$  per steady-state frame.

### **3.4 Acoustic Data used to Validate the Dichotomy**

In contrast with the acoustic data selected to first *unfold* the nature of vowel-speaker interactions, the two sets of *validating* data (PB and JB) embody a larger population of speakers, at the expense of a slightly reduced phonetic complexity. Indeed, these datasets comprise effectively two repetitions of vowels in /hVd/ context, spoken by 32 adult male speakers of American English, and 36 adult male speakers of Australian English, respectively. Furthermore, they each are in existence today (so far as the present study is concerned) in terms of a single formant-pattern measured at the steady-state of each vocalic nucleus. As summarised earlier in Figure 3.1, the two studies (Peterson and Barney (1952) and Bernard (1967), respectively) differ slightly in their method of formant frequency estimation; those methods will be reviewed in Section 3.4.1.

Subsequently, in Section 3.4.2, we will describe our method of obtaining so-called *simplified* LP cepstrum parameters from the available formant data, such that our investigations of vowel-speaker interactions manifest in the FC dataset can be paralleled using the PB and JB datasets, despite the unavailability of the actual speech waveforms of the latter.

### **3.4.1 Formant Estimation**

In the following two sections we briefly review the methods used by Peterson and Barney (1952) and by Bernard (1967), respectively, to obtain the invaluable sets of vocalic steady-state formant data which play a validating role in subsequent chapters of this thesis.

#### **3.4.1.1 Amplitude-Section Approach**

The approach adopted by Peterson and Barney (1952) to estimate the frequency locations of the first three formants of the recorded American English vowel sounds, involved first the identification of the “practically steady state” interval of each vowel, by visual examination of a broad-band spectrogram prepared (using a sound spectrograph) for each monosyllabic /hVd/ word. Lacking a more complete description of their procedure for selecting the steady-state portion of each vowel nucleus, we only know that they selected an interval between the influence of the initial /h/ (which presumably is quite negligible anyway) and the influence of the final /d/ consonant. Although they do not refer to any difficulties associated with non-steady or diphthongised vowel nuclei, their sample spectrograms shown (Peterson and Barney, 1952, Figure 2, p.176) for one of the female speakers (not used in our present study) suggest that they may indeed have encountered apparently transitional vowel nuclei, with which they must have dealt in a partly subjective but presumably systematic manner.

A so-called amplitude section (or harmonic magnitude-spectrum) was obtained at the selected steady-state portion of each vowel nucleus, and calibrated both on its vertical (dB) and horizontal (Hz) axes. Guided by the method proposed by Potter and Steinberg (1950), the frequency of each formant was then obtained by computing an

amplitude-weighted average of the main harmonic components which make up that formant peak, as follows:

$$F = \frac{\sum_i w_i f_i}{\sum_i w_i}, \quad (3.4)$$

where  $f_i$  is the frequency of the  $i^{\text{th}}$  spectral harmonic component, and  $w_i$  is its amplitude normalised with respect to the amplitude of the dominant component. It is interesting to note that Potter and Steinberg's (1950) approximation of the formant centre-frequency according to Equation 3.4, arose from their speculations regarding the human *perceptual* mechanisms in dealing with formant-like concentrations of spectral energy, rather than from any consideration of how those spectral components are shaped by the human speech *production* mechanism.

The main problems associated with the amplitude-section approach to formant estimation, include the often-encountered asymmetry in the amplitudes of spectral harmonic components of a formant, and the insufficient density of spectral harmonic components required to resolve a formant. The former problem was effectively dealt with by computing the weighted-average of spectral components (Equation 3.4), thus overcoming the ambiguity which might exist in having to select only a single component. The latter problem mainly occurred with high fundamental frequencies of voicing, as would be expected of the majority of adult female and child speakers in the group; one might therefore safely assume that the problem was not as prevalent for the subset of adult male speakers whose data will be used in our present study.

#### **3.4.1.2 Spectrogram Approach**

In contrast with the amplitude-section approach just reviewed, Bernard's (1967) formant measurements relied mainly on spectrograms of the /hVd/ words, which he prepared using a model Kay spectrograph. Thus, his visual acuity, manual dexterity, and acoustic-phonetic knowledge were all brought to bear on the task of tracing the first three formant trajectories by hand, and estimating their frequencies at the onset, target, and offset of each vocalic nucleus.

In that context, Bernard (1967, 1989) distinguishes at least two types of acoustic

realisations of the nominal monophthongs recorded. In the first type, the formant trajectories (those of  $F_1$  and  $F_2$  in particular) were found to conform with the ideal notion of a single, “well-defined target” (Bernard, 1989, p.191), located within the temporal region demarcated on the left by the onset of the vocalic nucleus (with or without a so-called onglide), and on the right by the start of the so-called offglide where the effect of the final /d/ consonant starts to bend particularly the  $F_2$  trajectory towards an offset value. In those cases, the measured formant-pattern of interest in our present study is that which Bernard took at a position mid-way through the steady-state of what he appropriately calls a “one-target” sound. In the second, so-called “two-target” type of sounds (which were found to occur predominantly for /i/ and /u/), he recorded formant measurements taken at the onset, the offset, and at the centres of each of two targets thus identified. For convenience, “only the F-patterns he estimated at the second target were taken as representative of the monophthongs realised with diphthongal quality” (Clermont, 1996, p.146), and thence used in our present study.

### **3.4.2 Simplified Cepstrum**

If we are to parallel our forthcoming unfolding of vowel-speaker interactions with similar investigations using the validating datasets (PB and JB), we therefore require (in addition to the formants) an acoustic parameterisation of those datasets in terms of whole-spectrum information which is carried by the cepstrum. However, as the original speech waveforms of the PB and JB datasets are either unavailable or would require inordinate efforts to acquire and segment in the same manner as in the original studies, we chose to make full use of the available formant data by creating from them corresponding sets of LP cepstra. By analogy with the notion of “simplified spectrum” put forward by Clermont (1991, p.110) in the context of his spectral-matching approach to formant estimation, an LP cepstrum thus obtained, may also be referred to as *simplified* because it embodies spectral-shape information derived only from (a limited number of) formants, and is therefore free of the influence of so-called spurious poles which usually accompany the true formant poles yielded by LP analysis of a frame of voiced speech data.

One way of obtaining the so-called simplified LP cepstrum from formant

measurements is first to convert the formants to LP autoregressive coefficients using the reverse conversion algorithm “FB2A” (ILS, 1983), which recursively computes the coefficients of the LP inverse polynomial by expanding the product of terms each corresponding to a resonance pole (Markel and Gray, 1976, p.7). The autoregressive coefficients can then be converted to LP cepstrum coefficients using the well-known recursive algorithm “A2CP” (Markel and Gray, 1976, p.230). The data processing involved in this method remains within the confines of the LP model, merely transforming between various LP parametric representations (i.e., from the subset of formant poles, to the autoregressive coefficients, to the cepstrum) of a simplified, all-pole spectrum.

An alternative but equivalent method of transforming formants to cepstra is the following, direct transformation (Itahashi, 1984, 1988):

$$c_k = \frac{2}{k} \sum_n e^{-\frac{k\pi B_n}{F_s}} \cos\left(\frac{2k\pi F_n}{F_s}\right), \quad (3.5)$$

where  $F_n$  and  $B_n$  are, respectively, the frequency and bandwidth of the  $n^{\text{th}}$  pole,  $F_s$  is the assumed sampling frequency, and  $c_k$  is the  $k^{\text{th}}$  cepstral coefficient. This formulation, which can be derived from the very definition of the cepstrum in terms of a minimum-phase, all-pole model (Oppenheim et al., 1968), provides some useful insights which might be gained only indirectly from the LP-based recursive routines. In particular, the decaying-exponential factor in Equation 3.5 suggests that the contribution of a broad-bandwidth pole to any of the cepstral coefficients is proportionately diminished in comparison with the contribution of a narrow-bandwidth pole with the same centre-frequency. Hence, the so-called spurious poles which are usually of much wider bandwidth than their true formant counterparts, already have a much smaller influence on the cepstrum obtained by LP analysis of a frame of voiced speech having clear formant structure.

Nevertheless, a so-called simplified cepstrum obtained using either Equation 3.5 or the LP recursive routines, is completely stripped of influence not only from the spurious poles, but also from the somewhat difficult-to-estimate higher formants which may still lie within the frequency-band extending to the half-sampling frequency.

In this regard, it may be noted that, using either of the two methods, the required parameter  $F_s$  can be chosen rather arbitrarily, subject only to the theoretical constraint that it exceed twice the frequency of the highest formant considered. However, in order to maintain consistency with the LP cepstra obtained from the FC dataset (as described earlier in Section 3.3.3), the sampling frequency is fixed at  $F_s = 10$  kHz when mapping the three available formants of the PB and JB datasets to corresponding sets of cepstra.

It may also be noted that formant bandwidths are required in the computation of the simplified cepstrum (using either of the two methods), as might intuitively be expected when the spectral shape is to be specified, and not just the frequency locations of the formant peaks. However, neither of the two classic datasets (PB and JB) include formant bandwidths, which are in any case notoriously difficult to estimate reliably. In this regard, Bogert's (1953) early attempt at measuring the formant bandwidths of 100 of the PB vowels recorded by the 33 male speakers, is certainly a landmark effort. On the harmonic amplitude spectra used by Peterson and Barney (1952) to obtain formant-frequency estimates (as described in Section 3.4.1.1), Bogert carefully drew a free-hand curve enveloping the amplitudes of the harmonic components of each formant peak, and thereby obtained an estimate of the bandwidth by measuring the width of the curve at 3 dB below its peak amplitude.

By replacing Bogert's free-hand curves with a number of template curves of different bandwidths obtained from an analytic relation, Dunn (1961) was able to improve on the accuracy of the bandwidth estimates, which he obtained on 400 samples of the PB vowels, comprising both repetitions of all ten /hVd/ words recorded by 20 of the male speakers. Notwithstanding the numerous sources of measurement error which include the sparse density of spectral harmonic components in each amplitude section, he was able to identify a likely trend. In a classic plot of per-vowel formant bandwidth versus formant frequency averaged over both repetitions and all 20 speakers, Dunn (1961, Fig.2, p.1739) suggested a piecewise-linear relationship on a logarithmic scale, with one line for each of the first three formants.

One way of supplementing the available formant frequencies of the PB and JB data with "synthetic" bandwidths prior to computing the cepstra, might therefore be to formulate an equation for each of the three linear relations suggested by Dunn, and thus

to determine the bandwidth of each formant on the basis of its centre-frequency. An alternative approach might be to adopt a fixed value for each of the first three formant bandwidths (regardless of formant frequency), based on the overall mean values of about 50 Hz, 64 Hz, and 115 Hz found by Dunn (1961) for the subset of the PB data. (Note that Fant's (1972) well-known equations which express each of the first three formant bandwidths in terms of combinations of the formant frequencies, and likewise Hawks and Miller's (1995) piecewise-polynomial expression which relates the bandwidth of a formant to its centre-frequency, are based on *closed-glottis* bandwidth data. In contrast, bandwidths measured by means of spectrographic analyses (Bogert, 1953; Dunn, 1961) or by LP analysis (the present study, as described earlier in Section 3.3.2) are based on frames of speech data which typically encompass at least two periods of the glottal cycle, and thus are presumably averaged over both open and closed glottis conditions).

Our final choice regarding the synthetic bandwidths required to convert the PB and JB sets of formant data into simplified cepstra, depends to a large extent on the nature of vowel-speaker interactions to be unfolded and later validated using those cepstra in Chapter 4, and on the robustness of our acoustic-phonetic findings in the face of varying formant bandwidth values. It would therefore seem pertinent to first assess the robustness of the acoustic-phonetic manifestations of vowel-speaker interactions using simplified cepstra generated from the FC dataset, for which we do have the original LP cepstra as a baseline for comparison. Indeed, it would be rather fortuitous to find those acoustic-phonetic manifestations relatively invariant despite the radical simplification of the cepstrum (and hence the spectrum) to carry information contained only in the first three formant frequencies.

### **3.5 Summarising Perspective**

In this chapter we described the three sets of recorded /hVd/ monosyllabic speech data, and the methods used to parameterise those data in terms of formants and cepstra measured at the vocalic steady-state (or acoustic vowel target). In the summarising perspective which follows, we shall briefly preview the role of the speech and speaker materials, and of the acoustic parameters derived from those datasets, in terms of their

usage in the forthcoming chapters of this thesis.

As summarised earlier in Figure 3.1, the FC dataset (Clermont, 1991) includes five repetitions of eleven vowels recorded by four adult, male speakers of Australian English, who have been determined (at least qualitatively, from a subjective, auditory-perceptual point of view) to form a sufficiently heterogeneous speaker group. The acoustic parameters measured from that dataset will therefore be used in Chapter 4 to *unfold* the phenomenon of vowel-speaker dichotomy. By contrast, the well-known and readily available PB dataset (Peterson and Barney, 1952) includes acoustic vowel data of 32 adult, male speakers of American English. Although it is limited to only two repetitions of each vowel per speaker, the more populous dataset offers the possibility of *validating* the phenomenon of dichotomy first unfolded using the data of only four speakers. However, as the dialectal heterogeneity of the PB dataset is known only qualitatively, we also consider the recently resurrected (Clermont, 1996) subset of data (Bernard, 1967) which includes effectively two repetitions of eleven vowels recorded by 14 Broad, 11 General, and 11 Cultivated, adult, male speakers of Australian English. Indeed, whilst the JB dataset will also play a validating role in Chapter 4, it provides the unique opportunity of separating the effects of *idiolectal* and so-called *intrinsic* speaker differences.

Acoustic parameterisation of the three datasets just reviewed, has yielded both a discrete and a continuous spectral representation in terms of the formant and the cepstral parameters, respectively. The salient features of those two types of spectral representation will be duly exploited in the following three chapters. First, as argued in our analysis of the literature (in Chapter 2), a whole-spectrum representation may lend itself more effectively to an investigation of the frequency-band dependence of vowel-speaker interactions. The LP cepstra obtained for the FC dataset, and later the simplified LP cepstra derived from the PB and JB formant data, will therefore play a vital role in Chapter 4, in unfolding and validating, respectively, the phenomenon of vowel-speaker interactions along the spectral continuum. The interpretive superiority of the formants will then be exploited in Chapter 4, in order to provide an acoustic-phonetic explanation of the phenomenon. Furthermore, the articulatory significance of the formants will be reasserted first in Chapter 5, where we present a new method of

vocal-tract shape parameterisation and estimation which maps the geometry of the vocal-tract from the formants, and then in Chapter 6, where we use those estimated vocal-tract shapes to provide an articulatory explanation of the phenomenon of vowel-speaker interactions, and of the speaker differences which underly those interactions.

## Chapter 4

### Acoustic-Phonetic Analysis of Speech-Speaker Dichotomy

#### 4.1 Introduction

Our review of the literature (in Chapter 2) has firmly established the importance and currency of the long-standing problem of speech-speaker dichotomy. Indeed, even for the relatively better understood, vocalic sounds of spoken language, we are still lacking a unified account of the separate influences of phonetic and speaker-specific attributes which are intertwined in the acoustic speech signal. A better understanding of the nature of those separate influences, and of their interactions, would have implications in both speech and speaker recognition by computer, and would also pave the way for more versatile, multi-speaker speech synthesis.

In this vein, the relatively more extensive literature on spoken vowel sounds is clearly replete with the time-honoured view that the most important, phonetic information is carried by the first two formants (or resonances of the vocal-tract)  $F_1$  and  $F_2$ . Not only are those parameters generally considered the best acoustic-phonetic descriptors of vowels, they are also strongly correlated with the traditional articulatory description of vowels in terms of the two dimensions tongue height and advancement, respectively. By contrast, there is mounting evidence in the literature that the third and higher formants, or indeed the spectral regions which encompass those higher resonances, carry more speaker-specific information, and therefore manifest greater amounts of inter-speaker variability.

If certain spectral regions of vowel sounds indeed contain predominantly either phonetic or speaker-specific information, then it would seem possible to observe this phenomenon in its duality, if machine classification of these sounds were attempted on an intra- and an inter-speaker basis. However, the issue of spectral manifestations of

vowel and speaker variability, raises the equally-important question of spectral representation. Despite the inherent difficulties in measuring the formant frequencies, they still hold a place of pride in acoustic-phonetics, and are perhaps the most articulatorily relevant, acoustic parameters. On the other hand, the cepstrum is easily obtained from the acoustic speech signal, and is perhaps better suited to an investigation of the frequency-band dependence of the vowel-speaker dichotomy, owing to its whole-spectrum representation. However, new methods of cepstrum and cepstral distance computation are required in order to enable the selection of frequency sub-bands in cepstrum-based vowel classification. In Section 4.2 we elaborate on the methodological aspects of our acoustic-phonetic investigations, which indeed use the LP cepstrum in vowel classification experiments designed to contrast the manifestations of phonetic and speaker influences along the spectral continuum.

In Section 4.3 we then apply our methodology to unfold the vowel-speaker dichotomy using the FC dataset described in Chapter 3. In our attempt to unfold the phenomenon with a greater certainty of having suppressed external or artefactual influences, we shall compare the results obtained using classifiers with hyper-planar (in Section 4.3.1) and hyper-quadratic (in Section 4.3.2) decision boundaries. However, in order to retain only the intrinsic manifestations of the vowel-speaker dichotomy, we must also address the related issues of data sufficiency and cepstrum dimensionality to which the latter, more sophisticated classifier is particularly sensitive.

Having thus unfolded the phenomenon of dichotomy using the ensemble of methods centred around the cepstrum, in Section 4.4 we then turn to the interpretively more superior formant parameters to provide an acoustic-phonetic explanation of the dichotomy. Indeed, the question of which vocalic, acoustic-phonetic subspaces are the most heavily influenced by vowel-speaker interactions, is best approached by interpreting the cepstrum-based classification results in terms of the speakers' vowel formant distribution.

Whilst the FC dataset provides sufficient phonetic richness and speaker differences to unfold the dichotomy, the question remains whether phonetic-speaker interactions will be equally manifest in a more populous, and perhaps more homogeneous dataset. As explained in Chapter 3, the PB dataset allows the possibility (in Section 4.4.4) of

validating the dichotomy from the point of view of speaker homogeneity. However, in order to be confident in applying our methodology to the so-called simplified cepstra generated from the formants of the PB dataset, we first examine (in Section 4.4.3) the dependence of the dichotomy on the simplified spectral representation using the FC dataset.

As the FC and PB datasets are known to contain either idiolectal or dialectal speaker differences, the further question arises whether the dichotomy is wholly reliant on that type of speaker heterogeneity, or whether it is also manifest when only within-idiolect, or so-called intrinsic speaker differences are present. In Section 4.4.5 we address that question using the JB dataset (as described in Chapter 3), in terms of (perceived) idiolectal homogeneity amongst adult, male speakers of Australian English.

Finally, in Section 4.5 we conclude with a summary of our acoustic-phonetic findings in regard to the vowel-speaker dichotomy, and point forward to our forthcoming, articulatory investigations.

## **4.2 Methodology for Unfolding the Dichotomy**

As described in Chapter 3, the dataset which is destined to unfold the phenomenon of vowel-speaker dichotomy comprises five repetitions of eleven vowels recorded in citation-form /hVd/ context by four adult, male, native speakers of Australian English (Clermont, 1991). The seven most steady-state, consecutive frames were carefully identified within each vowel nucleus according to criteria aimed at obtaining the best possible trajectories of the first four formants. Those seven frames were then re-analysed to obtain the 14<sup>th</sup>-order LP cepstrum parameters which we shall use to unfold the dichotomy, with the aid of the ensemble of methods herein described.

In Section 4.2.1 we provide an outline of our general approach, which consists of using the LP cepstra in vowel classification experiments designed to elicit the consequences of vowel and speaker influences across the spectral continuum. In Section 4.2.2 we provide justifications for considering the formant-enhanced spectral representation afforded by the quefrequency-weighted cepstrum (QCEP), and motivate the use of simplified cepstra derived from the FC formant data. In Section 4.2.3 we describe the two methods of vowel classification used in this chapter, which yield

hyper-planar and hyper-quadratic decision boundaries, respectively. In order to accommodate frequency sub-band selection in the classification experiments, new formulations of cepstral distance computation and cepstrum transformation are derived, which offer the flexibility of observing the behaviour of classification accuracy as a function of an increasing spectral range.

### **4.2.1 Approach**

In formulating an approach to unfold the vowel-speaker dichotomy, we must address two principal requirements. By definition, we require to contrast the acoustic manifestations of phonetic and inter-speaker differences. At the same time, we require to interrogate the frequency-band dependence of those differences. As presently described, our approach to unfolding the acoustic manifestations of vowel-speaker interactions, hinges on automatic (computer) vowel classification experiments designed specifically to lay bare those interactions along the spectral continuum.

First, the inherent duality of the vowel-speaker problem is heeded by contrasting the performance of vowel classification carried out alternately on an intra-speaker (or speaker-dependent), and on an inter-speaker (or speaker-independent) basis. As the phonetic dimension is present in both types of experiment, the vowel classification task is of equal relevance. However, whilst the former type of experiment is oblivious to inter-speaker differences and therefore embodies principally the phonetic dimension unimpeded, the latter type of experiment explicitly introduces the speaker dimension, and thus involves the more error-prone task of phonetic discrimination in the presence of inter-speaker variability. In that regard, it is generally accepted that, at least for normal, adult male speakers of sufficiently similar dialects of a given language, the acoustic-phonetic manifestations of inter-speaker differences are generally of a smaller magnitude than those of inter-vowel differences for a given speaker. Variability induced by inter-speaker differences can therefore be regarded as contaminating the acoustic vowel space, thereby degrading the performance of vowel classification according to the degree of manifestation of those differences. The contrast in performance of vowel classification first in the absence, and then in the presence of inter-speaker variability, can therefore be expected to shed some light on the relative strength of vowel and

speaker influences.

However, the question of the frequency-band dependence of those influences remains unanswered pending a method to contrast classification performance in various spectral regions. In that context, our analysis of the literature (in Chapter 2) has already suggested the potentially contrasting roles of the lower spectral regions which encompass the normal range of phonetic variation in the first two formants, and the less phonetically important (and perhaps more speaker-specific) higher spectral regions which contain the third and higher formants. If the acoustic manifestations of speaker differences are indeed more potent in the so-called higher than in the lower spectral regions, one might expect to observe their effect on vowel classification more clearly by initially considering only the lower, phonetically-relevant frequency bands, then using progressively greater amounts of spectral information recruited from the higher frequency bands.

Sadly, the traditional approach to frequency-band selection either by using filter-banks or by re-analysing or re-sampling the speech waveform, leaves much to be desired in terms of both flexibility and consistency in spectral representation. Indeed, the popular cepstral distance measures used in speech processing and recognition systems have to date been formulated with the inflexible constraint of comparing frames of speech over the entire available spectral range extending to half the sampling frequency. We shall attempt to remedy this deficiency in Section 4.2.3, where we propose a new transformation of LP cepstra, and a new formulation of a cepstral distance measure, both of which are expressed parametrically in terms of a lower and an upper spectral limit which demarcate any selected frequency band within the available spectral range. This new approach to cepstrum derivation and cepstral distance computation does provide the flexibility of performing vowel classification using spectral information contained in selected frequency sub-bands, without the restrictions normally imposed by a fixed bank of bandpass filters, and without the need to re-analyse or re-sample the speech waveform.

To recapitulate, the acoustic manifestations of vowel-speaker dichotomy will be unfolded by means of vowel classification experiments, which at once aim to resolve the separate influences of phonetic and speaker-specific attributes of spoken vowels, and to

determine the frequency-band dependence of those attributes and their interactions. In particular, the contrast between the performance of intra- and inter-speaker vowel classification will provide a basis for assessing the relative degree of acoustic manifestation of vowel and speaker influences. Determination of the frequency-band dependence of those influences will be facilitated by new formulations of cepstrum derivation and cepstral distance computation, which allow a more flexible approach to scanning the available spectral range incrementally from low to higher frequency bands.

### **4.2.2 Spectral Representations**

An important step which precedes the classification process and is well-known to play a major role in its performance, is the acoustic parameterisation which yields a particular type of spectral representation to be used in the classifier. In this vein, the cepstrum is widely regarded as one of the foremost parameters in state-of-the-art spoken language systems. Not only is the LP cepstrum obtained from the speech waveform robustly and with relatively low computational cost, it has also been found to outperform many alternative spectral representations in both speech and speaker recognition tasks. In Chapter 3 we gave a complete description of our methods of obtaining LP cepstra in the vocalic steady-state of /hVd/ monosyllables recorded by four speakers of Australian English (the FC dataset). In view of our planned experiments of vowel classification using those cepstra, we now re-examine some of our assumptions regarding analysis conditions and type of cepstral representation.

First, despite the variable analysis conditions adopted by necessity during the formant estimation procedure, the LP analysis order and pre-emphasis coefficient were subsequently fixed ( $M = 14$ ,  $PR = 0.98$ ) when obtaining the cepstrum in each frame, in order to secure consistency in spectral representation across repetitions, vowels, and speakers. A fixed order of analysis maintains consistency in the accuracy of spectral-shape information carried by the LP cepstrum, since potentially the same number of poles (a maximum of seven complex-conjugate poles) are made available to model the short-time spectral envelope of each vocalic frame. The order of analysis was chosen on the basis of guidelines put forward by Markel and Gray (1976, pp.154-156), who suggested that for voiced speech, a value equal to the sampling frequency in kHz would

offer most likely a sufficient number of poles to model the expected number of formant peaks in the spectrum, and that an additional number of terms (between two and five) should be added to model remaining spectral-shape characteristics, such as the spectral slope which results from the combined effects of glottal and lip radiation. Thus, we added four to the value of  $F_s = 10$  kHz to arrive at a reasonable order of analysis  $M = 14$ .

The next question concerns the dimensionality of the cepstrum itself. By definition, an exact representation of the LP spectrum is only obtained by having recourse to an infinite number of cepstral coefficients. Truncation of the cepstrum to a finite number of coefficients yields a smoothed representation of the all-pole spectrum. Nevertheless, Gray and Markel (1976) have shown that distances yielded by a cepstrum-based distance measure have a significantly high correlation (0.98) with root-mean-square (rms) distances between LP log-magnitude spectra computed over a large number of frames of speech, when the number of cepstral coefficients retained ( $NCC$ ) is equal to the analysis order  $M$ . This finding is of particular importance in recognition systems where the dimensionality of the feature space is required to be as small as possible, because it implies that the low-order cepstral coefficients retain the essential components of spectral-shape information. Indeed, state-of-the-art speech and speaker recognition systems do take advantage of this property and generally use only the low-order cepstrum (often supplemented with derivative parameters such as the delta-cepstrum). We are therefore justified in using the first  $NCC = M = 14$  cepstral coefficients  $\{c_1, c_2, \dots, c_{14}\}$ , as indicated earlier in Section 3.3.3, to represent each frame of vocalic speech data in our vowel classification experiments.

As also indicated in Chapter 3, the Negative Derivative of the LP Phase Spectrum (NDPS) offers an enhanced form of spectral representation which emphasises the formant peaks and minimises the interactions between neighbouring peaks. Those properties are captured by the root-power-sum (RPS) or the quefrency-weighted cepstrum (QCEP), where each cepstral coefficient is simply multiplied by its index (or quefrency) to yield the sequence  $kc_k$ . Whilst quefrency-weighting of LP cepstral coefficients has also been shown to improve their performance in vowel classification (Paliwal, 1982), our primary motivation for adopting the formant-enhanced spectral

representation afforded by the QCEP is to accentuate the acoustic-phonetic (and therefore perhaps the articulatory) relevance of the already quite powerful LP cepstrum.

In this vein, it is also of relevance to consider the *simplified* cepstrum (cf. Section 3.4.2), which carries only formant information, and is therefore cleansed of the influence of spurious and other spectral-shaping poles. In Chapter 3 we discussed the simplified cepstrum in the context of generating cepstral data for the PB and JB datasets which currently exist only in terms of the first three formant frequencies. Prior to performing vowel classification experiments on those datasets, it would indeed seem pertinent to first validate the results obtained with the original cepstra of the FC dataset, by using simplified cepstra generated from the measured formants.

### 4.2.3 Classification Methods and Distance Formulations

We next describe the vowel classification process itself, which uses the quefrequency-weighted LP cepstrum (QCEP) parameters discussed previously. We proceed from a discussion of data-partitioning strategies, to the classification rule, and finally to the two methods of classification with their respective entourages of cepstral distance formulations and methodological peculiarities.

A pattern-recognition approach is adopted, whereby the available data are divided into a *training* set which is used to form *prototypes* for each class (vowel), and a *test* set which is then used to arrive at an estimate of the probability of error in classification. There exist many such data-partitioning schemes, an excellent review of which is given by Toussaint (1974). Perhaps the simplest is the so-called re-substitution (or R) method where all of the available data are used for both training and testing. However, theory confirms intuition that the R-method provides an overly optimistic estimate of classification error, which therefore approaches the so-called true error-rate from below. On the other hand, estimates of misclassification yielded by the so-called holdout (or H) method have been shown to be overly pessimistic (thus approaching the true error-rate from above), as the available data are divided only once into mutually exclusive sets for training and testing, thereby stretching the capability of the classifier to generalise beyond the training data. A more reliable version of the H-method consists of averaging the estimates of misclassification obtained by randomly partitioning the

data several times into pairs of equal-sized training and test sets. However, the averaged estimate is known to be still overly pessimistic.

A less biased approach is given by the leave-one-out (or U) method, where a single pattern sample is withheld from the training data and used for testing, and the process is repeated over all such data-partitions to arrive at an average error-rate. A closely related and more general form of the U-method is the so-called rotation (or  $\Pi$ ) method, where a small subset (greater than or equal to a single sample, but where the ratio of total samples to left-out samples is an integer) of the available data is withheld from training and used for testing, and the error-rate is again determined by rotating the data and averaging the results.

In view of the reduction in bias achieved by the U-method and the  $\Pi$ -method, which therefore more closely approach the true rate of misclassification, we shall adopt a variant of these approaches to data-partitioning. In particular, our requirement to contrast the performance of vowel classification on an intra- and an inter-speaker basis, leads us to adopt a “leave-one-out” approach where the data of *one repetition at a time* are withheld in intra-speaker experiments, and the data of *one speaker at a time* are withheld in inter-speaker experiments. Thus for the FC dataset, intra-speaker vowel classification accuracy is averaged over five experiments per speaker (before being averaged over all four speakers’ results), whilst inter-speaker vowel classification accuracy is averaged over four experiments. This would appear to be the best approach to obtaining minimally-biased estimates of classification accuracy, while at the same time conforming to our methodological requirements aimed at unfolding the vowel-speaker dichotomy.

Having decided on a method of data-partitioning, we now turn to the classifier itself. In order to place the vowel-speaker dichotomy on a more solid footing, we consider two classifiers which differ mainly in the level of complexity used to represent each vowel prototype. In particular, the simpler of the two classifiers computes vowel prototypes using only first-order statistics (the mean) of the training data, whilst the more sophisticated classifier uses both first- and second-order statistics (the mean and covariance). As a result, the two classifiers may be distinguished on the basis of their decision boundaries: they use effectively *hyper-planar* and *hyper-quadratic* decision

boundaries, respectively.

A common feature of both classifiers is the well-known *minimum-distance classification rule*, whereby distances are computed from a given test sample to each of the class prototypes, and the test sample is then assigned to that class from whose prototype it was found to have the minimum distance. Implicit in this approach to classification is the requirement of a distance measure. Indeed, our plan to investigate the frequency-band dependence of vowel and speaker influences does require a method to compute the distance between two given frames of speech, not only across the entire available spectral range as implied in cepstral distance measures available to date, but within any selected frequency sub-band which lies in that range.

In the following two sections, we therefore describe the two classifiers which we later use to unfold the vowel-speaker dichotomy. In Section 4.2.3.1 we describe the classifier which uses hyper-planar decision boundaries, together with our new, parametric cepstral distance measure which allows selection of any frequency sub-band without requiring to re-compute the cepstrum. In Section 4.2.3.2 we then describe the classifier which uses hyper-quadratic decision boundaries, together with our new method of cepstrum computation which models the spectrum within any selected frequency sub-band, and which therefore allows interpretation of classification results in terms of those selected spectral regions.

#### 4.2.3.1 Classifier based on Hyper-Planar Decision Boundaries

The classifier first adopted to unfold the dichotomy, uses the training data to compute a single, per-class prototype which is the mean (or centroid) of the LP cepstrum vectors belonging to that class (vowel). For  $N_v$  vowels and  $S$  training samples per vowel, the  $i^{\text{th}}$  vowel prototype is therefore computed as follows:

$$\bar{\mathbf{c}}_i = \frac{1}{S} \sum_{l=1}^S \mathbf{c}_{il}, \quad i = \{1, 2, \dots, N_v\}, \quad (4.1)$$

where  $\mathbf{c}$  denotes an LP cepstrum vector with  $NCC = M = 14$  elements. Each cepstrum vector in the test data is then assigned to that class whose prototype is the “closest” (i.e., yields the smallest distance), and a correct or a false classification is tallied according to whether or not, respectively, the assigned class is the same as that of the

sample being tested. When the minimum-distance classification rule is thus employed with vowel prototypes constructed from only first-order statistics, the resulting decision boundaries which separate the classes are well-known to be hyper-planar (Duda and Hart, 1973; Tou and Gonzalez, 1974). (For a two-class problem, the hyper-plane is simply the perpendicular bisector of the line segment joining the two class means.)

The distance measure used to compare a given test sample with each vowel prototype, is based on the formant-enhanced NDPS representation given by the QCEP (cf. Section 4.2.2). We have already discussed the advantages of this spectral representation in Chapter 3, where we capitalised on the formant-sensitivity property of Yegnanarayana and Reddy's (1979) NDPS cepstral distance measure in our definition of inter-frame variance, used successfully to detect the vocalic steady-states of the recorded utterances which comprise the FC dataset. However, as with all cepstral distances proposed to date, the measure of spectral similarity is integrated effectively over the entire available spectral range which extends to half the sampling frequency (cf. Equation 3.1 in Chapter 3). In order to perform vowel classification using spectral information only within selected frequency sub-bands, direct selection of a sub-band in the distance measure itself would be of tremendous advantage both in terms of computational efficiency and consistency in spectral representation, as it would circumvent both a return to the speech waveform and a re-analysis with potentially different conditions.

Indeed, we have proposed a *parametric cepstral distance measure* (Clermont and Mokhtari, 1994) which satisfies the criteria just outlined, and thus affords greater flexibility in our search for the frequency-band dependence of the vowel-speaker dichotomy. The mathematical derivation of the distance measure begins by replacing the fixed limits of the integral in Equation 3.1, with the normalised-frequency variables  $\theta_1$  and  $\theta_2$  which represent, respectively, the lower and the upper limit of a selected frequency band within the entire spectral range  $[0, \pi]$ . As the NDPS itself is expressed in terms of a discrete cosine transform of the quefrequency-weighted LP cepstrum:

$$-\frac{d\phi(e^{j\theta})}{d\theta} \approx \sum_{k=1}^{NCC} kc_k \cos(k\theta), \quad (4.2)$$

(where the approximation tends to an exact equality as  $NCC \rightarrow \infty$ ), modification of the limits of integration in Equation 3.1 and substitution of Equation 4.2 yields the following expression:

$$d_{\text{QCEP}}^2(\mathbf{c}, \mathbf{c}', \theta_1, \theta_2) = \frac{1}{(\theta_2 - \theta_1)} \int_{\theta_1}^{\theta_2} \left[ \sum_{k=1}^{NCC} k(c_k - c'_k) \cos(k\theta) \right]^2 d\theta. \quad (4.3)$$

Owing to the parametric limits of integration, Parseval's theorem is rendered generally inapplicable, and Equation 4.3 is consequently decomposed then integrated, as follows:

$$d_{\text{QCEP}}^2(\theta_1, \theta_2) = \frac{1}{(\theta_2 - \theta_1)} [A(\theta_1, \theta_2) + B(\theta_1, \theta_2)], \quad (4.4)$$

where:

$$A(\theta_1, \theta_2) = \int_{\theta_1}^{\theta_2} \sum_{k=1}^{NCC} [k(c_k - c'_k) \cos(k\theta)]^2 d\theta, \quad (4.5)$$

$$B(\theta_1, \theta_2) = \int_{\theta_1}^{\theta_2} \left[ 2 \sum_{k=1}^{NCC-1} \sum_{l=k+1}^{NCC} kl(c_k - c'_k)(c_l - c'_l) \cos(k\theta) \cos(l\theta) \right] d\theta, \quad (4.6)$$

and where it is understood that the distance is computed between the two cepstral vectors  $\mathbf{c}$  and  $\mathbf{c}'$ . The results of carrying out the integrations in Equation 4.5 and 4.6, which themselves are the additive components of the parametric distance in Equation 4.4, are summarised as follows:

$$A(\theta_1, \theta_2) = \sum_{k=1}^{NCC} \alpha_k [k(c_k - c'_k)]^2, \quad (4.7)$$

where:

$$\alpha_k(\theta_1, \theta_2) = \frac{(\theta_2 - \theta_1)}{2} + \frac{\sin(2k\theta_2) - \sin(2k\theta_1)}{4k}, \quad (4.8)$$

and

$$B(\theta_1, \theta_2) = \sum_{k=1}^{NCC-1} \sum_{l=k+1}^{NCC} \beta_{kl} [kl(c_k - c'_k)(c_l - c'_l)], \quad (4.9)$$

where:

$$\beta_{kl}(\theta_1, \theta_2) = \frac{\sin((k-l)\theta_2) - \sin((k-l)\theta_1)}{(k-l)} + \frac{\sin((k+l)\theta_2) - \sin((k+l)\theta_1)}{(k+l)}. \quad (4.10)$$

Equations 4.4, and 4.7 through 4.10 describe a new, parametric cepstral distance measure (PCD) which at once takes advantage of the formant-enhancement property of the NDPS, and allows selection of any frequency sub-band  $[\theta_1, \theta_2]$  within the available spectral range.

Our formulation of a frequency-band selective cepstral distance has a number of salient features. First, it is computationally more efficient than having to re-analyse the speech waveform for every selection of a new frequency band. Second, it is more flexible than the traditional filter-bank approach to frequency sub-band analysis, as the only constraint on allowed values of the parameters which demarcate the selected spectral region is that they lie within the available spectral range extending to half the sampling frequency, i.e.,  $0 \leq \theta_1 < \theta_2 \leq \pi$ . Third, consistency in spectral representation is strictly maintained, as the mean-square difference is computed between the same two, cepstrally-smoothed NDPS, irrespective of the selected frequency band. This last property contrasts, for example, with the inevitably inconsistent spectral representation which would result either by re-sampling the speech waveform and re-analysing over a low-pass or a band-pass frequency range with the same or a different LP order of analysis, or by applying selective linear-prediction analysis (Makhoul, 1975) over the desired frequency band. By contrast with those methods, our parametric cepstral distance measure simply re-uses the cepstrum, which itself is computed only once, defined up to half the sampling frequency, and obtained with a fixed LP order of analysis  $M$ .

In Figure 4.1 we illustrate the PCD as it operates on selected frequency sub-bands of a pair of speech spectra measured in frames 60msec apart, in the diphthong nucleus of the monosyllabic word “hoy” recorded by speaker D of the FC dataset. The exact NDPS of those two frames are shown superimposed in the top graph, and reveal substantial differences in the frequency of the third formant  $F_3$ , and in the frequency of the fifth formant  $F_5$ . By contrast, the first, second, and fourth formants appear to have only smaller differences in their centre-frequencies, but larger differences in their bandwidths, which are known to be approximately (inversely) correlated with the heights of the formant peaks in the NDPS representation. As the PCD operates on the cepstrally-smoothed (rather than the exact) spectra, those are shown superimposed in

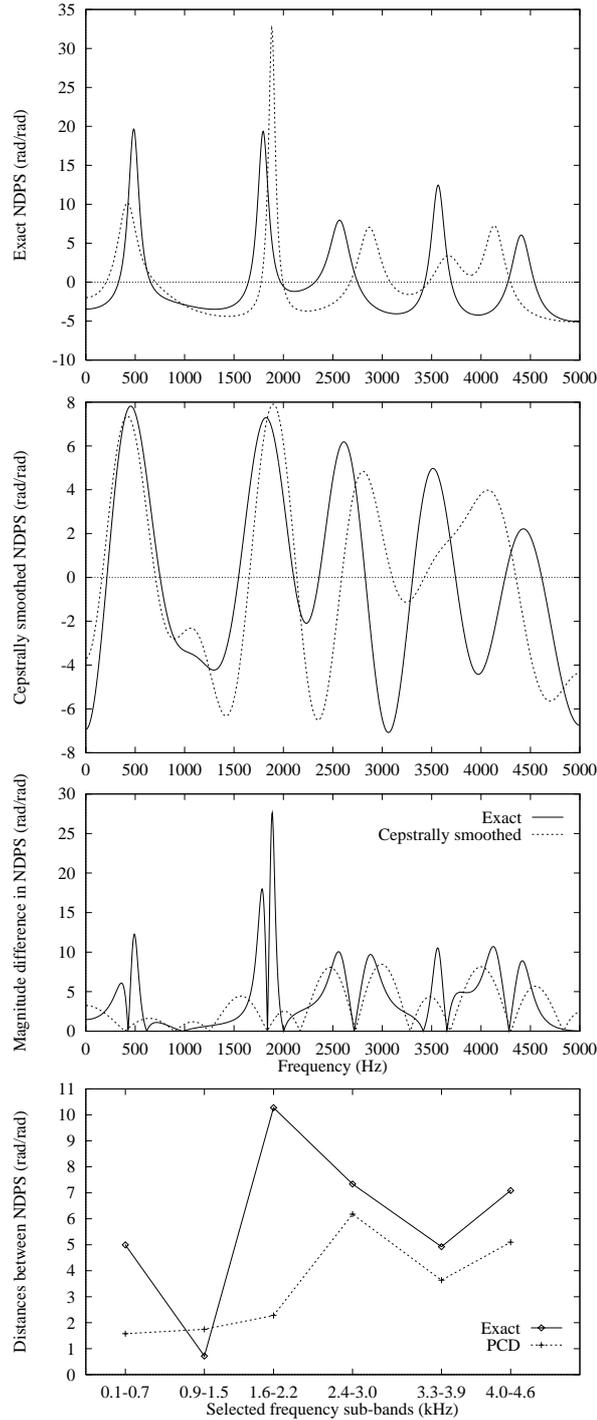


Figure 4.1: Illustration of *parametric cepstral distance measure* (PCD), applied to a pair of speech spectra measured in frames 60msec apart, in the diphthong nucleus of the monosyllabic word “hoy” recorded by speaker D of the FC dataset. *Top graph*: the two, exact NDPS, obtained via FFT operations on the LP autoregressive coefficients. *Second graph from top*: the two, cepstrally-smoothed NDPS, obtained by cosine-expansion of the truncated cepstral series ( $NCC=M=14$ ). *Third graph from top*: the magnitude of the differences between the pairs of exact and cepstrally-smoothed NDPS. *Bottom graph*: profiles of distances computed within selected spectral ranges — Euclidean distance between exact NDPS (diamond symbols joined by solid lines); PCD (plus symbols joined by dashed lines).

the second graph from the top. The effect of cepstral truncation to  $NCC = M = 14$  terms is clearly seen to normalise the differences observed in the formant bandwidths, and to retain the differences where they were observed to occur between the formant centre-frequencies. This observation is confirmed in the third graph from the top, which shows superimposed the magnitude of the differences between the exact pair of NDPS (solid curve), and between the cepstrally-smoothed pair of NDPS (dashed curve) — clearly, the former is more sensitive to the bandwidth-related difference in the range of the second formant. Finally, in the bottom graph we show the Euclidean distances computed within selected frequency sub-bands between the exact pair of NDPS (diamond symbols joined by solid lines), and the distances computed using the PCD in the same, selected frequency sub-bands (plus symbols joined by dashed lines). The large discrepancy between the two distances computed in the third sub-band (1.6–2.2 kHz) clearly illustrates the much lower sensitivity of the PCD to differences in formant bandwidths. Indeed, the largest distances yielded by the PCD occur in the two frequency sub-bands which encompass the third formants (in the range 2.4–3.0 kHz) and the fifth formants (in the range 4.0–4.6 kHz), respectively, which earlier were observed to exhibit the largest differences in their centre-frequencies.

The salient features of the PCD illustrated above therefore allow us to confidently pursue the question of frequency-band dependence of vowel and speaker influences, by appropriate selection of parameter values in our cepstral distance measure used in vowel classification experiments. In particular, our approach to scanning the available spectral range will be to fix the lower spectral limit at zero Hz ( $\theta_1 = 0$ ), and to increase the upper spectral limit up to the half-sampling frequency ( $\theta_2 = \pi$ ) in steps of 20 Hz. A complete set of intra- and inter-speaker vowel classification experiments will be performed at each step, and the accuracy profiles then plotted with the value of  $\theta_2$  on the abscissa. It is indeed the behaviour of classification accuracy as a function of the upper spectral limit with which we are primarily concerned in Section 4.3, as we unfold the vowel-speaker dichotomy.

#### **4.2.3.2 Classifier based on Hyper-Quadratic Decision Boundaries**

In order to gain more confidence in the behaviour of classification accuracy obtained

using the classifier based on hyper-planar decision boundaries, we also consider a more sophisticated classifier which is founded on the principles of Bayes decision theory. Both first- and second-order statistics are brought to bear in the computation of each vowel prototype, which comprises the mean and the covariance of the LP cepstra of that vowel in the training data. Assuming  $N_v$  vowels and  $S$  training samples per vowel, the  $i^{\text{th}}$  vowel prototype is therefore specified by the mean cepstral vector  $\bar{\mathbf{c}}_i$  given earlier in Equation 4.1, and by the following estimate of the  $NCC \times NCC$ , symmetric covariance matrix:

$$\mathbf{C}_i = \frac{1}{S} \sum_{l=1}^S (\mathbf{c}_{il} - \bar{\mathbf{c}}_i)(\mathbf{c}_{il} - \bar{\mathbf{c}}_i)^T, \quad i = \{1, 2, \dots, N_v\}, \quad (4.11)$$

where  $( )^T$  denotes the transpose of the vector or matrix enclosed in the parentheses.

The classifier discussed in the previous section makes no *a priori* assumptions about the distribution of the data, and uses effectively hyper-planar decision boundaries determined solely on the basis of the class means. On the other hand, it is well-known that if each class can be assumed to have a multivariate normal distribution which is completely specified by its mean vector and covariance matrix, Bayes decision theory offers an analytically tractable solution in terms of hyper-quadratic decision functions. Whilst under those ideal conditions the Bayes classifier is known to be optimal on an average basis (Tou and Gonzalez, 1974, p.121), only sub-optimal performance may be achieved in practice, owing to the necessarily finite amount of available training data, and to the likely departures of the per-class distributions from being strictly Gaussian. If, however, the vowel-speaker dichotomy first unfolded with the linear classifier is confirmed using the more powerful, statistical approach, we may then assume with greater confidence that the observed phenomenon is intrinsic in nature, and not an artefact of the classification methodology.

Although the usual formulation of the Bayes classification rule assigns a test sample to that class which yields the maximum (logarithmic) *probability*, negation of that probability will re-cast it in terms of a measure of *distance* which one seeks to minimise. Adopting the latter approach for the sake of consistency with the minimum-distance classification rule used in the previous section, the distance measure used to

compare a given test sample  $\mathbf{c}$  with the  $i^{\text{th}}$  vowel prototype  $(\bar{\mathbf{c}}_i, \mathbf{C}_i)$  is given by the following expression:

$$d_i(\mathbf{c}) = \frac{1}{2} \left[ (\mathbf{c} - \bar{\mathbf{c}}_i)^T \mathbf{C}_i^{-1} (\mathbf{c} - \bar{\mathbf{c}}_i) \right] + \frac{1}{2} \ln |\mathbf{C}_i| - \ln(p_i), \quad (4.12)$$

where  $(\ )^{-1}$  denotes the matrix inverse,  $|\ |$  denotes the matrix determinant, and  $p_i$  is the *a priori* probability of the  $i^{\text{th}}$  vowel. As the data-partitioning schemes described earlier yield equal numbers of samples per vowel in both the training and test sets, the *a priori* probabilities are therefore assumed to be vowel-independent and equal to the reciprocal of the number of classes, as follows:

$$p_i = \frac{1}{N_v}, \quad i = \{1, 2, \dots, N_v\}. \quad (4.13)$$

Hence, the last term of the distance measure in Equation 4.12 can be ignored, as it has no effect on the classification results. We note in passing, that if the covariance matrices of all classes are assumed to be equal (or if a single, pooled covariance matrix is used), Equation 4.12 reduces to the well-known Mahalanobis distance measure. Moreover, if the covariance matrices are replaced by an identity matrix, Equation 4.12 reduces to the unweighted, linear decision function described in the previous section.

Thus far we have assumed that the distance measure (Equation 4.12) operates on the original, quefrency-weighted cepstrum which represents the NDPS over the entire available spectral range extending to half the sampling frequency. However, as explained earlier, we would also like to perform vowel classification experiments where the distance between test and prototype cepstra is computed within selected frequency sub-bands. Owing to the use of the inverse and determinant of the covariance matrix, Equation 4.12 cannot easily be reformulated (as we did for the Euclidean distance measure in the previous section) to operate on a selected frequency sub-band while still using the original cepstra. To overcome this problem, we propose a new formulation of the cepstrum itself, whereby the original, cepstrally-smoothed NDPS is re-modelled over any selected frequency sub-band, and the unmodified Bayes distance measure can then be used while retaining consistency in spectral representation.

We begin our mathematical derivation of the *partial*, quefrency-weighted cepstrum (P-QCEP) by defining, analogously to Equation 4.2, the *cepstrally-smoothed*

negative derivative of the LP phase spectrum  $X(\theta)$ , which is obtained by the following cosine expansion of only the first  $NCC$  terms of the  $kc_k$  sequence (QCEP):

$$X(\theta) = \sum_{k=1}^{NCC} kc_k \cos(k\theta). \quad (4.14)$$

Those first  $NCC$ , non-zero terms of the QCEP can therefore be retrieved from the cepstrally-smoothed NDPS, by the following, inverse cosine transformation:

$$kc_k = \frac{\gamma}{\pi} \int_0^\pi X(\theta) \cos(k\theta) d\theta, \quad \gamma = \begin{cases} 1, & k = 0 \\ 2, & k > 0 \end{cases}. \quad (4.15)$$

In order to obtain a new cepstral sequence  $\tilde{c}_k$  which models the cepstrally-smoothed NDPS over a selected frequency sub-band  $[\theta_1, \theta_2]$ , it is only required to modify the limits of integration in Equation 4.15 accordingly. Substitution of Equation 4.14 into the modified, inverse cosine transform then yields the following expression:

$$\tilde{c}_k(\theta_1, \theta_2) = \frac{\gamma}{(\theta_2 - \theta_1)} \int_{\theta_1}^{\theta_2} \sum_{l=1}^{NCC} lc_l \cos(l\theta) \cos(k\theta') d\theta, \quad (4.16)$$

where  $\gamma$  is defined as in Equation 4.15, and the new, normalised frequency variable  $\theta'$  is defined by the following transformation:

$$\theta' = \frac{\pi(\theta - \theta_1)}{(\theta_2 - \theta_1)}, \quad (4.17)$$

which linearly maps the selected frequency interval  $\theta \in [\theta_1, \theta_2]$  to  $\theta' \in [0, \pi]$ .

Similarly to our decomposition of Equation 4.3 when deriving the parametric cepstral distance measure in the previous section, Equation 4.16 is first decomposed then integrated, as follows:

$$\tilde{c}_k(\theta_1, \theta_2) = \frac{\gamma}{(\theta_2 - \theta_1)} [A_k(\theta_1, \theta_2) + B_k(\theta_1, \theta_2)], \quad (4.18)$$

where:

$$A_k(\theta_1, \theta_2) = \cos(\lambda\theta_1) \sum_{l=1}^{NCC} lc_l \int_{\theta_1}^{\theta_2} \cos(l\theta) \cos(\lambda\theta) d\theta, \quad (4.19)$$

$$B_k(\theta_1, \theta_2) = \sin(\lambda\theta_1) \sum_{l=1}^{NCC} lc_l \int_{\theta_1}^{\theta_2} \cos(l\theta) \sin(\lambda\theta) d\theta, \quad (4.20)$$

and where:

$$\lambda = \frac{k\pi}{(\theta_2 - \theta_1)}. \quad (4.21)$$

Two distinct cases arise in Equations 4.19 and 4.20, depending on whether or not the index  $l$  is equal to  $\lambda$ . When  $l = \lambda$ , the two trigonometric functions in each integral have the same argument, and the following expressions are obtained:

$$A_k(\theta_1, \theta_2) = \frac{1}{2} \sum_{l=1}^{NCC} l c_l \cos(l\theta_1) \left[ (\theta_2 - \theta_1) + \frac{\sin(2l\theta_2) - \sin(2l\theta_1)}{2l} \right], \quad (4.22)$$

$$B_k(\theta_1, \theta_2) = \frac{1}{2} \sum_{l=1}^{NCC} l c_l \sin(l\theta_1) \left[ \frac{\cos(2l\theta_1) - \cos(2l\theta_2)}{2l} \right]. \quad (4.23)$$

When  $l \neq \lambda$ , the integral of Equation 4.19 is carried out to yield the following expression:

$$A_k(\theta_1, \theta_2) = \frac{1}{2} \cos(\lambda\theta_1) \sum_{l=1}^{NCC} l c_l x_{kl}(\theta_1, \theta_2), \quad (4.24)$$

where:

$$x_{kl}(\theta_1, \theta_2) = \frac{\sin((l-\lambda)\theta_2) - \sin((l-\lambda)\theta_1)}{l-\lambda} + \frac{\sin((l+\lambda)\theta_2) - \sin((l+\lambda)\theta_1)}{l+\lambda}, \quad (4.25)$$

and the integral of Equation 4.20 is carried out to yield the following expression:

$$B_k(\theta_1, \theta_2) = \frac{1}{2} \sin(\lambda\theta_1) \sum_{l=1}^{NCC} l c_l y_{kl}(\theta_1, \theta_2), \quad (4.26)$$

where:

$$y_{kl}(\theta_1, \theta_2) = \frac{\cos((l-\lambda)\theta_2) - \cos((l-\lambda)\theta_1)}{l-\lambda} - \frac{\cos((l+\lambda)\theta_2) - \cos((l+\lambda)\theta_1)}{l+\lambda}. \quad (4.27)$$

The two sets of Equations 4.22 and 4.23 ( $l = \lambda$ ), and 4.24 through 4.27 ( $l \neq \lambda$ ), respectively, are substituted back into Equation 4.18, and after some algebraic simplification we arrive at the following expression for the P-QCEP:

$$\tilde{c}_k(\theta_1, \theta_2) = \frac{\gamma}{2} \sum_{l=1}^{NCC} l c_l \zeta_{kl}(\theta_1, \theta_2), \quad (4.28)$$

where:

$$\zeta_{kl}(\theta_1, \theta_2) = \begin{cases} \cos(l\theta_1) + \frac{\sin(2l\theta_2 - l\theta_1) - \sin(l\theta_1)}{2l(\theta_2 - \theta_1)}, & l = \lambda \\ 2l(\theta_2 - \theta_1) \frac{(-1)^k \sin(l\theta_2) - \sin(l\theta_1)}{[l(\theta_2 - \theta_1)]^2 - [k\pi]^2}, & l \neq \lambda \end{cases}. \quad (4.29)$$

It is easy to verify that when  $\theta_1 = 0$  and  $\theta_2 = \pi$  (i.e., when the entire spectral range is selected), Equations 4.28 and 4.29 simply yield the QCEP without modification.

We note in passing that the derivation of the *partial cepstrum* (P-CEP), which represents the cepstrally-smoothed LP log magnitude spectrum (LMS) over a selected frequency sub-band, would yield the same result as above, except for the quefrequency-weight  $l$  which would be removed from Equation 4.28. However, as explained previously, the NDPS offers a formant-enhanced spectral representation which we therefore adopt consistently in our vowel classification experiments.

The P-QCEP just derived, and summarised in Equations 4.28 and 4.29, captures the shape of the cepstrally-smoothed NDPS within any selected frequency sub-band. Most importantly, our formulation of the P-QCEP circumvents re-analysis of the speech waveform, and merely defines it as a transformation of the cepstrum. Consequently, it maintains consistency in spectral representation, which other methods (such as selective LP analysis) do not. This point is aptly illustrated in Figure 4.2, where the P-QCEP is computed in two different frequency sub-bands of a single frame of speech (a vocalic steady-state frame taken from /hʌ:d/ recorded by speaker B), and used to re-compute the *partial* NDPS (via the discrete cosine transform in Equation 4.14) in those selected sub-bands. Clearly, the spectral representation is consistent with that of the original QCEP, and the selected portions of the cepstrally-smoothed NDPS are faithfully reproduced without excessive computational cost.

A relevant issue which arises in regard to the practical use of the P-QCEP, is the number of coefficients to retain. It is particularly important to address the question of P-QCEP dimensionality in the context of classification experiments using the quadratic distance measure in Equation 4.12, as the inadvertent inclusion of redundant parameters (i.e., those which do not sufficiently contribute to modelling the spectral shape in a selected sub-band) will no doubt tend to ill-condition the per-class covariance matrices. This, in turn, would lead to instability in the computation of the inverse and the

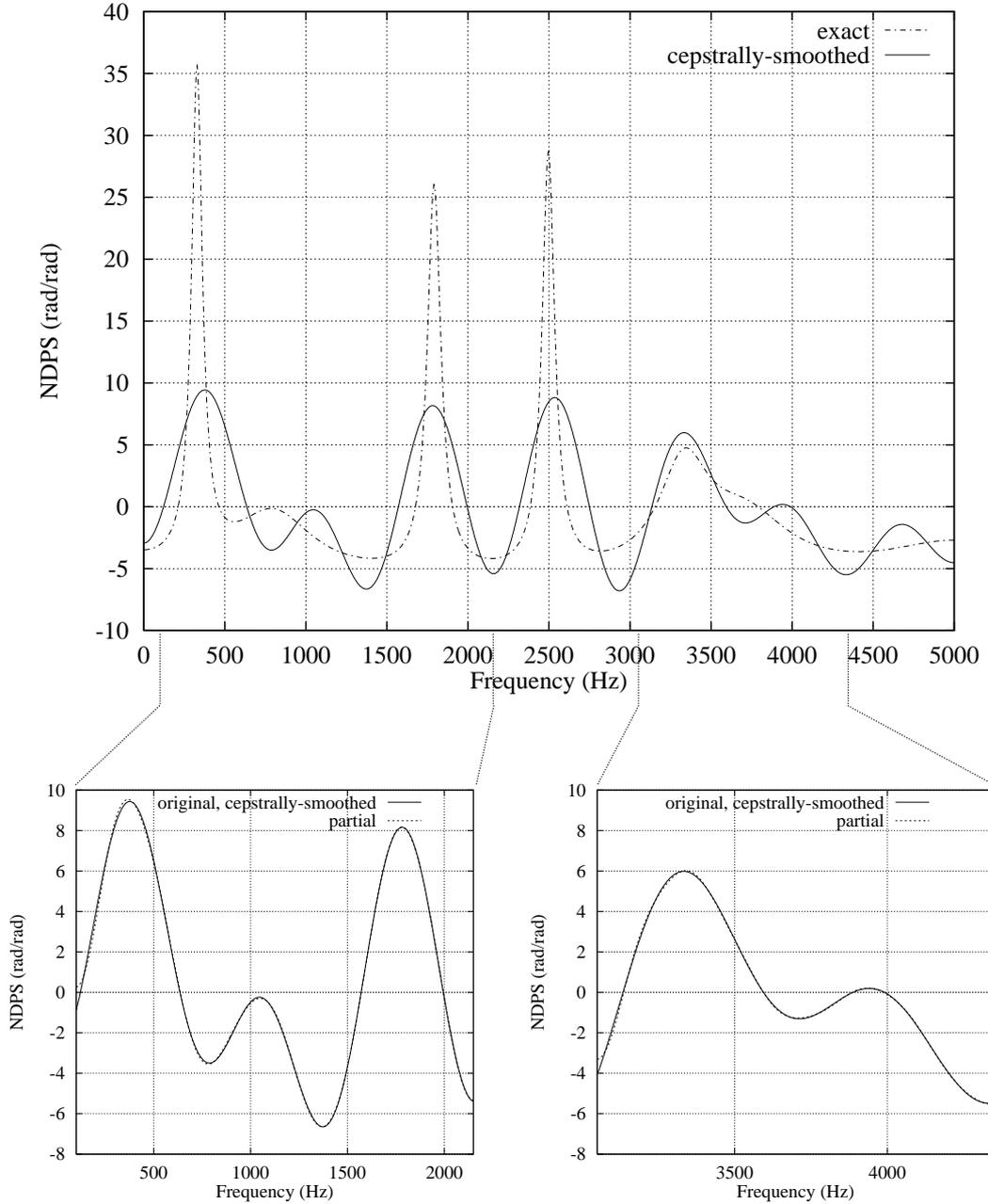


Figure 4.2: Illustration of *partial* (quefrency-weighted) cepstrum (P-QCEP), computed within two, selected spectral regions (*lower-left graph*: 100Hz to 2150Hz; *lower-right graph*: 3050Hz to 4350Hz) for a vocalic steady-state frame taken from /hʌ:d/ recorded by speaker B of the FC dataset. In the *top graph* are shown superimposed the exact (dashed-dotted curve) and the cepstrally-smoothed (solid curve), negative derivative of LP phase spectra (NDPS), obtained with  $M=NCC=14$ . As shown in the two *lower graphs*, cosine-expansion (using Equation 4.14) of the P-QCEP derived from the original cepstrum (using Equations 4.28 and 4.29) faithfully models the cepstrally-smoothed NDPS over the respective, selected frequency sub-bands.

determinant of those matrices, which then would yield meaningless or biased classification results. Naturally, over the entire spectral range  $[0, \pi]$ , one needs no more than the original *NCC* coefficients to exactly represent the cepstrally-smoothed NDPS. However, when only a portion of the entire spectrum is selected, the dimensionality of the P-QCEP is complicated in at least two ways.

First, as the mean spectral level in a frequency sub-band of the NDPS (or indeed of the LMS) is not necessarily equal to zero, the usual assumption  $c_0 = 0$  does not carry over to the P-QCEP (nor to the P-CEP). Consequently, the zeroth P-QCEP coefficient  $\tilde{c}_0$  must also be computed in order to retain the relative spectral amplitudes across frames, vowels, and speakers. Whilst this may appear to increase the dimensionality of the P-QCEP compared with the original cepstrum, a second, more serious complication occurs which tends to counteract that effect.

Indeed, selection of any frequency *sub*-band will inevitably reduce the number of spectral peaks and valleys encompassed, and is therefore expected to lead to a reduction in the number of P-QCEP coefficients required to model that portion of the spectral shape. Of course, the validity of that assumption will depend to a certain extent on the particular spectral range selected, and will also vary from one speech sound to another (e.g., the frequency band [500, 1000] Hz may contain two formant peaks for a mid-back vowel, but only part of a spectral valley for a high-front vowel, thus requiring perhaps a greater number of coefficients for the former than for the latter). However, it seems reasonable to assume that higher-order coefficients (beyond the original *NCC*) are not required to model a sub-band of the cepstrally-smoothed NDPS which itself is based on a truncated cepstral series. Ideally, one might hope to methodically reduce the number of P-QCEP coefficients retained, for example by selecting the number of coefficients in proportion to the range of the selected frequency sub-band.

In order to confirm the general validity of our assumptions, and perhaps to arrive at a reasonable method of determining the appropriate dimensionality of the P-QCEP, an empirical investigation was therefore carried out, in which the mean and standard deviation  $\sigma$  of each P-QCEP coefficient was computed over the entire FC dataset of vocalic steady-states, as a function of the upper limit  $\theta_2$  of the selected spectral range  $[0, \theta_2]$ . Each of the 15 graphs in Figure 4.3 shows the mean and  $\pm 1\sigma$  error-bar for a

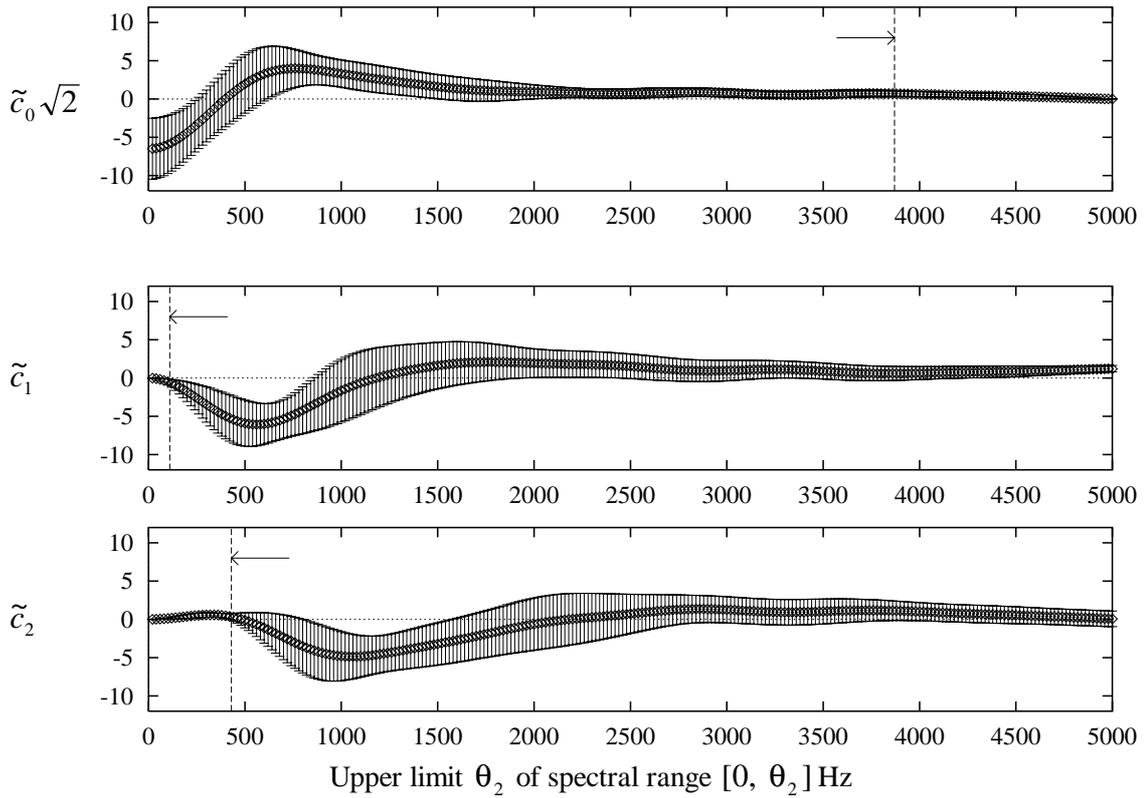


Figure 4.3: Mean (diamond symbols) and  $\pm 1\sigma$  (error-bars) in each *partial quefrency-weighted cepstrum* (P-QCEP) coefficient from the 0<sup>th</sup> to the 14<sup>th</sup> (continued on the next two pages), computed over all the FC dataset (7 steady-state frames, 5 repetitions, 11 vowels in /hVd/ context recorded by 4 adult, male speakers of Australian English), as a function of the upper limit  $\theta_2$  of the selected frequency sub-band  $[0, \theta_2]$ , incremented in steps of 20Hz. Also shown in each graph is a vertical (dashed) line which indicates: (i) for the zeroth coefficient, the *maximum* spectral range, and (ii) for the 1<sup>st</sup> through 14<sup>th</sup> coefficients, the *minimum* spectral range of acceptable variation in that coefficient, lest it becomes statistically redundant (as determined by a nominal threshold of 0.4 in the standard-deviation) and must therefore be excluded from the parameter set used in the classifier with quadratic decision functions (Equation 4.12). As discussed in the text, these graphs suggest a method of recruiting progressively higher-order P-QCEP coefficients in proportion to the selected spectral range. The suggested rule is implemented in our classification experiments by a step-wise linear function in the number of coefficients, from 2 at the narrowest spectral range, to all 14 coefficients at full spectral range. *N.B.*: the zeroth P-QCEP coefficient is weighted by the square-root of two before computing statistics and performing classification experiments, in order to account for its differential contribution in a weighted-Euclidean cepstral distance measure (as in the first term of Equation 4.12).

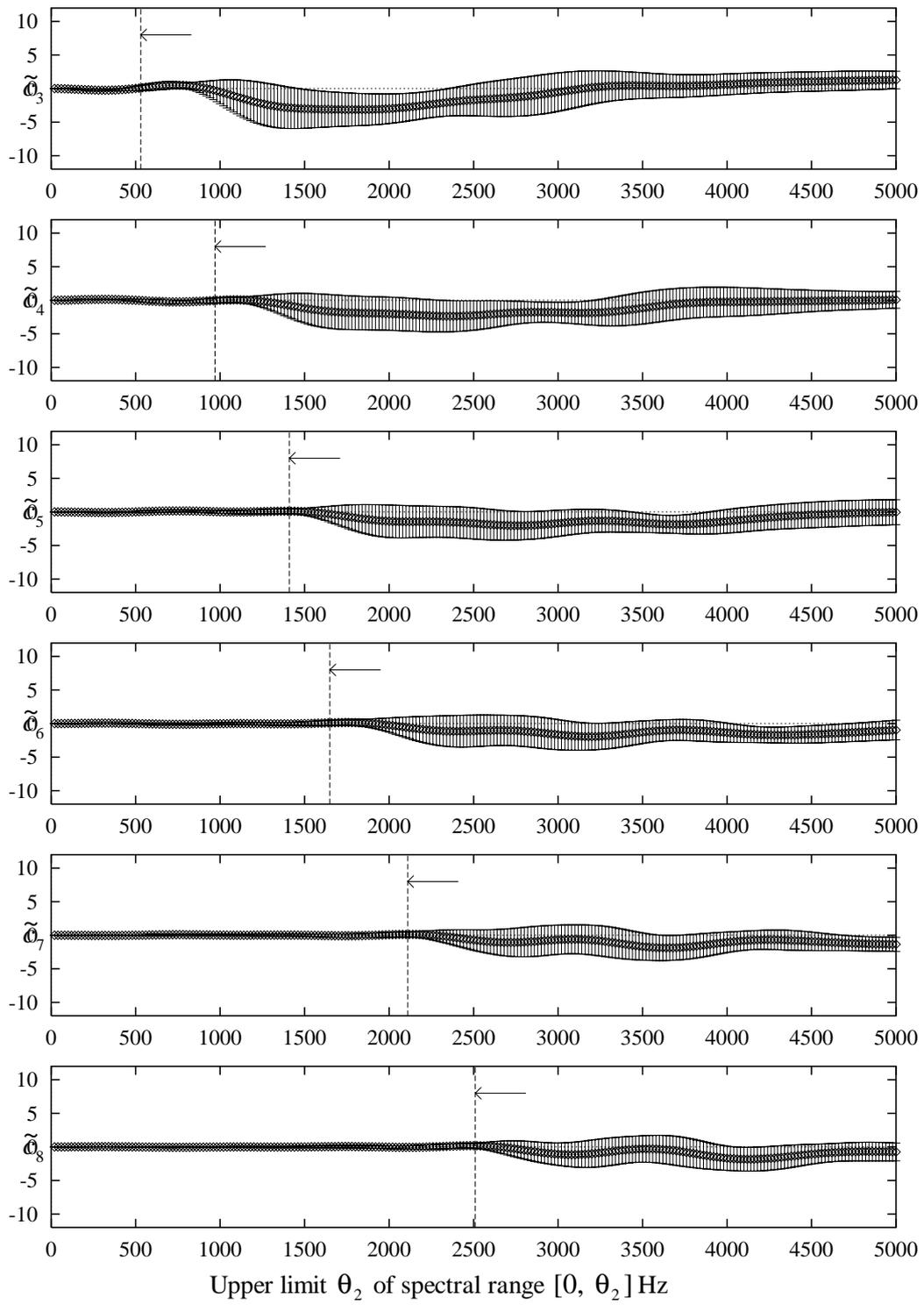


Figure 4.3: (continued from previous page).

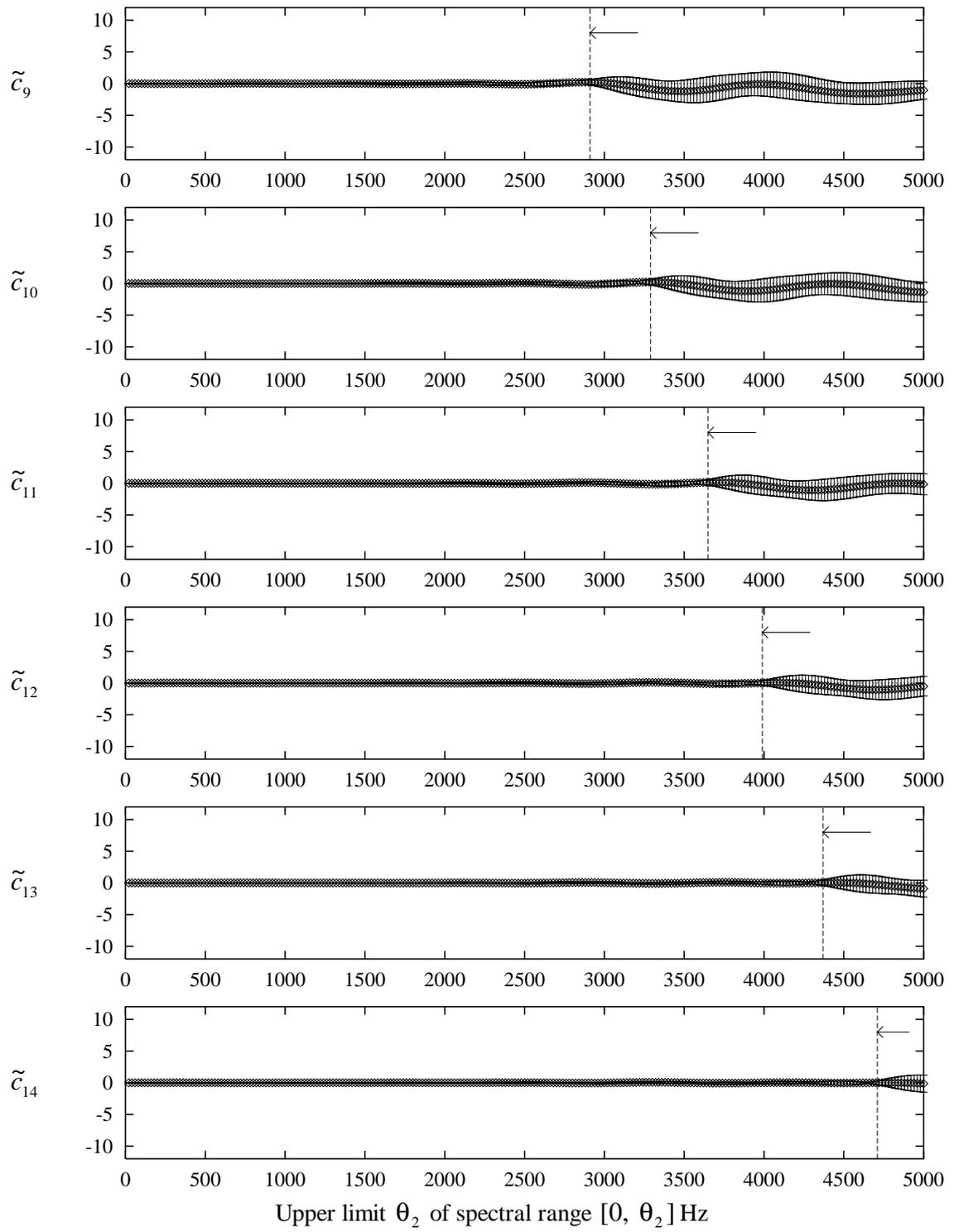


Figure 4.3: (continued from previous page).

single P-QCEP coefficient, obtained in steps of 20 Hz across the available spectral range. Clearly, our assumptions regarding the statistical redundancy of higher-order coefficients in narrower spectral ranges, is confirmed by the smaller standard-deviations in those parameters for low values of  $\theta_2$ . Furthermore, the minimum spectral range required for the variations in each coefficient to gain statistical significance (as determined qualitatively on each graph, in terms of the spectral range at which there appears a marked increase in  $\sigma$ ) clearly shows a gradual, upward progression as a function of the P-QCEP coefficient index. A nominal threshold of 0.4 on the minimum standard-deviation for any given coefficient to be deemed non-redundant, yields a nearly linear progression in the number of P-QCEP coefficients, from only 1 coefficient at about 100 Hz, to all 14 coefficients at full spectral range. On the other hand, both the mean and standard-deviation of the parameter  $\tilde{c}_0$  tend to zero as  $\theta_2$  approaches half the sampling frequency, indicating that the zeroth coefficient should not be used in the quadratic classifier when the selected frequency band is nearly the full spectral range.

These observations are summarised in Figure 4.4, where the highest allowed index of P-QCEP coefficients is indicated by a step-wise function of an increasing spectral range (solid lines); the required exclusion of the zeroth P-QCEP is indicated by the single step at 3870Hz, shown in the bottom right corner of the graph. Our empirical results therefore confirm the simple but reasonable assumption that *on average*, the number of P-QCEP coefficients retained (compared with the original *NCC*) be approximately determined by the ratio of the selected and the entire spectral range. For example, if the frequency sub-band extending to 2500Hz is selected, then it is reasonable to assume that on average, the first seven P-QCEP coefficients (in addition to  $\tilde{c}_0$ ) will suffice to faithfully represent a cepstrally-smoothed NDPS which originally is defined over a 5kHz range with 14 cepstral coefficients. Our results also suggest a much coarser step-size for  $\theta_2$  than proposed for the classification experiments described in the previous section. In order to avoid marked discontinuities in the behaviour of classification accuracy when an increment in  $\theta_2$  is accompanied by recruitment of a higher-order P-QCEP coefficient, the upper limit of the spectral range will be increased in steps of 200 Hz when using the quadratic classifier (in contrast with the step-size of 20 Hz proposed for the linear classifier). The vertical (dashed) lines in

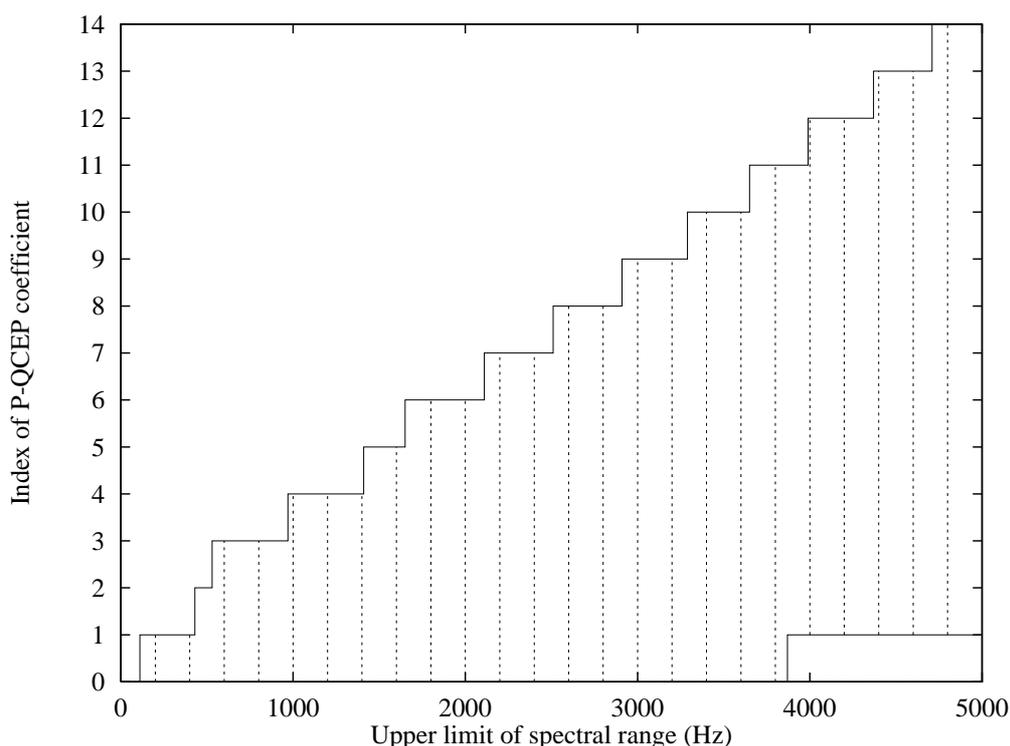


Figure 4.4: Summary of allowed P-QCEP coefficients as a function of upper spectral limit (increased in steps of 200Hz, as indicated by the dashed, vertical lines), based on the empirical results of P-QCEP mean and standard deviation as shown in Figure 4.3.

Figure 4.4 span the allowed number of P-QCEP coefficients to be used at each of the 25 steps across the entire spectral range.

Thus far we have discussed the issue of P-QCEP dimensionality only from a spectral representation point of view. However, we must also be wary of a phenomenon which has been brought to light mainly in the pattern-recognition literature, and which has come to be known as the “curse of dimensionality” (e.g., Meisel, 1972, p.13). In essence, a lower limit is imposed on the ratio of the number of training samples per class (i.e., the number of P-QCEP vectors per vowel) and the dimensionality of the parameter set (i.e., the number of coefficients in each P-QCEP vector). If that ratio is less than unity, the sample covariance matrix will be less than full rank, and computation of its inverse will require a method such as Singular Value Decomposition (SVD) to eliminate the inevitable nullity. Even if the ratio is greater than unity and the parameters themselves are reasonably uncorrelated and non-redundant, there is still the danger of obtaining heavily biased or nonsensical classification results, unless the training sample population is sufficiently greater than the intrinsic dimensionality of the

sample vectors. It has been suggested, for example, to observe a minimum ratio of between 3 and 5 (Foley, 1972; Sarma and Venugopal, 1977). A more conservative estimate is given by Piper (1992), who suggests a minimum number of training samples per class an order of magnitude greater than the dimensionality of the parameter set.

In our speaker-independent experiments the number of training samples per vowel is equal to 105 (7 frames, 5 repetitions, and 3 speakers in each round), which should perhaps be sufficiently large to avoid the problem of dimensionality (as the ratio is at worst equal to  $105 \div 14 = 7.5$ ). However, we may expect the problem to be manifest in our speaker-dependent experiments, where the number of training samples per vowel is reduced to only 28 (7 frames and 4 repetitions in each round, per speaker), and where the ratio is therefore as low as 2. One way of dealing with the problem is to impose an upper limit on the number of P-QCEP coefficients presented to the quadratic classifier. Unfortunately, by reducing the number of coefficients in order to counterbalance the effect of a small training sample size, we would also compromise on the accuracy of spectral representation.

Although our inter-speaker experiments may be more immune to the “curse of dimensionality”, we must also be wary of sub-optimal results caused by a violation of the assumption of normality. Whilst on a speaker-dependent basis the per-vowel distribution of the cepstrum or of the P-QCEP coefficients can perhaps be assumed to be reasonably multivariate Gaussian, this assumption is likely tenuous when the acoustic parameters of several speakers are pooled. On the other hand, if the number of speakers is sufficiently large, the per-vowel distribution of their acoustic parameters may tend to be sufficiently homogeneous that the assumption of normality is upheld. However, our inter-speaker experiments involve training the classifier on only three speakers’ data and testing it on the acoustic parameters of the remaining speaker. If the speakers are sufficiently consistent across repetitions of the same vowel, and sufficiently different from each other, then the assumption of normality which underpins the quadratic classifier may be violated, and the yielded classification accuracies therefore sub-optimal. It remains to be seen whether any of the potential difficulties outlined above will tend to undermine the phenomenon of vowel-speaker dichotomy, to which we now direct our attention.

## 4.3 The Dichotomy Unfolded

We shall now use the methodology developed in the previous section, to unfold the vowel-speaker dichotomy by way of vowel classification experiments performed using the FC dataset. The behaviour of intra- and inter-speaker vowel classification accuracy as a function of an increasing upper spectral limit, is first observed in Section 4.3.1 using the classifier with hyper-planar decision boundaries (described in Section 4.2.3.1). The validity of those observations is then tested in Section 4.3.2 using the classifier with hyper-quadratic decision boundaries which, as previously foreshadowed, must also be addressed in terms of its sensitivity to cepstrum dimensionality in relation to the amount of training data.

### 4.3.1 Behaviour of Classification Accuracy using Linear Classifier

Vowel classification experiments were first performed using the classifier described in Section 4.2.3.1, where each vowel prototype is characterised by the mean of the quefrency-weighted cepstral vectors (QCEP) for that vowel in the training data, and the classifier therefore uses hyper-planar decision boundaries. Our parametric cepstral distance measure (PCD) was used to perform a complete set of experiments at each selected frequency sub-band  $[0, \theta_2]$ , and the upper limit  $\theta_2$  of the spectral range was incremented in steps of 20 Hz.

First, a set of speaker-dependent experiments were performed, where one repetition at a time was left out of the training set and used as test data, and the results were averaged over those five experiments per speaker. The solid curve in Figure 4.5 shows the behaviour of classification accuracy averaged over all four speakers, as a function of an increasing upper spectral limit. Classification accuracy rises gradually to 91.6% as the spectral range is increased to 1600 Hz, beyond which it changes very little. The best performance of 92.6% is achieved at a spectral range of 2320 Hz (which encompasses nearly all of the first two formants of our four speakers, as we shall see later in Section 4.4), and the accuracy tapers off to 90.2% at full range (5000 Hz). Our intra-speaker classification curve is therefore in complete agreement with the well-established acoustic-phonetic theory (discussed in Chapter 2) that the first

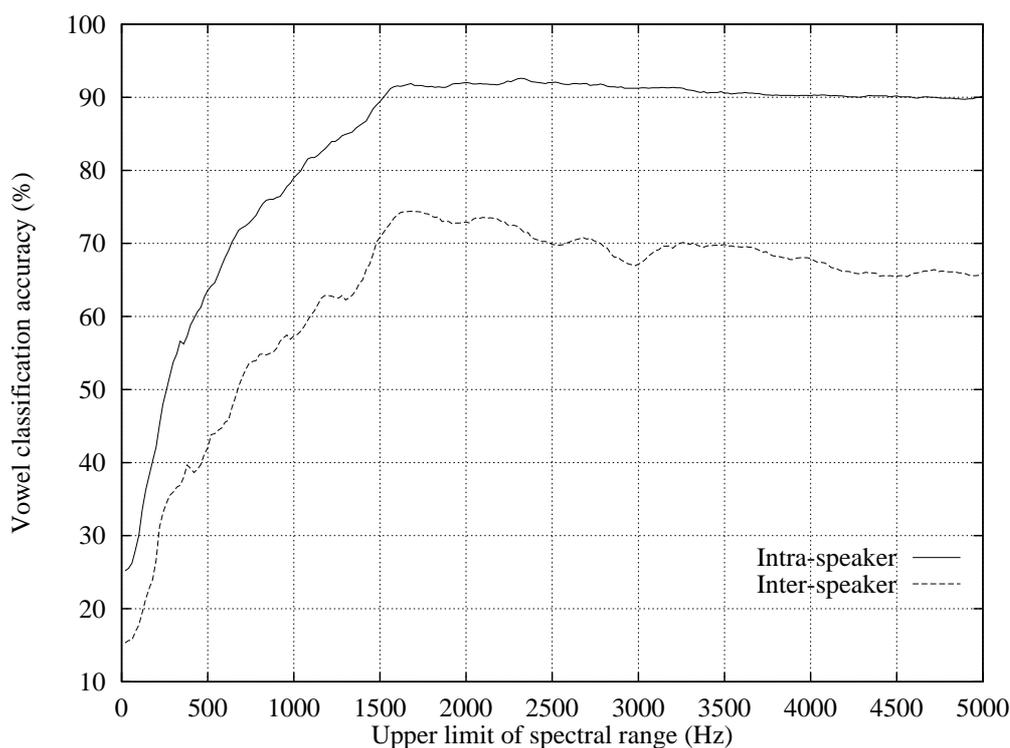


Figure 4.5: Intra- and inter-speaker vowel classification accuracy (solid curve and dashed curve, respectively) obtained with a linear classifier, as a function of the upper limit  $\theta_2$  of the spectral range  $[0, \theta_2]$ . Dataset (FC): 14<sup>th</sup>-order LP cepstra of 7 vocalic steady-state frames in 5 repetitions of 11 vowels recorded in /hVd/ context by 4 adult, male speakers of Australian English.

two formant peaks provide the most discriminating acoustic cues in the spectra of spoken vowels for any single speaker.

Next, a set of speaker-independent experiments were performed, where the cepstra of one speaker at a time were left out of the training set and used as test data. The dashed curve in Figure 4.5 shows the behaviour of classification accuracy averaged over all four experiments. Classification accuracy rises to a peak of 74.4% as the spectral range is increased to 1680 Hz, then deteriorates as the higher spectral regions are cumulatively recruited. The downward trend in accuracy appears to modulate through several falls and smaller rises (the most significant of which occurs at about 3 kHz), finally resting on 66.0% at full spectral range.

Consistent with our expectations regarding the effects of speaker differences on vowel classification, accuracies are unfailingly higher on an intra- than an inter-speaker basis. However, a more striking contrast is apparent in the behaviour of the two curves as a function of the spectral range. By comparison with the inter-speaker curve, the

behaviour of intra-speaker vowel classification accuracy might almost be viewed as *nearly asymptotic*, the asymptote of a little over 90 % having been reached just beyond 1500 Hz, and only minor changes in performance appearing across the higher spectral regions. By contrast, the inter-speaker curve has an unmistakable peak at 1680 Hz, marking the spectral boundary below which lies apparently useful phonetic information, and above which the consequences of speaker differences appear to dominate. Indeed, the detrimental influence of inter-speaker variability is perhaps most pronounced in the series of falls in accuracy from 74.4 % at 1680 Hz, to 66.9 % at 2980 Hz — a cumulative drop of 7.5 % across a 1300 Hz spectral region. The small shift in the intra-speaker curve from 91.9 % to 91.2 % across the same spectral region is, by comparison, an order of magnitude less significant.

This contrast in the behaviour of intra- and inter-speaker vowel classification accuracy does suggest a dichotomous relation between the spectral manifestations of vowel and speaker influences, which is not at all in disagreement with the extensive but largely unheeded body of literature reviewed in Chapter 2. On the contrary, our experimental results at once confirm the well-established phonetic importance of the low spectral regions of spoken vowels, and provide compelling evidence in support of the speaker potency of the higher spectral regions, particularly the frequency-band which extends from about 1.6 to 3.0 kHz. The vowel-speaker dichotomy is thus defined in terms of the two, juxtaposed spectral regions which we have identified by way of vowel classification experiments designed to reveal the consequences of vowel and speaker influences along the spectral continuum. An intriguing question then arises, whether these accuracy curves can be interpreted as embodying spectral regions of primary phonetic and speaker influence corresponding to particular formant ranges. However, before we venture to pursue that question (in Section 4.4), we first attempt to validate our findings using a more sophisticated classifier.

### **4.3.2 Behaviour of Classification Accuracy using Quadratic Classifier**

If the vowel-speaker dichotomy unfolded in the previous section is indeed an intrinsic phenomenon of speech, one might expect to observe a similar contrast in the behaviour of intra- and inter-speaker vowel classification accuracy obtained using a different

classifier. To test this hypothesis, we used the more sophisticated, quadratic classifier described in Section 4.2.3.2, where each vowel prototype is characterised by the mean vector and covariance matrix of the P-QCEP vectors for that vowel in the training data. The P-QCEP were computed from the original, 14<sup>th</sup>-order LP cepstra, and a complete set of experiments were performed at each selected frequency sub-band  $[0, \theta_2]$ , where the upper limit  $\theta_2$  of the spectral range was incremented in steps of 200 Hz. The number of P-QCEP coefficients retained at each selected spectral range, was dictated by our earlier, empirical findings regarding the statistical redundancy of each coefficient, as summarised in Figure 4.4. Indeed, whilst it may be tempting to select the subset of P-QCEP coefficients which yield the highest classification accuracy at each spectral range, our earlier results provide sufficient statistical evidence to contend and refute the adoption of that method.

The intra- and inter-speaker curves of vowel classification accuracy thus obtained, are superimposed in Figure 4.6, together with the earlier results yielded by the linear classifier. The behaviour of vowel classification accuracy averaged over the four separate, speaker-dependent experiments is shown by the diamond symbols (joined with short-dashed lines). Although the curve appears to be more noisy (owing to the step-wise recruitment of P-QCEP coefficients) especially at the lower frequency sub-bands, the overall trend is quite similar to the intra-speaker curve generated earlier with the linear classifier (solid line). Classification accuracy rises to 81.6% as the spectral range is increased to 1600 Hz, and the performance does not significantly improve (nor degrade) as the spectral range is increased further to about 3 kHz.

By contrast, the inter-speaker curve obtained using the quadratic classifier (shown in Figure 4.6 by the plus symbols joined with dash-dotted lines) rises to a peak of 74.9% as the spectral range is increased to 1600 Hz, and monotonically drops to 66.2% as the spectral range is increased further to 2800 Hz. The overall behaviour in classification accuracy is quite similar to the inter-speaker curve generated earlier with the linear classifier (dashed line), up to about 3.5 kHz. Although both curves tend to rise a little as the spectral range is increased just beyond 3 kHz, the quadratic classifier appears not to follow the subsequent, downward trend in accuracy exhibited by the linear classifier as the upper limit is increased to full spectral range. Nevertheless, the

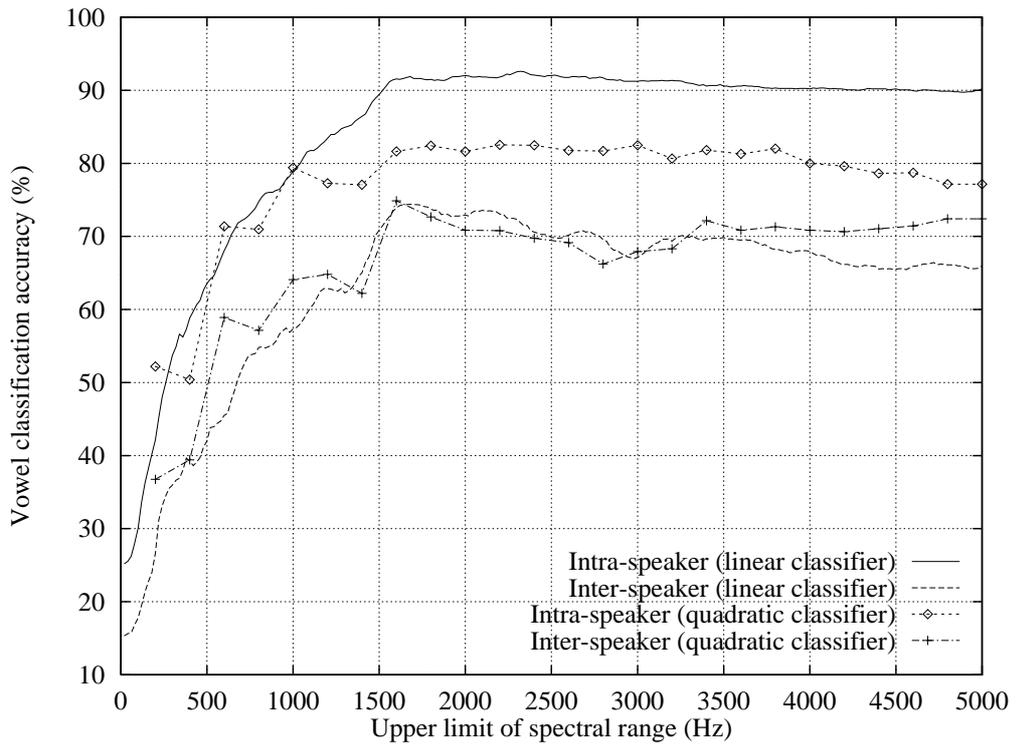


Figure 4.6: Intra- and inter-speaker vowel classification accuracy obtained with a linear classifier (solid and dashed curves, respectively, replicated from Figure 4.5) and with a quadratic classifier (diamond and plus symbols, respectively), as a function of the upper limit  $\theta_2$  of the spectral range  $[0, \theta_2]$ . Dataset (FC): 14<sup>th</sup>-order LP cepstra of 7 vocalic steady-state frames in 5 repetitions of 11 vowels recorded in /hVd/ context by 4 adult, male speakers of Australian English. The number of P-QCEP coefficients used in the quadratic classifier at each spectral range, was dictated by the step-wise function plotted in Figure 4.4.

accuracy at full spectral range (72.4 %) is still lower than the accuracy achieved at the aforementioned peak (74.9 %) with about half the number of P-QCEP coefficients. The quadratic classifier therefore does confirm the speaker potency of the spectral mid-range, defined earlier as approximately [1.6,3.0] kHz.

Notwithstanding this confirmation of the vowel-speaker dichotomy as portrayed in the overall *behaviour* of the accuracy curves yielded by the quadratic classifier, there appears a considerable discrepancy in the levels of accuracy attained by the two classifiers, in both the intra- and inter-speaker experiments. Contrary to the naive expectation that hyper-quadratic decision boundaries should perform *no worse* than hyper-planar decision boundaries, the intra-speaker curve yielded by the quadratic classifier is consistently *lower* by about 10 % when the spectral range is extended beyond 1000 Hz. Furthermore, it appears to taper off more rapidly as the spectral range

is increased beyond 4000 Hz, and attains only 77.1% at full spectral range. The levels of accuracy shown along the inter-speaker curve are also questionable, especially as the upper limit of the spectral range is increased from 1800 Hz to 2800 Hz, where the quadratic classifier performs consistently worse than the linear classifier.

This apparently sub-optimal performance was foreshadowed in Section 4.2.3.2 where, specifically in connection with the quadratic classifier which relies to a greater extent on the statistical properties of the data at hand, we cautioned against an excessive number of P-QCEP coefficients compared with the training sample size — thus to avoid the so-called “curse of dimensionality”. In this vein, Foley (1972) has shown that for a two-class problem with multivariate normal distributions, the performance of a quadratic classifier only begins to satisfactorily approach that of the optimum (Bayes) discriminant when the ratio of the number of training samples per class to the parameter-set dimensionality is greater than or equal to 3. Given a range of P-QCEP coefficients extending up to 14, and only 28 training samples per vowel (7 frames in each of 4 repetitions for the intra-speaker experiments), we may therefore infer that the sub-optimal accuracies yielded by the quadratic classifier in the *intra*-speaker experiments are indeed a consequence of the “curse of dimensionality”. On the other hand, we had earlier predicted that the *inter*-speaker experiments might be immune to the “curse of dimensionality”, owing to a sufficiently large training sample size. If the sub-optimal accuracies yielded by the quadratic classifier and shown along the inter-speaker curve in Figure 4.6 are rather a consequence of an ill-satisfied assumption of normality, they cannot be expected to improve simply by providing a greater number of training samples indiscriminately.

In order to test our hypotheses without compromising on the accuracy of spectral representation (i.e., without having to reduce the number of P-QCEP coefficients used at each selected spectral range), we turn to the more comprehensive version of the FC dataset which comprises 11 time-normalised frames in each of 9 vowel nuclei (the two vowels /ɔ/ and /ʊ/ are not included) recorded 5 times in /CVd/ context by the same, four speakers of Australian English, with the leading consonants  $C = /h, b, d, g, p, t, k/$ . As described in Chapter 3, the centre frames 5, 6, and 7 are used, on the basis that they are individually the three time-normalised frames which most closely match (in a

cepstral distance sense) the average steady-state frame of the corresponding vowel in /hVd/ context. Although this leads to a reduction in the number of frames per vocalic nucleus from seven to only three, the seven-fold increase in the number of leading consonants yields an overall advantage of *three times* the number of training (and test) samples over the smaller subset of /hVd/ vocalic steady-state data.

The top graph in Figure 4.7 shows the intra- and inter-speaker profiles of vowel classification accuracy obtained with the linear and the quadratic classifiers, using all of the /CVd/ data as outlined above. Our first observation is in regard to the phenomenon of vowel-speaker dichotomy itself, which does not seem to have suffered from the absence of the two, extreme back vowels, nor from the more relaxed assumptions regarding the exact location of the steady-state in each vowel nucleus. Despite those differences and the three-fold expansion of the data, the relative behaviour of the intra- and inter-speaker curves retains the imprint of vowel-speaker dichotomy which contrasts the low and the higher spectral regions.

More importantly in connection with our concerns regarding the relative performance of the two classifiers, the quadratic classifier yields an intra-speaker curve which is more nearly asymptotic than the one shown earlier in Figure 4.6, and which clearly equals or outperforms the linear classifier over all the spectral ranges considered. It would appear therefore, that the previously sub-optimal performance of the quadratic classifier, which degraded even further as the spectral range was extended and a greater number of coefficients recruited, can indeed be explained by an insufficient number of training samples per vowel, which has been increased from 28 to 84 in the present, intra-speaker experiments. By contrast, the relative level of accuracy obtained by the two classifiers in the inter-speaker experiments has not dramatically changed in comparison with the earlier, intertwined curves shown in Figure 4.6. This may imply that the previous training sample size of 105 was perhaps sufficient to avoid the “curse of dimensionality”, and that a different explanation of the sub-optimal performance is more likely.

In order to further substantiate our conclusions regarding the effect of training sample size, we explicitly reduce the amount of data by using only the first, third, and fifth repetitions, those being perhaps the most independent group of three repetitions

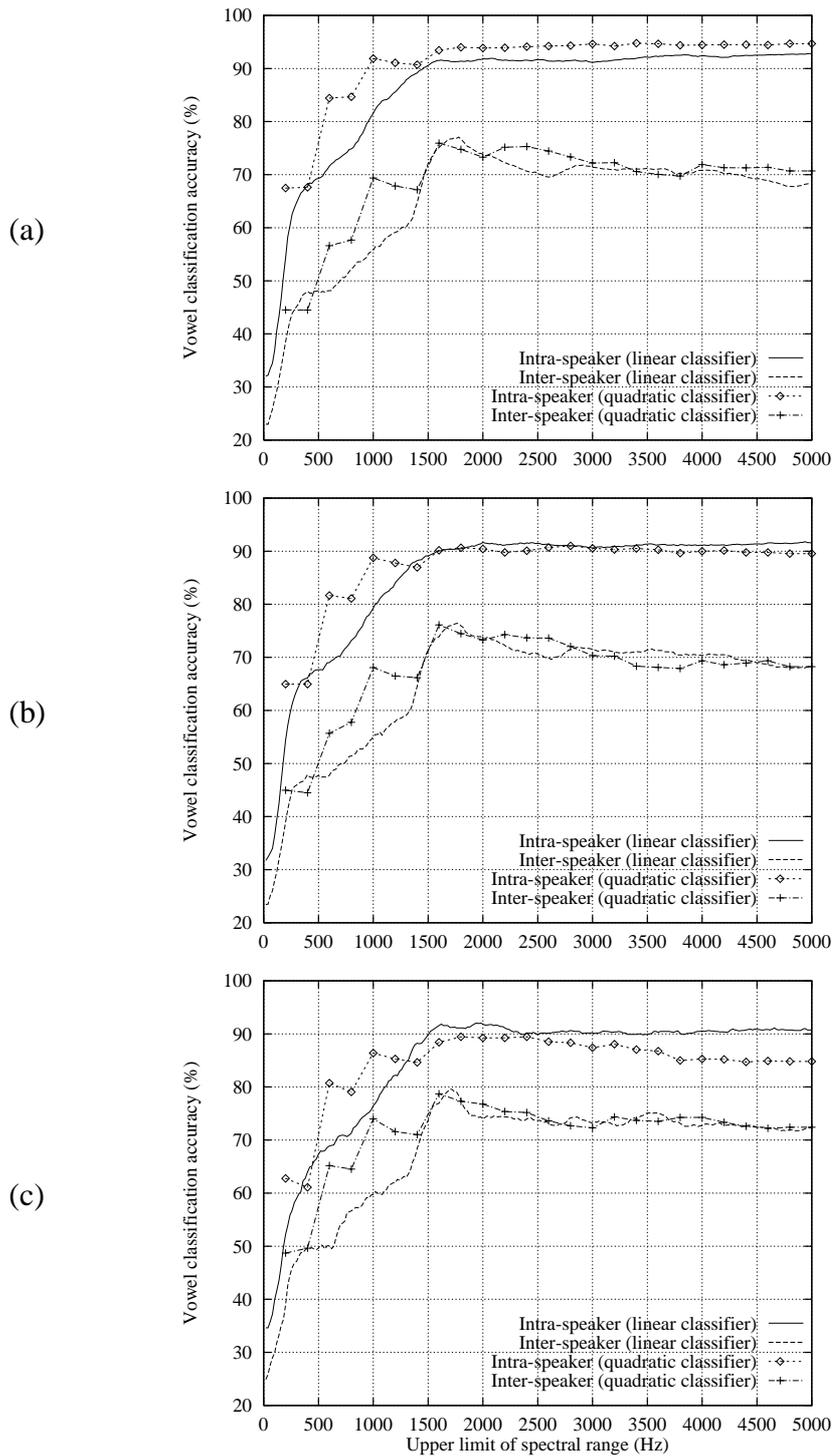


Figure 4.7: Intra- and inter-speaker vowel classification accuracy obtained with a linear classifier (solid and dashed curves, respectively) and with a quadratic classifier (diamond and plus symbols, respectively), as a function of the upper limit  $\theta_2$  of the spectral range  $[0, \theta_2]$ . Dataset (FC): 14<sup>th</sup>-order LP cepstra of the centre frames 5, 6, and 7 of 11 time-normalised frames in 5 repetitions of 9 vowels recorded in /CVd/ context by 4 adult, male speakers of Australian English. The number of P-QCEP coefficients used in the quadratic classifier at each spectral range, was dictated by the step-wise function plotted in Figure 4.4. *Panel (a)*: all data used in experiments (number of training samples per vowel TSPV: intra-speaker = 84; inter-speaker = 315). *Panel (b)*: data reduced to the three repetitions 1, 3, and 5 (TSPV: intra-speaker = 42; inter-speaker = 189). *Panel (c)*: data reduced to the two leading consonants /h, b/ (TSPV: intra-speaker = 24; inter-speaker = 90).

amongst the five. Having thus reduced the number of training samples per vowel from 84 to 42 in the intra-speaker experiments, the resulting accuracy curves in Figure 4.7(b) clearly show an overall degradation in the performance of the quadratic classifier relative to that of the linear classifier. By contrast, their relative performance in the inter-speaker experiments does not appear to have been altered by the reduction in the training sample size from 315 to 189.

An even further reduction of data is achieved (independently of the previous reduction in the number of repetitions) by considering the utterances with only the first two leading consonants  $C = /h, b/$ . Having thus reduced the per-vowel training sample size to only 24, the intra-speaker curves shown in Figure 4.7(c) indicate an even more striking degradation in the performance of the quadratic classifier, which continues the trend through Figures 4.7(a), (b), and (c), of an increasingly sub-optimal performance as the training sample size is nearly halved at each step. Our hypothesis is further supported by the accuracy curves shown in Figure 4.7(c), on the one hand by the gradually degenerating performance of the quadratic classifier as the spectral range is increased beyond about 3 kHz and the number of P-QCEP coefficients approaches the maximum of 14, and on the other hand by the generally superior performance of the quadratic classifier over the linear classifier in the very low spectral ranges up to about 1200 Hz where the number of P-QCEP coefficients is sufficiently small to avoid the “curse of dimensionality”.

Presumably owing to the smaller amount of consonant-induced variability in the selected centre frames of the vocalic nuclei, the inter-speaker curves shown in Figure 4.7(c) are slightly higher in accuracy compared with the earlier results using all seven leading consonants; those curves, however, are still intertwined and their relative levels of accuracy almost unaffected by the further reduction in training size to 90 samples per vowel. Indeed, even in Figure 4.7(a) where the inter-speaker experiments have recourse to 315 training samples per vowel, the quadratic classifier performs worse than the linear classifier over several spectral ranges. In order to test our hypothesis that the sub-optimal performance is caused by a weakening assumption of normality when the data of three, spectrally dissimilar speakers are pooled to train the classifier, we therefore reverse our method of data-partitioning such that the P-QCEP of only a single speaker

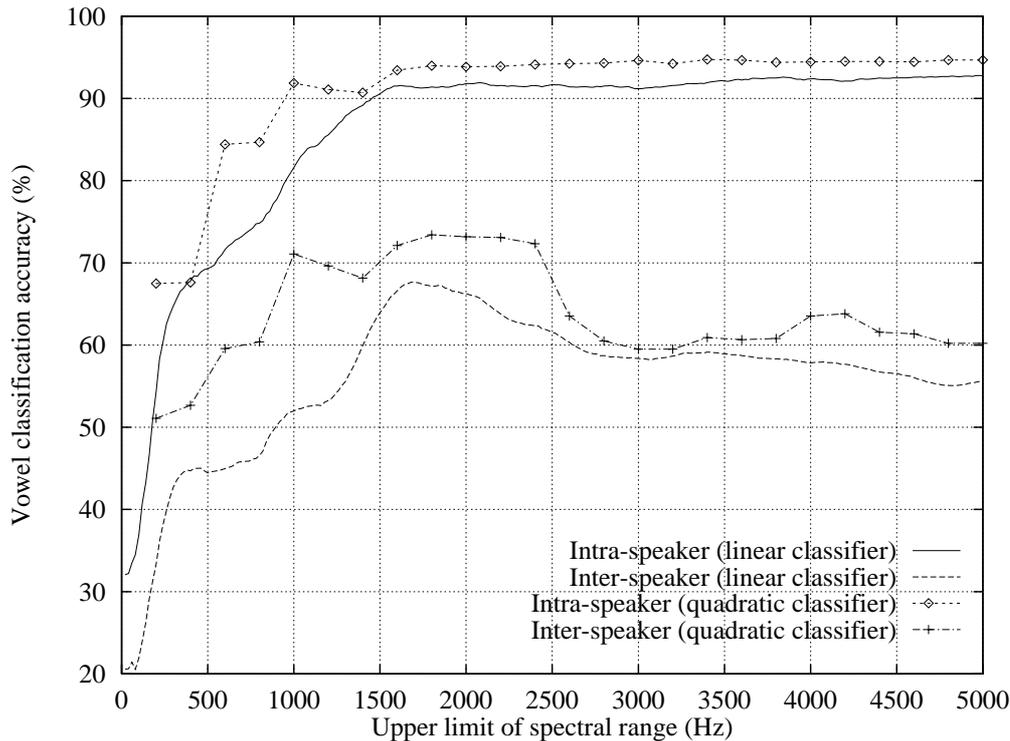


Figure 4.8: Intra- and inter-speaker vowel classification accuracy obtained with a linear classifier (solid and dashed curves, respectively) and with a quadratic classifier (diamond and plus symbols, respectively), as a function of the upper limit  $\theta_2$  of the spectral range  $[0, \theta_2]$ . Dataset (FC): 14<sup>th</sup>-order LP cepstra of the centre frames 5, 6, and 7 of 11 time-normalised frames in 5 repetitions of 9 vowels recorded in /CVd/ context by 4 adult, male speakers of Australian English. The number of P-QCEP coefficients used in the quadratic classifier at each spectral range was dictated by the step-wise function plotted in Figure 4.4. The *intra*-speaker curves are replicated from Figure 4.7(a), where for each speaker, one repetition at a time was left out of the training data and used as test data. The *inter*-speaker curves are here generated by training each classifier on one speaker’s data at a time, and using the remaining three speakers’ data as the test set — thus better satisfying the quadratic classifier’s underlying assumption of multivariate normality in the training parameter set.

at a time are used for training, the data of the remaining three speakers are used to test the classifier, and the results are averaged over the four separate experiments. Notwithstanding the inevitably pessimistic bias in the classification accuracies thus obtained, if the per-vowel distribution of the training data is indeed more nearly multivariate Gaussian on a speaker-dependent basis as we suspect, then the quadratic classifier should consistently perform better than the linear classifier.

In Figure 4.8 are shown the accuracy curves obtained by using the complete set of /CVd/ data in order to secure a sufficient training sample size (for the intra-speaker experiments in particular) and, in addition, by training the inter-speaker experiments on only one speaker at a time in order to better satisfy the quadratic classifier’s underlying

assumption of multivariate normality. Indeed, having finally met the criteria in regard to both dimensionality and normality, the quadratic classifier consistently performs *no worse* than the linear classifier, as expected both intuitively and theoretically. Furthermore, the phenomenon of vowel-speaker dichotomy is strongly conveyed in both sets of intra- and inter-speaker curves. The two, intra-speaker curves of vowel classification accuracy almost reach what might be regarded as their respective, asymptotic values as the spectral range is extended to a little over 1.5 kHz, and at full spectral range the linear and quadratic classifiers attain 92.8% and 94.7%, respectively. By contrast, the modified method of data-partitioning used with both the linear and quadratic classifier in the inter-speaker experiments, yields accuracy curves which attain a peak of 67.7% at 1680 Hz and 73.4% at 1800 Hz, respectively. As the spectral range is increased further to 3000 Hz, classification accuracy drops to only 58.4% and 59.5%, respectively, and at full spectral range the two curves attain only 55.7% and 60.2%, respectively.

The results presented in this section have not only confirmed the phenomenon of vowel-speaker dichotomy which we first unfolded in the previous section using the linear classifier, but have also upheld some inherent limitations of classification methods which rely on second- or higher-order statistics. In particular, our intra-speaker experiments have drawn attention to the importance of providing a sufficiently large number of training samples per class, compared with the dimensionality of each sample. Indeed, the “curse of dimensionality” was regarded by Rosenberg and Sambur (1975, p.173) as rendering the use of potentially more powerful, quadratic discriminant functions “impractical”, owing to the “sample size required to obtain reliable estimates of covariance matrices”. On the other hand, our inter-speaker experiments have underscored the frailty of embracing larger amounts of training data indiscriminately, without regard to the underlying assumptions of the classifier. In view of the inherent difficulties associated with the use of the quadratic classifier as evinced in this section, we shall retain only its emphatic confirmation of the vowel-speaker dichotomy, and henceforth revert to the more simple but stable classifier which uses hyper-planar decision boundaries.

## 4.4 The Dichotomy Explained

Thus far we have unfolded the phenomenon of vowel-speaker dichotomy by means of vowel classification experiments using the LP cepstrum. While those accuracy curves do provide compelling evidence of the dichotomous nature of vowel-speaker interactions across the spectral continuum, in this section we seek to explain the dichotomy in greater detail, and thereby gain a more insightful perspective on the phenomenon.

### 4.4.1 Spectral Regions of Vowel-Speaker Dichotomy

It is quite clear from our review of the literature (in Chapter 2) that the first two formant frequencies, and hence the spectral regions which include those two formants, carry the most important, acoustic information required to discriminate the vowels of any single speaker. From a vowel classification point of view, one would therefore expect the accuracies obtained in a speaker-dependent task to be relatively independent of the spectral regions which extend beyond the second formant. Indeed, the nearly-asymptotic behaviour of our intra-speaker accuracy curve (shown earlier in Figure 4.5) already confirms that implication.

However, a more informative perspective is offered in Figure 4.9 where, adjacent to the  $F_1F_2$  plane, are plotted those portions of the intra-speaker accuracy curve which encompass the  $F_1$  and the  $F_2$  ranges, respectively, of the four speakers' formant distribution. Classification accuracy rises to 65.7% as the upper limit of the spectral range is extended across the  $F_1$  range, to the lowest  $F_2$  (565Hz) of the formant distribution, and continues to rise gradually to 91.4% as the spectral range is further extended to the mid- $F_2$  of the four speakers' vowel formant distribution (hereafter defined as the mean  $F_2$  of the quasi-neutral vowel in /h3d/, at 1585Hz). By contrast, the inclusion of spectral information contained in the regions beyond the mid- $F_2$ , results in a nearly asymptotic increase of accuracy to 92.1% at the highest  $F_2$  (2391Hz), followed by a slight decrease to 90.2% at full range (5000Hz).

As noted earlier in Section 4.3.1, the highest accuracy (92.6%) is achieved at a spectral range of 2320Hz. The close proximity of that frequency value to the upper

limit of the  $F_2$  distribution renders our results in complete agreement with the traditional notion that the  $F_1F_2$  plane provides the best separation of vowels on an intra-speaker basis. However, our results further suggest that, at least for the data at hand, the most important spectral regions for intra-speaker vowel discrimination are those which extend to the mid- $F_2$  of the formant distribution. Indeed, the horizontal (dashed) line in Figure 4.9, which cuts through the centre of the ellipse drawn around the four speakers' formant distribution for /**ɜ**/ (the mid- $F_2$  defined earlier at 1585Hz), also intersects the intra-speaker accuracy curve at what might be regarded as its "knee". By contrast with the gradual improvement in accuracy as the spectral range is extended to the mid- $F_2$ , classification performance is relatively independent of the spectral regions which include those parts of the formant distribution lying above the mid- $F_2$ .

By analogy with Figure 4.9, a more enlightening perspective on the behaviour of our inter-speaker accuracy curve (first shown in Figure 4.5) is gained by plotting the relevant portions of that curve adjacent to the four speakers'  $F_2F_3$  plane, as shown in Figure 4.10. Classification accuracy rises to a maximum of 74.4 % as the spectral range is increased up to 1680Hz, which is only 95Hz higher than the mid- $F_2$  defined earlier. As shown by the vertical and horizontal (dashed) lines in Figure 4.10, the peak in accuracy is attained at a spectral range which is yet to embrace the majority of the  $F_2$  of the four speakers' front vowels, nor any of their  $F_3$  distribution. Indeed, the inclusion of spectral information contained in those regions beyond 1680Hz, results in a drop in accuracy to 70.0 % at the highest  $F_3$  (3240Hz), and to 66.0 % at full range (5000Hz). As noted earlier in Section 4.3.1, the detrimental influence of inter-speaker variability is most strongly manifest in the cascade of falls in accuracy from the peak to only 66.9 % at 2980Hz. Figure 4.10 clearly emphasises the acoustic-phonetic relevance of that higher spectral region, and shows that the cumulative drop of 7.5 % in accuracy occurs as the spectral range is extended to include the  $F_2$  of the speakers' front vowels, and nearly their entire  $F_3$  distribution.

If the dichotomy observed above is indeed a result of inter-speaker differences causing confusion in vowel classification, then one might expect a large concentration of speaker variance in those spectral regions where vowel classification accuracy has been observed to decrease. The methodology adopted to decompose the vowel and

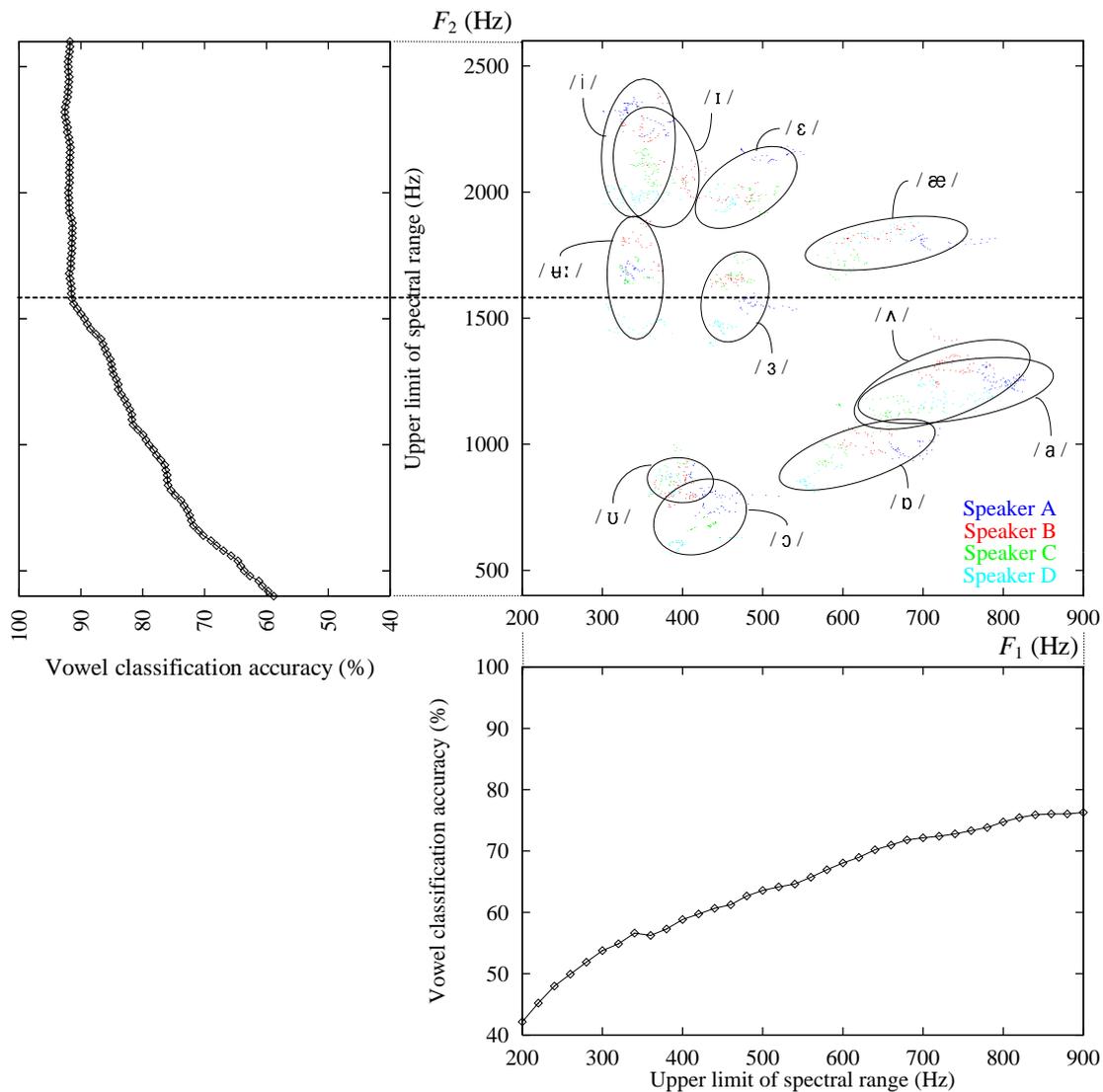


Figure 4.9:  $F_1F_2$  vowel space of all 4 male speakers (FC dataset), with a  $2\sigma$  ellipse drawn around each vowel cluster. Adjacent to the abscissa and ordinate are plotted the portions of the *intra*-speaker accuracy curve (from Figure 4.5) which span the  $F_1$  and the  $F_2$  ranges, respectively. The horizontal (dashed) line cuts through the centre of the ellipse for the quasi-neutral vowel /ɜ:/; the mid- $F_2$  thus defined, also intersects the accuracy curve at what might be regarded as its “knee”.

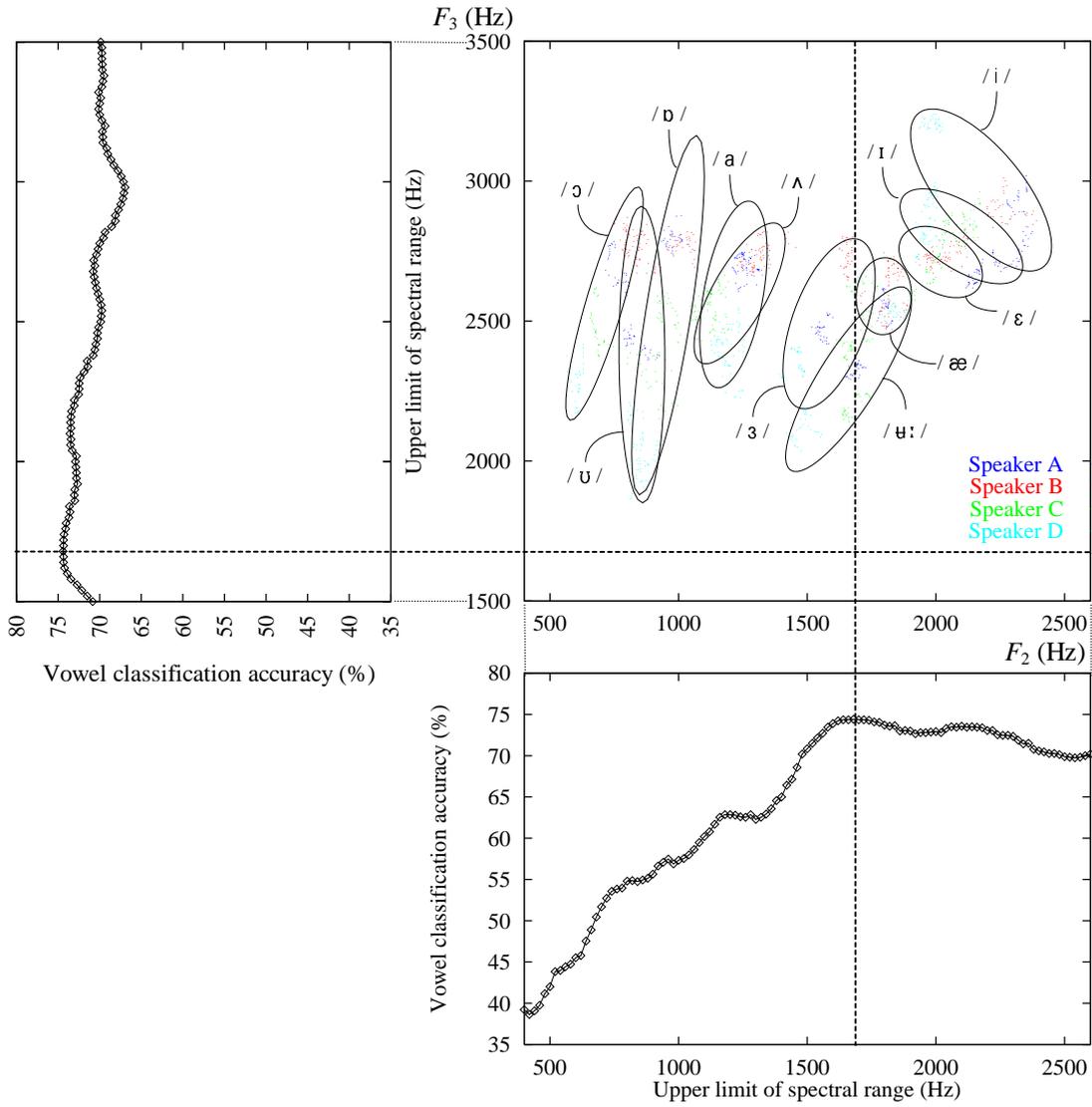


Figure 4.10:  $F_2F_3$  vowel space of all 4 male speakers (FC dataset), with a  $2\sigma$  ellipse drawn around each vowel cluster. Adjacent to the abscissa and ordinate are plotted the portions of the *inter*-speaker accuracy curve (from Figure 4.5) which span the  $F_2$  and the  $F_3$  ranges, respectively. The vertical and horizontal (dashed) lines intersect the accuracy curve at its peak (at 1680Hz), and cut across the formant plane in order to emphasise the acoustic-phonetic relevance of the spectral regions of primary phonetic or speaker influence.

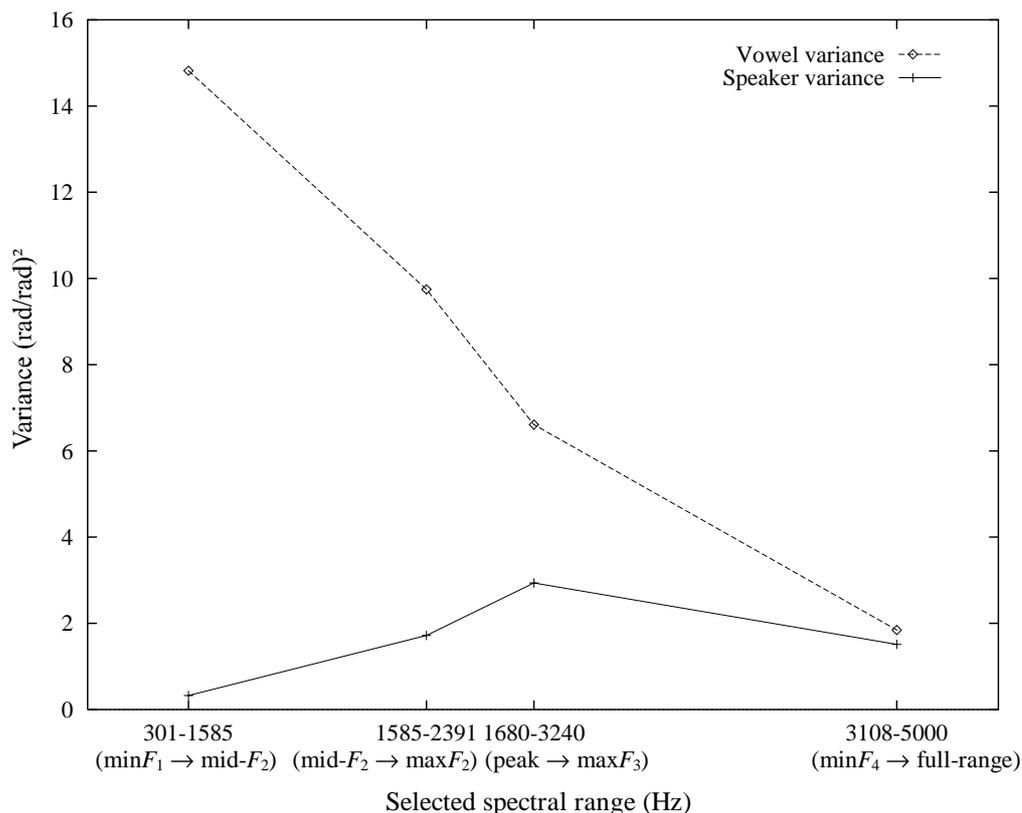


Figure 4.11: Profile of *speaker variance* and *vowel variance* in the FC dataset, computed using the parametric cepstral distance measure (PCD, see Section 4.2.3.1) in selected frequency sub-bands. The four sub-bands were selected on the basis of the formant ranges of the 4 speakers' vowel data which, as shown in Figures 4.9 and 4.10, correspond closely with the prominent features of the accuracy curves. The first sub-band extends from the lowest  $F_1$  to the mid- $F_2$  (which is close to the 'knee' of the intra-speaker accuracy curve). The second sub-band extends from the mid- $F_2$  to the highest  $F_2$ . The third sub-band extends from the peak in the inter-speaker accuracy curve (which is only 95Hz higher than the mid- $F_2$ ) to the highest  $F_3$ . The fourth sub-band extends from the lowest  $F_4$  to the full spectral range.

speaker contributions to the total variance is similar to that which Pols et al. (1973) and van Nierop et al. (1973) used in their frequency analysis of Dutch vowels. In particular, the vowel contribution is the variance in the 11 vowel centroids found by averaging over the data of all four speakers per vowel, while the speaker contribution is the variance in the 4 speaker centroids found by averaging over the data of all vowels per speaker. In our variance computations, however, we use the parametric cepstral distance measure (PCD, derived in Section 4.2.3.1), which allows greater flexibility than has hitherto been possible in the selection and analysis of frequency sub-bands.

Figure 4.11 shows the amount of speaker variance (plus symbols joined with solid lines) and vowel variance (diamond symbols joined with dashed lines) computed in each of four selected frequency sub-bands, which themselves were chosen on the basis of our

speakers' vowel formant distribution, and the corresponding, prominent features of the accuracy curves shown in Figures 4.9 and 4.10. The largest amount of vowel variance is indeed found in the two lowest frequency sub-bands which together span the entire  $F_1F_2$  range. In particular, the spectral region which we earlier found to have the most significant contribution to classification accuracy on an intra-speaker basis (that which extends to the mid- $F_2$  at 1585Hz), is here shown to have both the largest vowel variance and the smallest speaker variance. By contrast, vowel variance sharply drops and speaker variance rises to a maximum in the very spectral region where inter-speaker vowel classification accuracy was earlier observed to decrease, from its peak at 1680 Hz to the highest  $F_3$  at 3240Hz. In the highest frequency sub-band which extends from the lowest  $F_4$  (3108Hz) to the full range (5000Hz), vowel variance reduces even further to a level which is comparable to speaker variance in the same spectral range.

Whilst our analyses of variance can only provide an overall measure of spectral dispersion, and therefore cannot be expected to yield a more detailed explanation of the accuracy curves which themselves were determined by multi-speaker, vowel-to-vowel interactions, the profiles of spectral variance shown in Figure 4.11 do confirm the vowel and speaker-specific regions which we first unfolded by way of classification experiments. Indeed, the variance profiles computed over the broad spectral regions defined by the formant ranges which were also found to correspond with prominent features of the accuracy curves, do provide evidence for a high concentration of vowel variance in the low spectral regions, and for a high concentration of speaker variance together with a decreasing vowel variance in the higher spectral regions. Subtle differences in methodology and spectral representation notwithstanding, our conclusions regarding the spectral regions of primary vowel and speaker influence are in agreement with the spectral variance analyses reported for female speakers of Dutch (van Nierop et al., 1973, Fig.1), and more recently for male speakers of Japanese (Kitamura and Akagi, 1994, Fig.3).

#### **4.4.2 Acoustic-Phonetic Explanation of the Dichotomy**

Thus far in our explanation of the vowel-speaker dichotomy, we have considered only the gross features of the accuracy curves, which we related first to the speakers' vowel

formant distribution, then to the relative amount of vowel and speaker variance in broad frequency sub-bands. In particular, we have observed that vowel-speaker interactions are more strongly manifest in the spectral regions spanning the high- $F_2$  and the  $F_3$  of the speakers' vowel formant distribution. Indeed, Figure 4.10 clearly indicates that inter-speaker vowel classification accuracy begins to drop as the spectral range is extended to include the  $F_2$  of the speakers' front vowels and the  $F_3$  of their back vowels, the overall effect of which was identified earlier as the phenomenon of dichotomy. This apparent consistency between certain formant ranges and prominent features of the classification accuracy curves encourages a more detailed phonetic analysis of the vowel misclassifications which occur across the higher spectral regions, with a view towards identifying the vocalic subspaces which have caused the dichotomy.

To this end, first recall from our description (in Section 4.2.3) of the data-partitioning method adopted in our vowel classification experiments, that the inter-speaker accuracy curve is the mean of four curves, each generated by training the classifier on the data of three speakers, and testing on the data of the remaining, fourth speaker. Each of those four, per-speaker accuracy curves can be further decomposed in terms of the contribution to accuracy of each vowel tested. We therefore examined each of those 44 accuracy curves (11 vowels  $\times$  4 speakers), in order to determine the vowel and speaker contributions to the drop in accuracy observed across the higher spectral regions, and thus providing a more detailed acoustic-phonetic interpretation of the misclassifications which have caused the dichotomy.

The results of this manual decomposition of the inter-speaker accuracy curve are shown in Table 4.1, where the entry in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column lists: (i) the vowel(s) confused with the  $i^{\text{th}}$  vowel of the  $j^{\text{th}}$  speaker, (ii) the formant(s) responsible for the confusion, and (iii) the degree of contribution to the dichotomy as measured by the percentage drop in accuracy (to the nearest 5%) in each individual curve. Our aim being to provide an acoustic-phonetic explanation of the main drop in the inter-speaker accuracy curve across the higher spectral regions, only those misclassifications which directly contribute to that overall drop across the mid- $F_2$  and  $F_3$  ranges are considered. The formant(s) deemed responsible for the confusions in each case, are determined by

Vowel	Speaker			
	A	B	C	D
/i/	/ɔ, ʊ:/ (F <sub>2</sub> &F <sub>3</sub> , 10%)	/ɪ/ (F <sub>3</sub> , 50%)	/ɪ/ (F <sub>2</sub> , 15%; F <sub>3</sub> , 40%)	/ɪ/ (F <sub>2</sub> , 30%)
/ɪ/	/i/ (F <sub>2</sub> , 5%)	/ɛ/ (F <sub>3</sub> , 10%)	/i/ (F <sub>2</sub> , 10%; F <sub>3</sub> , 5%) /ɛ/ (F <sub>3</sub> , 30%)	/ɛ/ (F <sub>2</sub> , 80%)
/ɛ/	/i/ (F <sub>2</sub> , 100%)			
/æ/		/ɛ, ʊ/ (F <sub>2</sub> , 10%) /ɛ/ (F <sub>3</sub> , 15%)		
/a/				/ʌ/ (F <sub>3</sub> , 90%)
/ʌ/	/a/ (F <sub>3</sub> , 20%)	/a/ (F <sub>3</sub> , 5%)		
/ɒ/		/a/ (F <sub>3</sub> , 40%) /ʊ/ (F <sub>3</sub> , 5%)	/a/ (F <sub>3</sub> , 15%) /ʊ/ (F <sub>3</sub> , 35%)	/æ/ (F <sub>2</sub> &F <sub>3</sub> , 5%) /æ, ɔ/ (F <sub>3</sub> , 30%)
/ɔ/	/æ/ (F <sub>3</sub> , 15%)		/ʊ/ (F <sub>3</sub> , 5%)	/ɪ, ɛ/ (F <sub>2</sub> &F <sub>3</sub> , 15%) /ɜ/ (F <sub>3</sub> , 10%)
/ʊ/	/ɜ/ (F <sub>3</sub> , 15%)			/ɔ, æ/ (F <sub>3</sub> , 20%)
/ʊ:/	/ɜ/ (F <sub>3</sub> , 10%)	/ɜ/ (F <sub>3</sub> , 15%) /ɛ/ (F <sub>3</sub> , 20%)		/ɜ/ (F <sub>2</sub> , 10%) /i/ (F <sub>2</sub> &F <sub>3</sub> , 10%) /i/ (F <sub>3</sub> , 15%)
/ɜ/		/ʊ:/ (F <sub>2</sub> , 25%) /ɔ/ (F <sub>3</sub> , 60%) /ɛ, æ/ (F <sub>3</sub> , 10%)	/ʊ:/ (F <sub>2</sub> , F <sub>3</sub> , 60%)	/i, ʌ/ (F <sub>2</sub> &F <sub>3</sub> , 10%)

Table 4.1: Acoustic-phonetic decomposition of the dichotomy in inter-speaker vowel classification behaviour (Figure 4.5), in terms of the vowel misclassifications that contribute to the drop in accuracy across the higher spectral regions which encompass the high- $F_2$  and the  $F_3$  of the speakers' vowel formant distribution. ' $F_2$ & $F_3$ ' indicates misclassifications caused by overlapping  $F_2$  and  $F_3$  ranges.

relating the spectral region of the drop in accuracy, to the speakers'  $F_1F_2$  and  $F_2F_3$  formant distributions shown, respectively, in Figures 4.9 and 4.10.

Misclassifications are observed to arise in those higher spectral regions, from confusions: (i) between the  $F_3$  of back vowels and the  $F_2$  of front vowels, (ii) amongst the  $F_3$  of front vowels, (iii) amongst the  $F_3$  of back vowels, and (iv) amongst the  $F_2$  of front vowels, in this increasing order of detrimental effects. An example of the first category of misclassifications is the drop of about 15% for the back vowel /ɔ/ of speaker D as the upper limit of the spectral range is increased from about 2.1kHz to 2.3 kHz, owing to confusions with the front vowels /ɪ/ and /ɛ/. Referring to the  $F_2F_3$  formant distribution, we infer that the confusions are most likely caused by the  $F_3$  of the back vowel overlapping with the  $F_2$  of the (other speakers') front vowels.

Confusions amongst the  $F_3$  of front vowels include the drop in accuracy of about

50% in /i/ of speaker B across the spectral region from about 2.7 kHz to 2.9 kHz, owing to misclassifications with the neighbouring vowel /ɪ/. Examples of confusions amongst the  $F_3$  of back vowels include the drop of 40% in /ɒ/ of speaker B owing to misclassifications with the vowel /a/, and the drop of 90% in /a/ of speaker D caused by misclassifications with the highly overlapping vowel /ʌ/. In addition to the confusions between and amongst clearly front or back vowels, are the misclassifications of the quasi-neutral vowel /ɜ/ of speakers B and C, which occur respectively across the  $F_3$  range with /ɔ/, and across both the  $F_2$  and  $F_3$  ranges with /ɝ:/.

However, the largest single contribution to the dichotomy occurs across the spectral region from about 1.8 kHz to 2.4 kHz, where the accuracy curve for the vowel /ɛ/ of speaker A drops by 100% owing to confusions with /i/ in the  $F_2$  range of those front vowels. Similarly across the same spectral region, classification accuracy of the front vowel /ɪ/ of speaker D drops by 80% owing to confusions with the neighbouring vowel /ε/ in the  $F_2$  range.

The vowel confusions shown in Table 4.1 and discussed above, are a direct consequence of vowel-speaker interactions in the higher spectral regions, and as such, they embody the specific types of speaker differences that are manifest in our spoken vowel data. In this context, it is pertinent to recall from our review of the literature (in Chapter 2) that Sambur (1975) found the  $F_2$  of front vowels and the  $F_3$  of the back vowel /u/ to be the most speaker-discriminating formant parameters for those spoken vowels of American English. More recently, Mella's (1994) study of French spoken vowels has also shown that  $F_3$  of /u/ and  $F_2$  of front vowels are the best formant parameters for speaker identification. These findings are further supported by van den Heuvel et al.'s (1993) study which showed that for the Dutch vowels /i/ and /a/, the largest amount of speaker variability is contained in the spectral regions near the  $F_2$  and the  $F_3$  of those vowels, respectively. Whilst these studies were concerned either with speaker identification or with direct measures of inter-speaker variability on a per-vowel basis, they support our own conclusions regarding the relative speaker-specificity of different regions of the acoustic-phonetic vowel space, which we have here determined by decomposing the vowel-speaker interactions which caused a drop in inter-speaker vowel classification accuracy across the higher spectral regions.

### 4.4.3 Dependence on Spectral Representation

Thus far we have unfolded, and provided an acoustic-phonetic explanation of the vowel-speaker dichotomy, by way of vowel classification experiments performed using either the original, 14<sup>th</sup>-order LP (quefrequency-weighted) cepstrum, or the partial but equivalent spectral representation afforded by the P-QCEP. In both cases, the cepstrum, and hence the spectral shape of each vocalic frame of speech, is determined by a combination of all the poles (which include typically six or seven complex-conjugate pairs) yielded by LP analysis. Amongst those poles are usually found the formants, or resonances of the vocal tract which, owing to their relatively smaller bandwidths, generally have the strongest influence in shaping the prominent, local peaks in the spectral envelopes of spoken vowels.

In view of the well-known articulatory relevance of the formants, an important question is therefore whether our methodology of unfolding the vowel-speaker dichotomy can be applied with equal success using so-called *simplified cepstra* which, as defined earlier (in Section 3.4.2), contain only formant information. This question is of particular relevance to our forthcoming validation of the dichotomy in Sections 4.4.4 and 4.4.5 using the PB and JB datasets, which contain only the centre-frequencies of the first three formants of each vowel token. It is of equally vital concern to establish the dependence of the vowel-speaker dichotomy on the formant parameters, in view of our forthcoming articulatory explanation in Chapter 6. In this section, we therefore test the robustness of the dichotomy by performing vowel classification experiments using simplified cepstra computed from the formants which, as described in Chapter 3, were carefully measured in each of the seven, vocalic steady-state frames of the recorded /hVd/ monosyllables which make up the FC dataset.

If we are to maintain consistency with the number of acoustic parameters afforded by the PB and JB datasets, then we must restrict ourselves to only the first three, measured formant frequencies, and use fixed formant bandwidths. Rather than use arbitrary bandwidths, however, we take advantage of our formant measurements and fix the bandwidths to their respective mean values computed over all eleven vowels and four speakers ( $\bar{B}_1 = 99$  Hz,  $\bar{B}_2 = 128$  Hz, and  $\bar{B}_3 = 218$  Hz). Similarly in the following

section where the PB dataset will be used to validate the dichotomy, we are fortunate in having recourse to Dunn's (1961) measurements of the mean formant bandwidths computed over a subset of 20 of the adult male speakers.

With only three formants, the order of the LP polynomial is perforce  $M = 6$ , and thus only the first six, simplified cepstral coefficients are required to fully specify the model. In view of the analysis conditions used to derive the original LP cepstra ( $NCC = M = 14$ ), an important question then arises regarding the number of simplified cepstral coefficients required to secure a *spectral representation* as close as possible to the original. This question is easily answered by computing the distance between corresponding pairs of original (14<sup>th</sup>-order) and simplified cepstra, for a range of values of  $NCC$  selected in generating the latter. The solid curve (joining the diamond symbols) in Figure 4.12 shows the mean of those distances computed over all 1540 frames of the FC dataset, using the PCD selected over the full spectral range [0, 5000]Hz. Despite the much lower LP order of  $M = 6$ , the global minimum of that curve is obtained when the number of simplified cepstral coefficients is equal to the original  $NCC = 14$ .

As the simplified cepstra thus generated do not contain any poles above the highest  $F_3$  (3240Hz), an inevitable mismatch occurs between the original and the simplified cepstra in the highest available spectral regions which encompass the fourth and possibly the fifth formants. Distances obtained using the PCD would therefore be expected to be smaller, if the selected spectral range is restricted to exclude those highest regions. Indeed, the dashed curve (joining the plus symbols) in Figure 4.12 shows that when the selected spectral range is limited to extend only to the highest  $F_3$ , smaller mean distances are obtained at values of  $NCC$  not far from the global minimum, while the minimum itself remains at  $NCC = 14$ . It is also interesting to note from the curves in Figure 4.12, that as the simplified spectral representation is either smoothed or enhanced beyond a certain level (in particular, when  $NCC \leq 7$  or  $NCC \geq 25$ , respectively), the mean distances computed over the narrower spectral range begin to exceed those computed over the full spectral range. This result is to be expected if the excessive blurring or sharpening, respectively, of the formant peaks in the simplified spectra can be assumed to yield distances which are, on average, larger than those obtained over the mismatched spectral regions lying above the highest  $F_3$ .

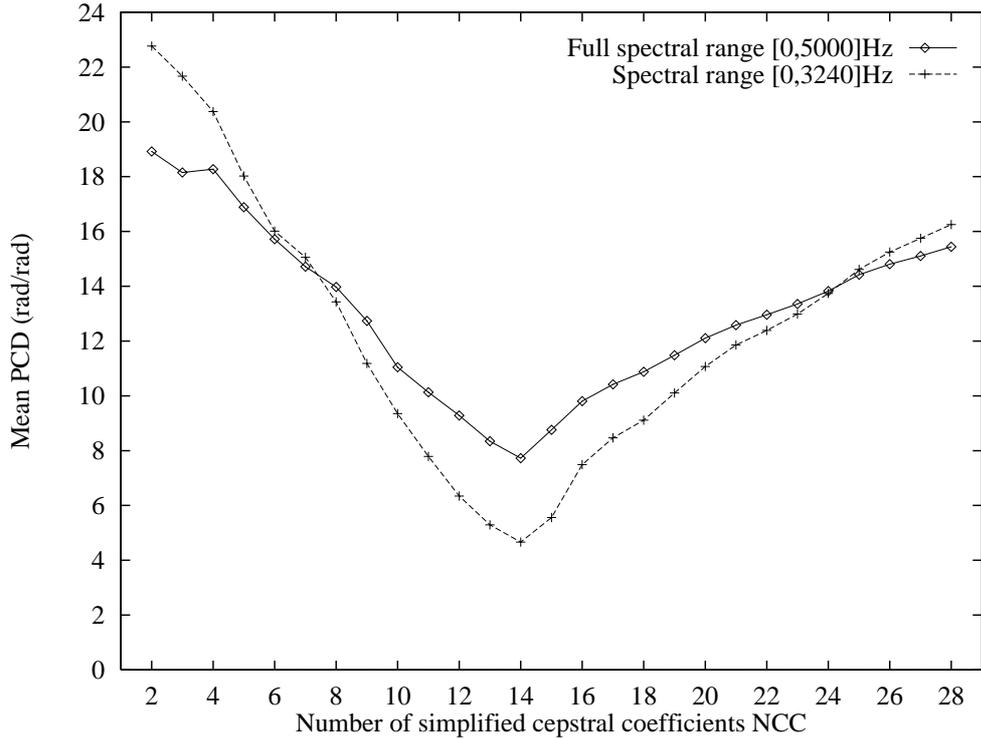


Figure 4.12: Mean distances computed over the entire FC dataset, between *original cepstra* ( $NCC=M=14$ ) and corresponding, *simplified cepstra* ( $M=6$ ,  $NCC=2,3,\dots,28$ ) generated from the first three, measured formant frequencies, with bandwidths fixed to mean values ( $B_1=99\text{Hz}$ ,  $B_2=128\text{Hz}$ ,  $B_3=218\text{Hz}$ ), and sampling frequency  $F_s=10\text{kHz}$ . Quefrequency-weighted cepstral distances were computed using the PCD, first over the full spectral range  $[0,5000]\text{Hz}$  (diamond symbols joined with solid lines), then over the spectral range  $[0,3240]\text{Hz}$  which extends only to the highest  $F_3$  (plus symbols joined with dashed lines).

In sum, the mean cepstral distances shown in Figure 4.12 strongly suggest that if the simplified cepstra are generated using the original sampling frequency  $F_s = 10\text{kHz}$ , then the closest spectral representation to the original is obtained by retaining the same number of simplified cepstral coefficients  $NCC = 14$ , despite the much lower LP order of  $M = 6$  which inevitably results from using only the first three formants.

Nevertheless, even with  $NCC = 14$  and the PCD computed over the reduced spectral range, a non-negligible, mean distance of about  $4.7\text{ rad/rad}$  is obtained, which is due to at least four main differences between the original and the simplified spectral representations. The first and perhaps most obvious difference is that the simplified cepstra are explicitly cleansed of the influence of spurious poles. Whilst the NDPS spectral representation is relatively insensitive to real poles (e.g., Yegnanarayana, 1978) and to complex-conjugate, spurious poles which are usually of much wider bandwidth

than their true formant counterparts, part of the mean distance noted above may be due to the existence of narrower-bandwidth spurious poles in the original cepstra, especially those which tend to distort the spectral valleys between formant peaks.

The cause of the second type of difference lies in our use of fixed formant bandwidths in generating the simplified cepstra. Although the degree of mismatch is perhaps lessened by using the means of the measured bandwidths, and although the NDPS-based cepstral distance measure has been observed to be relatively less sensitive to bandwidth variations than to differences in formant centre-frequency (Clermont and Mokhtari, 1994), the amplitude of each formant peak is known to be inversely proportional to its bandwidth in the NDPS spectral representation (e.g., Yegnanarayana, 1978), and large differences in bandwidth might therefore contribute to at least a portion of the mean distance noted above.

Perhaps a more consequential difference between the original and the simplified cepstra, is due to the fixed LP analysis conditions used to obtain the former, which in many cases differed from the analysis conditions finally adopted in measuring the formants themselves (as explained in Chapter 3). It can therefore be expected that the formant peaks in the simplified spectral representation are not always as clearly defined in the original cepstrum. This might indeed contribute a major portion of the mean distance noted above.

The fourth and perhaps least consequential difference, is caused by the higher parts of the speakers'  $F_3$  range overlapping with the lower parts of their  $F_4$  range. Indeed, the lowest  $F_4$  was measured at 3108Hz, which implies an overlapping spectral region of width 132Hz. Any  $F_4$  peaks which are present in the original spectra might contribute to larger distances within that spectral range.

Bearing in mind these potential differences in spectral representation, a new set of intra- and inter-speaker vowel classification experiments were performed as a function of an increasing spectral range, using the linear classifier and methods of data-partitioning and cepstral distance computation described earlier in Section 4.2.3, but using the simplified cepstra with  $NCC = 14$ . The resulting accuracy curves are compared in Figure 4.13 with those obtained earlier using the original cepstra. Although the intra-speaker curves are quite similar in their overall behaviour, the simplified

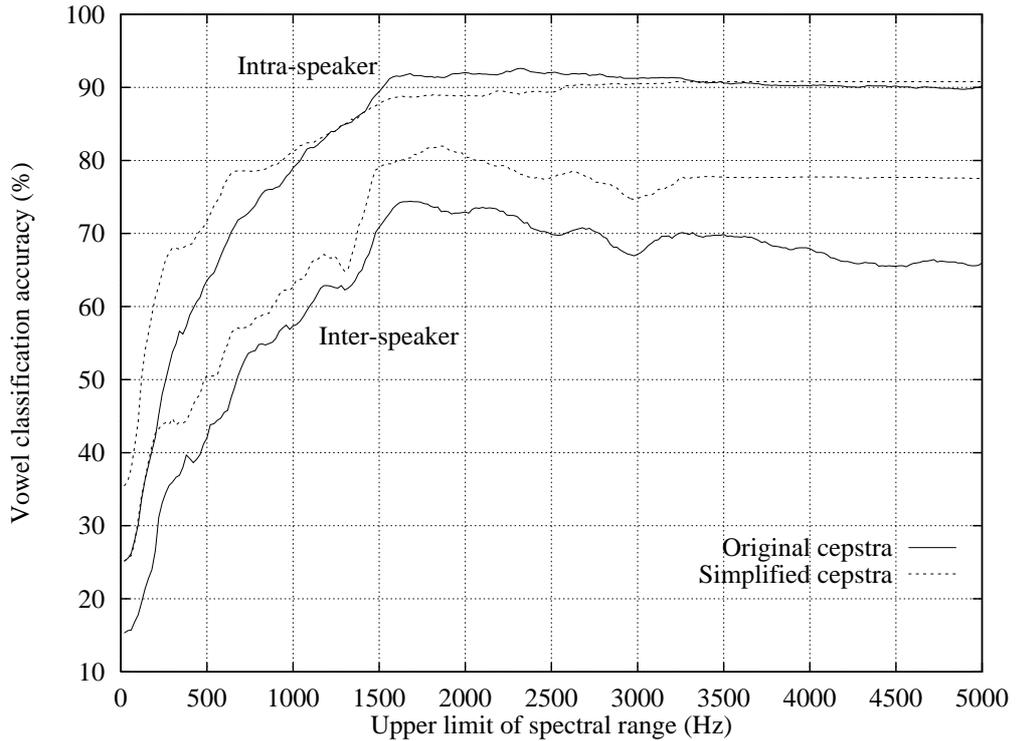


Figure 4.13: Intra- and inter-speaker vowel classification accuracy yielded by a linear classifier, as a function of the upper limit  $\theta_2$  of the spectral range  $[0, \theta_2]$ , using either the *original cepstra* (solid curves) or the *simplified cepstra* (dashed curves) generated from the first three, measured formant frequencies, with bandwidths fixed to mean values ( $B_1=99\text{Hz}$ ,  $B_2=128\text{Hz}$ ,  $B_3=218\text{Hz}$ ), sampling frequency  $F_s=10\text{kHz}$ , and  $NCC=14$ . Dataset (FC): LP cepstra of 7 vocalic steady-state frames in 5 repetitions of 11 vowels recorded in /hVd/ context by 4 adult, male speakers of Australian English.

cepstra yield a more nearly-asymptotic curve, which more closely matches our expectations regarding the well-known tendency of the  $F_3$  and higher spectral regions to provide only minor improvements to vowel discrimination on an intra-speaker basis.

By contrast, the inter-speaker curve yielded by the simplified cepstra is consistently higher in accuracy than that obtained using the original cepstra. Clearly, the removal of spurious poles, the more consistent definition of formant peaks, and the effective normalisation of formant bandwidths in the simplified spectral representation, lead to an improved performance of vowel discrimination on an inter-speaker basis. Nevertheless, that improvement appears to be manifest across the entire spectral range, and the phenomenon of vowel-speaker dichotomy which is signalled by the drop in inter-speaker vowel classification accuracy across the high- $F_2$  and  $F_3$  spectral regions, is therefore retained. Indeed, classification accuracy rises to a peak of 81.9% as the spectral range is extended to 1860Hz, and subsequently drops to 74.7% at 2980Hz.

Vowel	Speaker			
	A	B	C	D
/i/		/ɪ/ (F <sub>3</sub> , 20%)	/ɪ/ (F <sub>2</sub> , 55%; F <sub>3</sub> , 15%)	/ɪ/ (F <sub>2</sub> , 10%)
/ɪ/	/i/ (F <sub>2</sub> , 30%)	/ɛ/ (F <sub>3</sub> , 25%)	/i/ (F <sub>2</sub> , 5%)	/ɛ/ (F <sub>2</sub> , 15%) /ʉ:/ (F <sub>2</sub> &F <sub>3</sub> , 5%)
/ɛ/	/i/ (F <sub>2</sub> , 100%)			
/æ/				
/a/				/ʌ/ (F <sub>3</sub> , 100%)
/ʌ/	/a/ (F <sub>3</sub> , 5%)	/a/ (F <sub>3</sub> , 10%)		
/ɒ/	/a/ (F <sub>3</sub> , 5%)	/a/ (F <sub>3</sub> , 5%) /ʌ/ (F <sub>3</sub> , 5%)	/a/ (F <sub>3</sub> , 20%) /ʊ/ (F <sub>3</sub> , 45%)	/ʊ/ (F <sub>3</sub> , 70%) /æ/ (F <sub>2</sub> &F <sub>3</sub> , 5%)
/ɔ/				/ɛ/ (F <sub>2</sub> &F <sub>3</sub> , 5%)
/ʊ/	/ɔ/ (F <sub>3</sub> , 45%)	/ɒ/ (F <sub>3</sub> , 55%) /ɔ/ (F <sub>3</sub> , 20%)		
/ʉ:/		/ɪ/ (F <sub>3</sub> , 25%) /ɜ/ (F <sub>3</sub> , 15%)		/ɜ/ (F <sub>2</sub> , 15%)
/ɜ/		/ʉ:/ (F <sub>2</sub> , 5%) /æ/ (F <sub>3</sub> , 5%)	/ʉ:/ (F <sub>2</sub> , F <sub>3</sub> , 15%)	

Table 4.2: Acoustic-phonetic decomposition of the dichotomy in inter-speaker vowel classification behaviour for the FC dataset using *simplified cepstra* derived from the first three formant frequencies with bandwidths fixed to mean values ( $B_1=99\text{Hz}$ ,  $B_2=128\text{Hz}$ ,  $B_3=218\text{Hz}$ ), in terms of the vowel misclassifications that contribute to the drop in accuracy across the higher spectral regions which encompass the high- $F_2$  and the  $F_3$  of the speakers' vowel formant distribution. ' $F_2\&F_3$ ' indicates misclassifications caused by overlapping  $F_2$  and  $F_3$  ranges.

Although the frequency location of the peak in accuracy is nearly 200Hz higher than the peak obtained using the original cepstra, a comparison with the four speakers'  $F_1F_2$  and  $F_2F_3$  vowel formant distributions shown earlier in Figures 4.9 and 4.10 reveals that the turning point in the accuracy curve still occurs as the spectral range is extended to include the  $F_2$  of front vowels and the  $F_3$  of back vowels.

In view of the general similarities observed in the behaviour of vowel classification accuracy across the available spectral range using either simplified or original cepstra, the question then arises whether the drop in the inter-speaker accuracy curve across the higher spectral regions is caused by speaker-induced vowel confusions similar to those which we listed earlier in Table 4.1. The results of our vowel-speaker decomposition of the inter-speaker curve yielded by the simplified cepstra, shown in Table 4.2, indeed confirm that the majority of so-called dichotomous vowel confusions which occur

across those higher spectral regions, are caused by the same combinations of formants, vowels and speakers found earlier. In particular, the two most significant contributions to the dichotomy still occur, respectively, across the  $F_2$  range as the vowel / $\epsilon$ / of speaker A is confused with / $i$ /, and across the  $F_3$  range as the vowel / $a$ / of speaker D is confused with / $\Lambda$ /.

Despite those and other similarities, the magnitude of the percentage drops in accuracy shown in the entries of Table 4.2 are often different to those listed earlier in Table 4.1, and certain vowel misclassifications appear to replace others listed earlier. For example, there are three relatively large drops in accuracy which involve the  $F_3$  range of the vowel / $U$ /, none of which were observed earlier using the original cepstra: a drop of about 70% occurs as / $\mathfrak{D}$ / of speaker D is confused with / $U$ /; a drop of 55% occurs as / $U$ / of speaker B is confused with / $\mathfrak{D}$ /; and a drop of 45% occurs as / $U$ / of speaker A is confused with / $\mathfrak{C}$ /.

In the high front vowels, a redistribution of the drops in accuracy caused by misclassifications of / $i$ / as / $\mathfrak{I}$ / occurs, which deemphasises the contribution of the  $F_3$  range of speaker B, and emphasises instead the  $F_2$  range of speaker C. Also, misclassifications of the quasi-neutral vowel / $\mathfrak{Z}$ / of those two speakers either no longer occur, or are much reduced in effect.

These examples serve to illustrate the inaccuracy of the naive assumption, that the LP cepstrum measured with fixed analysis conditions in the steady-state of vocalic nuclei will clearly and consistently reflect the formant structure of those speech sounds. Even the formant-enhancement property of the NDPS cannot entirely compensate for the slight aberrations in the frequency-location of formant peaks, their bandwidths, and the largely unknown and undocumented effects of spurious poles yielded by the usual methods of LP analysis. As a result, our detailed decomposition of the individual vowel confusions contributing to the dichotomy has revealed certain differences compared with the dichotomous confusions obtained using the original cepstra, and those differences can be attributed directly to the types of differences in spectral representation discussed earlier. Indeed, one might even expect that the dichotomous misclassifications which occur using simplified cepstra, are either more readily, or more directly related to the speakers' formant ranges, and that a more accurate, acoustic-phonetic explanation of the dichotomy will therefore follow from the simplified spectral

representation.

In this light, it is encouraging to note that the vowel-speaker decomposition of the dichotomy shown in Table 4.2 confirms our previous results which implicate confusions primarily amongst the  $F_2$  of front vowels, amongst the  $F_3$  of back vowels, and amongst the  $F_3$  of front vowels. It is indeed this broad acoustic-phonetic explanation of the dichotomy which is to be of long-lasting value in our interpretation of vowel-speaker interactions, and which now permits us to consider using simplified cepstra to further explore the phenomenon of dichotomy with the PB and JB datasets.

#### **4.4.4 Dependence on Speaker Homogeneity**

The clearly dichotomous, inter-speaker accuracy curves presented thus far, have been generated using a population of speakers who differ quite significantly in the spectral properties of their vowels. An intriguing question then arises, whether vowel-speaker interactions will yield a similar dichotomous behaviour in inter-speaker classification of vowels spoken by a more densely populated, and perhaps more homogeneous group of speakers. To answer this question, we now turn to the PB dataset which, as described in Chapter 3, comprises two repetitions of ten vowels spoken in /hVd/ context by 32 adult, male speakers of American English. Although very little is known about the dialectal composition of those speakers, we have argued (in Chapter 3) that the homogeneity of the group as a whole is probably enhanced by the sheer number of speakers. If that assumption is indeed correct, then the PB dataset offers the possibility of assessing the dependence of the vowel-speaker dichotomy on the degree of speaker homogeneity.

In order to perform vowel classification experiments using the methodology presented earlier, we first need to convert the available formant data into cepstra. In this context, it is encouraging to recall our results of the previous section which confirmed that the dichotomous behaviour of inter-speaker vowel classification accuracy is indeed preserved when the spectral representation is simplified to contain only formant information. We therefore proceed to generate simplified cepstra from the first three formant frequencies of the PB dataset, using the LP method described in Section 3.4.2. Analogously to our methodology used in the previous section, the formant bandwidths

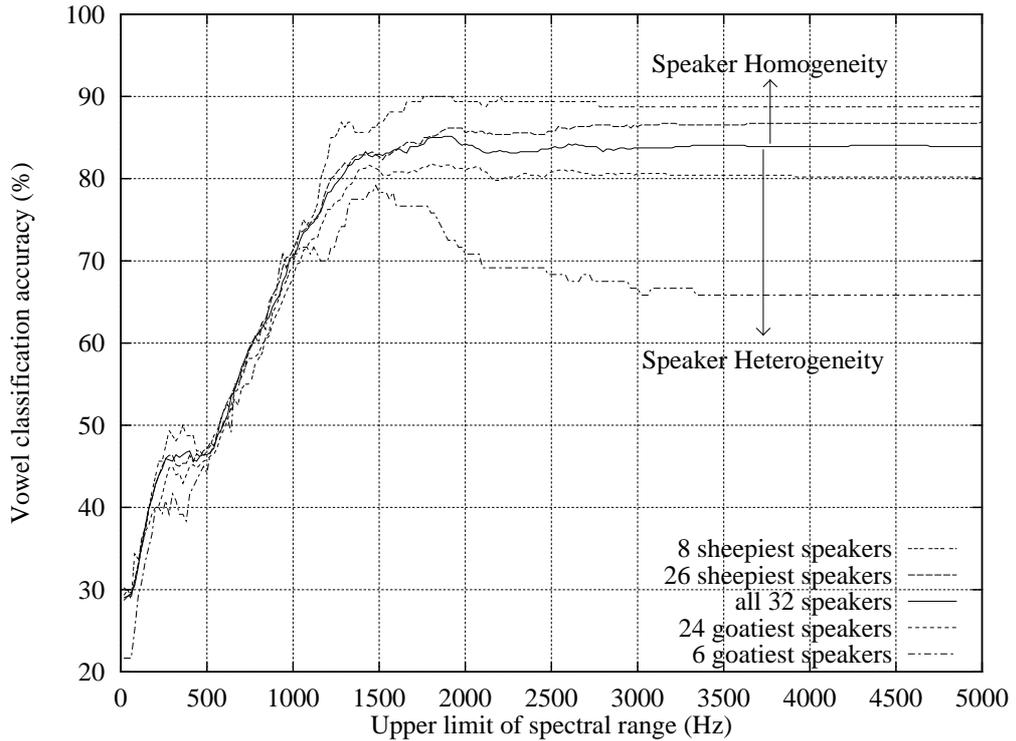


Figure 4.14: Trend towards a more *clear dichotomy* or a more *nearly asymptotic* behaviour in inter-speaker vowel classification accuracy, as a function of an *increasing* or a *decreasing proportion*, respectively, of “goat” to “sheep” speakers. Dataset (PB): 2 repetitions of 10 vowels spoken by 32 adult, male speakers of American English; simplified cepstra were generated from the first three formant frequencies, with sampling frequency  $F_s=10\text{kHz}$ ,  $NCC=14$ , and formant bandwidths fixed to mean values ( $B_1=50\text{Hz}$ ,  $B_2=65\text{Hz}$ , and  $B_3=114\text{Hz}$ ) as measured by Dunn (1961).

are fixed to their respective mean values, as measured by Dunn (1961) for a subset of 20 of the adult male speakers ( $\bar{B}_1 = 50\text{ Hz}$ ,  $\bar{B}_2 = 64\text{ Hz}$ , and  $\bar{B}_3 = 115\text{ Hz}$ ). Our results of the previous section further suggest that  $NCC = 14$  simplified cepstral coefficients are required with the sampling frequency set to  $F_s = 10\text{kHz}$ , despite the lower LP-model order of  $M = 6$ .

The solid curve in Figure 4.14, representing inter-speaker vowel classification accuracy as a function of upper spectral limit, exhibits a maximum of 85.2% by including spectral information up to 1900Hz, then drops to 83.1% for the spectral range extending to 2300Hz. Although this small dip is reminiscent of our earlier inter-speaker curves obtained for the FC dataset, the drop of only 1.3% from the peak (1900Hz) to full range (5000Hz) suggests that a potentially dichotomous behaviour has been blurred as a result of averaging across the individual accuracy curves of all 32 speakers.

If the spectral dichotomy observed earlier in Figure 4.13 is indeed an intrinsic phenomenon, then one could speculate that the blurring effect illustrated by the solid curve in Figure 4.14 has occurred as a result of an imbalance towards a greater degree of homogeneity amongst the speakers, and that any significant amount of speaker differences which have caused the small dip in accuracy in the higher spectral regions is due to only a small subset of the 32 speakers, the “goats” as might be referred to. The more clearly dichotomous, accuracy curves obtained earlier for our four very dissimilar speakers of Australian English may therefore be used as typical evidence for identifying the so-called “goat” speakers amongst the 32. Indeed, our “leave-one-speaker-out” approach to data-partitioning allows us to identify those individual classification curves which show strong evidence of a spectral dichotomy, and which therefore would characterise the “goat” speakers. By contrast, the more nearly asymptotic individual accuracy curves would identify the spectrally more homogeneous, “sheep” speakers.

This criterion can be quantified in terms of a measure of the degree of dichotomy, by computing the difference between the highest accuracy in each individual classification curve, and the accuracy attained at full spectral range. The score thus obtained is equal to zero for nearly-asymptotic curves where the maximum accuracy is achieved at full-range. By contrast, a more dichotomous classification curve with a clearly defined peak followed by a drop in accuracy, would yield a higher score, up to a maximum of 100%. The scores thus obtained for each speaker are listed in the second column of Table 4.3, and the individual accuracy curves of the two speakers with the equal highest score of 20% (speakers 16 and 20) are shown superimposed with the overall accuracy curve in Figure 4.15(a).

Our examination of the 32 speakers’ individual accuracy curves has revealed that in a large number of cases, the accuracy peaks then drops, but subsequently rises again as the spectral range is extended further to 5000Hz. A mild example is the individual curve of speaker 16 shown in Figure 4.15(a); more emphatic examples are the accuracy curves of speakers 1 and 22, shown superimposed with the overall accuracy curve in Figure 4.15(b). In those cases, the measure of dichotomy defined earlier does not capture the significant drop in accuracy which occurs across a frequency sub-band between the peak and full-range. It is therefore necessary to compute the extent of that

Speaker	max – full range (%)	largest drop (%)	error at full range (%)	Mean (%)
<b>16</b>	<b>20.0</b>	<b>25.0</b>	<b>45.0</b>	<b>30.0</b>
<b>20</b>	<b>20.0</b>	<b>15.0</b>	<b>30.0</b>	<b>21.7</b>
<b>26</b>	<b>15.0</b>	<b>15.0</b>	<b>30.0</b>	<b>20.0</b>
<b>9</b>	<b>10.0</b>	<b>15.0</b>	<b>25.0</b>	<b>16.7</b>
<b>5</b>	<b>10.0</b>	<b>10.0</b>	<b>25.0</b>	<b>15.0</b>
<b>23</b>	<b>10.0</b>	<b>5.0</b>	<b>30.0</b>	<b>15.0</b>
12	10.0	5.0	25.0	13.3
15	0.0	0.0	40.0	13.3
8	0.0	5.0	30.0	11.7
22	0.0	10.0	20.0	10.0
25	5.0	5.0	20.0	10.0
28	10.0	10.0	10.0	10.0
33	10.0	10.0	10.0	10.0
19	5.0	10.0	10.0	8.3
29	5.0	10.0	10.0	8.3
1	0.0	10.0	10.0	6.7
11	5.0	5.0	10.0	6.7
14	0.0	5.0	15.0	6.7
18	0.0	0.0	20.0	6.7
27	0.0	10.0	10.0	6.7
32	0.0	10.0	10.0	6.7
3	5.0	5.0	5.0	5.0
6	0.0	0.0	15.0	5.0
13	0.0	5.0	10.0	5.0
<b>7</b>	<b>0.0</b>	<b>0.0</b>	<b>10.0</b>	<b>3.3</b>
<b>17</b>	<b>0.0</b>	<b>0.0</b>	<b>10.0</b>	<b>3.3</b>
<b>24</b>	<b>0.0</b>	<b>0.0</b>	<b>10.0</b>	<b>3.3</b>
<b>30</b>	<b>0.0</b>	<b>0.0</b>	<b>10.0</b>	<b>3.3</b>
<b>4</b>	<b>0.0</b>	<b>5.0</b>	<b>0.0</b>	<b>1.7</b>
<b>10</b>	<b>0.0</b>	<b>5.0</b>	<b>0.0</b>	<b>1.7</b>
<b>21</b>	<b>0.0</b>	<b>0.0</b>	<b>5.0</b>	<b>1.7</b>
<b>31</b>	<b>0.0</b>	<b>0.0</b>	<b>5.0</b>	<b>1.7</b>

Table 4.3: Rank-ordered list of the 32 adult, male speakers of the PB dataset (speakers numbered 1,3,4,5,...,33 in *Column 1*) in terms of the following three attributes of their individual classification curves: *Column 2*: the difference between the maximum accuracy and that which is attained at full spectral range (5000Hz); *Column 3*: the largest drop in accuracy across the higher spectral regions; *Column 4*: the classification error at full spectral range. The rank-ordering is based on the mean of those three scores, as listed in *Column 5* (speakers who have the same mean score are listed according to their speaker number). Highlighted in bold-font at the top of the table are the 6 “goat” speakers with the highest mean scores (15% or greater); highlighted in bold-font at the bottom of the table are the 8 “sheepiest” speakers with the lowest mean scores (less than 5%).

largest, continuous drop in accuracy, and thus to take into account the potency of the higher spectral regions without being bound by the value at full-range. Those scores are listed for each speaker in the third column of Table 4.3.

Our selection of “goat” and “sheep” speakers would not be complete without also

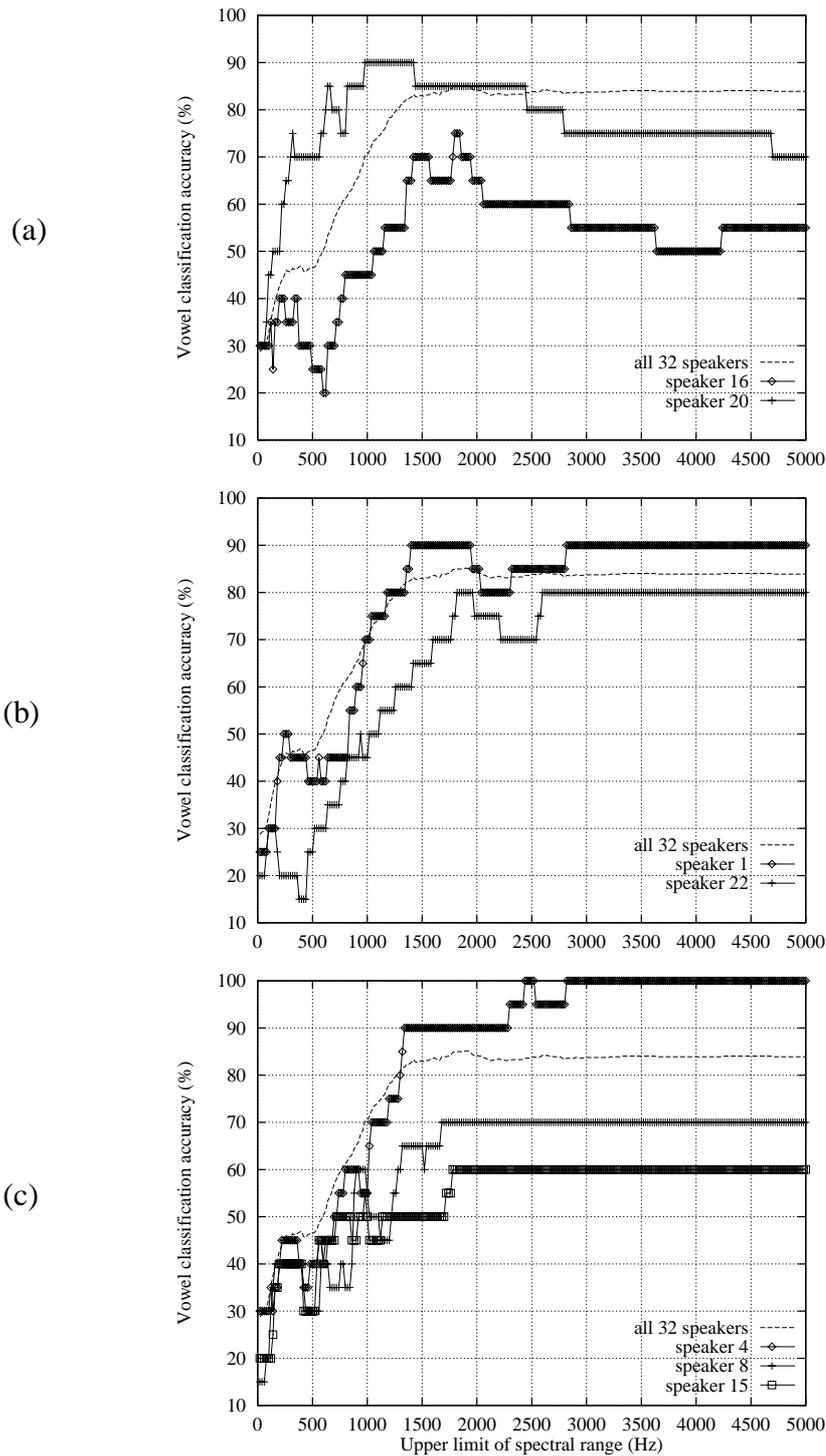


Figure 4.15: A few individual speakers' accuracy curves superimposed with the overall inter-speaker curve obtained using all 32 speakers of the PB dataset. *Panel (a)*: the two most clearly dichotomous, individual accuracy curves ("goat" speakers 16 and 20). *Panel (b)*: a drop in accuracy underscores the potency of the higher spectral regions, despite a further rise in accuracy towards full spectral range (speakers 1 and 22). *Panel (c)*: relatively low accuracies are indicative of outliers (speakers 8 and 15) notwithstanding the absence of a spectral dichotomy; by contrast, a nearly asymptotic curve which attains 100% at full spectral range is indicative of an extreme "sheep" (speaker 4).

considering what might be regarded as a more conventional definition of the former as outliers in the group, and of the latter as representative or typical speakers in the group. If the measure of dissimilarity which is implicit in this definition can be taken in the broadest sense to operate over the entire available spectral range, then it can be quantified in terms of the classification *error* at full-range (i.e., the difference between 100% and the classification accuracy at 5000Hz). Those errors are listed for each speaker's individual accuracy curve in the fourth column of Table 4.3; the greater the error (or the lower the accuracy at full-range), the more "outlier" (hence the "goatier") the speaker. We have already observed two examples of this type of contrast in Figures 4.15(a) and (b), where the lower accuracy at full-range for speakers 16 and 22, respectively, indicate that they are the more "outlier" of each pair of speakers. However, two extreme cases are the individual curves of speakers 8 and 15 shown superimposed in Figure 4.15(c): although they do not show any evidence of a spectral dichotomy (as might be signalled by a drop in accuracy over the higher spectral regions), the relatively low accuracies observed both across the entire spectral range and at 5000Hz suggest that those two speakers are "outliers" in the group. By contrast, the individual accuracy curve of speaker 4 (shown in the same Figure) is amongst the least dichotomous, and attains 100% at full-range.

To summarise, we propose three measures to quantify the degree of "goatiness" or "sheepiness" of each speaker in terms of the following attributes of their individual classification curves: (i) the difference between the maximum accuracy and that which is attained at full spectral range, (ii) the largest drop in accuracy across the higher spectral regions, and (iii) the classification error at full-range. Linear coefficients of correlation computed between those scores listed in Table 4.3 in columns 2 and 3 ( $r = 0.75$ ), columns 2 and 4 ( $r = 0.58$ ), and columns 3 and 4 ( $r = 0.44$ ), indicate that they each would contribute non-redundantly to a combined measure of "goatiness". Although in general an optimally weighted combination of those scores (each of which can take on a value between 0% and 100%) may be found to be superior, we have found that their mean (listed in the fifth column of Table 4.3) will suffice to rank-order the 32 speakers in terms of their quantified degree of "goatiness" (higher mean score) or "sheepiness" (lower mean score).

Indeed, the list of speakers in Table 4.3 is sorted in decreasing order of the final, mean score. That rank-ordered list is then used to categorise the speakers (as indicated by the broken lines in Table 4.3) into the following, three groups: (i) the 6 “goatiest” speakers (with the highest mean scores, and the most clearly dichotomous, individual accuracy curves), (ii) the 8 “sheepiest” speakers (with the lowest mean scores, and the most nearly asymptotic, individual accuracy curves), and (iii) the remaining 18 speakers who are neither extreme “goats” nor extreme “sheep”.

In order to assess the influence of speaker homogeneity on the phenomenon of vowel-speaker dichotomy, we performed a new set of inter-speaker vowel classification experiments using the generated, simplified cepstra of the speakers taken from various combinations of those three groups. First, the 8 “sheepiest” speakers (listed at the bottom of Table 4.3) are treated as a separate group, and their overall curve of classification accuracy is shown by the triple-dashed line in Figure 4.14. Next, the 18 speakers listed in the middle of Table 4.3 are added to the group, and the resulting classification curve for the 26 “sheepiest” speakers is shown by the long-dashed line in Figure 4.14. Those 18 speakers are then pooled with the 6 speakers listed at the top of Table 4.3, and the resulting accuracy curve for the 24 “goatiest” speakers is shown by the short-dashed line in Figure 4.14. Finally, the 6 “goatiest” speakers are treated as a separate group, and their overall classification curve is shown by the dash-dotted line in Figure 4.14.

Those curves of inter-speaker classification accuracy shown superimposed with the original, solid curve in Figure 4.14, clearly show a trend towards *higher accuracies* and a more *nearly asymptotic* behaviour as the population is made to comprise a higher proportion of “sheep” speakers; and a trend towards a *clearer dichotomy* as the proportion of “goat” speakers is increased. This striking progression in the behaviour of the accuracy curves can therefore be taken to suggest that even amongst themselves, the “sheepiest” speakers are spectrally a more *homogeneous* group, and the “goatiest” speakers are spectrally a more *heterogeneous* group. Before assessing the validity of that assumption more directly, we seek to gain a more informative perspective on the acoustic-phonetic relevance of the spectral regions over which inter-speaker vowel classification accuracy is observed to either rise or fall. In particular, the relevant

portions of the two accuracy curves obtained, respectively, for the 6 “goat” and the 8 “sheepiest” speakers, are replotted in Figure 4.16 adjacent to the  $F_1F_2$  vowel formant space of the two groups of speakers, and in Figure 4.17 adjacent to their  $F_2F_3$  vowel space.

As shown in Figure 4.16, the two classification curves rise in accuracy as the upper limit of the spectral range is increased across the speakers’ entire  $F_1$  distribution. The two curves then begin to diverge as the spectral range is extended beyond about 1 kHz — the curve for the 6 “goat” speakers dips slightly then rises to a peak of 79.2 % at 1480Hz, while the curve for the 8 “sheep” speakers continues to rise to 86.3 % at the same range. However, the most striking difference in the behaviour of the two curves occurs as the spectral range is extended beyond 1480Hz to include the  $F_2$  of the speakers’ front vowels, whereupon the nearly asymptotic behaviour of the “sheep” speakers’ accuracy curve contrasts with a marked drop in the “goat” speakers’ accuracy curve. As shown in Figure 4.17, the clearly detrimental influence of speaker differences in the higher spectral regions continues to elicit a downward trend in the accuracy curve of the “goat” speakers across their entire  $F_3$  range, to only 65.8 % at 3500Hz. By contrast, the accuracies obtained for the “sheep” speakers change very little over the spectral ranges which encompass their high- $F_2$  and  $F_3$ , attaining 88.8 % at 3500Hz. (As shown previously in Figure 4.14, the accuracy displayed on each curve at full spectral range is the same as that which is attained at 3500Hz, thus confirming that the simplified spectral representation contains no formants higher than the  $F_3$ .)

The sharp contrast in the behaviour of the accuracy curves for the “goat” and the “sheep” speakers does suggest, as noted earlier, that those two groups of speakers should differ markedly in their degree of spectral homogeneity. In so far as the simplified cepstrum carries, by definition, spectral information pertaining only to the formants, it is reasonable to expect a reflection of those differences in homogeneity, in the speakers’ vowel formant space. Indeed, the  $F_1F_2$  and  $F_2F_3$  vowel formant distributions shown in Figures 4.16 and 4.17 both for the 6 “goat” speakers (solid ellipses) and for the 8 “sheepiest” speakers (dashed ellipses) would seem to offer a more direct interpretation of the acoustic-phonetic implications of the presumed heterogeneity and homogeneity, respectively, of those two groups of speakers. Clearly,

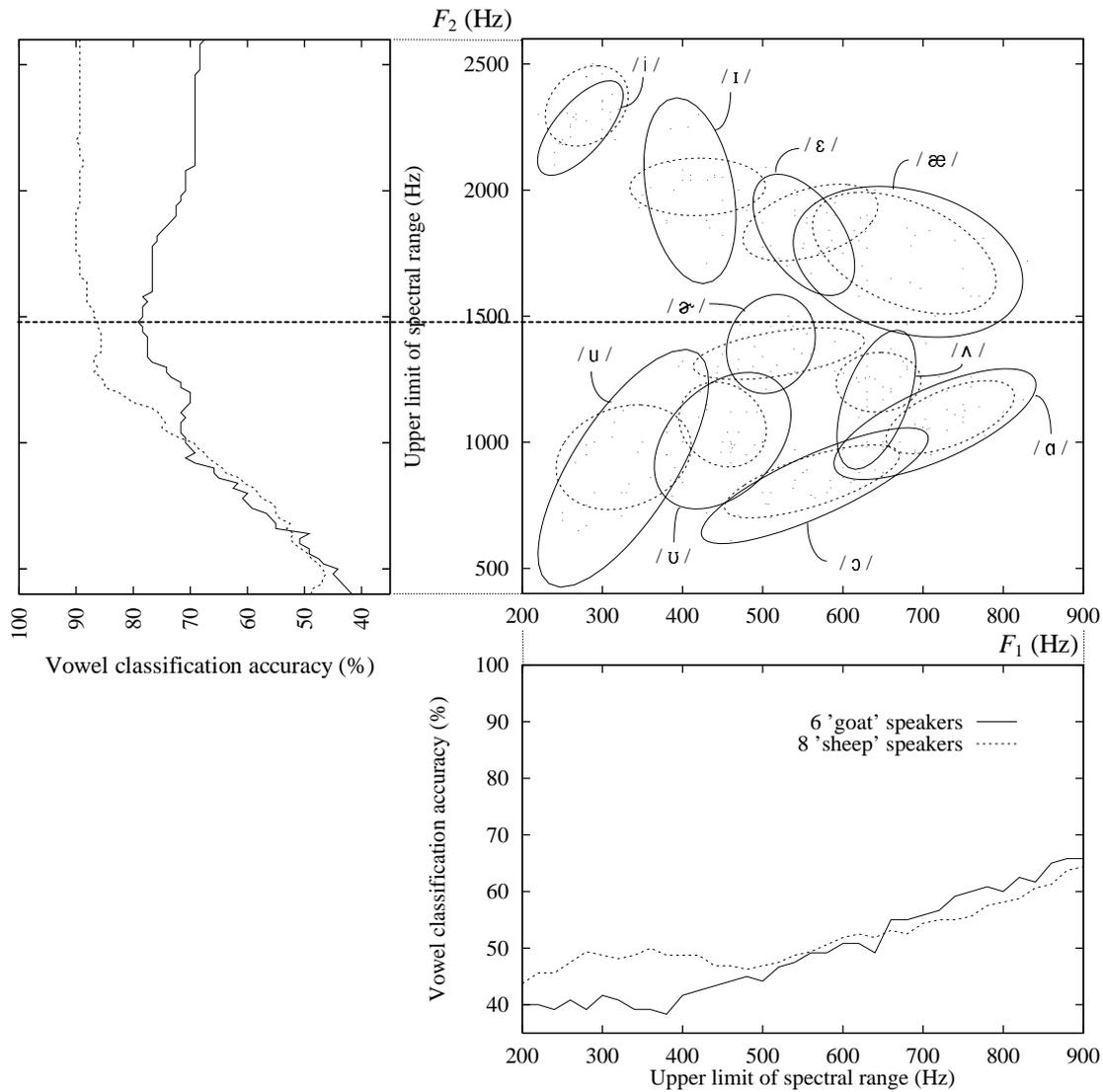


Figure 4.16:  $F_1F_2$  vowel space of the 6 “goat” and the 8 “sheepiest” male speakers (PB dataset), with a  $2\sigma$  ellipse drawn around each vowel cluster (solid and dashed lines, respectively). Adjacent to the abscissa and ordinate are plotted the portions of the *inter*-speaker accuracy curves (from Figure 4.14) which span the  $F_1$  and the  $F_2$  ranges, respectively. The horizontal (heavy dashed) line cuts through the formant plane at 1480Hz, which corresponds to the peak of the clearly dichotomous accuracy curve obtained for the “goat” speakers.

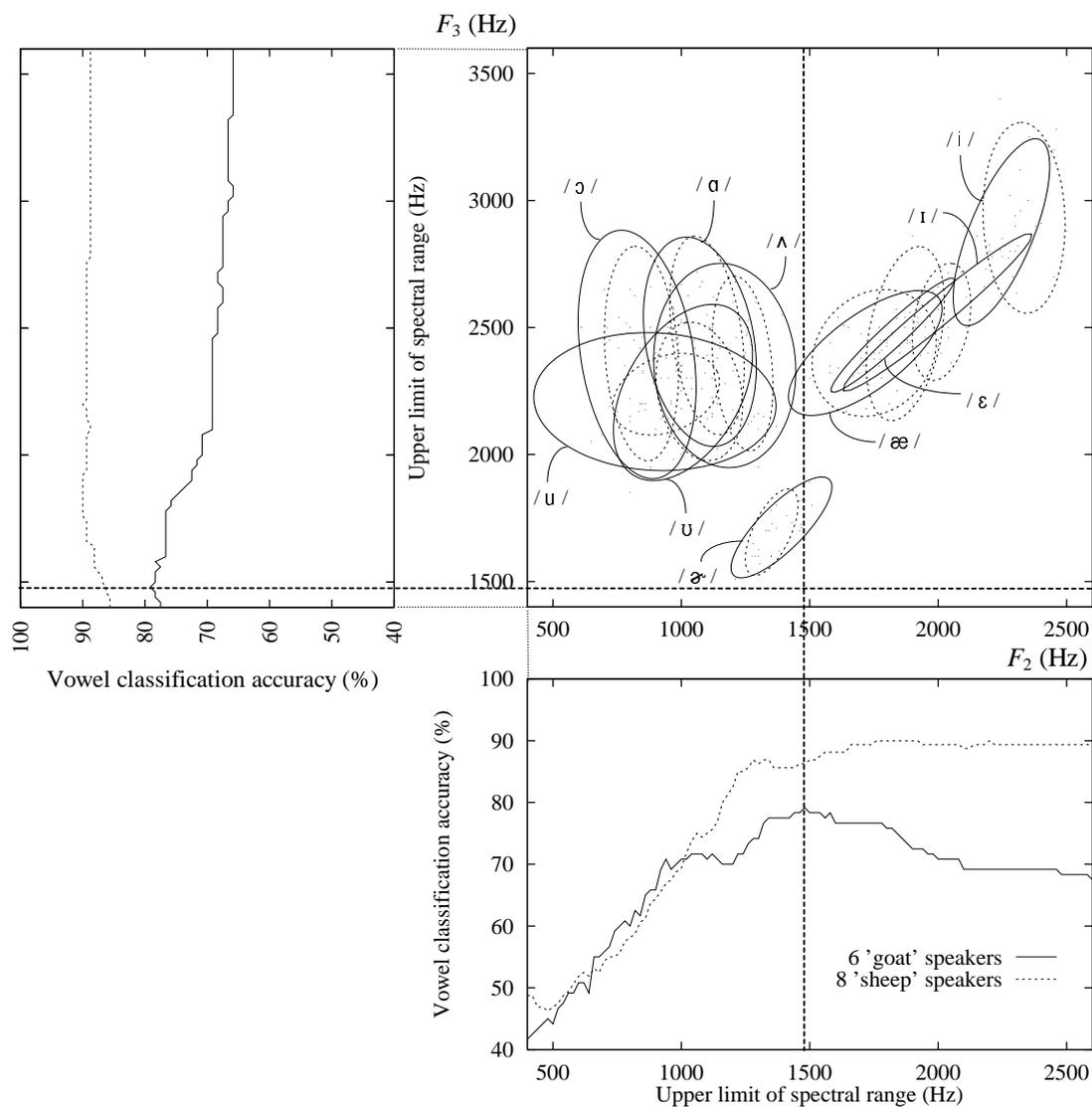


Figure 4.17:  $F_2F_3$  vowel space of the 6 “goatiest” and the 8 “sheepiest” male speakers (PB dataset), with a  $2\sigma$  ellipse drawn around each vowel cluster (solid and dashed lines, respectively). Adjacent to the abscissa and ordinate are plotted the portions of the *inter-speaker* accuracy curves (from Figure 4.14) which span the  $F_2$  and the  $F_3$  ranges, respectively. The vertical and horizontal (heavy dashed) lines intersect the clearly dichotomous accuracy curve at its peak (at 1480Hz), and by cutting across the formant plane they emphasise the acoustic-phonetic relevance of the spectral regions of primary phonetic or speaker influence.

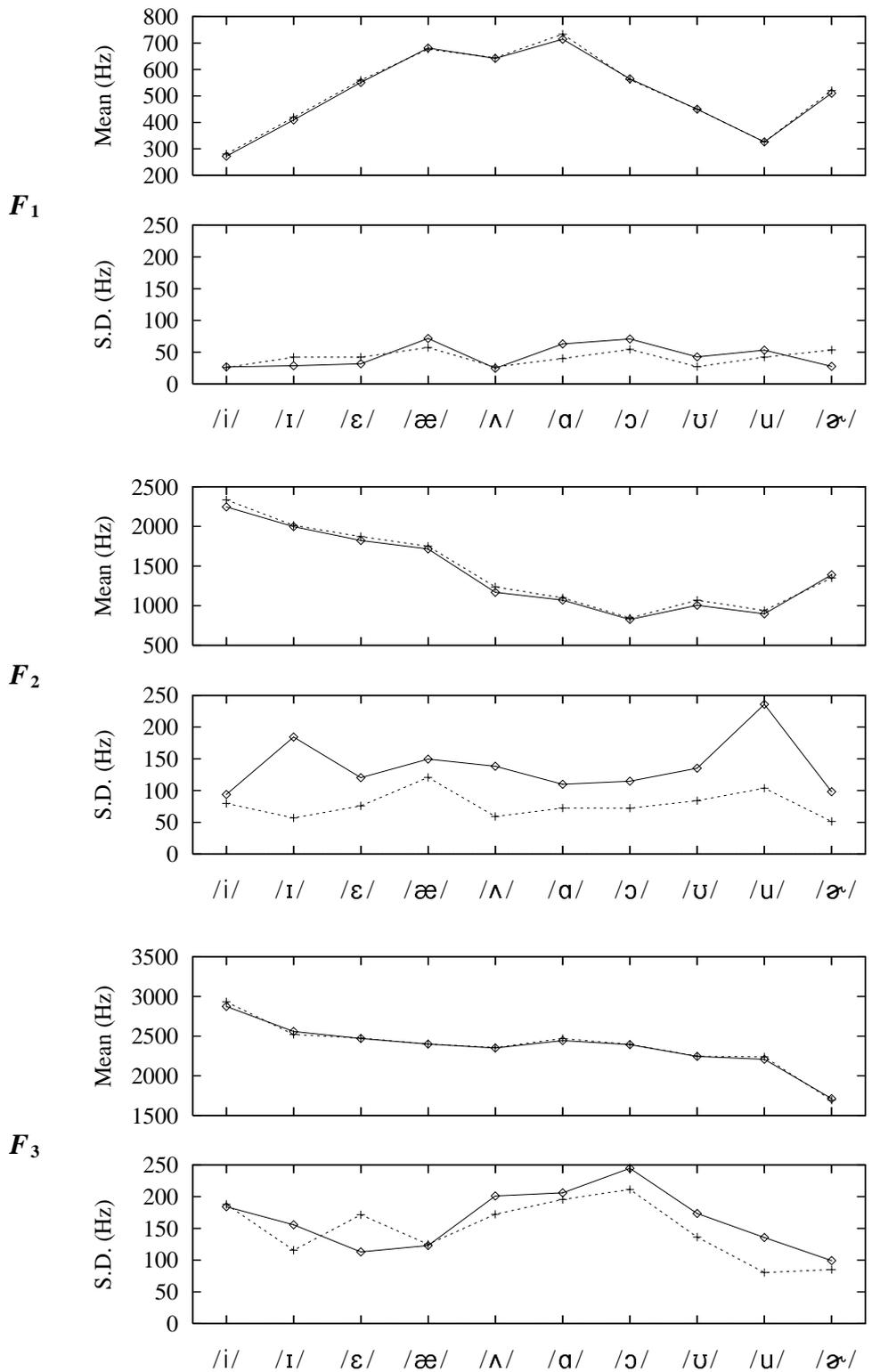


Figure 4.18: Mean (upper graph) and standard deviation (lower graph) in the first three formant frequencies  $F_1$ ,  $F_2$ , and  $F_3$  for each vowel, computed for the 6 “goatiest” speakers (diamond symbols joined by solid lines) and for the 8 “sheepiest” speakers (plus symbols joined by dashed lines) of the PB dataset, as determined by the rank-ordered list in Table 4.3.

the two groups differ not only in the orientation of their per-vowel ellipses, but also in the size of those ellipses — as might be expected of a more homogeneous group, the “sheep” speakers are characterised by a generally smaller dispersion of formant values per vowel. Perhaps the two most noticeable differences in the orientation of the ellipses are for the mid-front vowels /ɪ/ and /ɛ/ which, for the “sheep” speakers, have a much narrower and well-separated distribution along the  $F_2$  dimension; by contrast, the “goat” speakers appear to have a much greater degree of overlap amongst their front vowels along the  $F_2$  dimension, and in particular in the  $F_2F_3$  plane.

Differences in the formant distributions of the two groups of speakers can be further quantified by a comparison of their per-vowel mean and standard deviation in each of the first three formant frequencies. As shown in Figure 4.18, the profiles of per-vowel *mean* formant frequency are nearly indistinguishable across the two groups of speakers; the most consistent trend appears to be the slightly lower mean  $F_2$  (by 45Hz on average) of the “goat” speakers, in all of the vowels except /æ/. By contrast, the profiles of standard deviation shown in Figure 4.18 do indicate substantial differences, which confirm our earlier observations regarding the “sheep” speakers’ relatively smaller vowel ellipses. While the standard deviations in  $F_1$  are comparably low for both groups of speakers, the “goat” speakers have a consistently larger standard deviation in  $F_2$  across all the vowels (by 60Hz on average), and a consistently larger standard deviation in  $F_3$  of all the back vowels (by 33Hz on average).

Our measurements of formant dispersion are thus consistent with the notion of inter-speaker variability causing the drop in vowel classification accuracy across the higher spectral regions, which clearly exhibit a more potent manifestation of vowel-speaker interactions for the “goat” speakers. Even greater insights are gained by decomposing the vowel and “goat” speaker contributions to that drop in accuracy, the results of which are listed in Table 4.4. For example, the high front vowel /i/ of speaker 16 (who was ranked the “goatiest” speaker in Table 4.3, and whose individual accuracy curve was shown earlier in Figure 4.15(a) to be the most clearly dichotomous in the context of all 32 speakers) is confused with /ɪ/ in the  $F_3$  range, and the front vowel /æ/ of speaker 20 (ranked second in Table 4.3) is confused in the  $F_2$  range with the  $F_3$  of the back vowel /ʌ/. However, Table 4.4 reveals that the main contributions to the

Vowel	Speaker					
	5	9	16	20	23	26
/i/			/ɪ/ (F <sub>3</sub> )			
/ɪ/		/ɛ/ (F <sub>2</sub> )	/i/ (F <sub>2</sub> )		/ɛ/ (F <sub>2</sub> )	/i/ (F <sub>2</sub> )
/ɛ/			/ɪ/ (F <sub>2</sub> )	/æ/ (F <sub>2</sub> )		
/æ/			/ɛ/ (F <sub>2</sub> )	/ʌ/ (F <sub>2</sub> &F <sub>3</sub> )		
/ʌ/	/ʊ/ (F <sub>3</sub> )	/ʊ/ (F <sub>3</sub> )				
/ɑ/						
/ɔ/						
/ʊ/					/u/ (F <sub>3</sub> )	
/u/						/ɔ/ (F <sub>3</sub> )
/ə/						

Table 4.4: Acoustic-phonetic decomposition of the dichotomy in inter-speaker vowel classification behaviour observed (dash-dotted curve in Figure 4.14) for the 6 “goatliest” of the 32 adult, male speakers of the PB dataset, in terms of the vowel misclassifications that contribute to the drop in accuracy across the higher spectral regions which encompass the high- $F_2$  and the  $F_3$  of the 6 speakers’ vowel formant distribution. ‘ $F_2$ & $F_3$ ’ indicates misclassifications caused by overlapping  $F_2$  and  $F_3$  ranges.

drop in accuracy across the higher spectral regions are caused by confusions amongst the  $F_2$  of front vowels, and amongst the  $F_3$  of back vowels.

In this vein, it is remarkable to note the general similarity of the misclassifications listed here for the 6 “goatliest” speakers of the PB American English dataset, to those listed earlier (in Tables 4.1 and 4.2) for our four speakers of Australian English. Admittedly, the persistent trends in acoustic-phonetic manifestations of the dichotomy in two different datasets of spoken English, reinforce our definition of vowel-speaker interactions, and further underscore the speaker-related potency of the higher spectral regions which encompass the high- $F_2$  and the  $F_3$  of spoken vowel sounds.

#### 4.4.5 Dependence on Idiolectal Differences

Thus far we have unfolded and provided an acoustic-phonetic explanation of the phenomenon of vowel-speaker dichotomy using two datasets of vowels recorded by speakers known to be of mixed idiolectal or dialectal backgrounds. To what extent do the observed manifestations of the dichotomy depend on the types of speaker differences which exist between idiolects (of Australian English, as in the FC dataset) or between dialects (of American English, as in the PB dataset)? The JB dataset offers a

unique opportunity to address that question, in terms of the idiolectal variations found between speakers of Australian English.

However, as the speaker population of the JB dataset is comparable in size to that of the PB dataset, one can almost predict that the dichotomy manifest in the former would be as “blurred” as that obtained in our investigations at the start of the previous section. Although our method of speaker rank-ordering could then be recalled to identify the “goat” speakers and thereby obtain a clearer dichotomy, we herein disregard the issue of speaker homogeneity as it was elucidated in the previous section, and assume that a blurred dichotomy is sufficient to indicate the presence of vowel-speaker interactions. Indeed, in this section we seek to gain insights regarding the *idiolectal* dependence of the dichotomy, despite potential blurring of the phenomenon which might be caused by other factors (such as the number of speakers) contributing to speaker homogeneity.

As described in Chapter 3, the JB dataset comprises formants measured in the steady-state of effectively two repetitions of 11 vowels, spoken in /hVd/ context by 36 speakers of Australian English. Bernard’s (1967) own auditory judgments during an informal interview with each speaker, form the basis of our idiolectal labelling of 14 Broad (numbered 1 through 14), 11 General (numbered 15 through 25), and 11 Cultivated (numbered 26 through 36) speakers of Australian English. Speakers within each of those three groups can therefore be regarded as *idiolectally homogeneous*, at least as determined from a perceptual point of view by an expert phonetician.

In preparation for our vowel classification experiments, we convert the formant frequencies to LP simplified cepstra using the method outlined in Section 3.4.2, with  $NCC = 14$ ,  $F_s = 10$  kHz, and assuming fixed formant bandwidths equal to the mean values measured earlier for the same set of 11 vowels recorded by our four speakers of Australian English ( $\bar{B}_1 = 99$  Hz,  $\bar{B}_2 = 128$  Hz, and  $\bar{B}_3 = 218$  Hz). The dependence of the vowel-speaker dichotomy on idiolectal speaker differences is then assessed, by performing inter-speaker vowel classification as a function of an increasing upper spectral limit, in three, distinct experiments. First, *idiolect-independent* vowel classification is performed by training the classifier on the data of 35 of the speakers at a time, and testing on the data of the remaining speaker. Next, *inter-idiolect* vowel

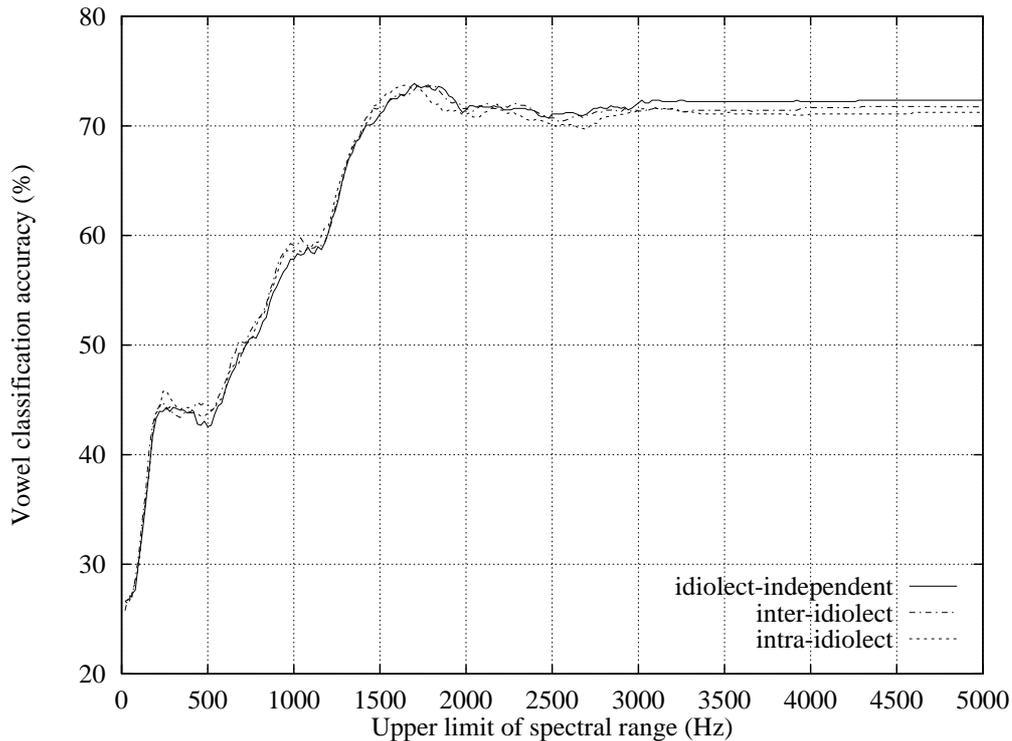


Figure 4.19: Dichotomy in *idiolect-independent*, *inter-idiolect*, and *intra-idiolect* vowel classification behaviour on an inter-speaker basis, using the JB dataset of 14 “Broad”, 11 “General”, and 11 “Cultivated”, adult male speakers of Australian English.

classification is performed where, rather than leave out one speaker at a time, the training set comprises the data of all the speakers in two idiolectal groups, and the classifier is tested using the data of all the speakers in the remaining, third idiolectal group. Finally, *intra-idiolect* vowel classification is performed by adopting the leave-one-speaker-out approach within each of the three idiolectal groups at a time.

The solid curve in Figure 4.19 shows the behaviour of *idiolect-independent*, inter-speaker vowel classification accuracy using the data of all 36 speakers. It exhibits a peak of 73.9 % at 1700Hz, followed by a drop in accuracy to 70.7 % at 2480Hz, and a smaller rise to 72.3 % at full spectral range. As predicted, this behaviour is quite similar to the blurred dichotomy observed earlier (in Figure 4.14) for the 32 speakers of the PB dataset, and therefore suggests that the effect of a more clear dichotomy has been diminished by averaging over a large number of speakers.

Admittedly a more extreme form of eliciting the effects of idiolectal speaker differences, is to explicitly acknowledge the three separate groups of speakers by performing *inter-idiolect* vowel classification as described above. The mean of the three

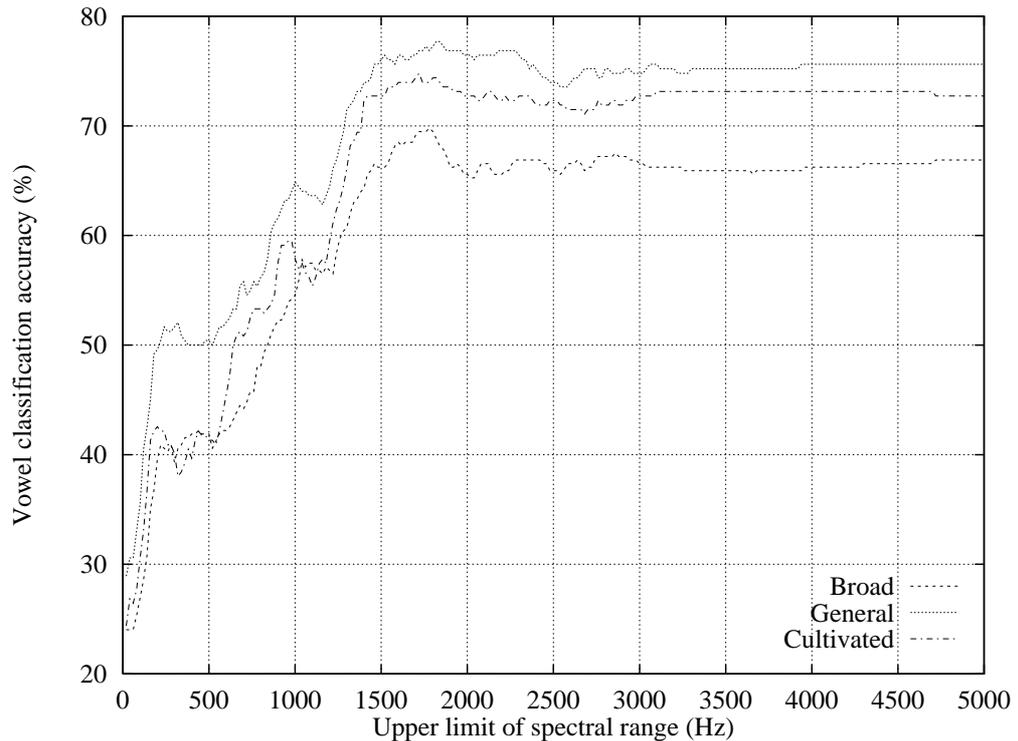


Figure 4.20: Idiolectal decomposition of the mean, *inter-idiolect* vowel classification accuracy curve (dash-dotted line in Figure 4.19), in terms of the three accuracy curves obtained by training on two idiolects and testing the data of all the speakers in the third idiolect (as noted in the legend).

accuracy curves thus obtained is shown in Figure 4.19 by the dash-dotted curve, which exhibits a peak of 73.7% at 1720Hz, followed by a drop in accuracy to 70.3% at 2540Hz and a small rise to 71.7% at full spectral range. This behaviour is very similar to that of the idiolect-independent accuracy curve shown in the same graph and discussed earlier. Indeed, if vowel-speaker interactions are assumed to be manifest even in a blurred dichotomy, then those curves together confirm the phenomenon in the presence of idiolectal speaker differences.

Further insights are gained by decomposing the mean, inter-idiolect accuracy curve in terms of its three constituent curves, as shown superimposed in Figure 4.20. Although the three curves exhibit comparable degrees of dichotomous behaviour (i.e., similar drops in accuracy across the higher spectral regions), the highest accuracies are attained along the entire curve for the speakers of the General idiolect, and the lowest accuracies are achieved for the Broad speakers. These results certainly confirm the findings of Kumar (1996) who performed inter-idiolect vowel classification on the same dataset using the formant frequencies, and found that when a Bayes classifier is trained

using the data of speakers in the remaining two idiolects, the best performance is achieved for the General speakers, and the worst performance for the Broad speakers. This is not entirely unexpected as it seems reasonable to assume that the General speakers exhibit acoustic-phonetic characteristics which are, almost by definition, intermediate between the other two, more extreme idiolects; and that the Broad speakers have characteristics which are on the whole more distant, and perhaps marked with greater variation compared with the other two idiolectal groups. Indeed, this interpretation of the relative levels of the three accuracy curves also increases our confidence in Bernard's (1967) idiolectal categorisation of the speakers, which itself was based purely on auditory-perceptual judgments rendered during an informal interview with each speaker.

The idiolect-independent and inter-idiolect vowel classification accuracy curves presented thus far, are consistent with our earlier results using the FC and the PB datasets, and show that the vowel-speaker dichotomy is indeed manifest in the presence of idiolectal speaker differences. We now address the question of whether the dichotomy is preserved despite the *absence* of idiolectal differences, by performing inter-speaker vowel classification experiments separately for the speakers within each idiolectal group. The dashed curve in Figure 4.19 is the mean of the three, *idiolect-dependent* or *intra-idiolect* accuracy curves thus obtained. It exhibits a rise in accuracy to a peak of 73.7 % at 1640Hz, followed by a drop in accuracy to 69.7 % at 2700Hz and a small rise to 71.2 % at full spectral range. The remarkable similarity of the behaviour of this curve to that of the idiolect-independent (solid) and the inter-idiolect (dash-dotted) curves, immediately suggests that the phenomenon of vowel-speaker dichotomy does not necessarily rely on idiolectal differences between speakers. In fact, if we were to rank-order the three curves shown in Figure 4.19 in terms of their overall degree of dichotomy, we should rank the intra-idiolect curve highest, on the basis of its slightly larger drop in accuracy across the higher spectral regions (a drop of 4.0 % as the upper limit of the spectral range is extended from 1640Hz to 2700Hz), and its overall lower accuracy at full spectral range. Although this comparison reveals only minor contrasts in the behaviour of the three accuracy curves, it does suggest that the detrimental influence of *within-idiolect*, or *intrinsic speaker differences* on vowel

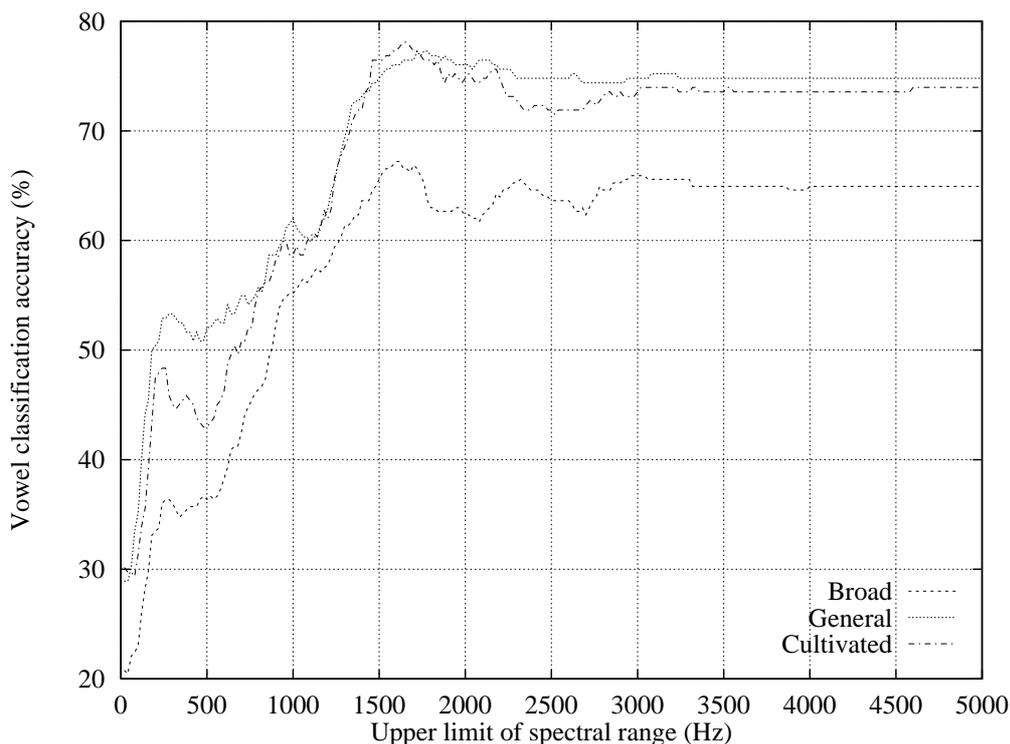


Figure 4.21: Idiolectal decomposition of the mean, *intra-idiolect* vowel classification accuracy curve (dashed line in Figure 4.19), in terms of the three accuracy curves obtained by performing inter-speaker vowel classification using only the speakers within each idiolectal group.

classification accuracy in the higher spectral regions, is *no less significant* than the influence of speaker differences between idiolects.

This implication is upheld by the dichotomous behaviour of the three individual, *intra-idiolect* accuracy curves superimposed in Figure 4.21, all of which exhibit a drop in accuracy across the higher spectral regions. Similarly to the three, *inter-idiolect* curves shown earlier in Figure 4.20, the curve pertaining to the speakers of the Broad idiolect is consistently the lowest in accuracy; and the curve obtained for the General speakers is highest in accuracy both across the higher spectral regions and at full spectral range. However, the highest accuracy itself is attained along the Cultivated speakers' curve, which exhibits a peak of 78.1% at 1640Hz, followed by a large drop in accuracy to 71.5% at 2520Hz. By comparison, the accuracy curve obtained for the speakers of the General idiolect exhibits a less dramatic drop, from a peak of 77.3% at 1720Hz, to 74.4% at 2680Hz.

Whilst the dichotomous behaviour of all the accuracy curves generated thus far using the JB dataset does show that the detrimental influence of speaker differences in

Vowel	Idiolectal group (Australian English)		
	“Broad”	“General”	“Cultivated”
/i/	/ɪ/ (F <sub>2</sub> , 10%; F <sub>3</sub> , 15%) /ɛ/ (F <sub>2</sub> , 15%; F <sub>3</sub> , 5%)	/ɪ/ (F <sub>3</sub> , 10%)	/ɪ/ (F <sub>2</sub> , 5%)
/ɪ/	/i/ (F <sub>2</sub> , 15%) /ɛ/ (F <sub>2</sub> , 15%)	<b>/i/ (F<sub>2</sub>, 40%)</b> /ɛ/ (F <sub>2</sub> , 5%)	<b>/ɛ/ (F<sub>2</sub>, F<sub>3</sub>, 35%)</b>
/ɛ/	/ɪ/ (F <sub>2</sub> , 10%)	/ɪ/ (F <sub>2</sub> , 20%)	/ɪ/ (F <sub>2</sub> , 5%) /ʊ:/ (F <sub>2</sub> , 5%)
/æ/		/ɛ/ (F <sub>2</sub> , 5%)	
/a/			
/ʌ/		/a/ (F <sub>3</sub> , 20%)	/a/ (F <sub>2</sub> , 5%)
/ɔ/	/a/ (F <sub>3</sub> , 5%) /ʊ/ (F <sub>3</sub> , 5%)		/ɔ/ (F <sub>3</sub> , 5%)
/ɔ/	/ʊ/ (F <sub>3</sub> , 5%)	/ʊ/ (F <sub>3</sub> , 10%)	/æ/ (F <sub>2</sub> &F <sub>3</sub> , 10%)
/ʊ/	/ɔ/ (F <sub>3</sub> , 5%)	/ɔ/ (F <sub>3</sub> , 20%)	
/ʊ:/	/ɜ/ (F <sub>2</sub> , 15%)		
/ɜ/	/ʊ:/ (F <sub>2</sub> , 15%)		/ʊ:/ (F <sub>3</sub> , 5%)

Table 4.5: Acoustic-phonetic decomposition of the dichotomy in *inter-idiolect* vowel classification behaviour (dash-dotted curve in Figure 4.19), in terms of the vowel misclassifications that contribute to the drop in accuracy across the higher spectral regions which encompass the high- $F_2$  and the  $F_3$  of the JB speakers’ vowel formant distribution. ‘ $F_2$ & $F_3$ ’ indicates misclassifications caused by overlapping  $F_2$  and  $F_3$  ranges.

the higher spectral regions is independent of (perceived) idiolectal differences, we are compelled to ask whether those detrimental influences are caused by similar types of vowel confusions. Our acoustic-phonetic decomposition of the contributions to the drop in accuracy in the *inter-idiolect* curve (the dash-dotted curve in Figure 4.19, whose three constituent curves were shown in Figure 4.20) is listed in Table 4.5. As highlighted in that Table, the two largest contributions are caused, respectively, by confusions of the General speakers’ front vowel /ɪ/ with /i/ in the  $F_2$  range of those vowels, and by confusions of the Cultivated speakers’ front vowel /ɪ/ with /ɛ/ in the  $F_2$  and  $F_3$  ranges. The two largest contributions amongst the back vowels are caused, respectively, by confusions of the General speakers’ /ʌ/ with /a/, and by confusions of their /ʊ/ with /ɔ/, both occurring in the  $F_3$  range of those vowels.

For the sake of comparison, the results of our acoustic-phonetic decomposition of the vowel and speaker contributions to the drop in accuracy observed in the higher spectral regions for each of the three *intra-idiolect* curves (in Figure 4.21), are listed in

Vowel	Speaker (of “Broad” Idiolect)													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
/i/														
/ɪ/	/i/ (F <sub>2</sub> )	/ɛ/ (F <sub>2</sub> )	/i/ (F <sub>2</sub> )	/i/ (F <sub>3</sub> )			/ɛ/ (F <sub>2</sub> ) /i/ (F <sub>3</sub> )		/i/ (F <sub>2</sub> )	/i/ (F <sub>2</sub> )	/ɛ/ (F <sub>2</sub> )		/i/ (F <sub>2</sub> , F <sub>3</sub> )	/i/ (F <sub>2</sub> )
/ɛ/	/æ/ (F <sub>2</sub> )						/ʌ: (F <sub>2</sub> )	/ɪ/ (F <sub>2</sub> )			/ɪ/ (F <sub>2</sub> )	/ɪ/ (F <sub>2</sub> )		/ɪ/ (F <sub>2</sub> , F <sub>2</sub> &F <sub>3</sub> )
/æ/	/ʌ: (F <sub>2</sub> )	/ɛ/ (F <sub>3</sub> <sup>+</sup> )												
/a/		/ʌ/ (F <sub>3</sub> )							/ʌ/ (F <sub>3</sub> )		/ʌ/ (F <sub>3</sub> )	/ʌ/ (F <sub>3</sub> )	/ʌ/ (F <sub>3</sub> )	
/ʌ/	/a/ (F <sub>2</sub> )		/a/ (F <sub>2</sub> )	/a/ (F <sub>3</sub> )		/a/ (F <sub>2</sub> )	/a/ (F <sub>3</sub> <sup>+</sup> )				/ɒ/ (F <sub>3</sub> <sup>+</sup> )		/a/ (F <sub>2</sub> , F <sub>3</sub> )	
/ɒ/			/ʊ/ (F <sub>3</sub> )						/ɔ/ (F <sub>3</sub> )					
/ɔ/		/ʊ/ (F <sub>3</sub> )								/ʊ/ (F <sub>3</sub> )			/ʊ/ (F <sub>3</sub> )	
/ʊ/		/ɔ/ (F <sub>3</sub> )		/ɔ/ (F <sub>3</sub> )										/ɔ/ (F <sub>3</sub> )
/ʌ: (F <sub>2</sub> )	/ɜ/ (F <sub>2</sub> )					/ɜ/ (F <sub>2</sub> )							/ɜ/ (F <sub>2</sub> )	
/ɜ/ (F <sub>3</sub> )			/ʌ: (F <sub>3</sub> )					/ʌ: (F <sub>3</sub> <sup>+</sup> )		/ʌ: (F <sub>2</sub> )		/ʌ: (F <sub>2</sub> )		/ʌ: (F <sub>3</sub> )

Table 4.6(a): Acoustic-phonetic decomposition of the dichotomy in *intra-idiolect*, inter-speaker vowel classification behaviour for the 14 “Broad” speakers of the JB dataset (dashed curve in Figure 4.21), in terms of the vowel misclassifications that contribute to the drop in accuracy across the higher spectral regions which encompass the high- $F_2$  and the  $F_3$  of those speakers’ vowel formant distribution. ‘ $F_2$ & $F_3$ ’ indicates misclassifications caused by overlapping  $F_2$  and  $F_3$  ranges; ‘ $F_3^+$ ’ indicates misclassifications which occur in spectral regions higher than the speakers’  $F_3$  range.

Tables 4.6(a), (b), and (c), for the 14 Broad, 11 General, and 11 Cultivated speakers, respectively. Consistent with the less dramatic drop in accuracy noted earlier for the General speakers’ accuracy curve in Figure 4.21, there appear to be relatively fewer misclassifications listed in Table 4.6(b) than in Tables 4.6(a) and (c). Table 4.6(b) is particularly sparse in confusions amongst the General speakers’ back vowels — confusions which appear profusely for both the Broad and Cultivated speakers. Despite the relatively large contributions to the dichotomy in *inter-idiolect* vowel classification by the General speakers, they appear to have only a relatively smaller contribution to the dichotomous behaviour of the *intra-idiolect* accuracy curve.

These idiolect-specific characteristics notwithstanding, the vowel confusions listed in Tables 4.5 and 4.6 convey much the same trend observed earlier both for the FC

Vowel	Speaker (of “General” Idiolect)										
	15	16	17	18	19	20	21	22	23	24	25
/i/			/ɪ/ (F <sub>3</sub> )					/ɪ/ (F <sub>2</sub> ,F <sub>3</sub> )			
/ɪ/				/i/ (F <sub>3</sub> )		/ɛ/ (F <sub>2</sub> )	/ɛ/ (F <sub>2</sub> )	/ɛ/ (F <sub>2</sub> )			
/ɛ/		/ɪ/ (F <sub>2</sub> ,F <sub>3</sub> )		/ɪ/ (F <sub>2</sub> )							/ɪ/ (F <sub>2</sub> )
/æ/											
/a/											
/ʌ/					/a/ (F <sub>3</sub> )						
/ɒ/											
/ɔ/							/ʊ/ (F <sub>3</sub> )				
/ʊ/											
/ɜ:/					/ɛ/ (F <sub>2</sub> )					/ɜ/ (F <sub>3</sub> )	
/ɜ/											

Table 4.6(b): Acoustic-phonetic decomposition of the dichotomy in *intra-idiolect*, inter-speaker vowel classification behaviour for the 11 “General” speakers of the JB dataset (dotted curve in Figure 4.21), in terms of the vowel misclassifications that contribute to the drop in accuracy across the higher spectral regions which encompass the high- $F_2$  and the  $F_3$  of those speakers’ vowel formant distribution.

Vowel	Speaker (of “Cultivated” Idiolect)										
	26	27	28	29	30	31	32	33	34	35	36
/i/		/ɪ/ (F <sub>2</sub> ,F <sub>3</sub> )					/ɪ/ (F <sub>3</sub> )			/ɪ/ (F <sub>2</sub> )	/ɪ/ (F <sub>2</sub> )
/ɪ/	/i/ (F <sub>2</sub> )	/ɛ/ (F <sub>2</sub> )		/ɛ/ (F <sub>2</sub> )	/ɛ/ (F <sub>2</sub> )	/i/ (F <sub>2</sub> )		/ɛ/ (F <sub>2</sub> )		/i/ (F <sub>2</sub> )	/i/ (F <sub>2</sub> )
/ɛ/	/ɪ/ (F <sub>2</sub> )	/ɜ:/ (F <sub>2</sub> )	/ɪ/ (F <sub>2</sub> )							/ɪ/ (F <sub>2</sub> )	
/æ/							/ɛ/ (F <sub>2</sub> ,F <sub>3</sub> )				
/a/											
/ʌ/	/a/ (F <sub>3</sub> )		/a/ (F <sub>3</sub> )			/a/ (F <sub>3</sub> )					
/ɒ/											
/ɔ/		/ʊ/ (F <sub>3</sub> )		/ʊ/ (F <sub>3</sub> )	/ʊ/ (F <sub>3</sub> )				/æ/ (F <sub>2</sub> &F <sub>3</sub> )		
/ʊ/								/ɔ/ (F <sub>3</sub> <sup>+</sup> )			
/ɜ:/								/ɜ/ (F <sub>3</sub> )			
/ɜ/		/ɪ/ (F <sub>2</sub> )	/ɜ:/ (F <sub>3</sub> )								/ɜ:/ (F <sub>2</sub> )

Table 4.6(c): Acoustic-phonetic decomposition of the dichotomy in *intra-idiolect*, inter-speaker vowel classification behaviour for the 11 “Cultivated” speakers of the JB dataset (dash-dotted curve in Figure 4.21), in terms of the vowel misclassifications that contribute to the drop in accuracy across the higher spectral regions which encompass the high- $F_2$  and the  $F_3$  of those speakers’ vowel formant distribution. ‘ $F_2$ & $F_3$ ’ indicates misclassifications caused by overlapping  $F_2$  and  $F_3$  ranges; ‘ $F_3^+$ ’ indicates misclassifications which occur in spectral regions higher than the speakers’  $F_3$  range.

dataset (in Tables 4.1 and 4.2) and for the “goat” speakers of the PB dataset (Table 4.4) — contributions to the drop in accuracy across the higher spectral regions arise mainly from confusions amongst the  $F_2$  of front (and mid-) vowels, and amongst the  $F_3$  of back vowels. Those manifestations of vowel-speaker interactions have been shown in this section to appear in their basic form, irrespective of (perceived) idiolectal speaker differences. That the dichotomy should thus be manifest even for idiolectally homogeneous groups of speakers, is yet further validation of its fundamental nature as a basic phenomenon of speech.

## 4.5 Concluding Summary

In this chapter we have approached the long-standing problem of speech-speaker dichotomy from a novel perspective, which involves contrasting the behaviour of vowel classification accuracy as a function of an increasing upper spectral limit, first on an intra-speaker basis where the phonetic dimension is unimpeded by speaker differences, then on an inter-speaker basis where phonetic and speaker-specific influences were expected to interact. A minimum-distance, linear classifier was used to first unfold the dichotomy using the phonetically-rich FC dataset of vocalic steady-states recorded in /hVd/ context by four adult, male speakers of Australian English. Progressive recruitment of higher-frequency spectral information in vowel classification, was achieved by way of a new, parametric cepstral distance measure (PCD), which allows selection of any frequency sub-band within the available spectral range. The dichotomy was then validated using a more sophisticated, quadratic classifier, for which frequency sub-band selectivity was achieved by a recursive transformation of the cepstrum to obtain a so-called partial (quefrequency-weighted) cepstrum (P-QCEP). Owing to the quadratic classifier’s sensitivity to data dimensionality, its role was limited to validating the general, contrastive behaviour of the accuracy curves, whereas all subsequent analyses were carried out using the less sensitive linear classifier.

The dichotomy itself is embodied in the contrast between the *nearly-asymptotic* behaviour of intra-speaker vowel classification accuracy, and the *dichotomous* behaviour of inter-speaker accuracy across the spectral continuum. Whilst the former underscores the importance of the low spectral regions for the purposes of vowel

discrimination, the latter evinces the speaker-related potency of the higher spectral regions where classification accuracy is observed to drop. The acoustic-phonetic relevance of those spectral regions of primary phonetic and speaker influence was determined by referring to the speakers' vowel formant distribution, which first confirmed the well-known phonetic salience of the spectral regions which encompass the two lowest formants or resonances of the vocal tract  $F_1$  and  $F_2$ . By contrast, a detailed acoustic-phonetic decomposition of the vowel and speaker contributions to the drop in inter-speaker vowel classification accuracy across the higher spectral regions, clearly implicated speaker-induced confusions mainly amongst the  $F_2$  of front vowels, and amongst the  $F_3$  of back vowels. Apart from clearly substantiating previous, largely unacknowledged works regarding the relative speaker-specificity of those acoustic-phonetic, vocalic subspaces, we have provided a new methodology for examining vowel-speaker interactions in the acoustic-phonetic domain.

Prior to using the PB and JB datasets to validate the dichotomy, its dependence on spectral representation was investigated by performing vowel classification using *simplified cepstra* generated from the carefully measured formants of the FC dataset. The simplified spectral representation was shown to preserve the dichotomy, and to yield the same, general trend of speaker-induced vowel confusions in the higher spectral regions. Indeed, one might expect the explicit removal of the so-called spurious poles, together with the effective normalisation of formant bandwidths, to have yielded a more direct, and perhaps a more truthful, acoustic-phonetic explanation of the vowel-speaker dichotomy in terms of the speakers' formant ranges.

Simplified LP cepstra were then created from the formants of the PB dataset, and used both to validate the dichotomy and to explore its dependence on speaker homogeneity. Indeed, the relatively larger, and evidently more homogeneous population of 32 adult, male speakers, only yielded a so-called blurred dichotomy in the behaviour of inter-speaker vowel classification accuracy across the spectral continuum. A new method of rank-ordering the speakers in terms of their degree of "sheepiness" or "goatiness" with respect to the given population, then led to a progression of accuracy curves, from the most nearly asymptotic (for the "sheepiest" speakers), to the most clearly dichotomous (for the "goatiest" speakers). The spectral manifestation of the

dichotomy was thus clearly shown to depend on the degree of heterogeneity amongst the speakers. Moreover, a comparison of the per-vowel formant statistics of the “goat” and the “sheep” speakers showed the heterogeneity of the former to be manifest mainly in terms of a larger dispersion in the  $F_2$  of all vowels, and in the  $F_3$  of the back vowels. Differences between the American English (PB) and our Australian English (FC) dataset notwithstanding, an acoustic-phonetic decomposition of the drop in accuracy observed across the higher spectral regions for the “goatiest” (PB) speakers corroborated the speaker-induced vowel confusions noted earlier amongst the  $F_2$  of front vowels and amongst the  $F_3$  of back vowels, thus raising our confidence in the basic nature of the vowel-speaker dichotomy.

Finally, the JB dataset of 14 Broad, 11 General, and 11 Cultivated, adult male speakers of Australian English, was used to determine the dependence of the dichotomy on idiolectal speaker differences. Our results first confirmed the overall levels of accuracy expected of each idiolectal group, thus raising our confidence in the idiolectal labelling of each speaker, which was based on purely auditory-perceptual impressions. The behaviour of inter-speaker vowel classification accuracy across the spectral continuum was then shown to be equally dichotomous, regardless of the presence or absence of (perceived) idiolectal speaker differences. Moreover, an acoustic-phonetic decomposition of the vowel and speaker contributions to the drop in accuracy observed across the higher spectral regions, confirmed our earlier findings using the FC and the PB datasets, implicating the speaker-related potency of the  $F_2$  of front (and mid-) vowels and the  $F_3$  of back vowels.

The consistency of our acoustic-phonetic explanation of the dichotomy across three datasets of spoken English vowels, urges further investigations, as set out in our Introduction (in Chapter 1), aimed at providing a physical, or articulatory explanation of the phenomenon. This relatively uncharted area of speech research deserves careful considerations, unhampered by computational burdens imposed by excessively large amounts of data of a large number of speakers. Our articulatory explanation will therefore focus on the four-speaker, Australian English (FC) dataset already used in this chapter to unfold the dichotomy.

Lacking direct articulatory measurements of those speakers, we shall attempt a

physical explanation in terms of vocal-tract shapes estimated from the measured acoustic parameters. However, as reviewed in Chapter 2, acoustic-to-articulatory mapping is itself a highly problematic area of research which has not yet converged to a commonly-accepted solution. The following chapter is therefore devoted to gaining a fresh outlook on the problem of vocal-tract shape parameterisation and estimation, driven by specific requirements which emerge from our current, acoustic-domain investigations. In Chapter 6 we then apply the area-function parameterisation and estimation methods developed in Chapter 5, and thereby examine the phenomenon of vowel-speaker dichotomy from a speech production point of view.

## Chapter 5

### Vocal-Tract Shape Parameterisation and Estimation

#### 5.1 Introduction

The spectral manifestations of the interactions between phonetic and speaker-specific attributes of steady-state vowels of spoken Australian and American English, were unveiled in the preceding chapter, by way of classification experiments based upon the linear-prediction (LP) cepstrum. The flexibility afforded by our parametric cepstral distance measure (PCD, derived in Section 4.2.3.1) indeed complemented the whole-spectrum representation of the LP cepstrum, and allowed direct examination of the spectral ranges which predominantly contain either phonetic or speaker-specific influences. A more detailed examination of the vowel-speaker dichotomy was then performed by having recourse to the spectral resonance (or formant) frequencies. Although the task of formant-estimation is considerably more problematic than that of obtaining the LP cepstrum, the interpretive superiority of the formants led to a more revealing, acoustic-phonetic explanation of the dichotomy. As set out in our Introduction (Chapter 1), we now seek an even more fundamental interpretation of the vowel-speaker interactions which have thus far been elucidated only in the acoustic domain, by extending our investigations into the domain of speech production. A first step towards this end is taken here by considering the intertwined problems of vocal-tract shape estimation and parameterisation.

As suggested in our literature review (Chapter 2), difficulties in acquiring directly-measured physiological data have to a large extent prohibited a detailed, or a large-scale investigation of the articulatory sources of speaker variability. Acoustic-to-articulatory mapping, on the other hand, is potentially a more efficient and practical method, which may therefore facilitate investigations of the articulatory correlates of the many types of

variability in the acoustic speech signal. Sadly, the prospect of utilising estimated (rather than directly measured) articulatory data seems to have suffered from want of a “perfect” solution to the so-called inverse problem, which should be capable of recovering, with a large degree of confidence, the actual vocal tract configurations of any given speaker. Our review of the relevant literature (in Section 2.4.3.2) has already underscored the crucial role of both the articulatory parameterisation and the vocal-tract acoustic model in determining the intrinsic nonuniqueness of the inverse mapping. In our pursuit of an articulatory explanation of the phenomenon of dichotomy, we therefore seek a method of acoustic-to-articulatory mapping which employs a vocal-tract acoustic model with the least degree of intrinsic nonuniqueness.

In addition, we seek a vocal-tract shape parameterisation which not only inherits the uniqueness properties of the chosen vocal-tract acoustic model, but also permits a direct relation with the acoustic parameters used in the previous section to provide an acoustic-phonetic explanation of the dichotomy. In this vein, it is hardly surprising that the formants, which were ideally suited to providing an acoustic-phonetic explanation, are also the acoustic parameters which have been traditionally, and remain to this day, the most directly associated with the physical properties of the vocal-tract. It follows that, in order to obtain an articulatory interpretation of the dichotomy analogous to the acoustic-phonetic explanation offered in Chapter 4, it would be of considerable benefit to adopt a vocal-tract shape parameterisation which is related as directly as possible to the formants.

Our aim in this chapter is therefore to develop an approach to the estimation of vocal-tract area-functions from the acoustic speech waveform of non-nasalised vowels, which satisfies two main criteria: (1) by necessity, the model adopted for the inverse mapping must be capable of yielding inherently unique vocal-tract shapes; and (2) the area-function parameterisation should be expressible directly as a function of the acoustic resonances of the vocal-tract.

In Section 5.2 we address these two issues of resonance-based parameterisation and uniqueness, and thereby provide a bipartite rationale for our proposed approach. Theoretical and empirical results are brought to bear in Section 5.3, on the problem of determining the resonance-based parameters of unique, LP-derived area-functions. In

Section 5.4 we then describe our hybrid method of area-function estimation, and use it to evaluate our proposed method of vocal-tract shape parameterisation. In Section 5.5 we first introduce a new method of quantifying inter-repetition variability amongst vocal-tract shapes, which then allows an evaluation of our hybrid method of inversion. We conclude in Section 5.6 with a summary of the contributions arising from our work in vocal-tract shape estimation and parameterisation, and with a discussion of its significance in the context of our articulatory explanation of the dichotomy which is to follow in Chapter 6.

## 5.2 Rationale

Our criteria for an appropriate method of acoustic-to-articulatory mapping, as stated above, incite a twofold rationale concerning *parameterisation* and *uniqueness* of estimated vocal-tract shapes, which we therefore consider, respectively, in the next two sections.

### 5.2.1 Resonance-Based Parameterisation of Area-Functions

The seminal work of Schroeder and Mermelstein (1965), subsequently expanded by each author separately (Schroeder, 1967; Mermelstein, 1967), is an acoustically-motivated and theoretically-derived parameterisation of the vocal-tract area-function. In contrast with the many physiologically-based models of the vocal-tract which have appeared in the literature, the parameterisation derived by these authors is rooted in the acoustic, and more specifically in the resonance, properties of the vocal-tract. It is therefore re-examined here in some detail, with a view to highlighting its strengths and limitations, and placing it in the context of our forthcoming, articulatory investigation of the vowel-speaker dichotomy.

In order to better appreciate the implications of the modelling paradigm pioneered by those two authors, we have included in Appendix B, a mathematical re-derivation of what we hereafter will refer to as the Schroeder-Mermelstein (SM) model. Indeed, as re-derived therein, the complete model is summarised in the following two equations:

$$\ln A(x) = \ln A_0 + \sum_{m=1}^M a_m \cos\left(\frac{m\pi x}{L}\right), \quad (5.1)$$

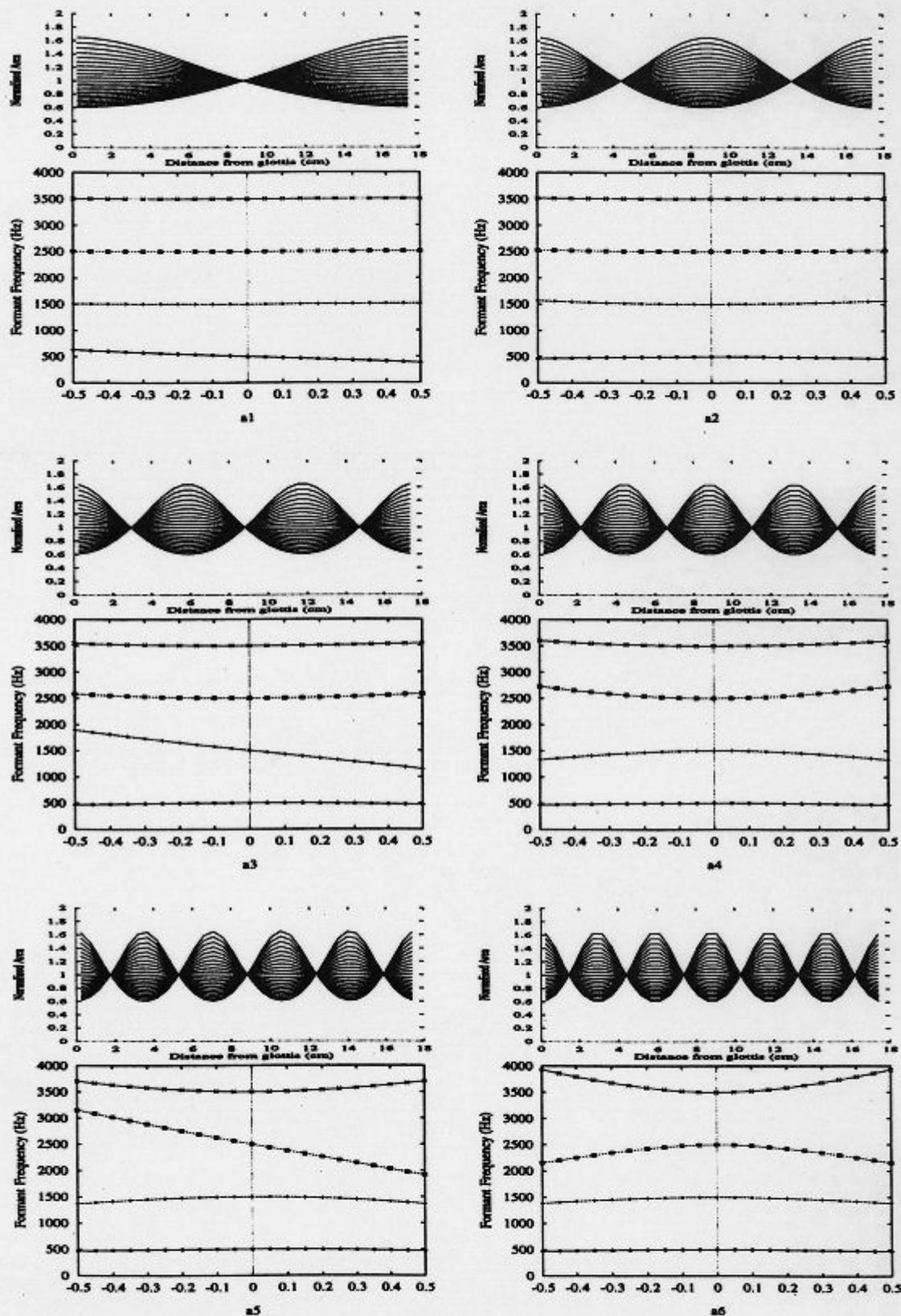


Figure 5.1: Nomograms depicting changes in the first four formant frequencies, as a function of perturbations in each of the first six, vocal-tract shape-parameters of the SM model. Each area-function is of length  $L=17.65\text{cm}$ , and the area scaling factor is  $A_0=1.0$ ; consistently with the assumptions of the SM model, the formant frequencies are synthesised using a completely lossless vocal-tract acoustic model (see Appendix C).

which defines the parameterisation of the vocal-tract (logarithmic) area-function  $A(x)$  of length  $L$ , in terms of the coefficients  $a_m$  of its Fourier cosine series; and

$$\frac{\delta F_n}{F_n} = -\frac{1}{2} a_{2n-1}, \quad (5.2)$$

which defines the theoretical relation between the acoustic- and the articulatory-domain parameters. In particular, as illustrated in the nomograms on the left in Figure 5.1, antisymmetric perturbation of the area-function of a uniform, completely lossless acoustic tube according to each of the odd-indexed shape parameters, does induce a quasi-linear perturbation of a unique formant frequency about its neutral value. By contrast, as shown in the nomograms on the right in Figure 5.1, perturbation of an even-indexed parameter  $a_{2n}$  has no first-order<sup>1</sup> effect on the formant frequencies. The SM model thereby establishes an acoustically-important distinction between *symmetric* and *antisymmetric* perturbations of a uniform area-function (where the symmetry is defined with respect to the mid-point along the length of the area-function).

Not long after its inception, the SM model’s potential in yielding articulatory insights into acoustic speech phenomena was demonstrated by Broad (1972, p.418), who combined Equations 5.1 and 5.2, then differentiated the result with respect to time to obtain a “measure of the average acoustically detectable movement throughout the vocal tract.” The coarse, phonetic segmentation (of an all-sonorant utterance) thus afforded, was then refined by integrating the expression only across a selected length of the vocal-tract, thereby obtaining a segmentation based on the acoustically-inferred velocity profile of, for example, the tongue dorsum. The SM model was also advocated by Broad and Wakita (1977), who showed that it aptly predicts the  $F_2$ -dependence of vocal-tract shapes, in terms of the distinction between *front* and *back* places of constriction; similarly, the top-left panel in Figure 5.1 illustrates the  $F_1$ -dependence, which is aptly predicted in terms of the *open* versus *closed* distinction.

However, despite the interpretive potential of the SM model, it is known to suffer the limitation of *nonuniqueness*. In particular, while the antisymmetric-shape parameters  $a_{2n-1}$  are each associated with a unique formant frequency  $F_n$ , the

---

<sup>1</sup> By “first-order”, we refer not only to the *magnitude* of the induced perturbation, but also to its *monotonicity* (or *quasi-linearity*).

symmetric-shape parameters  $a_{2n}$  remain ambiguous in the inverse mapping. In this vein, Borg (1946) is credited (by Schroeder, 1967; Mermelstein, 1967; Schroeter and Sondhi, 1994) to have proven that not a single, but a *doubly*-infinite set of eigenvalues, which correspond to two different sets of boundary conditions, is both necessary and sufficient to uniquely determine the state of a completely lossless resonant system. The *lip impedance function*, which is defined as the driving-point impedance of the vocal tract as seen from the lips, is particularly useful in this context, because it embodies knowledge of both sets of eigenvalues; in terms of the complex frequency variable  $s = j\omega$ , it is expressed as follows:

$$Z_{\text{in}}(s) = \frac{\prod_{n=1}^{\infty} (s - \omega_n)}{\prod_{n=1}^{\infty} (s - \omega_n^{(\text{c-c})})}, \quad (5.3)$$

where the *zeros* correspond to the well-known *formants*, and the *poles* of the lip impedance function correspond to the resonances under “closed-closed” (c-c) boundary conditions (implying an infinite acoustic impedance at both ends of the acoustic tube). Analogously to Equation B.13 (in Appendix B), that second set of eigenvalues is given by the following expression:

$$\lambda_n^{(\text{c-c})} = \frac{\omega_n^{(\text{c-c})2}}{c^2} = \left( \frac{n\pi}{L} \right)^2, \quad n = 1, 2, 3, \dots, \quad (5.4)$$

and the “c-c” resonance frequencies themselves interleave with the formant frequencies, as follows:

$$\omega_1 < \omega_1^{(\text{c-c})} < \omega_2 < \omega_2^{(\text{c-c})} < \omega_3 < \dots \quad (5.5)$$

Both Schroeder (1967) and Mermelstein (1967) have pointed out that the poles of the lip impedance function are related to the even-indexed parameters  $a_{2n}$  of the SM model in the same way that the formants are related to the odd-indexed parameters (in Equation 5.2). However, those poles can only be measured using a special apparatus known as a “lip impedance tube” (Schroeder, 1967; Gopinath and Sondhi, 1970) which has a mouthpiece sealed to the speaker’s lips, and which requires articulation of sustained vowels without phonation. The nonuniqueness problem then stems from the fact that the acoustic speech signal contains resonance information pertaining to only *one* of the required, two sets of boundary conditions — namely, the formants (or the

zeros of the lip impedance function, but not its *poles*); hence, the symmetric components of vocal-tract shapes remain undetermined in the inverse mapping, unless additional constraints are imposed.

Indeed, Mermelstein (1967) imposed the constraint of *antisymmetry* (which is, after all, a corollary of the SM model itself), simply by setting the undetermined, even-indexed parameters to zero. While this turned out to be a reasonable constraint for a number of vocalic configurations, it led to a poor re-estimate of the X-ray measured area-function (Fant, 1960) for the Russian vowel /u/ in particular, which has a strong symmetric component. More recently, Yehia and Itakura (1994, 1996) proposed to estimate the values of the first *eight* parameters of the SM model, using only the first *three* formant frequencies; in particular, the ambiguity associated with the values of the five shape-parameters  $a_2$ ,  $a_4$ ,  $a_6$ ,  $a_7$ , and  $a_8$ , was resolved by a combination of static and dynamic constraints on the estimated shapes themselves. Although, as reviewed much earlier (in Chapter 2), the nonuniqueness problem is exacerbated when the number of vocal-tract shape parameters exceeds the number of acoustic parameters, the SM model is *inherently nonunique* if only the recorded speech signal is available.

In that context, it is important to emphasise that the SM model embodies the nonuniqueness of a completely lossless vocal-tract, perhaps in its most fundamental form. Whilst other, more physiologically-relevant parameterisations of the vocal-tract may afford more direct constraints on individual articulators such as the tongue body, the jaw, and the lips, the SM model captures the essential, acoustically-relevant manifestations of nonuniqueness in the entire shape of a completely lossless vocal-tract. It is also interesting to note that the acoustic-theoretic principles on which the SM model is founded, are equally the basis of the more recent, Distinctive Regions and Modes (DRM) model of Mrayati et al. (1988). Admittedly, the DRM model has already been used more extensively than the SM model: for example, in /V-V/ and /V-C-V/ trajectory modelling (Carré and Mrayati, 1991; Carré et al., 1992), in automatic recognition of place of articulation of plosives (Soquet and Saerens, 1994), in automatic recognition of vowels and /V-V/ sequences (Candille and Meloni, 1995), in real-time articulatory speech synthesis (Hill et al., 1995), and in acoustic-to-articulatory mapping itself (Richards et al., 1995). However, as the DRM model is founded on the same

acoustic principles (of a completely lossless acoustic tube) which are embodied in the SM model, it does share the nonuniqueness properties of the latter, albeit without some of the distinct advantages of the SM model's parameter-set.

Indeed, one of the advantages of the SM model, is its inherently *smooth* representation of vocal-tract area-functions, which therefore allows (cf. Equation 5.1) computation of the cross-sectional area at any position  $x$  along the length of the vocal-tract. The degree of smoothness itself is determined by the number of formants used in the inversion; thus, as conjectured by Mermelstein (1967), the low-order formants can be used to obtain a *bandlimited* version of an area-function. In that vein, the efficiency of the SM model is manifest in its *minimal* set of parameters, which describe the vocal-tract shape in terms of an *orthogonal* set of basis functions (the cosine terms in Equation 5.1).

However, one of the disadvantages of the SM model<sup>2</sup>, is that the estimated vocal-tract *shape* depends on the assumed vocal-tract *length* (VTL); Mermelstein (1967) has already illustrated the potential sensitivity of the shape-length dependence. Based on Zue's (1969) thesis, Paige and Zue (1970) then proposed to optimise VTL such that the resulting shape is least eccentric compared with a uniform area-function. Within the analytical framework of the SM model, they obtained a *quadratic* expression for the error criterion with respect to vocal-tract length, thus firmly establishing the existence of a unique, global minimum for any given area-function. Furthermore, they were able to re-estimate the vocal-tract lengths of Fant's (1960) six Russian vowels, to within the tolerance prescribed by Fant's own, X-ray measurements ( $\pm 5\%$ ). This combination of theoretical elegance and empirical support has encouraged further use of the underlying principle, which has since been referred to as "minimal articulatory antagonism" (Lindblom and Sundberg, 1971), "minimal articulatory difference" (Bonder, 1983b), or "minimum of muscle work" (Sorokin, 1992).

To summarise our preceding discussion, it is apparent that the great strength of the SM model lies in its functional representation of vocal-tract shapes in terms of orthogonal parameters, one-half of which are acoustic-phonetically significant by virtue

---

<sup>2</sup> Indeed, of any articulatory model which retains the vocal-tract length as an explicit parameter.

of their quasi-linear, one-to-one mapping with the formant frequencies. However, the absence of resonance information (in the acoustic speech signal) pertaining to the required, second set of vocal-tract boundary conditions, implies that only the antisymmetric shape components can be inferred in practice. We therefore direct our attention in the next section, to the linear prediction (LP) vocal-tract model, which does inherently resolve the nonuniqueness problem without recourse to further, articulatory constraints.

### 5.2.2 Inherent Uniqueness in LP-derived Area-Functions

Indeed, as briefly reviewed in Chapter 2, the LP model does theoretically guarantee uniqueness in acoustic-to-articulatory mapping. In view of the inherent nonuniqueness of the completely lossless vocal-tract model reviewed in the previous section, we are therefore compelled to ask *how* the LP model secures uniqueness in vocal-tract shapes. In search of an answer to that question, we first consider the uniqueness issue as dealt with by the model's two main proponents, namely Atal and Wakita.

Atal (1970) is widely credited (e.g., by Markel and Gray, 1976) for being the first to derive an acoustic tube model of the vocal-tract directly from the speech waveform, and to show that the formant *frequencies* and *bandwidths* are sufficient to uniquely determine the area-function in terms of a finite number of equal-length sections. In Atal's version of the LP vocal-tract model, the formant bandwidths are brought about by a single, frequency-independent source of loss at the lip-end of an otherwise lossless acoustic tube. Invoking electrical-circuit analysis theory, Atal then proved (see, for example, Strube, 1977) that a discrete,  $M$ -section area-function is uniquely obtained through knowledge of the first  $M$ , discrete-time samples of the autocorrelation function of the pressure developed across the resistive lip termination, in response to a unit volume-velocity impulse at the glottis.

Atal and Hanauer (1971, Appendix F) then showed that an  $M^{\text{th}}$ -order, all-pole transfer function with all its poles inside the unit circle in the  $z$ -plane, can always be interpreted as that of a lossless acoustic tube with  $M$  equal-length sections, terminated by a unit acoustic resistance at the lips. Furthermore, they defined a procedure for determining the area-function of such an acoustic tube, using the *covariance* method of

LP analysis. However, as pointed out by Markel and Gray (1976, p.79), they could make no claim “as to the applicability of the method for estimating *vocal tract* area functions”, owing to the potential instability of the covariance method of analysis, which may cause poles to lie on or outside the unit circle in the  $z$ -plane.

On the other hand, Wakita (1973) showed that the inverse filter obtained by the *autocorrelation* method of LP analysis, whose stability is theoretically guaranteed (Markel and Gray, 1976, p.103), is an equivalent representation of a lossless acoustic tube model with a resistive termination at the *glottal* end, and a short-circuit at the lips. Furthermore, he demonstrated that plausible vocal-tract shapes can be obtained directly from the speech waveform, provided certain analysis conditions (such as appropriate pre-emphasis) are applied.

Wakita and Gray (1975) then shed some light on the *uniqueness* of area-functions estimated by that method. In particular, they derived an expression for the *lip impedance function* of the  $M$ -section acoustic-tube model, both for a lossless and a lossy glottal termination. First, consistently with Borg’s (1946) proof reviewed in the previous section, Wakita and Gray (1975, p.579) showed that for a lossless model, the numerator and the denominator polynomials of the lip impedance function are *independent* of each other, such that “both of them are needed for the determination of a unique tube shape”; indeed, the roots of the two polynomials thus obtained are merely the resonance frequencies which satisfy, respectively, the two different sets of boundary conditions. By contrast, upon reinstatement of the lossy, resistive glottal termination, the two polynomials were found to be *inter-dependent*; it was shown that either of those polynomials can be obtained from the other, simply by changing the sign of all of the reflection coefficients. Indeed, while for the completely lossless model the two sets of eigenvalues (the roots of the numerator and denominator polynomials, respectively) are independent and *real-valued* (i.e., they have frequencies only), for the LP model with a resistive glottal termination they are dependent and *complex-valued* (i.e., they have both frequencies and bandwidths).

Whilst the formant *bandwidths* are thus clearly implicated in both Atal’s and Wakita’s explanations of the uniqueness of LP-derived area-functions, we sadly remain ignorant of the vocal-tract *shape*-related manifestations of uniqueness. In particular,

despite the implications clearly foreshadowed by the SM model, the literature apparently offers no insights regarding the roles of the antisymmetric and the symmetric components of LP-derived vocal-tract shapes, and their possible relations with the formant frequencies and bandwidths — we only know that by specifying the first  $M/2$  formants, a discrete or step-wise,  $M$ -section area-function is obtained uniquely.

In that regard, it is important to note that the *shape* of an LP-derived area-function (similarly to that of a completely lossless model, as reviewed in the previous section) depends on its *length*. The principle of *minimal articulatory effort* which was applied first by Paige and Zue (1970), was then adapted to the LP method of inversion by Wakita (1977), who proposed to determine the vocal-tract length by minimising the eccentricity of the discrete LP area-function with respect to a uniform tube. Indeed, Wakita was able to re-estimate the lengths of five of Fant's (1960) area-functions of six Russian vowels, with errors of the order of  $\pm 5\%$  using the first four formant frequencies and bandwidths. Furthermore, he obtained reasonable estimates of the vocal-tract length of vowels recorded by several, adult male and female speakers of American English.

However, owing to the fact that the effective half-sampling frequency is inversely proportional to the length of each vocal-tract section, there remains the analysis artefact of an upper limit on allowed vocal-tract length (Wakita, 1977):

$$L_{\max} = \frac{cM}{4F_{\text{hi}}}, \quad (5.6)$$

where  $F_{\text{hi}}$  is the frequency of the highest formant considered,  $M$  is the number of sections, and  $c = 35300$  cm/sec is the velocity of sound propagation in the vocal-tract. If for a given set of formant frequencies and bandwidths, there cannot be found an optimum vocal-tract length less than  $L_{\max}$ , Wakita offers no alternative but to accept that upper limit (and the vocal-tract shape thus obtained).

Another limitation which is common to both the completely lossless and the LP-based inversion method, concerns the absolute value of estimated vocal-tract areas. For a completely lossless acoustic tube, the formants are independent of the value of the area scaling factor  $A_0$  (cf. Equation 5.1); similarly for the LP acoustic tube model, which is completely specified by its reflection coefficients. Theoretically,  $A_0$  should not

be so large as to violate the assumption of acoustic plane-wave propagation in the tract, and it should not be so small as to imply friction at the place of constriction for a vowel; physiologically, constraints on tongue-body volume suggest that  $A_0$  should not have an excessive, phonetic range of variation. In practice, the area scaling factor is usually determined such that either the area at the glottal end, or the average area, or the vocal-tract volume, remains fixed.

Whilst both the LP and the SM model are thus capable of yielding only *normalised*<sup>3</sup>, or *relative* vocal-tract area-functions, there appears to be no evidence in the literature to suggest that this limitation presents an obstacle in using such estimated shapes to gain physiological insights into acoustic speech phenomena. On the contrary, a very small number of studies have even attempted to gain such insights using the LP method of inversion — these include Crichton and Fallside’s (1974) experimental system for deaf speech training; Gath and Yair’s (1988) successful identification of lingual “tremor” in the sustained sound /l/ recorded by Parkinsonian subjects; and Hansen and Womack’s (1996) plausible descriptions of the articulatory differences in the vowel /e/, in the word “help” recorded in *neutral* and *angry* states of emotional stress. Although the estimated shapes cannot be claimed to be exactly those produced by the speakers, the plausibility of the insights gained in such studies does suggest that the consequences of the LP model’s well-known limitations (as described, e.g., by Wakita, 1979; Sondhi, 1979) may not be as severe as the literature generally portrays.

In that context, perhaps the most misleading impression conveyed in the literature on LP-based inversion, is that area-functions can be estimated directly by LP analysis of the speech waveform, after applying appropriate pre-emphasis. Indeed, Wakita (1973, p.422) does suggest that even a simple pre-emphasis of 6 dB/octave is “essential”, in order to approximately equalise the combined spectral-slope effects of the glottal source and lip radiation (Fant, 1960), and thereby reduce the possibility of obtaining “unusual tract shapes”. More elaborate methods of counteracting the influences of source and radiation characteristics by *flattening* the spectral slope and *enhancing* the

---

<sup>3</sup> Throughout this chapter, “*normalised area-function*” implies that the scale on the ordinate is dimensionless, owing to the fact that the area scaling factor ( $A_0$  in Equation 5.1) is acoustically inconsequential; this should not be confused with *speaker normalisation* of area-functions, to which we refer in Chapter 6.

formant peaks, include the adaptive inverse filtering approach developed by Nakajima et al. (1973), the adaptive enhancing filter proposed by Tanaka and Nakajima (1975), and Fuchi's (1977) use of the negative derivative of the LP phase spectrum (NDPS) to obtain an idealised, "stop/pass bands" spectrum. According to Sondhi (1979), "the only *tenable* conclusion is that the area recovery is *very strongly* dependent on the assumed source and radiation characteristics."

Apart from the fact that the influences of source and radiation characteristics on the formants can sometimes be appreciable, it is rarely acknowledged that LP analysis usually yields a set of poles, amongst which may be found not only the formants, but also the so-called *spurious poles*. Whilst the LP spectrum may not be adversely affected by the presence of spurious poles (especially those of wide bandwidth), it is fair to assume that the shape of the LP area-function itself might be significantly affected — as far as the LP vocal-tract model is concerned, a spurious pole which lies in between two "true" formants is just another resonance of the acoustic tube. Implicit in Wakita's (and Atal's) descriptions of the LP-based method of inversion, is therefore the retention of only those poles which can be considered as the "true" *formants*.

However, the requirement of formants (estimation of which is itself non-trivial) does not resolve the issue of the *relevance* of measured formants to the LP model. Indeed, one of the major criticisms of the LP model is that it "lumps" all the sources of loss into a single, resistive termination (at the glottal end in Wakita's model). By contrast, the formants measured from the acoustic speech waveform presumably include the effects of all losses which naturally occur in the human vocal-tract. For example (Fant, 1960; Flanagan, 1972), the glottal inductance tends to raise the centre-frequencies of formants which are more strongly affiliated with the pharyngeal cavity; the viscous and heat-conduction losses which are manifest along the surface of the vocal-tract walls, contribute mainly to the bandwidths of the higher formants; vocal-tract wall-vibrations tend to increase both the bandwidths and centre-frequencies of mainly the lower formants; and the lip-radiation impedance tends to increase the bandwidths of mainly the higher formants, and also to lower the centre-frequencies of formants which are more strongly affiliated with the oral cavity.

Wakita (1979) proposed to compensate for the LP model's lack of sufficiently

realistic losses, by way of a “formant frequency conversion chart” based on formants synthesised with a more “realistic” (i.e., a more lossy) vocal-tract model. Similarly, Hafer and Coker (1975) had earlier attempted to correct measured formant frequencies prior to inversion, in order “to account for inductive yielding walls of the vocal tract”. However, it remains questionable to what extent such *model*-based formant correction procedures are applicable to real, measured data; and how viable it is to formulate a similar correction procedure for formant *bandwidths* (which are, after all, more significantly affected than the formant frequencies by the LP model’s simplified assumptions in regard to vocal-tract losses).

Despite these limitations, the LP-based method of inversion is at an advantage compared with many other such methods, owing to the fact that the LP model is an acoustic *analysis* model, whereby conversion of formants into an area-function is direct (or non-iterative) and computationally inexpensive. However, as suggested by Broad and Shoup (1975, quoted in Section 2.4.3.2), the potential of the LP vocal-tract model to yield physiological insights into acoustic speech phenomena has to date been shamefully under-exploited. Perhaps a determining factor in the overwhelming neglect of the LP model over the past two decades, is its very coarse, step-wise representation of area-functions, from which it is admittedly difficult to extract crucial articulatory landmarks such as the place of constriction. More importantly, the coarse representation inhibits quantitative comparisons of area-functions of potentially different lengths, whether of different vowels or different speakers.

Those types of problems might conceivably be resolved by smoothing the discrete LP area-function, and thereby obtaining a continuous and parameterised representation. Indeed, Wakita and Gray (1975, Figure 4) used a Chebyshev polynomial approximation (also used earlier by Nakajima et al., 1973) to illustrate the LP-derived area-functions of five American English vowels. Although Wakita and Gray (1975, p.578) thus obtained “visually satisfying results” with the places of constriction more precisely defined, their particular choice of smoothing function was admittedly “somewhat arbitrary”.

On the other hand, as reviewed in the previous section, the SM model does afford a smooth representation of vocal-tract shapes, and more importantly, a resonance-based parameterisation. Whilst it may therefore seem advantageous to use the SM model to

parameterise LP-derived area-functions, it is not immediately obvious whether *the acoustically-relevant components of LP-derived vocal-tract shapes are equivalent to those of a completely lossless vocal-tract on which the SM model is based*. In the next section we seek to answer that question, by attempting to identify the parameters of unique, LP-derived area-functions.

### **5.3 Parameters of Unique LP-derived Area-Functions**

In Section 5.1 we defined the two main criteria which our method of inversion would need to satisfy — namely, uniqueness and formant-based parameterisation. Our bipartite rationale then addressed these issues by invoking two classic, but sadly over-neglected models, which do promise to fulfil our requirements. First, we reviewed the SM model, which does indeed provide an acoustically-meaningful parameterisation of the vocal-tract area-function. However, its underlying assumption of a completely lossless vocal-tract, implies that a unique area-function can only be obtained if both the zeros (formants) and the poles of the lip impedance function are known. This fundamental limitation then led us to consider the LP vocal-tract model, which is the only one to guarantee uniqueness of area-functions estimated using acoustic parameters, all of which can be measured from the recorded speech signal. However, whilst both the formant frequencies and the bandwidths are implicated in the LP model's uniqueness, it has never been shown how LP-derived vocal-tract *shapes* exploit the information contained in the formants, and of how those shapes might best be parameterised in an acoustically meaningful way.

As far as we are aware, Wakita and Gray (1975) were the first and only researchers to attempt to relate the LP model with the earlier work of Schroeder (1967) and Mermelstein (1967). Indeed, they suggested that the LP and the SM models could be used interchangeably to obtain a unique vocal-tract shape, once the appropriate acoustic parameters had been identified by LP analysis. In particular, they proposed that the frequencies of the zeros and poles of the lip impedance function (as required by the SM inversion method) could be found after LP analysis of the speech waveform, by setting the glottal reflection coefficient to unity in order to render the LP vocal-tract model completely lossless, then finding the required zeros of the lip impedance function

by solving for the roots of the modified LP polynomial, and similarly the required poles, after changing the sign of all of the reflection coefficients. It is interesting to note that the LP line spectrum pair (LSP) parameters (Itakura, 1975), which are often regarded merely as an alternative set of LP acoustic parameters, are indeed the inter-leaving frequencies of the zeros and poles of the lip impedance function of a completely lossless LP vocal-tract model, and are therefore precisely the necessary acoustic parameters which Wakita and Gray (1975) suggested for obtaining a unique vocal-tract shape using the SM inversion method.

Does this imply that the poles of the lip impedance function are contained in the acoustic speech signal after all, and that the SM model can therefore be used to obtain unique vocal-tract shapes without resort to the lip impedance-tube method proposed by Schroeder (1967)? To answer that question, it is important to note that Wakita and Gray's (1975) proposed method of uniting the LP and the SM models, relies on *post-analysis modification* of the LP glottal reflection coefficient; subsequent conversion to LSP frequencies does not change the shape of the LP area-function, which itself is uniquely determined by the frequencies and *bandwidths* of the LP poles yielded in the original analysis. The poles of the lip impedance function thus obtained, are therefore only indirectly inferred from the LP area-function, rather than directly measured from the acoustic speech signal.

It emerges from the preceding discussion, that our knowledge of the uniqueness properties of LP-derived vocal-tract shapes is still rather limited. Although in their respective formulations of the LP-based inversion method, Atal and Wakita certainly provide evidence of the uniqueness of LP-derived area-functions, they stop short of describing how the components of estimated vocal-tract shapes themselves are involved in the uniqueness. By contrast, the SM model lends direct insights into those fundamental components of vocal-tract shapes which are of greatest relevance from acoustic-phonetic, perceptual, and articulatory points of view. Notwithstanding the differences between the respective vocal-tract models, it would therefore seem appropriate to use the SM model as a basis for identifying the parameters of unique LP-derived area-functions.

As reviewed in the previous section, the underlying difference between the SM

model and the LP acoustic-tube model is that the latter includes a resistive element at the glottal end. The acoustic consequences of a purely resistive glottal termination have been studied both theoretically (e.g., Fant, 1960; Flanagan, 1972) and empirically (e.g., Badin and Fant, 1984). As a result, it is well-established that a resistive glottal termination influences the formant frequencies far less than the bandwidths; furthermore, that the resulting losses are *frequency-independent*. We might therefore expect that the LP model shares the distinctive articulatory-acoustic relations embodied in the SM model — in particular, that it inherits the property of a quasi-linear relation between formant frequencies and the corresponding, antisymmetric shape components of the logarithmic area-function. We address this issue in Section 5.3.1, by presenting new results which shed light on the formant frequency-dependence of LP-derived vocal-tract shapes. In Section 5.3.2 we then address the question of fundamental concern, whether the formant bandwidths, which provide the crucial second-half of acoustic information required to secure uniqueness of LP-derived vocal-tract shapes, are at all related to the undetermined, symmetric components of those shapes.

### **5.3.1 Dependence of LP-derived VT-Shapes on Formant Frequencies**

Our first step towards marrying the SM model and the LP model, is to prove their equivalence in regard to the dependence of vocal-tract shapes on the formant frequencies. This proof is first considered from a theoretical point of view (in Section 5.3.1.1), with the simplifying assumption of a two-section vocal-tract area-function. It is then extended (in Section 5.3.1.2) with empirical results which substantiate our hypothesis of the equivalence of the two models, and provide the most extensive validation of the SM model ever to appear.

#### **5.3.1.1 Partial Theoretical Proof**

One of the distinguishing features of the completely lossless SM model reviewed in Section 5.2.1, is its *continuous* representation of vocal-tract area functions, in terms of the variable  $x$ . Hence, each parameter of the SM model specifies a component of the vocal-tract shape, which is independent of the number of sections used to implement the model in practice (as would be required, for example, to synthesise the acoustic

resonances of a given area-function using numerical procedures). The LP model, on the other hand, not only represents the area-function in terms of a discrete number of lossless sections  $M$  (which, in the absence of real-valued poles, is equal to twice the number of LP poles effectively used), but is also characterised by a resistive (i.e., a lossy) termination at one end of the vocal-tract. In order to confirm our hypothesis that LP-derived vocal-tract shapes are governed by the same basic principles that define the SM model, it is therefore necessary first to recast the main result of the SM model in terms of a discrete-sectioned, lossless acoustic tube.

Bonder (1983a) has shown that Webster's Horn Equation (Equation B.3 or B.4, in Appendix B) can be used to derive an analytical solution for the resonances of a lossless acoustic tube with up to 10 equal-length sections. For simplicity, we will consider a single-resonance acoustic tube which, according to the LP model, comprises only two sections of equal length  $L/2$ . The simplified version of Bonder's (1983c) so-called " $n$ -tube formula" then reduces to the following expression (cf. also Fant, 1960, p.65; Flanagan, 1972, p.70):

$$\tan^2\left(\frac{\pi L F_1}{c}\right) = \frac{A_1}{A_2}, \quad (5.7)$$

where  $A_1$  and  $A_2$  are the cross-sectional areas of the front and the back vocal-tract sections, respectively, and  $F_1$  is the frequency of the single resonance.

According to the SM model (Equation 5.2), the most efficient way to perturb  $F_1$  from its neutral value is to perturb the uniform area-function according to the antisymmetric shape parameter  $a_1$ . For a two-section acoustic tube, Equation 5.1 then yields the following area-perturbations which would induce a positive perturbation in the first formant frequency:

$$A_m = e^{-a_1 \cos((2m-1)\pi/4)}, \quad m = 1, 2, \quad (5.8)$$

where it is assumed that the area of each section takes on the value given by Equation 5.1 at the centre of that section (i.e., at  $x = 3L/4$  and  $x = L/4$  for the front and the back sections, respectively), and that  $A_0 = 1$ . Equations 5.8 and 5.7 then together yield the following expression for the single resonance-frequency  $F_1$  in terms of the vocal-tract shape parameter  $a_1$ :

$$F_1 = \frac{c}{\pi L} \tan^{-1} \sqrt{e^{-a_1 \sqrt{2}}}. \quad (5.9)$$

This last result, which is the homologue of Equation 5.2 for  $n=1$ , can be regarded as an *exact* version of the SM model for a lossless two-tube. It is exact in that it correctly predicts the value of the single resonance frequency as it would be obtained by a sufficiently accurate numerical procedure to solve Webster's Horn Equation for a two-section area-function, independently of the degree of eccentricity of that area function with respect to a uniform tube. Assuming  $c = 35300$  cm/sec and  $L = 17.65$  cm as in our earlier illustration of the SM model (Figure 5.1), the solid curve in Figure 5.2 illustrates the result just derived, showing the predicted variations in  $F_1$  about its neutral value of 500 Hz, for area-function perturbations over a fairly wide range  $a_1 \in [-2, +2]$ . The exactness of Equation 5.9 is indeed evidenced by the way in which it correctly predicts the  $F_1$  asymptotes of 0Hz and 1000Hz, respectively, for large positive and large negative perturbations of  $a_1$ . By contrast, the linearity of the SM model (Equation 5.2 for  $n=1$ , shown here by the dotted curve) is strictly only valid for relatively small perturbations about a uniform area-function.

Having established the exact form of the SM model for a completely lossless area-function with two sections, we now proceed to derive an analogous expression for the LP model. According to Wakita & Gray's (1975) formulation, the single resonance of a two-tube LP-model is given by the complex-conjugate roots of the following numerator-polynomial of the lip impedance function:

$$z^2 + \mu_1(1 + \mu_2)z + \mu_2 = 0, \quad (5.10)$$

where  $\mu_1 = (A_1 - A_2) / (A_1 + A_2)$  is the reflection coefficient which describes the shape of the two-section area-function,  $\mu_2 = \rho c / A_g^{(LP)}$  is the glottal reflection coefficient which determines the effective (LP) glottal opening area  $A_g^{(LP)}$  and hence the losses introduced by the glottal resistance, and  $z = e^{j2\pi LF/c}$ . Solving for the roots of the quadratic Equation 5.10, the single resonance frequency is then expressed in terms of the two reflection coefficients, as follows:

$$F_1 = \frac{c}{2\pi L} \tan^{-1} \left( \frac{\sqrt{4\mu_2 - \mu_1^2(1 + \mu_2)^2}}{-\mu_1(1 + \mu_2)} \right), \quad (5.11)$$

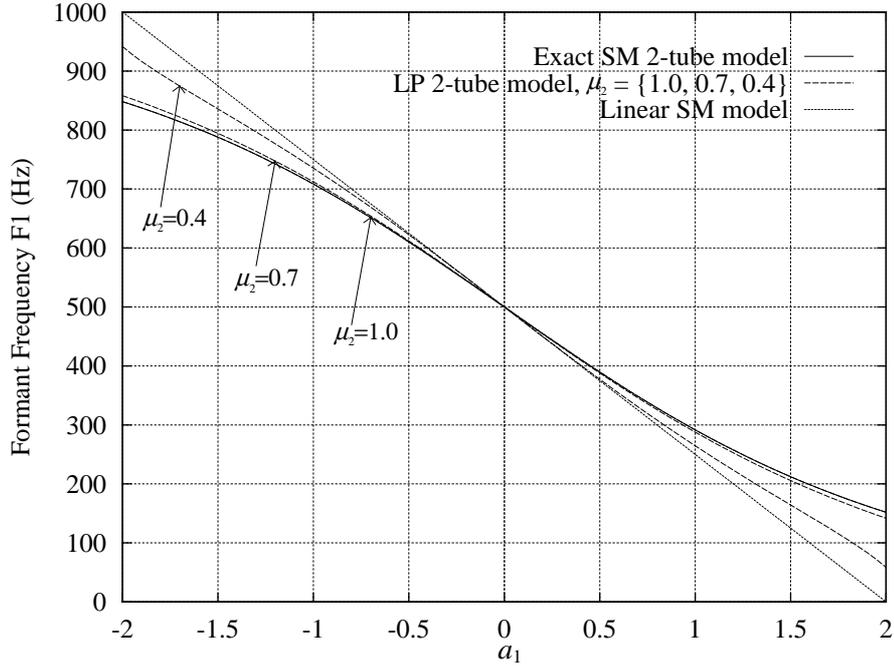


Figure 5.2: Theoretical behaviour of the resonance frequency  $F_1$  of a 2-tube, as a function of the shape-perturbation parameter  $a_1$ . *Solid curve*: exact relation (Equation 5.9) derived from a completely lossless 2-tube. *Dashed curves*: exact relation (Equations 5.11 and 5.12) derived from a 2-section LP model, shown here for three values of the glottal reflection coefficient  $\mu_2 = \{1.0, 0.7, 0.4\}$ . *Dotted curve*: linear relation predicted by the SM model (Equation 5.2).

where the range of the inverse-tangent function is taken as  $[0, \pi]$  by adding  $\pi$  to the result whenever the argument is negative-valued, in order to ensure that only the pole lying in the top half of the  $z$ -plane is considered.

Finally, to obtain an expression analogous to Equation 5.9, we need to derive the relation between the reflection coefficient  $\mu_1$  and the area-function parameter  $a_1$ . Making use of Equation 5.8 and the definition of the reflection coefficient in terms of the individual areas, we obtain the following:

$$\mu_1 = -\tanh\left(\frac{a_1}{\sqrt{2}}\right). \quad (5.12)$$

Substitution of Equation 5.12 into Equation 5.11 then yields the desired relation which is analogous to the earlier-derived Equation 5.9, with the extra flexibility afforded by the resistive loss-element at the glottal end of the two-tube LP-model — an infinite glottal impedance (i.e. a closed glottis) can be obtained by setting  $\mu_2 = 1$ , while finite and increasing losses are introduced as  $\mu_2$  decreases towards zero.

The dashed curves in Figure 5.2 illustrate the LP-based equations just derived, for three different values of the glottal reflection coefficient  $\mu_2 = \{1.0, 0.7, 0.4\}$ . As expected, when  $\mu_2 = 1$  (completely lossless, two-tube LP-model) the dashed curve coincides exactly with the solid curve, which was obtained earlier from the exact version of the two-tube SM model (Equation 5.9). The dashed curve for  $\mu_2 = 0.7$  shows a very similar behaviour of the resonance frequency, indicating that the inclusion of a moderate amount of loss at the glottal end does not significantly alter the relation between  $F_1$  and the shape of a two-tube. However, it is interesting to note that as the glottal resistance is made to play a more significant role (dashed curve for  $\mu_2 = 0.4$ ), the behaviour of  $F_1$  in the two-tube LP-model begins to approach the quasi-linear behaviour (dotted curve) predicted by the original SM model (Equation 5.2), which is based on a completely lossless acoustic-tube, and which itself makes no assumptions as to the number of vocal-tract sections.

We should emphasise, however, that the range over which  $a_1$  has been perturbed in Figure 5.2, is much wider than would normally be expected for vocalic configurations (for example, Mermelstein's (1967) analysis of Fant's (1960) six Russian vowels yielded values for  $a_1$  spanning approximately  $[-1, 1]$ , i.e., about half of the range shown here). It is also quite clear in Figure 5.2 that the small but systematic differences between the theoretically-derived models, are only of significance for such large perturbations as are shown at the positive and the negative extremes along the abscissa. Within the more restricted range of perturbations (e.g.,  $[-1, 1]$ ) which are typical of vocal-tract shapes for vowels, Figure 5.2 therefore confirms that the behaviour of the resonance frequency  $F_1$  as a function of the shape-perturbation parameter  $a_1$  is essentially common to both models. Our results concerning the two-section vocal-tract, thereby support a partial theoretical proof of the validity of the SM model within the LP modelling framework.

### 5.3.1.2 Empirical Validation

We now aim to further substantiate the partial theoretical proof presented thus far, by extending our validation to vocal-tract shapes with more than two sections. Two methods are used to illustrate empirically the relation between formant frequencies and

antisymmetric (and symmetric) perturbations of an LP area-function. In particular, we show that this relationship is essentially the same as that predicted by the SM model for a completely lossless acoustic tube.

The first method is similar to our earlier illustration of the SM model (Figure 5.1, in Section 5.2.1), and involves nomograms depicting formant frequency variations as a function of perturbations in a single SM-model parameter at a time. However, in contrast to the earlier results which were generated using a completely lossless vocal-tract model, the resonances of computed vocal-tract shapes are here synthesised using the LP vocal-tract model, which not only involves a discrete number of equal-length sections (strictly equal to twice the number of formants), but also includes a resistive termination at the glottal end. Hence, for each perturbation of an SM-model parameter  $a_n$ , an area-function with  $M$  equal-length sections is first computed, with the area of each section taking on the value given by Equation 5.1 at the centre of that section, and assuming  $A_0 = 1$ . The discrete and normalised area-function  $A_m$ ,  $m = 1, \dots, M$  is then converted to a set of reflection coefficients  $\mu_m$ ,  $m = 1, \dots, M - 1$ , and the glottal (the  $M^{\text{th}}$ ) reflection coefficient is chosen such that  $0 < \mu_M \leq 1$ . A well-known recursive algorithm (Markel and Gray, 1976) is then used to transform the set of  $M$  reflection coefficients to a set of  $(M + 1)$  LP autoregressive coefficients, which defines the  $M^{\text{th}}$ -order polynomial of the LP inverse filter. Finally, the  $M/2$  formant frequencies are obtained by solving for the roots of the LP inverse-filter polynomial<sup>4</sup>.

Figure 5.3 shows a set of nomograms obtained for LP area-functions of  $M = 8$  sections, resulting in the synthesis of only the first four formants. As in Figure 5.1, the panels on the left-hand side depict formant-frequency nomograms for perturbations of each of the first three odd-indexed parameters ( $a_1$ ,  $a_3$ , and  $a_5$ ) separately, while those on the right-hand side show perturbations of each of the first three even-indexed parameters ( $a_2$ ,  $a_4$ , and  $a_6$ ). Also shown in each panel are the sets of superimposed, eight-section LP area-functions, each with a fixed total length  $L = 17.65$  cm, whose shapes are determined according to the single parameter being perturbed about its neutral (zero) value. Rather than the usual, step-wise depiction of LP area-functions,

---

<sup>4</sup> This well-known sequence of procedures for obtaining the formants of a given, discrete area-function, will henceforth be referred to as “LP synthesis”.

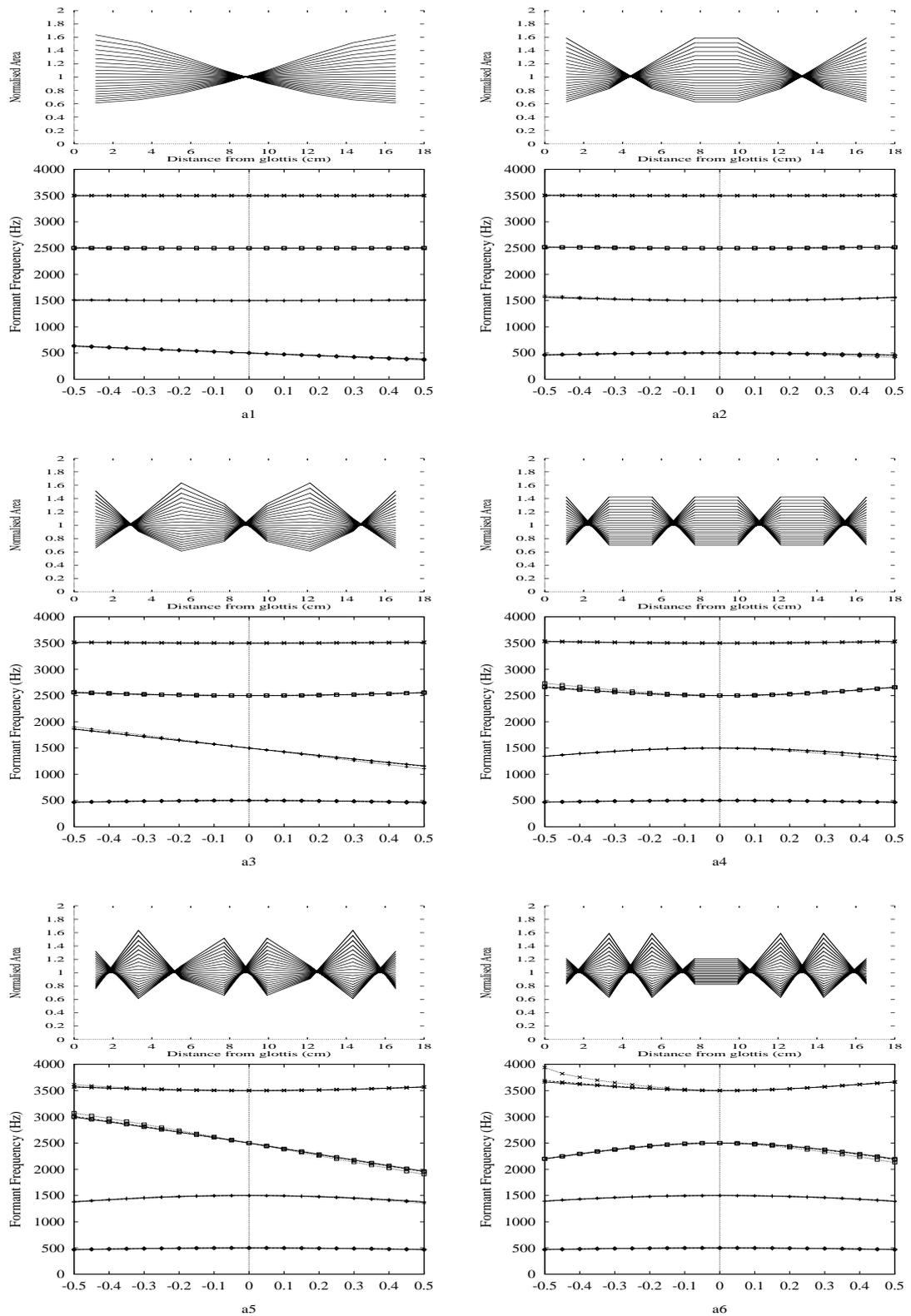


Figure 5.3: Nomograms depicting changes in the first four formant frequencies, as a function of perturbations in each of the first six, vocal-tract shape-parameters of the SM model. Each  $M=8$ -section area-function is of length  $L=17.65$ cm, and the area scaling factor is  $A_0=1.0$ ; the formants are synthesised using the LP vocal-tract acoustic model, with the glottal reflection coefficient  $\mu_M=1.0$  (solid curves), 0.7 (dashed curves), and 0.4 (dotted curves).

we emphasise their shapes by plotting a piecewise-linear representation obtained by joining the section-centres; these area-functions are therefore merely a more coarsely-sampled version of those shown earlier in Figure 5.1.

For each formant in Figure 5.3, are shown three curves which have been obtained, respectively, using a completely lossless implementation of the LP vocal-tract model ( $\mu_M = 1$ , shown by the solid curve), and two other configurations with increasing amounts of loss ( $\mu_M = 0.7$ , shown by the dashed curve; and  $\mu_M = 0.4$ , shown by the dotted curve). It is evident that in most cases, the dashed and the solid curves are indistinguishable — only in a few cases, mainly in the higher formants and for extreme values of perturbation, are the dotted curves visually separable from those pertaining to a more tightly-closed glottis. Figure 5.3 thus confirms that the resistive glottal termination has only a very small influence on the LP-synthesised formant frequencies. In particular, as the equivalent glottal opening area is increased by decreasing the value of  $\mu_M$ , a slightly steeper (more negative) slope is obtained for the first-order relation between each  $F_n$  and the corresponding shape parameter  $a_{2n-1}$ . Nevertheless, the general behaviour of the formant-frequency nomograms is similar to that observed in Figure 5.1, with each of the odd-indexed Fourier cosine coefficients claiming first-order control of a unique formant frequency, and with the even-indexed parameters exerting only second-order influence by comparison. We therefore conclude that these results are in agreement with the theoretical predictions of formant frequency behaviour based on the SM model (Equation 5.2).

In our second, empirical validation of the equivalence of the SM and LP models, we aim to reveal the strength and the inherent structure of the relations between shape and resonance parameters, by considering a more extensive range of shape-parameter perturbations about a neutral tube. Retaining the eight parameters ( $a_n$ ,  $n = 1, \dots, 8$ ) which are implied by the SM model (Equation 5.1) for an 8-section area-function, a total of 6561 different vocal-tract shapes are generated by all possible combinations of ternary perturbations of each parameter ( $-0.1, 0.0, +0.1$ ). The area-functions are found using Equation 5.1, and assuming a fixed vocal-tract length  $L = 17.65$  cm and a normalised area-scaling factor  $A_0 = 1$ ; a completely lossless acoustic tube is assumed ( $\mu_8 = 1$ ), and the first four formant frequencies are obtained using the LP forward

	$a_1$	$a_3$	$a_5$	$a_7$
$F_1^{(\text{rel})}$	<b>-0.996</b>	-0.001	0.000	0.000
$F_2^{(\text{rel})}$	-0.001	<b>-0.991</b>	-0.004	-0.001
$F_3^{(\text{rel})}$	0.000	-0.004	<b>-0.986</b>	-0.010
$F_4^{(\text{rel})}$	0.000	-0.001	-0.007	<b>-0.997</b>

Table 5.1: Linear correlation coefficients (*Pearson’s r*) between each of the first four (relative) formant frequencies LP-synthesised from 6561 perturbations of an 8-section, uniform area-function, and each of the first four odd-indexed VT-shape parameters from which those area-functions were generated.

routines. Finally, the coefficient of correlation (or *Pearson’s r*) is computed across all 6561 configurations, between the shape parameters and the *relative* formant frequencies  $F_n^{(\text{rel})} = (F_n - F_n^{(\text{neut})}) / F_n^{(\text{neut})}$  (i.e., each formant frequency normalised with respect to its neutral value  $F_n^{(\text{neut})} = (2n - 1)c / 4L$  for the given vocal-tract length, as implied in Equation 5.2).

Table 5.1 shows the correlations thus obtained. As expected, the table is heavily dominated by the entries along the main diagonal. Whilst the SM model theoretically establishes the negatively-sloped and quasi-linear relation between the  $n^{\text{th}}$  relative formant frequency and the corresponding shape parameter  $a_{2n-1}$ , the high, negative correlations along the main diagonal of Table 5.1 summarise the strength of that linearity. In addition, the relatively insignificant correlations in all of the off-diagonal entries of Table 5.1, support the notion of a predominantly one-to-one mapping between each odd-indexed Fourier cosine parameter and the corresponding formant frequency. As far as we are aware, this evidence provides the most extensive, empirical validation of the SM model ever reported. More importantly, our results further validate the equivalence of the SM and the lossless LP model<sup>5</sup>.

Indeed, the theoretical and empirical evidence presented in this section supports our hypothesis that the LP model and the SM model are *isomorphic* in regard to the fundamental relation between formant frequencies and the antisymmetric components of the logarithmic area-function. This mutual congruence in the basic acoustic properties

<sup>5</sup> A complementary proof of the equivalence of those two models, is to show that the LP method of inversion can be “stepped down” such as to yield area-functions which would have been obtained using the SM model, starting from the same set of formant frequencies. The interested reader is referred to Appendix D.

of the two models can therefore be used to advantage, by adopting a parameterisation of LP-derived area-functions in terms of the Fourier cosine coefficients of the SM model (cf. Equation 5.1). Insofar as this type of parameterisation allows a smooth and continuous representation of vocal-tract shapes, it immediately resolves the difficulties of accurately identifying crucial articulatory landmarks such as the place of constriction, and suggests ways to approach the problem of quantitatively comparing two or more LP-derived area-functions (elaborated in Section 5.5.1). However, the key strength of the SM-model parameterisation lies in its acoustically-meaningful description of vocal-tract shapes, which has hitherto never been attempted within the LP modelling framework.

### 5.3.2 Dependence of LP-derived VT-Shapes on Formant Bandwidths

Thus far, we have been fortunate to be able to recall the SM model and unfold it as far as elucidating the acoustic determinants of the antisymmetric components of LP-derived vocal-tract shapes. However, despite the acknowledged importance of the formant *bandwidths* in securing uniqueness in the LP-based method of inversion, it has never been explained *how* these acoustic parameters contribute to the *shape* of LP-derived area-functions. In the next two sections, we therefore consider this problem from a theoretical and an empirical point of view, respectively.

#### 5.3.2.1 Theoretical Motivation

In this section we present a theoretical derivation of the relation between the formant bandwidths and the glottal reflection coefficient of the LP vocal-tract model. Whilst a similar derivation was presented briefly by Kasuya and Wakita (1979), we herein present a more detailed derivation which leads to an expression involving the mean bandwidth, and which then allows us to foreshadow the bandwidth-dependence of LP-derived vocal-tract shapes.

According to the well-known recursive algorithms for converting between the polynomial (autoregressive) coefficients and the reflection coefficients (e.g., Markel and Gray, 1976), the glottal reflection coefficient  $\mu_M$  is equal to the  $M^{\text{th}}$  coefficient of the inverse-filter polynomial (see for example Equation 5.10 which shows that, for a

second-order polynomial in  $z$ , the constant term is equal to the highest-indexed reflection coefficient  $\mu_2$ ). In the same vein, it is relevant to recall Wakita's (1973) statement that, as the LP inverse-filter polynomial has only real or complex-conjugate roots, the coefficient of its highest-order term ( $z^{-M}$ ) is determined only by the bandwidths of those roots. Therefore it is not surprising that the bandwidths should determine the value of the glottal reflection coefficient, which itself defines the effective glottal opening area  $A_{M+1}$ , and hence the amount of losses dissipated through the characteristic resistance of the glottal termination.

Indeed, a well-known theorem of algebra (e.g., Chrystal, 1964, p.432) states that, for a monic polynomial (such as that of the LP inverse-filter), the constant term (or the  $M^{\text{th}}$  autoregressive coefficient, which is equal to the glottal reflection coefficient  $\mu_M$ ) is proportional to the product of all  $M$  roots of the polynomial, as follows:

$$\mu_M = (-1)^M \prod_{m=1}^M z_m. \quad (5.13)$$

Without loss of generality, the LP order of analysis  $M$  can be assumed to be even, thereby reducing the constant of proportionality to positive unity. Further, since the roots of the LP polynomial occur in complex-conjugate pairs, the product of all of the complex roots is equal to the product of their magnitudes, such that Equation 5.13 can be rewritten as follows:

$$\mu_M = \prod_{m=1}^M |z_m|. \quad (5.14)$$

It is also well known that the bandwidth of the  $m^{\text{th}}$  root of the LP inverse-filter polynomial is proportional to the natural-logarithm of the magnitude of that root in the  $z$ -plane, and is given by the following:

$$B_m = -\frac{F_s}{\pi} \ln|z_m|, \quad (5.15)$$

where  $F_s$  is the sampling frequency. The sum of the bandwidths of all  $M$  roots can therefore be expressed as follows:

$$\sum_{m=1}^M B_m = -\frac{F_s}{\pi} \ln\left(\prod_{m=1}^M |z_m|\right). \quad (5.16)$$

Substituting for the product of the magnitude of the roots from Equation 5.14, and noting also that  $F_s = cM / 2L$  (which is a property common to all vocal-tract models with a discrete number of sections, such as the LP model), Equation 5.16 can finally be rewritten as follows:

$$\bar{B} = -\frac{c}{2\pi L} \ln(\mu_M), \quad (5.17)$$

where  $\bar{B}$  denotes the mean resonance bandwidth. According to this last result, the mean bandwidth of the  $M/2$  independent poles of an  $M^{\text{th}}$ -order LP analysis, is proportional to the logarithm of the glottal reflection coefficient, and is inversely proportional to the assumed vocal-tract length.

More importantly, Equation 5.17 states that, for a fixed vocal-tract length, the glottal reflection coefficient is the sole determinant of the *mean* bandwidth of the formants generated by an LP area-function. This implicates the remaining reflection coefficients  $(\mu_1, \dots, \mu_{M-1})$  which define the *shape* of the vocal-tract, in determining the value of the individual formant bandwidths *relative to the mean bandwidth*  $\bar{B}$  for the given LP area-function. In the next section we demonstrate the truth of this implication, and as a result, empirically derive a new set of vocal-tract shape parameters which relate distinctively to the *relative formant bandwidths*, analogously to the SM model (Equation 5.2) which relates each odd-indexed Fourier cosine coefficient to a unique, *relative formant frequency*.

### 5.3.2.2 Empirical Justifications

With this clear objective in mind, we now endeavour to explore the dependence of LP-derived vocal-tract shapes on the formant bandwidths. In particular, we embrace the theoretically-motivated framework of the SM model, and seek to gain insights into the bandwidth-dependence of shapes perturbed about a neutral tube. However, rather than venture a hypothesis on the type of perturbation which might be necessary to induce a distinctive change in only a single formant bandwidth, we use the LP inverse method to obtain the vocal-tract shapes corresponding to such acoustic perturbations.

In order to obtain vocal-tract shapes perturbed from the neutral configuration according to variations in a single bandwidth at a time, we first need to identify the

formant frequencies and bandwidths of a *uniform* LP area-function. A cursory examination of the LP-based nomograms presented earlier (in Figure 5.3) reveals that the neutral-tube formant frequencies of the LP model are basically the same as those of a completely lossless uniform tube of the same length  $L$ , as given by the quarter-wavelength formula; i.e.,  $F_n = (2n - 1)c / 4L$ , regardless of the value of the glottal reflection coefficient. As the resistive glottal termination of the LP acoustic-tube model induces energy losses which are frequency-independent, the values of the  $M/2$  bandwidths of a uniform tube represented by  $M$  equal-length sections are identical to the mean bandwidth, for which we have derived an exact relation (Equation 5.17). For example, an  $M$ -section, uniform LP area-function of total length  $L = 17.65$  cm and glottal reflection coefficient  $\mu_M = 0.7$ , has the following set of formant parameters:  $F_n = (2n - 1)500$  Hz and  $B_n \approx 113.5$  Hz, for  $n = 1, \dots, M/2$ .

The bandwidth-dependence of LP-derived vocal-tract shapes can now be assessed by perturbing a selected bandwidth, while keeping all other formant parameters (i.e., the frequencies and the remaining bandwidths) fixed at their respective, neutral values. Assuming a nominal, neutral bandwidth of 100 Hz, each formant bandwidth is chosen in turn, and perturbed two steps below (50 Hz and 80 Hz) and two steps above (125 Hz and 200 Hz) the neutral value. The LP inverse method is then used to obtain a vocal-tract shape for each set of formant parameters, by the well-known transformations through LP autoregressive coefficients and LP reflection coefficients. As our aim is to discern an underlying pattern in the vocal-tract shapes obtained in this way, the first seven formants are specified in order to secure  $M = 14$  vocal-tract sections, which does permit a satisfactory spatial resolution for visually identifying the general form of the area-function perturbation.

Figure 5.4 shows the 14-section, LP-derived vocal-tract shapes obtained by the procedure just described, for changes in each of the first three formant bandwidths. It is immediately apparent that perturbation of only the  $n^{\text{th}}$  formant bandwidth, induces an approximately *sinusoidal* perturbation of the area-function, with  $2n - 1$  half-cycles within the length of the vocal-tract from the glottis to the lips. Recall from Section 5.2.2 that the area scaling factor  $A_0$  is acoustically inconsequential in both the completely lossless and the LP vocal-tract model; hence, the area-functions are plotted in such a

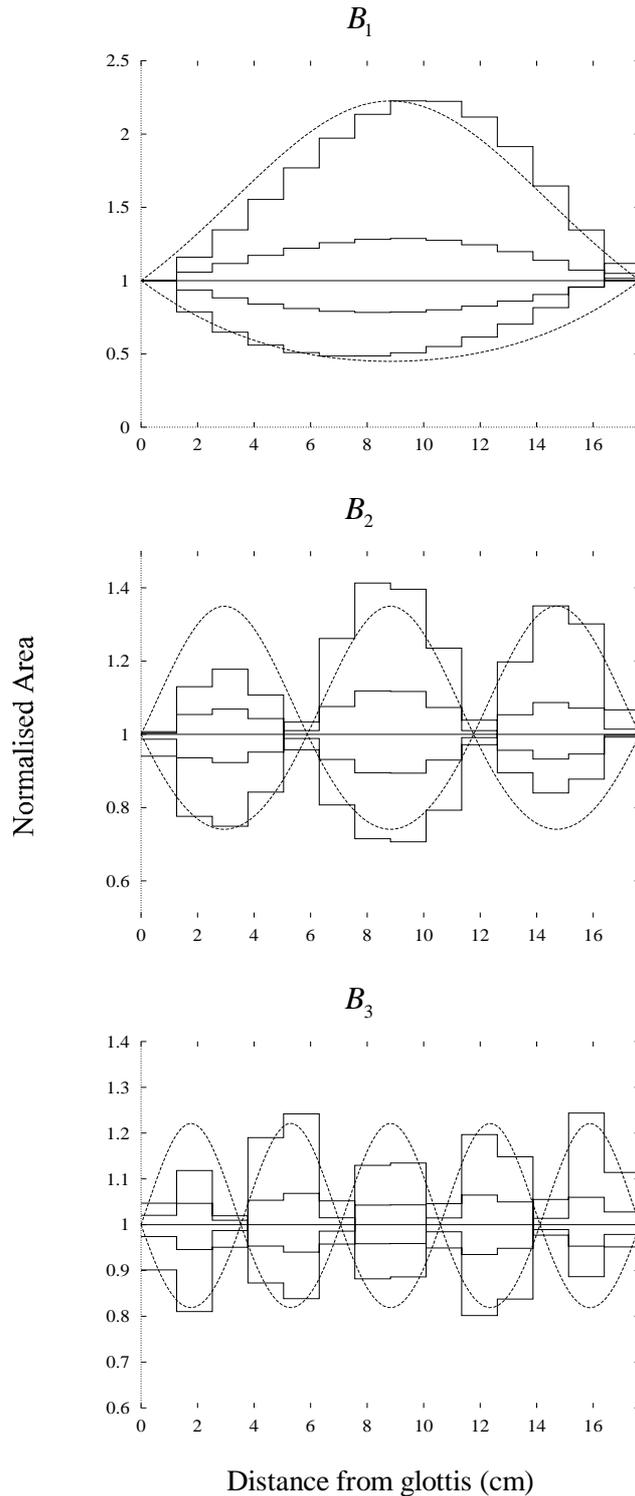


Figure 5.4: *Bandwidth-dependence of LP-derived vocal-tract shapes.* Each formant bandwidth  $B_n$  is perturbed over five values {50, 80, 100, 125, 200} Hz, with all other bandwidths fixed at the nominal, neutral value 100Hz; the formant frequencies are fixed at neutral values  $F_n=(2n-1)c/4L$ , and the vocal-tract length is fixed at  $L=17.65\text{cm}$ . The *top*, *middle*, and *bottom* graphs show the LP-derived area-functions for perturbations in  $B_1$ ,  $B_2$  and  $B_3$ , respectively. Seven formants are used in order to obtain 14-section LP area-functions (*solid lines*); at about the mid-length of each area-function, the largest through smallest area corresponds to the increase in bandwidth from 50Hz to 200Hz as listed above. *Dashed curves*: envelopes of the 1st, 3rd, and 5th sinusoidal components for the given  $L$ , superimposed on the step-wise area-functions in order to emphasise the functional form of the shape perturbations.

way as to visually emphasise the sinusoidal component, by setting the mean logarithmic area to zero in the bottom two graphs, and by setting a constant area for the first section in the top graph.

The significance of the LP-based result portrayed in Figure 5.4, is perhaps best appreciated in the context of the SM model. In Section 5.3.1 we established a link between the SM model and the LP model, by showing that the latter does inherit the fundamental relation between antisymmetric shape components and perturbations in the formant frequencies. In Figures 5.1 and 5.3 we then showed that for both the SM and LP models, respectively, the formant frequencies are relatively independent of the symmetric shape components. Remarkably, Figure 5.4 reveals that when a completely lossless vocal-tract model is augmented with a resistive termination at the glottal end (as in Wakita's formulation of the LP model), *symmetric* (sinusoidal) perturbations of the uniform area-function are associated with distinctive variations in the resulting formant *bandwidths*. Our results therefore elucidate the bandwidth-dependence of LP-derived area-functions, and furthermore suggest a functional form of the required, symmetric shape perturbations (as shown in Figure 5.4 by the superimposed, dashed curves), in terms of *the odd-indexed coefficients of the Fourier sine (rather than the SM model's original Fourier cosine) series of the logarithmic area-function*.

Having thus identified the shape components which fundamentally relate to the individual bandwidths, a more complete assessment of the implied, shape-resonance relations is afforded by nomograms prepared using the LP model. Considering only the first four formants, LP area-functions are each represented by  $M = 8$  sections of equal length. Each of the shape parameters (i.e., the odd-indexed Fourier cosine coefficients  $a_{2n-1}$  and the newly-implicated, odd-indexed Fourier sine coefficients, henceforth denoted as  $b_{2n-1}$ ) is perturbed across the range  $[-0.8,+0.8]$  in steps of 0.05, while keeping all other parameters fixed at zero value. Either discrete cosine expansion (for the  $a_{2n-1}$ ) or discrete sine expansion (for the  $b_{2n-1}$ ) is then used to obtain 8-section area-functions of fixed length  $L = 17.65$  cm, from which the first four formant frequencies and bandwidths are LP-synthesised (assuming a nominal value for the glottal reflection coefficient  $\mu_M = 0.7$  which, according to Equation 5.17, yields a reasonably realistic, mean bandwidth of  $\bar{B} \approx 113.5$  Hz).

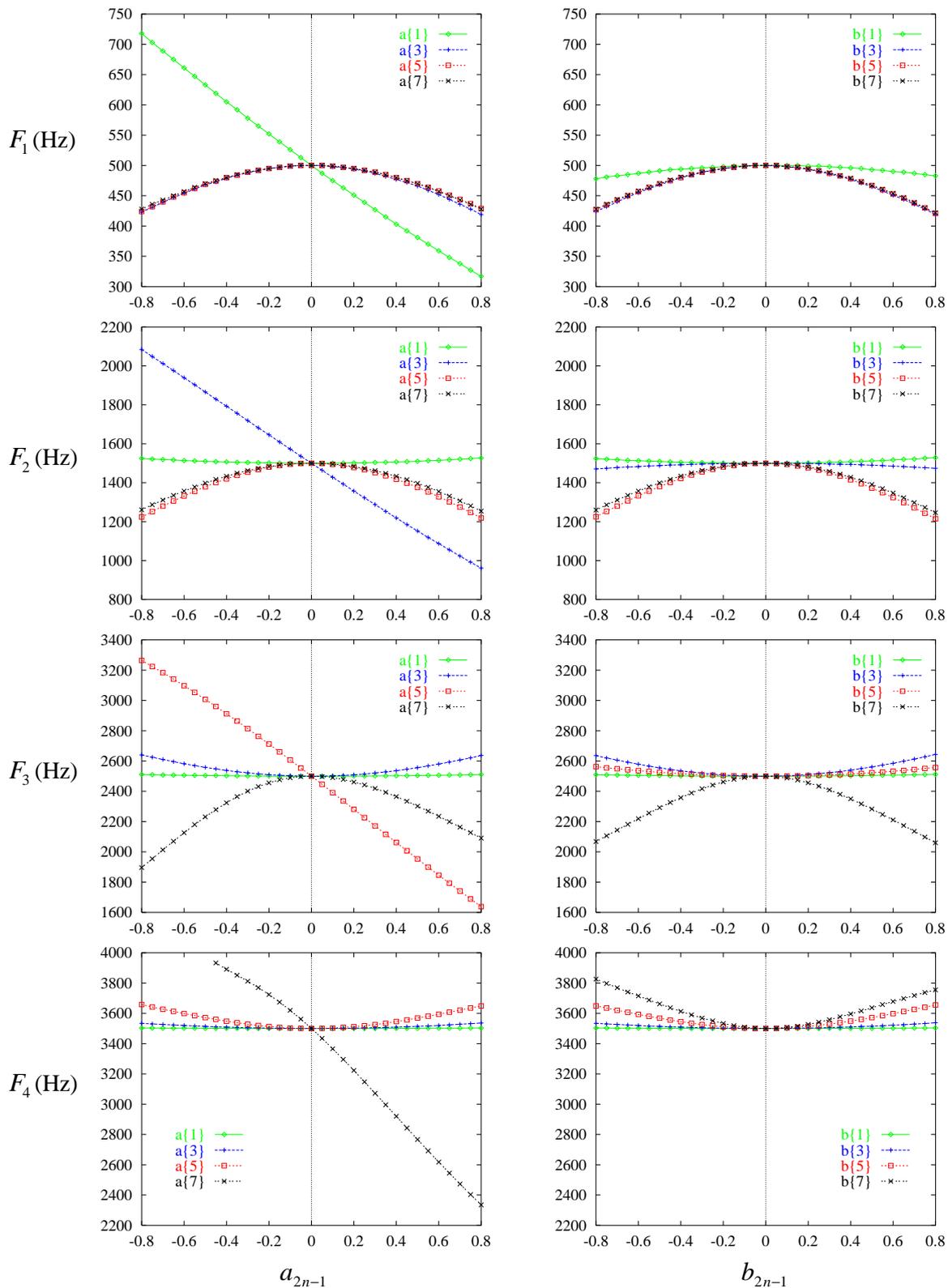


Figure 5.5(a): *Formant-frequency nomograms* generated by perturbing a uniform, 8-section LP area-function, according to each of the first four, odd-indexed Fourier cosine ( $\Delta a_{2n-1}$ , *left graphs*) and sine ( $\Delta b_{2n-1}$ , *right graphs*) shape-components, for  $n=1$  (diamond symbols joined by green lines),  $n=2$  (plus symbols joined by blue lines),  $n=3$  (square symbols joined by red lines), and  $n=4$  (cross symbols joined by black lines).

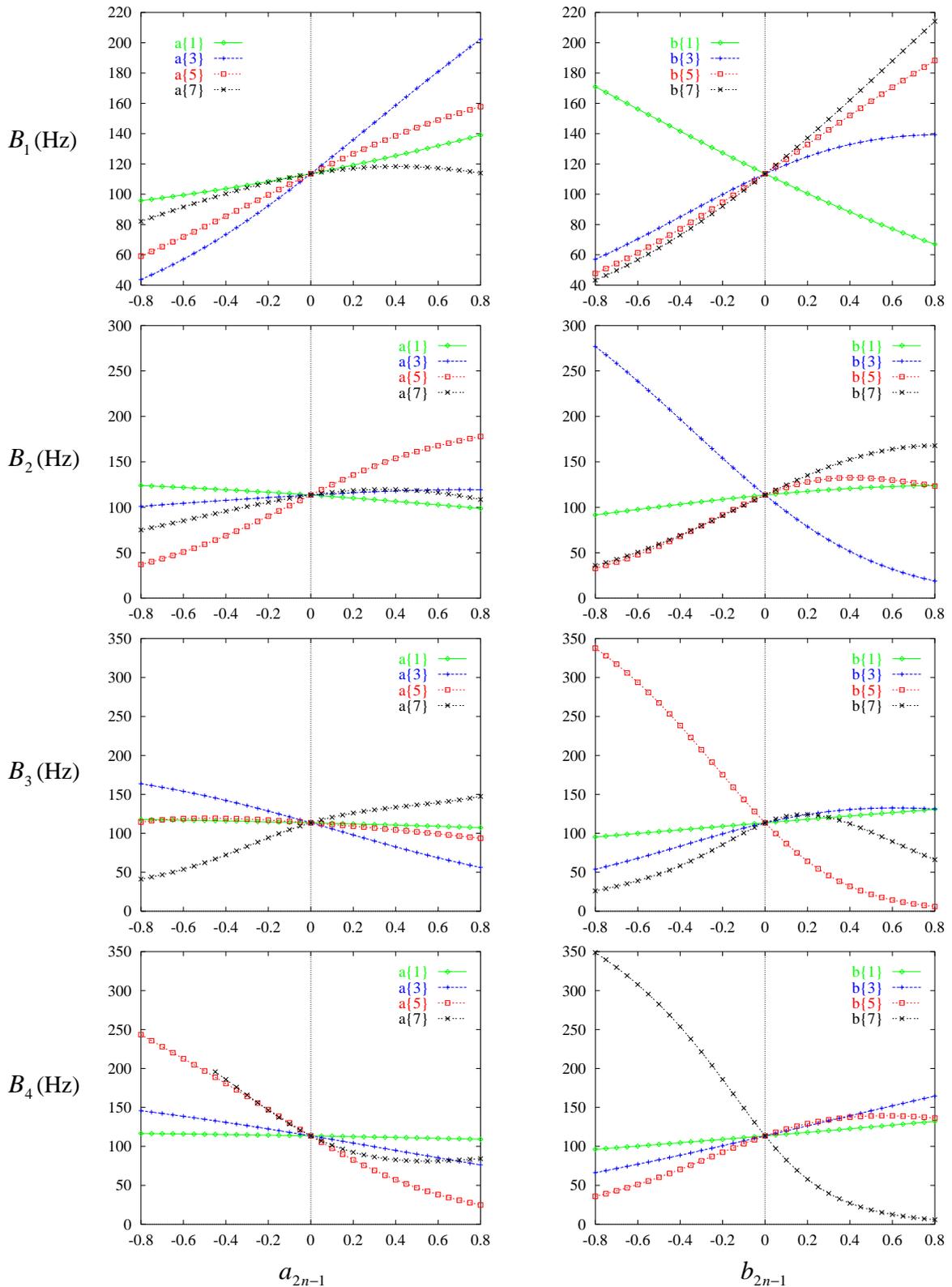


Figure 5.5(b): *Formant-bandwidth nomograms* generated by perturbing a uniform, 8-section LP area-function, according to each of the first four, odd-indexed Fourier cosine ( $\Delta a_{2n-1}$ , left graphs) and sine ( $\Delta b_{2n-1}$ , right graphs) shape-components, for  $n=1$  (diamond symbols joined by green lines),  $n=2$  (plus symbols joined by blue lines),  $n=3$  (square symbols joined by red lines), and  $n=4$  (cross symbols joined by black lines).

Figure 5.5(a) shows the influence of each of the first four parameters  $a_{2n-1}$  (left graphs) and  $b_{2n-1}$  (right graphs) on each of the first four formant *frequencies*. It is quite clear that the first-order influence of each parameter  $a_{2n-1}$ , is a *negatively-sloped, quasi-linear* variation in the corresponding formant frequency  $F_n$ . By comparison, the influence of each of the remaining Fourier-cosine shape parameters on that formant frequency, as indeed of each of the Fourier-sine shape parameters shown in the right-hand graphs, is relatively minor. The largest of these second-order effects appear in the  $F_3$  nomograms for perturbations in  $a_7$  (left graph) and in  $b_7$  (right graph), which describe the antisymmetric and the symmetric shape-component, respectively, of highest spatial resolution for the 8-section LP model. An interesting phenomenon is observed in the bottom-left graph of Figure 5.5(a), where there appears a discontinuity in the nomogram for  $F_4$ , as  $a_7$  is reduced below about  $-0.5$ . This is merely an artefact of the 8-section vocal-tract model, which prohibits measurement of formants higher than the effective upper spectral limit (or half-sampling frequency), equal to 4kHz for the given vocal-tract length. Notwithstanding such analysis artefacts, the nomograms in Figure 5.5(a) further support our earlier evidence concerning the distinctive influence of antisymmetric vocal-tract shape perturbations on the formant frequencies. Furthermore, they confirm that *symmetric* shape perturbations caused by variations in the odd-indexed Fourier *sine* components (which we earlier associated with the formant bandwidths), indeed have relatively little effect on LP-synthesised formant frequencies.

The analogous set of formant *bandwidth* nomograms for the 8-section LP model is shown in Figure 5.5(b), where the graphs on the left and on the right again pertain to perturbations in the parameters  $a_{2n-1}$  and  $b_{2n-1}$ , respectively. These nomograms do support our earlier hypothesis of a distinctive relation between each odd-indexed Fourier sine parameter  $b_{2n-1}$  and the corresponding bandwidth  $B_n$ . In particular, they show that the relation is *negatively-sloped* and *quasi-linear*. However, a comparison of Figure 5.5(b) with Figure 5.5(a) reveals that the so-called second-order influences on the bandwidths are relatively greater than those on the formant frequencies. For example, the variations in  $B_1$  caused by perturbations in  $a_3$  appear to be even larger in magnitude than the so-called first-order effect of  $b_1$  on that bandwidth. Nevertheless, similarly to the first-order formant-frequency nomograms shown on the left in Figure

5.5(a), the first-order nomograms shown on the right in Figure 5.5(b) are each characterised by a *more strongly negative slope* than any of the other curves for the same bandwidth.

Our earlier, theoretical motivations (in Section 5.3.2.1), together with the empirical evidence just presented, strongly suggest that the odd-indexed Fourier sine components of LP-derived vocal-tract shapes are distinctively related to the formant bandwidths, whose very existence secures uniqueness in the LP-based method of inversion. Our results thereby elucidate the uniqueness property of LP-derived area-functions, for the first time in terms of their underlying shape components. Together with the earlier results presented in Section 5.3.1, we have indeed identified the intrinsic parameters of unique LP area-functions.

## **5.4 Method of Area-Function Parameterisation and Estimation**

Our results presented in the previous section, were concerned with the formant dependence of LP-derived vocal-tract shapes. In particular, in Section 5.3.1 we confirmed the validity of the SM model within the LP modelling framework, and in Section 5.3.2 we then provided a theoretical motivation and empirical justifications regarding the components of LP-derived area-functions which depend on the formant bandwidths. In order to fulfill our requirements as set out in Section 5.1, in the next section we unite the newly-identified parameters of unique, LP-derived area-functions, and thereby propose a hybrid, LP-SM method of inversion. In Section 5.4.2 we then use the inversion method to evaluate our proposed parameterisation of vocal-tract shapes.

### **5.4.1 Description**

From our theoretical and empirical results presented in Section 5.3, follows our proposed extension of the SM model (cf. Equation 5.1) to include two important sets of acoustic parameters. In particular, our results concerning the formant *frequency*-dependence of LP-derived area-functions suggest that the *asymmetric* shape components  $a_{2n-1}$  of the original SM model be indeed retained; and our results concerning the *bandwidth*-dependence of LP-derived area-functions suggest that the

*symmetric* shape components, originally represented (Schroeder, 1967; Mermelstein, 1967) by the even-indexed Fourier *cosine* coefficients  $a_{2n}$ , be replaced with the odd-indexed Fourier *sine* coefficients  $b_{2n-1}$ . Our proposed method of area-function parameterisation is therefore given by the following expression:

$$\ln A(x) = \ln A_0 + \sum_{n=1}^{M/2} a_{2n-1} \cos\left(\frac{(2n-1)\pi x}{L}\right) + \sum_{n=1}^{M/2} b_{2n-1} \sin\left(\frac{(2n-1)\pi x}{L}\right), \quad (5.18)$$

where  $M$  is assumed to be even, without loss of generality.

Perhaps the most distinctive feature of this new parameterisation, is that it provides a mathematically *complete* and *minimal* description of vocal-tract shapes (up to the desired degree of smoothness  $M$ ). Indeed, the sine and cosine terms in Equation 5.18 form a mutually *orthogonal* set of basis functions which describe the vocal-tract shape in terms of the acoustically-relevant, spatial components. Furthermore, the completeness afforded by this set of basis functions does confirm both the *necessity* and the *sufficiency* of the formant frequencies and bandwidths.

As the principle of orthogonality precludes compensatory relations amongst our chosen parameters, the uniqueness property which is inherent to the LP vocal-tract model is thereby retained. However, whilst it is easily shown that all of the trigonometric terms in Equation 5.18 are mutually orthogonal, it is important to note that the symmetric components do have a non-zero mean, spatial level, and that the  $b_{2n-1}$  coefficients are therefore not strictly orthogonal to the area scaling factor  $A_0$ . This implies that in parameterising a given area-function according to Equation 5.18, the sine coefficients can only be uniquely determined if a consistent method is adopted to first determine the area scaling factor.

Towards this end, our proposed method takes advantage of the boundary conditions of each of the odd-indexed Fourier sine terms which, by definition, have zero value both at  $x = 0$  and at  $x = L$ . In particular, we first determine the *antisymmetric* components of the given, logarithmic area-function (by odd-indexed, discrete cosine transformation), and subtract them from the original vocal-tract shape in order to yield a primarily *symmetric* area function. (If the antisymmetric shape components could be determined *completely* by identifying an *infinite* number of odd-indexed Fourier cosine

coefficients, then the area function obtained after subtraction would be *purely* symmetric). The area scaling factor is then determined such that  $\ln A_0$  equals the mean of the logarithmic areas at the glottal and at the lip ends. Having thus defined  $A_0$ , the symmetric shape parameters  $b_{2n-1}$  are found uniquely by odd-indexed, discrete sine transformation. This simple and consistent method also ensures that the first  $M/2$  of the  $a_{2n-1}$  and the  $b_{2n-1}$  parameters found for a given value of  $M$ , are exactly replicated when a higher value of  $M$  is used to parameterise the same area-function.

The block diagram of Figure 5.6 shows how our proposed method of area-function parameterisation is used within the LP-based inversion method. Central to the method is the well-known set of recursive routines to convert, in turn, from formant frequencies and bandwidths (and the effective LP sampling frequency, as determined by the number of formants  $M/2$  and the assumed vocal-tract length  $L$ ), first to LP autoregressive coefficients, then to LP reflection coefficients, and finally to a step-wise, LP area-function. The raw,  $M$ -section LP area-function thus obtained for a particular value of the vocal-tract length, is then parameterised according to Equation 5.18 (as described in the preceding paragraph) in order to determine the shape parameters  $\{a_{2n-1}, b_{2n-1}\} n = 1, \dots, M/2$ . This entire procedure is repeated for a range of vocal-tract lengths, and that length is finally chosen which minimises the eccentricity of the vocal-tract shape (the so-called minimal articulatory distance, or MAD) with respect to a uniform area-function. Analogously to Paige and Zue's (1970) derivation which was based on the original SM model, it can easily be shown (e.g., using Parseval's theorem) that the MAD criterion to be minimised when using our parameterisation method (Equation 5.18), is given by the following expression:

$$MAD^2 = \frac{1}{L} \int_0^L [\ln A(x)]^2 dx = \frac{1}{2} \sum_{n=1}^{M/2} [a_{2n-1}^2 + b_{2n-1}^2] \quad (5.19)$$

Consistently with the SM model, logarithmic areas are used in the articulatory distance measure in order to place greater emphasis on differences in the acoustically-salient locations of constriction along the vocal tract.

Although our method of area-function estimation illustrated in Figure 5.6 is similar to that proposed by Wakita (1977), who was indeed the first to combine the LP-based

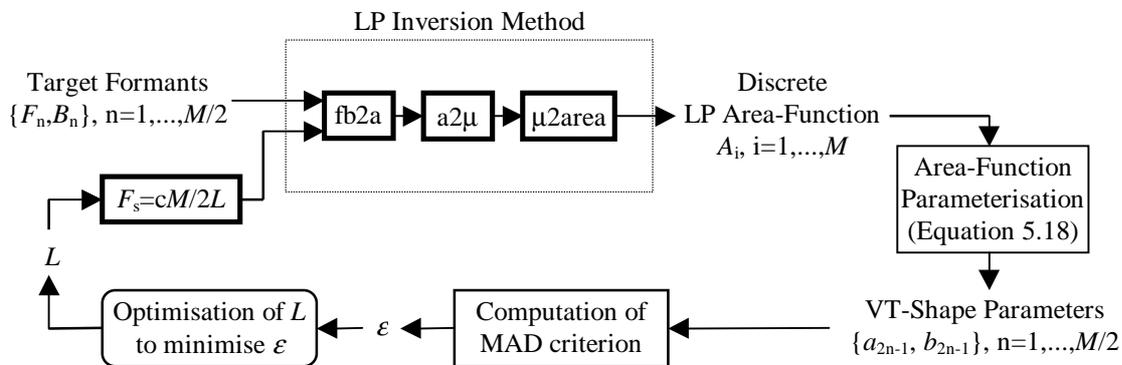


Figure 5.6: Block-diagram of proposed area-function estimation method, valid for  $L < L_{\max}$ .

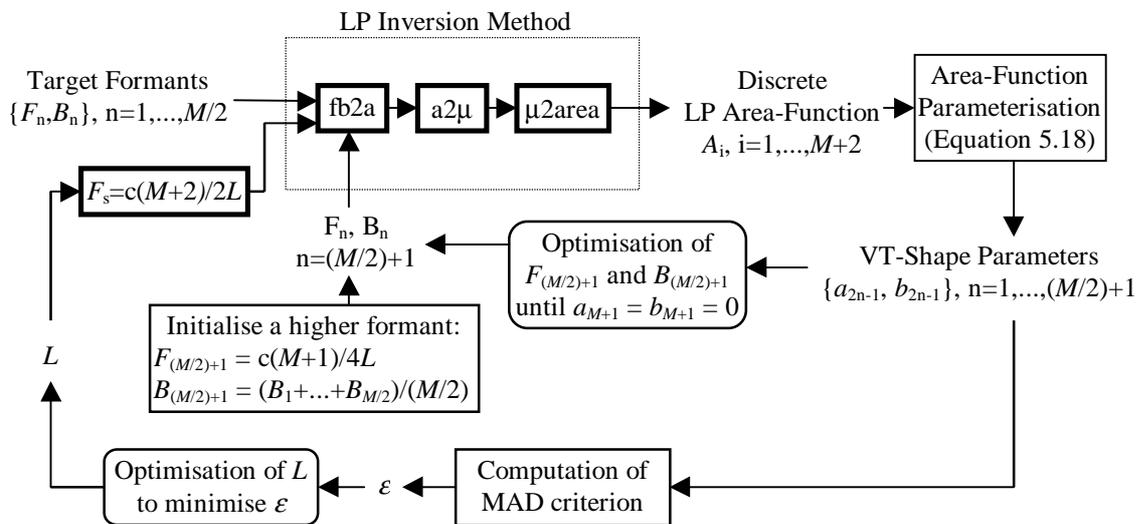


Figure 5.7: Block-diagram of proposed area-function estimation method, to be used when an optimum vocal-tract length  $L < L_{\max}$  could not be found using the method of Figure 5.6.

inversion method with the vocal-tract length optimisation criterion of Paige and Zue (1970), our proposed method does extend the work of these authors. First, the restrictively coarse, step-wise representation of the vocal-tract shape yielded by the LP model, is replaced by a smooth outline from which crucial articulatory landmarks such as the place of constriction can be determined more accurately. Note that the degree of smoothness depends on the number of shape parameters used (cf. Equation 5.18), and ultimately, on the number of formants used in the inversion. We shall return to this important aspect of the proposed model in Section 5.4.2.2, where we evaluate the representational fidelity of the vocal-tract shape parameters, as a function of the degree of smoothness  $M$ .

As a result of the smooth representation, the computation of the MAD criterion used to optimise the vocal-tract length is no longer restricted to the entire interval  $[0, L]$ . Indeed, if so desired, the integral in Equation 5.19 can be numerically computed across any selected interval  $[x_1, x_2]$  along the length of the vocal-tract, and different, physiologically-motivated sections of the area-function can thus be weighted differently. In that context, Yehia and Itakura (1994) determined from area-functions measured on mid-sagittal X-ray images of a single speaker of Japanese, that the two regions near the mid-length of the vocal-tract and close to the glottis, respectively, are less likely than other regions to exhibit large phonetic variations, and should therefore be given extra weight in the computation of MAD with respect to that speaker's average area-function.

Our proposed method of inversion also overcomes the limitation of an upper bound  $L_{\max}$  on allowed vocal-tract lengths which, as discussed in Section 5.2.2, is inherent to all vocal-tract models which have only a limited number of sections. If the MAD criterion is not found to have a minimum for a vocal-tract length less than  $L_{\max}$  (which itself depends on the value of the highest formant frequency; cf. Equation 5.6), Wakita (1977) offers no alternative but to accept this upper limit as the final solution. However, it is interesting to note that Zue (1969) had already addressed this issue in his own inversion method, which was based on a completely lossless acoustic-tube model, and which therefore made use of the zeros (formants) and poles of the lip impedance function. Indeed, Zue (1969) found that the frequency of the highest singularity often

exceeded the upper spectral limit (i.e., the effective half-sampling frequency), which itself is determined by the number of sections and by the length of the vocal-tract. In order to overcome this limitation while retaining a fixed number of vocal-tract sections, he augmented the low-order poles and zeros with a number of higher-order singularities whose frequencies were chosen to be those of a uniform tube of the same length. This solution to the problem is particularly attractive, since it is founded on the theoretical result that the higher-order singularities of a non-uniform area-function converge to those of a neutral tube of the same length. It can also be shown (simply by substituting the well-known quarter-wavelength formula for the highest given formant frequency, into Equation 5.6) that if the highest singularity is equal to its neutral value for a given vocal-tract length, then that length is always less than the upper limit  $L_{\max}$ , and the problem is thus resolved.

Our own solution to this problem in the context of the LP-based inversion method of Figure 5.6, is indeed founded on the principle of adding a higher formant to the given list of target formants. However, we transcend the simple assumption of a *neutral* higher formant, by invoking the very property of our resonance-based parameterisation which predicts that the additional, higher formant will primarily influence the vocal-tract shape components of correspondingly higher resolution. Whilst a neutral value for the higher formant only *aspires* towards affecting the resulting vocal-tract shape as little as possible, our method of area-function parameterisation can be used to explicitly *enforce* this outcome.

Figure 5.7 shows the block-diagram of our method of area-function estimation which is used in case the earlier method outlined in Figure 5.6 fails to minimise the MAD criterion for a vocal-tract length less than  $L_{\max}$ . At each value of  $L$  higher than that upper limit, a single, higher formant is introduced, with a frequency which is initially set to its neutral value for the given  $L$ , and with a bandwidth which is initially set to the mean of the target bandwidths (following our results of Section 5.3.2.1). As a result of the additional formant, the effective sampling frequency is increased by  $c / L$  Hz; the number of vocal-tract sections is increased by 2; and the discrete, LP area-function is then parameterised with two more shape parameters ( $a_{M+1}$  and  $b_{M+1}$ ) than had previously been used. However, in order to maintain consistency with the vocal-

tract shapes obtained at the original (lower) shape-resolution, we insist that the higher formant must not itself contribute to additional, higher-resolution components of the resulting vocal-tract shape. In particular, as shown by the inner loop in Figure 5.7, we iteratively optimise both the frequency and bandwidth of the additional, higher formant (using steepest-descent optimisation), such that the resulting, parameterised LP area-function retains its *original* degree of smoothness, i.e.,  $F_{(M/2)+1}$  and  $B_{(M/2)+1}$  are jointly optimised until  $a_{M+1} = 0$  and  $b_{M+1} = 0$ , to within a specified tolerance — only then do we proceed to compute the MAD criterion for the area-function obtained at that particular vocal-tract length. As indicated by the outer loop in Figure 5.7, this entire procedure is repeated in search of an optimum vocal-tract length which minimises the MAD criterion, and which then yields the final, parameterised area-function.

## 5.4.2 Evaluation of Proposed Area-Function Parameterisation

Whilst our method of area-function parameterisation (Equation 5.18) follows naturally from our earlier derivation of the parameters of unique LP-derived area-functions, it still warrants an evaluation of its effectiveness. We shall therefore evaluate our proposed parameterisation, first (in Section 5.4.2.1) by using our hybrid method of inversion to obtain the *correlations* between synthetic formant and estimated shape parameters; and then (in Section 5.4.2.2) by testing the ability of our shape parameters to capture the spatial characteristics of directly-measured area-functions obtained from the literature.

### 5.4.2.1 Inter-parameter Correlations

We herein aim to evaluate the acoustic-phonetic relevance of our proposed set of vocal-tract shape parameters, in the context of the inversion method described in the previous section. In particular, we would like to confirm that the statistical correlation between each formant parameter and the corresponding shape parameter, is indeed as significant as suggested in our results of Section 5.3. This hypothesis is tested by first synthesising formant values from a large number of different area-functions, then applying our hybrid LP-SM method of inversion to re-estimate the shape parameters and evaluate their correlation with the synthetic formants.

First, area-functions are generated by all combinations of ternary perturbations ( $-0.1$ ,  $0.0$ , and  $+0.1$ ) of the first four cosine parameters ( $a_{2n-1}$ ,  $n = 1, \dots, 4$ ) and of the first four sine parameters ( $b_{2n-1}$ ,  $n = 1, \dots, 4$ ), thus yielding a total of 6561 distinct shapes, with a fixed vocal-tract length  $L = 17.65$  cm and a normalised area scaling factor  $A_0 = 1.0$ . Assuming a nominal value for the glottal reflection coefficient ( $\mu_g = 0.7$ , for which the mean of the first four bandwidths is predicted by Equation 5.17 to be  $\bar{B} \approx 113.5$  Hz), the formant frequencies and bandwidths of all 6561 perturbed vocal-tract shapes are LP-synthesised. Those synthetic formants are then used to re-estimate the area-functions, using our hybrid LP-SM method of inversion.

Consistently with the SM model (and similarly to our earlier investigation in Section 5.3.1.2), correlations between vocal-tract shape and resonance parameters are computed using the *relative* formant frequencies  $F_n^{(\text{rel})} = (F_n - F^{(\text{neut})}) / F^{(\text{neut})}$ ; consistently with our results of Section 5.3.2 where we found that the *shape* of the LP area-function determines only the formant bandwidth *pattern* about the mean value, correlations are computed using the *relative* bandwidths  $B_n^{(\text{rel})} = (B_n - \bar{B}) / \bar{B}$ . The first four, relative formant frequencies and bandwidths are then paired with each of the eight vocal-tract shape parameters in turn, and the coefficient of correlation (or *Pearson's r*) is computed across all 6561 configurations.

Indeed, the high correlations along the main diagonal of Table 5.2 do provide support for a strongly linear relation between each relative acoustic resonance and the corresponding vocal-tract shape parameter. The weakest of these correlations ( $-0.917$ ) summarises the strength of the linear relation between the relative second formant bandwidth and the corresponding shape parameter  $b_3$ , while the strongest correlation ( $-0.996$ ) holds between the relative  $F_4$  and the corresponding shape parameter  $a_7$ . By comparison, the most significant, off-diagonal correlation ( $+0.550$ ) is that which describes the strength of the relation between the relative third formant bandwidth and the highest-indexed Fourier sine coefficient  $b_7$ . As already foreshadowed in our nomograms of Figure 5.5, the magnitude of the off-diagonal elements in Table 5.2 suggest that the influence of certain asymmetric shape parameters  $a_{2n-1}$  on the formant bandwidths, is not entirely insignificant. However, as we have already discussed, the first-order effect retains its distinctiveness owing to the *negative sign* of induced

	$a_1$	$a_3$	$a_5$	$a_7$	$b_1$	$b_3$	$b_5$	$b_7$
$F_1^{(rel)}$	<b>-0.995</b>	0.143	0.170	0.213	0.245	0.092	-0.005	-0.175
$F_2^{(rel)}$	0.150	<b>-0.990</b>	0.173	0.225	0.249	0.119	-0.012	-0.189
$F_3^{(rel)}$	0.174	0.178	<b>-0.983</b>	0.241	0.274	0.129	0.004	-0.224
$F_4^{(rel)}$	0.216	0.229	0.249	<b>-0.996</b>	0.334	0.154	0.008	-0.274
$B_1^{(rel)}$	0.124	0.528	0.333	0.113	<b>-0.922</b>	-0.165	0.174	0.377
$B_2^{(rel)}$	-0.066	0.041	0.461	0.135	-0.199	<b>-0.917</b>	0.299	0.464
$B_3^{(rel)}$	-0.020	-0.243	-0.062	0.263	0.236	0.327	<b>-0.925</b>	0.550
$B_4^{(rel)}$	-0.007	-0.114	-0.426	-0.357	0.432	0.421	0.449	<b>-0.938</b>

Table 5.2: Linear correlation coefficients (Pearson’s  $r$ ) between each of the first four, relative formant frequencies and bandwidths LP-synthesised from 6561 perturbations of an 8-section, uniform area-function (with the glottal reflection coefficient fixed at a nominal value  $\mu_8 = 0.7$ ), and each of the first eight VT-shape parameters re-estimated from those formant data using the hybrid LP-SM method of inversion described in Section 5.4.1.

perturbations in the acoustic parameters, which is here reflected in the strongly negative correlations along the main diagonal of Table 5.2.

These correlations do lend strong support to our proposed method of vocal-tract shape parameterisation. Indeed, our proposed modification and extension of the SM model is here justified quantitatively, by the strength of the one-to-one mapping between the odd-indexed Fourier sine parameters and the LP formant bandwidths. It is also interesting to note that the strong correlations obtained between the first-order pairs of formant and shape parameters, do indirectly underscore the appropriateness of the MAD criterion which was used in the inversion.

#### 5.4.2.2 Representation of Directly Measured Area Functions

Our evaluation thus far has established the quasi-linear, one-to-one mapping between vocal-tract shape parameters and the LP-based formant frequencies and bandwidths. Equally relevant to this, acoustically-motivated evaluation of the parameterisation, is to consider its accuracy in merely *representing* area-functions obtained from direct measurements of the human vocal-tract. In that context, it is important to first evaluate the inherent smoothing of directly-measured area-functions, and then to assess the relative contribution of each parameter in representing those vocal-tract shapes.

The vocal-tract shapes which we shall use for this purpose, are taken from published works in which the authors have tabulated the area-functions measured using either X-ray imaging or magnetic resonance imaging (MRI) techniques. Whilst such directly-measured data are themselves prone to various, non-negligible sources of error (e.g., the necessary assumptions on the lateral width of vocal-tract sections when mapping mid-sagittal cross-dimensions to cross-sectional areas), it is generally accepted that they do provide, simply by virtue of their having been acquired by direct observation, the most accurate information that one might currently obtain on the entire shape of the human vocal-tract during vowel production.

The six, X-ray measured area-functions of Russian vowels given by Fant (1960), are perhaps the most celebrated and widely used data on vocalic area-functions, and for many years remained the only, relatively complete and easily-accessible of such data. More recent, MRI methods apparently have none of the potential hazards to which subjects are exposed during X-radiography, and in addition they can be used to reconstruct a volumetric profile of the vocal-tract airway. Nevertheless, they still suffer from a number of disadvantages, which include the requirement of repeatedly sustained articulations due to the long image-acquisition time, and the poor detection of bone structures such as the teeth, due to their low concentrations of hydrogen. Notwithstanding these limitations, a number of recent studies have used MRI to obtain invaluable area-function measurements which they have also generously published. Baer et al. (1991) provide area-functions of the four point vowels for one adult, male speaker of American English and one adult, male speaker of British English; Yang and Kasuya (1994) list area-functions of five vowels produced by two adult speakers (one male and one female) and one child, male speaker of Japanese; and Story et al. (1996) tabulate the area functions of 11 sustained vowels of a single, male speaker of American English. In addition, Beautemps et al. (1995) list the area-functions of the three cardinal vowels which were initially obtained from mid-sagittal X-ray images of a single, male speaker of French, and which were then refined by an optimisation procedure (applied to the mapping from cross-dimensions to areas) in order to minimise the error between the measured and resynthesised formants. In total, these five studies offer 33 directly-measured, vocalic area-functions produced by six adult, male speakers of four different

languages, including two dialects of English.

As the three area-functions of French vowels given by Beautemps et al. (1995) are originally listed with unequal section lengths, we first re-sample them by linear interpolation of the logarithmic areas in order to obtain equal-length sections not longer than the smallest original section-length (approximately 0.3 cm). Similarly, Yang and Kasuya's (1994) five area-functions of Japanese vowels are re-sampled at equal-length intervals no greater than 0.5 cm, owing to their treatment of the area at the lip opening as a separate section of unspecified length, which we here assume to have a nominal value of 0.01 cm. Each of the 33 directly-measured (and possibly re-sampled) area-functions is then re-scaled (or normalised) to a mean logarithmic value of zero, consistently with the acoustical insignificance of the area-scaling factor in the LP vocal-tract model.

According to the well-known Nyquist sampling theorem, the number of independent parameters used to model a given area-function must not exceed the total number of vocal-tract sections. Each of the 33, re-scaled area-functions is therefore parameterised (as described in Section 5.4.1) with  $M$  equal to the highest, even integer no greater than the number of vocal-tract sections. The area-functions with the fewest number of sections (18) are those given by Baer et al. (1991); consequently, each of the 33, re-scaled and parameterised area-functions is re-expanded according to Equation 5.18, with  $M/2$  ranging from 0 to 9, the former yielding a uniform area-function with only an area-scaling factor, and the latter yielding a smoothed vocal-tract shape with the first 18 Fourier-series terms. A quantitative measure of the goodness of fit at each degree of smoothness (integer-value of  $M/2$ ), is obtained by computing the root-mean-square (rms) error between each logarithmic area-function and its smoothed version, computed at the centre of each original, equal-length section.

The diamond symbols (joined by solid lines) in Figure 5.8 show the average rms error in smoothing the 33 directly-measured area functions, at each value of  $M/2$ . As expected, the rms error is a monotonically decreasing function of the number ( $M$ ) of shape parameters used to represent each area-function. Our results clearly show that the first two, sine and cosine terms are the most important spatial components of directly-measured, vocalic area-functions, and that only small, incremental improvements are

gained by including terms at  $(M/2) = 3$  and higher. This observation agrees with the conclusions drawn by Liljencrants (1971), who found it necessary to use only the first two (even-indexed) Fourier sine and cosine terms to represent X-ray measured, mid-sagittal tongue-contours mapped onto a Cartesian coordinate system.

As the degree of spatial resolution  $M/2$  is identical to the number of formants used in the inversion, the diamond symbols in Figure 5.8 provide an optimistic benchmark for the rms error to be expected when using the inversion method. For example, if we assume that only the first four formants are available in general, the solid curve in Figure 5.8 suggests an average, lower-bound rms error of about 0.22 when comparing estimated vocal-tract shapes with so-called exact, or directly-measured area-functions. Errors in excess of this benchmark could then be ascribed to the estimation method itself, rather than the parameterisation.

In order to assess the relative contribution of antisymmetric and symmetric components of directly-measured area functions, the rms errors were also computed with respect to the smoothed vocal-tract shapes obtained by re-expanding only the Fourier cosine terms, and only the Fourier sine terms, respectively, at each value of  $M/2$ . The averaged results shown in Figure 5.8, clearly illustrate the nearly equal contribution of antisymmetric components (plus symbols, joined by dashed lines) and symmetric components (square symbols, joined by dotted lines). These results also suggest that on average, the first two cosine components ( $a_1$  and  $a_3$ , whose contributions are shown along the dashed lines as  $M$  is increased from 0 to 1, and from 1 to 2, respectively) and the first two sine components ( $b_1$  and  $b_3$ , whose contributions are shown along the dotted lines as  $M$  is increased from 0 to 1, and from 1 to 2, respectively) are the most important spatial components of the 33 area-functions.

The results shown in Figure 5.8 and discussed above, were concerned with the rms errors averaged over all of the 33 area-functions. However, as foreshadowed by Schroeder (1967) and Mermelstein (1967), individual vocal-tract shapes of certain vowels can be strongly associated with either symmetric or antisymmetric components. For example, Schroeder and Mermelstein found that the area-function of the Russian vowel /u/ given by Fant (1960), is very poorly represented by antisymmetric shape components alone, and can not therefore be adequately estimated by the SM model,

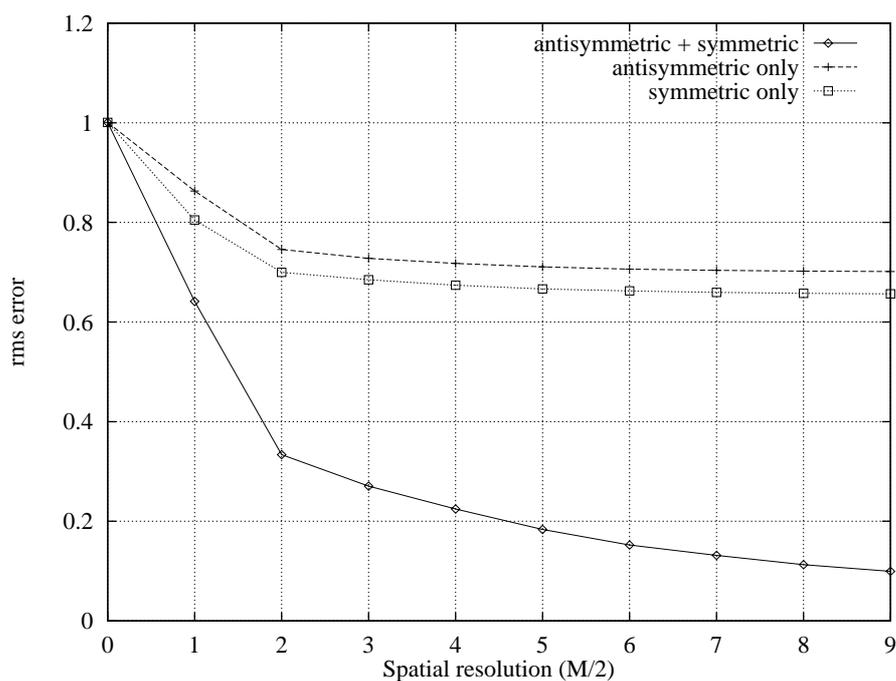


Figure 5.8: Average root-mean-square (rms) error in *representing* 33 directly-measured area-functions obtained from the literature (see text). Each area-function is first parameterised according to Equation 5.18, with the total number of shape-parameters ( $a_{2n-1}$  and  $b_{2n-1}$ ) equal to the highest even integer not greater than the number of equal-length vocal-tract sections. Representations at increasingly higher degrees of spatial resolution ( $M/2$  in Equation 5.18) are then obtained by truncating the parameters at  $M/2=\{0,1,2,\dots,9\}$ , and the rms error of fit is computed between the smoothed and the original, logarithmic area-functions. *Diamond symbols*: average rms error in representation of each area-function using both antisymmetric ( $a_{2n-1}$ ) and symmetric ( $b_{2n-1}$ ) shape components. *Plus symbols*: average rms error in representation using only antisymmetric ( $a_{2n-1}$ ) shape components. *Square symbols*: average rms error in representation using only symmetric ( $b_{2n-1}$ ) shape components.

using only the formant frequencies. Their findings are confirmed by our own results (shown in Appendix E) of the individual rms profiles for each area-function, which indicate that mid- to high-, back vowels (e.g., Fant’s (1960) Russian /u/; Baer et al.’s (1991) British English /u/ and American English /u/; Yang and Kasuya’s (1994) Japanese /u/ and /o/; Beautemps et al.’s (1995) French /u/; and Story et al.’s (1996) American English /o/, /ʊ/, and /u/) most consistently exhibit vocal-tract shape *symmetry*, owing presumably to the location of the linguo-velar constriction at approximately midway along the length of the vocal-tract.

By contrast, the two point vowels /i/ and /a/, which are usually associated with extremes of lingual articulation in the high-front and in the low-back parts of the vocal tract, are the most consistently *antisymmetric* of the directly-measured area-functions

(except for Beautemps et al.'s (1995) French /a/, which admittedly has a more symmetric shape, owing to the much smaller and decreasing areas towards the lip end). From our quantitative evidence may be inferred the physiological point of view that asymmetry in vocal-tract shapes implies conservation of the mass of the tongue body — for example, when the tongue moves forwards and upwards to form a constriction for a high-front vowel, the pharyngeal cavity is endowed with a proportionately larger volume; and when it moves backwards and downwards to form a constriction for a low-back vowel, larger areas are naturally attained in the front, oral cavity.

Our results can also be interpreted from an acoustical point of view, as a consequence of the close relationships which we have shown to exist between antisymmetric shape components and the formant frequencies, and between symmetric shape components and the formant bandwidths. Indeed, the relative dominance of the first two Fourier-series shape components, does confirm the relative importance of the first two formants in vowel production. The significant contribution of the two *antisymmetric* shape parameters  $a_1$  and  $a_3$  in particular, provides indirect proof (gleaned here only from results of shape parameterisation) that  $F_1$  and  $F_2$  are indeed the most important acoustic determinants of the phonetic identity of vowels. At the same time, however, the significance of the *symmetric* shape parameters in representing directly-measured area-functions, is consistent with their importance in securing uniqueness in LP-derived vocal-tract shapes.

## 5.5 Evaluation of Estimation Method

Evaluation of any proposed method of inversion has always been a major weakness in studies concerned with acoustic-to-articulatory mapping. Whilst confidence is usually raised if acoustic and articulatory data are simultaneously measured, existing sources of error in directly measured articulatory data (some of which were discussed earlier, in Section 5.4.2.2) do suggest that they should not be given *carte blanche* as the ultimate reference. On the other hand, there seems to have been some over-reliance on indirect evaluations, such as perceptual or spectrographic fidelity of resynthesised speech — it has rightly been asserted (e.g., Sondhi, 1979) that, owing mainly to the problem of non-uniqueness, resynthesis alone is not sufficient for evaluating estimated articulatory data.

In this vein, perceptual judgements are probably even more inconclusive, owing to the suggested “many-to-one relationship between articulatory/acoustic and perceptual targets” (Gay et al., 1991, p.445) for vowel sounds.

These limitations notwithstanding, directly-measured area functions are admittedly the best reference that we have regarding data on real vocal-tract shapes. However, owing to technical obstacles, simultaneously-measured acoustic data are rare, and even when they are documented, the formant bandwidths are invariably missing. It would therefore seem expedient to evaluate our method of inversion first using acoustic data obtained by synthesis which, although admittedly biased towards a more favourable assessment, does afford a systematic definition of the various sources of error. To this end, in Section 5.5.2 we will take as a point of departure the 33 published, directly-measured vocal-tract area-functions which we used earlier (in Section 5.4.2.2), and impress certain assumptions which enable us to evaluate our hybrid LP-SM method of inversion first under *model-matched* (i.e., LP-matched), then under presumably *more realistic* conditions.

In Section 5.5.3 we then proceed to estimate vocal-tract shapes from the measured formant data of our four adult, male speakers of Australian English. In particular, we will first use our measurements of the first four formant frequencies and bandwidths, to quantify the formant-induced variability of LP-estimated area-functions, and thus arrive at a prescription for dealing with the large amounts of variability normally encountered in measured bandwidths. We will then consider the well-known, but ill-treated problem of the incompatibility of measured formant bandwidths with those ideally required by the LP vocal-tract model.

However, a necessary requirement in the forthcoming evaluation is to be able to quantify the variability in estimated vocal-tract shapes. As area-functions are generally of different lengths, and as variations in the horizontal position of the lips and in the vertical position of the larynx preclude them from being considered as fixed articulatory landmarks, the problem of area-function alignment is highly non-trivial; nor has it been addressed to a sufficient degree in the literature. In Section 5.5.1 we therefore propose a method of aligning parameterised area-functions, which we shall then use both in our subsequent evaluation of our estimation method, and also later in our articulatory

investigation of the speech-speaker dichotomy in Chapter 6.

### **5.5.1 Inter-repetition Alignment of Area-Functions**

As already discussed in our review of the literature (in Chapter 2), the problem of directly comparing two or more vocal-tract shapes and quantifying their differences, has rarely been addressed in previous studies. A major obstacle which has impeded progress in that regard, is presumably the absence of an accepted method of aligning vocal-tract area-functions (whether directly measured or estimated from acoustics) such as to render physically meaningful comparisons. The problem is exacerbated by the well-known tendency for the position of the lips and of the larynx to vary from one articulatory configuration to another, thus prohibiting the use of either end of area-functions as an invariant positional reference. As the overall lengths of different area-functions are also different in general, one is faced with the non-trivial problem of how best to align area-functions prior to quantifying their shape-related differences.

One might suppose that if certain, fixed articulatory landmarks could be located along each area function, then they might be aligned by piecewise-linear expansion or contraction along the length dimension, anchored at those landmarks. Unfortunately, this attractive solution is not only unrealistic in the context of area-function estimation, but would also be quite difficult and laborious to achieve with a high degree of precision even with direct articulatory measurements. Previous approaches to this problem include Harshman et al.'s (1977) elaborate definition of a speaker-specific, mid-sagittal reference grid; Wood's (1979) alignment of area functions (measured from X-ray films) by anchoring them at the single point of reference defined by the position of the central incisors; and Högberg's (1995) inter-gender normalisation of the total lengths of X-ray measured area functions, by a piecewise-linear rescaling which relies on the definition of the boundary between the oral and the pharyngeal parts of the vocal tract. Clearly, such methods cannot be applied to *estimated* vocal-tract shapes, which lack the convenience of physiological landmarks.

However, if one requires to quantify the variability, or the dispersion in a group of area-functions which are variants of the same, basic vocal-tract shape, then one might

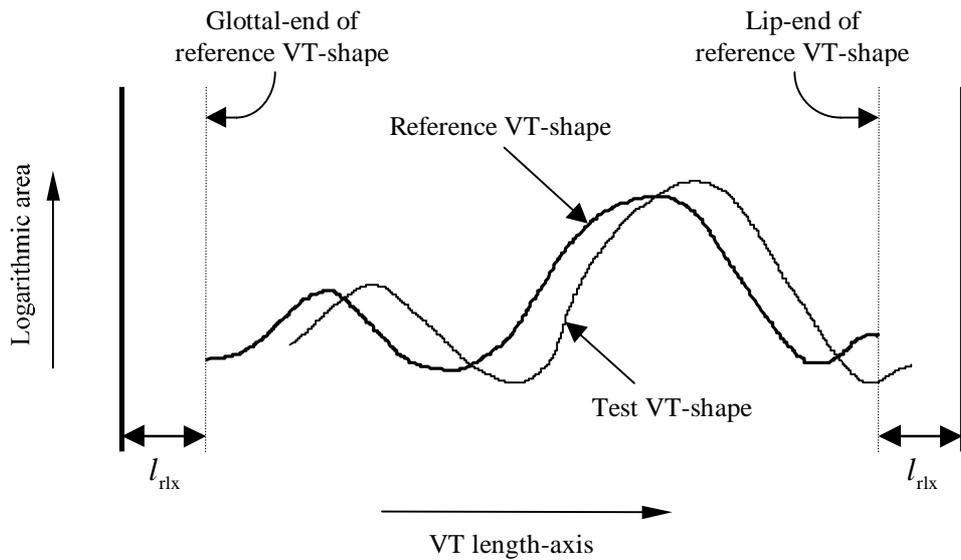


Figure 5.9: Schematised illustration of our method of aligning similarly-shaped vocal-tract area-functions. The longest area-function is selected as the so-called *reference shape*, and the *test shape* is then shifted along the length-axis in steps of  $\Delta x$ , within the limits set by the relaxation intervals of length  $l_{rlx}$  at both ends of the reference shape. At each step, the root-mean-square (rms) error between the two shapes (logarithmic area-functions) is computed within their mutually-overlapping regions, and the two area-functions are finally aligned at that position which yielded the minimum rms error.

take advantage of the fact that those shapes are assumed to share a common set of articulatory characteristics. This condition may arise, for example, when area-functions are to be compared across a number of analysis frames and repetitions of a given vowel of a single speaker. One way of overcoming the lack of a fixed frame of reference, is then to allow each area function a positional degree of freedom along the length axis, subject to the constraint of a minimal amount of variation about the typical, or *prototype* vocal-tract shape which they all are assumed to embody. The problem of aligning the given set of (presumed) phonetically equivalent area functions, thus reduces to a minimisation of the root-mean-square (rms) dispersion of the vocal-tract shapes relative to each other.

As initially proposed and demonstrated by Clermont (1991; 1993) who successfully time-aligned repetitions of formant-contours of diphthongs, the easiest approach (shown schematically in Figure 5.9) is to slide each shape (formant-contour or, in the present study, vocal-tract area-function) along the length axis, until the rms difference is minimised with respect to a so-called *reference shape*, which itself is

chosen to be the longest one of the set. As the shapes are, by definition, not too dissimilar, a global minimum in the rms error can be found by sliding each, so-called *test* area-function, only within certain limits which are defined by so-called *relaxation intervals* placed at both ends of the reference area-function.

In practice, this is achieved by first expanding the Fourier series terms in Equation 5.18 in order to obtain the smoothed versions of the parameterised area-functions themselves, sampled at small, equal-length intervals  $\Delta x$ . The area-function with the longest length is then selected as a reference, and a sufficient number of sections is allowed at either end to accommodate the relaxation intervals, each of length  $l_{rx}$ . Each of the test area-functions is then shifted within these outer limits, and at each step the rms distance is computed between the test and reference, *logarithmic* area-functions (consistently with the SM model), over that part of the overall length which is common to both (the *mutually-overlapping region*, as hereafter referred to). The global minimum in the rms error is thus found to within an accuracy of  $\pm \frac{1}{2} \Delta x$  for each area-function.

As a result of the procedure outlined above, each of the parameterised vocal-tract shapes is endowed with an extra parameter which identifies its position along the  $x$ -axis, relative to a fixed frame of reference defined by the longest area-function in that set. A so-called *prototype* area-function can then be found simply by computing the average of the (logarithmic) areas at every interval  $\Delta x$ . The total length of the prototype is determined by the mean length of the area-functions in the set; and its end-points are located equi-distantly about the mid-point of the overall mutually-overlapping (MOL) region. Finally, the prototype area-function thus found, can be parameterised according to Equation 5.18, and regarded as the basic vocal-tract shape of which the given set of area-functions were assumed to be variants.

As we shall see in Chapter 6, the procedure just outlined will prove to be very useful in reducing the multi-frame and multi-repetition area-functions, to yield a single, prototype vocal-tract shape for each vowel of each speaker. In the following sections, we shall use it primarily to quantify the variability in vocal-tract shapes estimated under various conditions, by retaining the minimum rms error yielded after each alignment.

## 5.5.2 Re-estimation of Directly Measured Area-Functions

The first part of our evaluation concerns re-estimation of vocal-tract area-functions which have already been directly measured using X-ray imaging or MRI techniques — in particular, the 33 area-functions obtained from the literature and used earlier in Section 5.4.2.2. The formant data required as input to our hybrid LP-SM method of inversion, will be synthesised first under model-matched conditions (in Section 5.5.2.1), then under more realistic conditions (in Section 5.5.2.2). Our method of inter-repetition alignment will be used to quantify the differences between the original and re-estimated area-functions, thereby assessing the detrimental influences of using model-mismatched acoustic data. We then appraise (in Section 5.5.2.3) the effectiveness of using such model-based results in correcting the (presumed) “more realistic” formant data prior to inversion.

### 5.5.2.1 Model-matched Conditions

The LP model which lies at the heart of our inversion method, makes certain unrealistic assumptions (as discussed earlier in Section 5.2.2), the two most drastic of which are the following: (i) the number of equal-length vocal-tract sections is equal to twice the number of formants considered; and (ii) the only source of losses in the vocal-tract is a frequency-independent, resistive element at the glottal end. We herein conform to the first of these assumptions, by representing each of the 33 directly-measured area-functions in terms of 8 equal-length sections, obtained by parameterisation (with  $M = 8$  in Equation 5.18) and re-expansion of the areas at the centre of each section (i.e., effectively by interpolating the smoothed area-function at equal-length intervals). Assuming a nominal value  $\mu_g = 0.8$  for the glottal reflection coefficient, the first four formant frequencies and bandwidths of each of the smoothed, 8-section area-functions are then synthesised using a vocal-tract acoustic-simulation model (see Appendix C for details) with all losses removed except for the glottal resistance. The synthetic formants thus obtained under model-matched conditions (and listed in Table 5.3), are then used to re-estimate each of the area-functions, using our hybrid LP-SM method of inversion described earlier in Section 5.4.1.

Study	Vowel	Formant Frequency (Hz)				Formant Bandwidth (Hz)			
		$F_1$	$F_2$	$F_3$	$F_4$	$B_1$	$B_2$	$B_3$	$B_4$
Fant (1960)	/i/	232.4	2271.5	3125.1	3909.7	20.2	70.0	35.4	178.1
	/e/	440.5	2025.2	2952.2	3912.1	19.6	55.9	107.3	121.2
	/a/	708.3	1232.3	2606.9	3858.7	128.7	55.8	32.8	77.7
	/o/	564.7	1028.8	2457.9	3435.1	85.2	22.9	19.5	143.5
	/u/	298.2	665.6	2362.9	3397.2	45.6	24.4	19.0	168.3
	/ɪ/	294.6	1897.3	2367.4	3310.3	34.6	19.8	86.2	123.3
Baer et al. (1991) (Speaker TB)	/i/	271.6	2144.9	2899.3	3946.8	23.4	68.6	67.2	142.3
	/æ/	641.7	1529.3	2475.2	4019.7	96.6	99.0	36.1	70.0
	/a/	524.3	1165.8	2483.4	3864.2	128.1	36.2	29.7	92.3
Baer et al. (1991) (Speaker PN)	/u/	302.4	1335.7	2288.5	3399.3	76.9	9.2	112.1	74.7
	/i/	282.9	2230.1	3157.0	3826.3	24.3	74.0	82.5	120.7
	/æ/	694.0	1656.9	3116.3	4264.0	67.5	64.7	40.3	145.8
	/a/	600.5	1337.6	3002.7	4099.9	164.9	43.3	27.6	82.5
Yang & Kasuya (1994)	/u/	318.3	1806.5	2495.8	3514.5	24.3	10.1	60.5	177.8
	/i/	282.4	2103.7	2646.6	3530.4	42.4	75.9	79.1	77.5
	/e/	541.7	1844.9	2506.8	3598.4	48.8	49.5	75.2	112.1
	/a/	710.9	1211.5	2583.3	3571.6	91.0	49.2	63.3	79.6
Beautemps et al. (1995)	/o/	578.9	1067.2	2607.5	3702.2	37.8	11.2	26.6	200.3
	/u/	415.0	1581.2	2378.9	3390.3	69.5	22.6	90.6	100.4
	/i/	232.9	2013.0	2977.6	3741.7	21.7	74.4	69.6	120.9
Story et al. (1996)	/a/	522.4	1338.7	2465.1	3578.8	35.4	75.5	42.6	124.2
	/u/	362.5	1030.4	2312.9	3589.4	34.3	10.1	166.9	56.2
	/i/	236.4	2441.2	3345.2	3836.9	8.8	61.2	60.9	169.6
	/ɪ/	501.3	2189.5	2868.7	3749.6	7.4	40.7	86.7	164.3
	/ɛ/	670.4	2196.3	2910.8	3996.8	8.6	48.7	87.2	171.2
	/æ/	761.7	1971.9	2755.5	3746.9	8.2	40.4	98.5	152.7
	/ʌ/	754.3	1379.8	2656.0	3533.5	17.4	41.2	78.3	148.2
	/ɑ/	835.7	1232.0	2811.5	3734.3	48.1	29.7	19.9	187.0
	/ɔ/	673.7	1156.1	2780.4	3417.8	13.8	46.2	74.4	152.0
	/o/	477.3	1019.7	2540.1	3647.8	17.5	7.9	92.2	169.4
Story et al. (1996)	/u/	487.2	1003.8	2675.7	3588.3	28.0	38.3	72.1	148.8
	/u/	281.0	1395.8	2421.3	3526.5	12.2	0.8	77.4	184.4
	/ɚ/	570.3	1662.8	2970.9	3569.3	19.4	78.2	98.9	90.0

Table 5.3: Formant frequencies and bandwidths of 33 directly-measured area-functions, synthesised using a transmission-line analog of the vocal-tract (see Appendix B) with a glottal resistance as the only source of loss. Each area-function is first parameterised with  $M = 8$  shape parameters, which are then used in Equation 5.18 to obtain 8 equal-length sections. The glottal reflection coefficient is fixed at a nominal value  $\mu_8 = 0.8$ .

As the LP-based inversion is theoretically reversible, the 8-section area-functions yielded by constraining the inversion to use the original vocal-tract lengths are almost identical to the original, 8-section area-functions. Indeed, the average rms error obtained after alignment (with  $\Delta x = 0.05$  cm and  $l_{rx} = 2.0$  cm), was found to be only 0.04. However, the MAD criterion can only be tested by allowing *it* to determine the optimum length of each area-function under these, so-called model-matched conditions.

In particular, the method depicted in Figure 5.6 is first used in search of a minimum in the MAD criterion, for a range of vocal-tract lengths starting from 12 cm and incrementing in steps of 0.4 cm up to the limit  $L_{max}$ , which is determined by the highest (the fourth) formant frequency (cf. Equation 5.6 in Section 5.2.2). A second search is then carried out over a much smaller range of vocal-tract lengths which brackets the identified minimum in the MAD criterion, to determine the optimum length at a finer resolution of 0.05 cm. Of the 33 area-functions considered, only two (Fant's (1960) Russian /u/ and Story et al.'s (1996) American English /u/) were found *not* to have a minimum in the MAD criterion for a length less than their respective  $L_{max}$ . The optimum lengths of these area-functions were therefore obtained using the method depicted in Figure 5.7, with an additional, fifth formant used to overcome the  $L_{max}$  limitation (as described in Section 5.4.1).

In Figure 5.10 we illustrate typical curves of MAD versus  $L$  obtained in finding an optimum vocal-tract length, shown here only for the six area-functions of Fant (1960). Our results are indeed comparable to the curves obtained by Paige and Zue (1970) for the same area-functions, even though their method of inversion was somewhat different from ours. Similarly to their results, Figure 5.10 indicates that the high-front vowel /i/ is the most eccentric (relative to a uniform tube) of the six vocal-tract shapes, even after optimisation of the vocal-tract length to minimise eccentricity; the optimised area-function of the mid-front vowel /e/ is not only the shortest, but also the least eccentric of the six vocal-tract shapes; the re-estimated area-function of the lip-rounded vowel /u/ is found to be the longest. However, contrary to the results of Paige and Zue (1970, Table III) and of Wakita (1977, Figure 6), who both slightly underestimated the length of this high-back vowel, our LP-based method clearly overestimates its length, yielding the largest, positive error (13.6 %) amongst the re-estimated lengths of the 33 area-

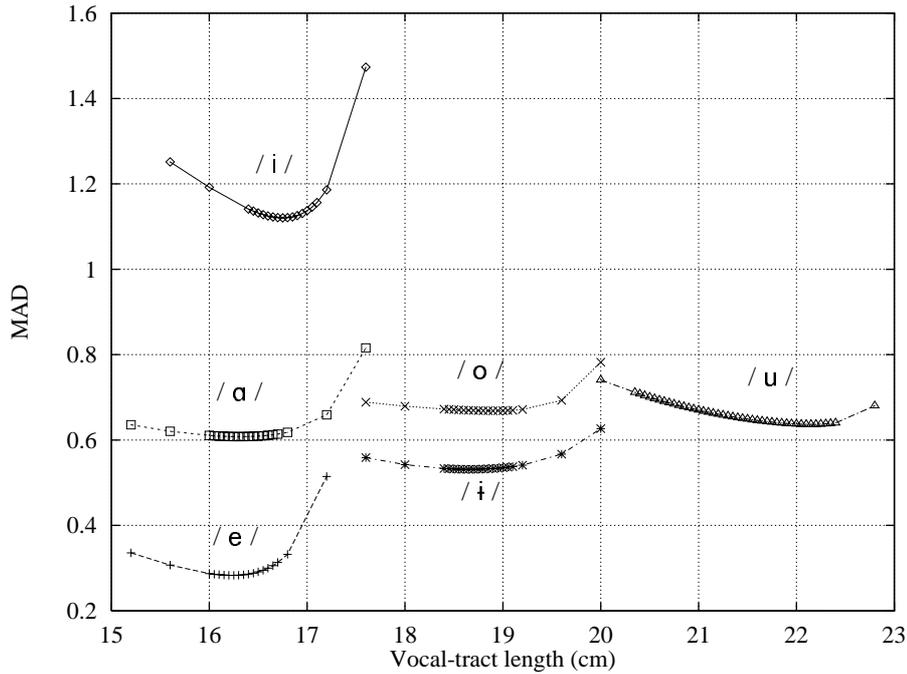
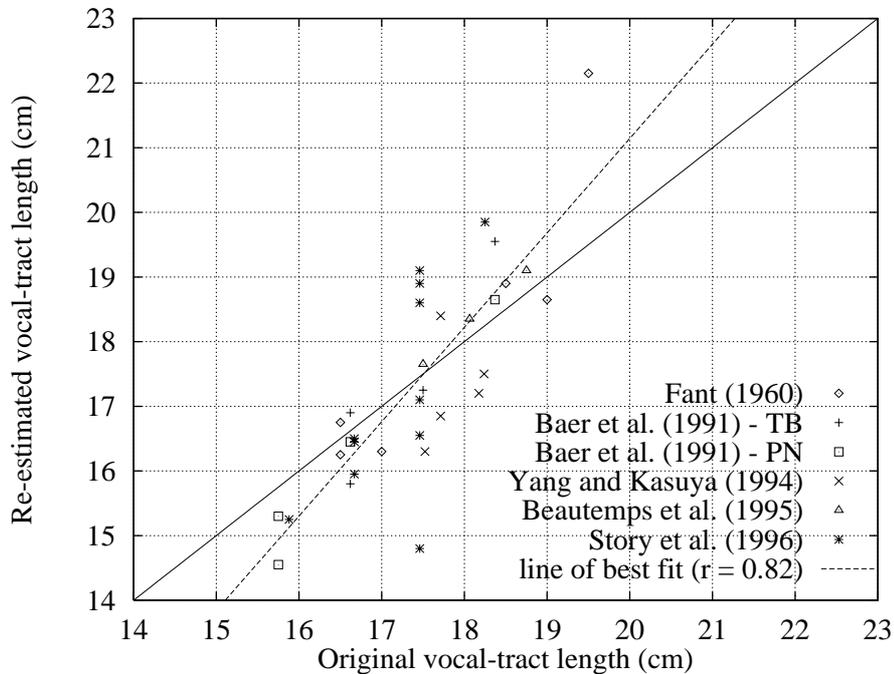


Figure 5.10: Curves of MAD versus  $L$ , obtained in determining the optimum vocal-tract length of each of Fant's (1960) six Russian vowels. The vocal-tract shapes are re-estimated using the method given in Section 5.4.1 (cf. Figures 5.6 and 5.7), with the MAD criterion defined in Equation 5.19. The first four, synthetic formant frequencies and bandwidths (cf. Table 5.3) are used as input to the inversion method.



Figures 5.11: Scatter-plot of re-estimated versus original vocal-tract lengths, for the 33 directly-measured area-functions obtained from the literature. The vocal-tract shapes are re-estimated from the first four formant frequencies and bandwidths (cf. Table 5.3) synthesised under model-matched conditions. *Solid line*: 45-degree line drawn for reference only. *Dashed line*: result of a linear regression analysis (correlation coefficient 0.82; cf. Equation 5.20).

functions.

The scatter-plot in Figure 5.11 shows the relation between the re-estimated and original vocal-tract lengths for all 33 area-functions considered. The distribution of the scatter suggests that our method of area-function estimation tends to underestimate the lengths of originally shorter area-functions, and to overestimate the lengths of originally longer area-functions, thus *exaggerating* the normal variation in vocal-tract length from vowel to vowel. Indeed, a linear regression analysis of the 33 pairs of vocal-tract lengths, yields a correlation coefficient of 0.82, and the following relation:

$$L_{\text{re-estimated}} = 1.46L_{\text{original}} - 8.06, \quad (5.20)$$

which is shown by the dashed line in Figure 5.11. It is interesting to note that the line of best fit intersects the solid line (which represents an ideal condition where the re-estimated and the original lengths are identical) at about  $L = 17.5$  cm, which is known to be approximately the average vocal-tract length of an adult male.

The obvious exaggeration of inter-vowel variations in vocal-tract *length* notwithstanding, it is important to note that the vocal-tract *shapes* themselves are re-estimated remarkably well. Indeed, as illustrated by the dashed curves superimposed on the graphs in Appendix F, articulatory features such as the place of lingual constriction, and the place of labial constriction for the lip-rounded vowels, are preserved fairly accurately. For example, whilst the total length of the re-estimated area-function of Fant's (1960) /u/ is clearly overestimated by the largest amount (as discussed earlier), the re-estimated vocal-tract shape itself preserves (after alignment) not only the places of linguo-velar and labial constriction at approximately 7-8 cm and 19 cm from the original position of the glottis, respectively, but also retains the shape and volume of the back and the front cavities. The overestimated length appears to be therein manifested mainly by an exaggeration in the amount of lip-protrusion, and also partly by the position of the larynx, which is effectively lowered by 0.55 cm compared with original. Similarly, although the re-estimated length of Story et al.'s (1996) area-function for /ɑ/ is underestimated by the largest amount (-15.2%), its shape retains the main place of constriction at approximately 6-6.5 cm from the glottis, and appears to be adversely affected mainly towards the front part of the oral cavity.

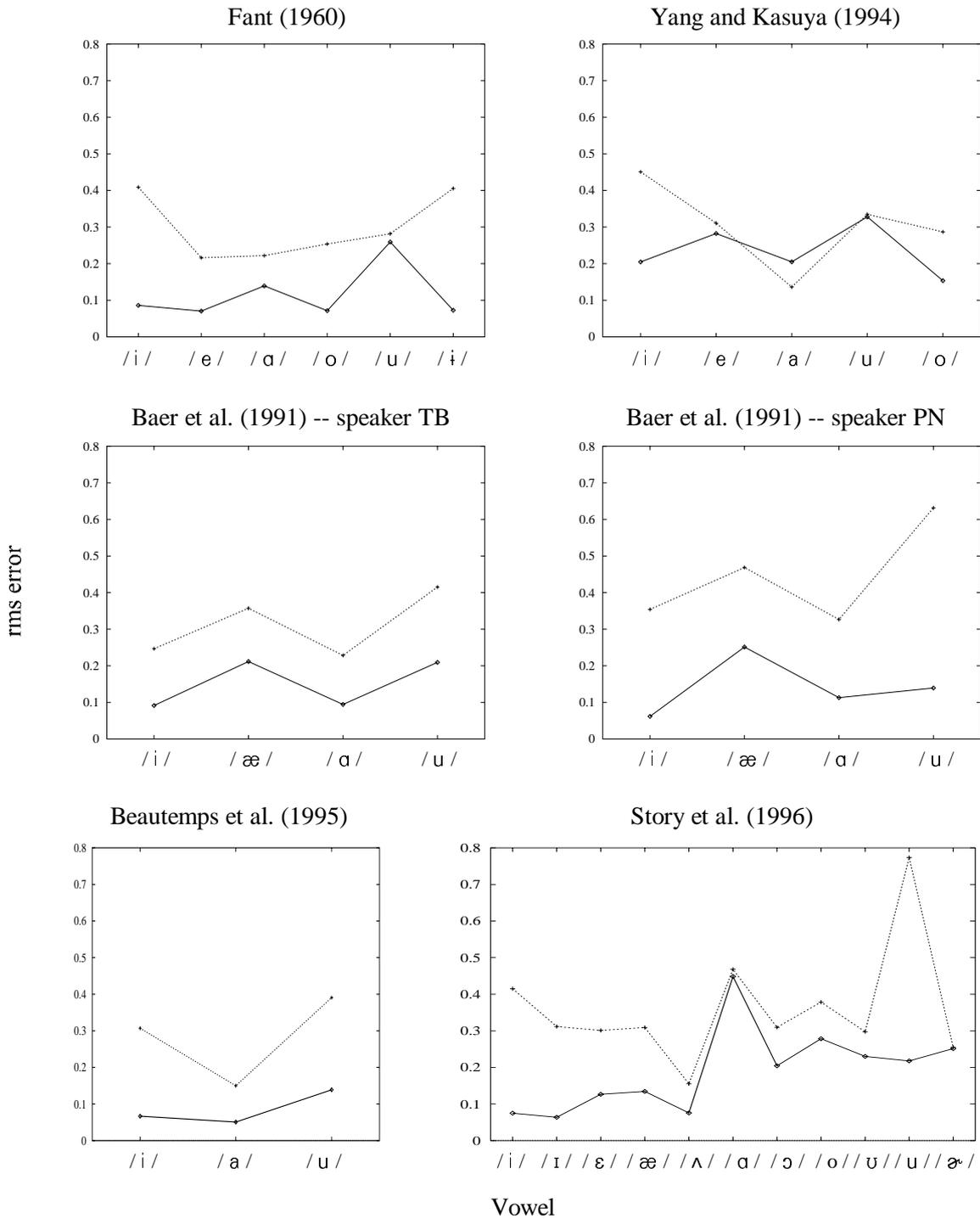


Figure 5.12: Root-mean-square (rms) errors computed using the alignment procedure (cf. Section 5.5.1), between vocal-tract shapes *re-estimated* from the first four formants synthesised under various conditions, and the shapes which *represent* each original area-function with the same degree of smoothness ( $M/2=4$ ), for each of 33 directly-measured area-functions obtained from the literature. *Diamond symbols, joined by solid lines*: 8-section area-functions; LP synthesis (cf. Table 5.3). *Plus symbols, joined by dashed lines*: original section-lengths and areas; a lossy vocal-tract model (cf. Table 5.4).

In drawing conclusions regarding the appropriateness of the MAD criterion for estimating the vocal-tract length, we are compelled, in light of the evidence discussed above, not only to compare the re-estimated and the original lengths, as earlier studies have done, but also to quantify the dissimilarity between the original and the re-estimated vocal-tract *shapes*. As discussed earlier in Section 5.2.2, quantification of differences between raw, step-wise LP area-functions of different lengths is non-trivial (and indeed has apparently never been attempted!), owing partly to the fact that the original and the re-estimated section boundaries would no longer coincide. By contrast, our method of area-function parameterisation yields a smoothed representation of each area-function, and the alignment method described in Section 5.5.1 can then be used to quantify the differences between original and re-estimated vocal-tract shapes.

The rms distances thus computed after alignment of each re-estimated vocal-tract shape and the smoothed, parameterised version of the corresponding, original area-function, are shown by the diamond symbols (joined by solid lines) in Figure 5.12. Whilst they all are, as expected, numerically larger than the mean rms distance (0.04) computed earlier for the vocal-tract shapes re-estimated using their original lengths, the mean of the 33 distances (0.12) is still lower than the mean rms error (0.22) incurred in *representing* those area-functions using the same number of shape parameters (as obtained in Section 5.4.2.2). The largest, anterior shift required to align the pairs of vocal-tract shapes was found to be 0.65 cm, for Story et al.'s (1996) American English /ɑ/, the length of which was the most underestimated; the rms distance computed for that area-function (0.45) is also the largest, and is mainly attributable to the inaccurate re-estimation of the size of the oral cavity and of the lip opening area. The largest alignment-shift towards the glottal end was found to be -0.55 cm, for Fant's (1960) Russian /u/, the length of which was the most overestimated. The remaining vocal-tract shapes were found to require shifts along the length-axis by amounts intermediate between those two extremes, and they generally also yielded lower, rms distances. It is encouraging, therefore, to note that the close resemblance of the re-estimated shapes to the original area-functions (as shown in the graphs of Appendix F), is quantitatively confirmed by the rms errors yielded by our automatic method of alignment.

The results presented above, collectively provide quantitative support for the

inversion method used under model-matched conditions, and with only a finite number of formants. Thus, they also lend credence to the MAD criterion, by which the vocal-tract length (and effectively the vocal-tract shape) is determined. However, in reality, the formant frequencies and bandwidths measured from the acoustic speech signal are not “matched” with the LP vocal-tract model. We therefore continue our evaluation of the inversion method in the next section, using somewhat more realistic, “model-mismatched” conditions.

### 5.5.2.2 More Realistic Conditions

In order to introduce more realistic conditions in our evaluation of the inversion method, we herein relax the constraints arising from two critical assumptions underlying the LP vocal-tract model referred to in the previous section. The first of those assumptions concerns the rarely acknowledged fact, that the formants synthesised using any type of transmission-line analog of the vocal-tract, depend to a certain extent on the *number of sections* used to represent a given area-function. Indeed, it was to avoid mismatch in synthesised formants which led Wakita (1977, 1979) to first reduce Fant’s (1960) area-functions to 8 sections prior to using the LP method (with four formants) to re-estimate those vocal-tract shapes. As it is well known (e.g., Fant, 1960) that a more accurate acoustic simulation of the vocal-tract is afforded by using a greater number of sections, we introduce more realistic conditions (and thus a model-mismatch) by synthesising the formants using the original section-lengths and areas of the 33 area-functions used previously.

The second, crucial assumption concerns the distribution of acoustic energy losses in the vocal-tract. More realistic conditions are simulated by using a transmission-line analog which does include not only the glottal resistance (once again assuming a nominal value of 0.8 for the glottal reflection coefficient), but also a glottal inductance; a viscosity factor, heat-conduction and wall-vibration losses at each section; and a radiation impedance at the lips (see Appendix C for a more detailed description). The first four formant frequencies and bandwidths thus synthesised (and listed in Table 5.4) are then used to re-estimate each of the 33 area-functions using our hybrid LP-SM method of inversion.

Study	Vowel	Formant Frequency (Hz)				Formant Bandwidth (Hz)			
		$F_1$	$F_2$	$F_3$	$F_4$	$B_1$	$B_2$	$B_3$	$B_4$
Fant (1960)	/i/	293.1	2311.5	3106.9	3804.9	87.9	81.7	290.6	192.5
	/e/	461.1	1997.6	2859.1	3670.1	56.3	105.5	245.6	559.3
	/a/	686.6	1108.4	2480.3	3625.2	167.6	136.4	103.7	214.4
	/o/	546.4	888.7	2399.8	3536.2	116.9	80.3	55.8	146.8
	/u/	301.3	631.3	2393.1	3774.6	89.4	52.2	42.7	101.9
	/ɨ/	339.8	1524.3	2361.3	3407.8	72.6	120.7	68.0	99.9
Baer et al. (1991) (Speaker TB)	/i/	322.3	2185.3	2852.2	3963.9	66.6	69.4	110.5	148.9
	/æ/	643.8	1442.5	2278.6	3645.5	88.7	114.7	153.3	386.9
	/ɑ/	540.2	1089.3	2457.1	3777.3	104.8	65.7	72.4	201.1
Baer et al. (1991) (Speaker PN)	/u/	341.4	1097.0	2374.9	3536.3	99.2	48.8	64.7	65.7
	/i/	333.2	2293.1	2997.6	3727.1	67.6	77.6	349.2	191.5
	/æ/	695.5	1595.0	2916.9	3815.7	62.2	88.5	390.2	580.6
Yang & Kasuya (1994)	/ɑ/	622.8	1269.6	2894.5	4034.7	154.1	87.5	153.1	219.4
	/u/	351.5	1465.4	2454.9	3570.0	51.5	62.6	50.6	73.1
	/i/	329.3	2082.5	2481.0	3240.9	82.0	114.4	165.6	205.3
	/e/	545.2	1598.8	2280.3	3437.9	63.1	95.1	107.3	117.1
Beautemps et al. (1995)	/a/	670.4	1080.9	2515.5	3438.9	76.3	81.8	86.5	123.3
	/o/	508.4	858.7	2598.7	3728.7	46.3	44.5	57.0	66.4
	/u/	431.0	1252.3	2341.4	3421.8	93.7	63.0	88.5	98.6
	/i/	292.9	2070.4	3054.0	3720.7	87.0	85.8	138.8	205.9
Story et al. (1996)	/a/	518.7	1304.9	2380.6	3548.5	60.1	90.4	77.8	136.1
	/u/	358.8	867.4	2339.3	3444.9	77.9	47.9	118.3	67.1
	/i/	290.1	2521.1	3196.2	3949.2	76.7	89.5	340.9	257.0
	/ɪ/	501.4	2031.6	2602.8	3594.8	40.5	95.1	153.5	244.1
	/ɛ/	648.5	2009.0	2612.6	3595.0	42.8	133.5	230.6	525.6
	/æ/	730.9	1840.2	2570.2	3470.0	41.3	126.3	229.3	345.0
	/ʌ/	692.2	1302.1	2626.4	3371.6	55.3	107.2	178.2	391.8
	/ɑ/	808.9	1147.9	2807.1	3550.6	126.5	124.3	155.5	532.2
	/ɔ/	632.6	1069.7	2370.1	3007.6	77.0	182.9	723.5	107.3
	/o/	403.7	868.4	2521.0	3802.8	58.7	49.8	86.8	134.1
Story et al. (1996)	/ʊ/	407.3	943.1	2741.3	3817.6	63.9	63.1	77.6	104.5
	/u/	315.1	1126.0	2484.8	3787.1	67.7	55.7	76.4	193.4
	/ə/	535.2	1572.9	2274.1	3072.9	56.2	121.8	164.5	279.1

Table 5.4: Formant frequencies and bandwidths of 33 directly-measured area-functions, synthesised using a transmission-line analog of the vocal-tract (see Appendix C) which includes glottal impedance, viscosity, heat-conduction, wall-vibration, and lip-radiation losses. The original, published section-lengths and areas are used, and the glottal reflection coefficient is fixed at a nominal value  $\mu_{\text{glott}} = 0.8$ .

The re-estimated area-functions are shown in Appendix F (dotted curves), superimposed on each of the original area-functions (step-wise, solid lines) from which the formants were synthesised, and the area-functions (dashed curves) re-estimated in the previous section under the model-matched conditions. It is immediately apparent from these graphs, that despite the deliberate mismatch in the number of sections and in the vocal-tract acoustic model used to synthesise the formants, the re-estimated shapes still capture most of the essential, articulatory features such as constriction locations and cavity sizes, often with remarkable fidelity. Indeed, there do not appear to be any catastrophic errors, as for example a front vowel having a back place of constriction.

On the other hand, a number of area-functions are not so well re-estimated. The worst cases are the area-functions of the vowel /u/, for which a comparison of Tables 5.3 and 5.4 reveals not only a change in the bandwidth pattern, but also a drop, often more than 200Hz, in the second formant frequency; this has been found to be caused primarily by the mismatch in the number of vocal-tract sections. However, this does not appear to affect Fant's (1960) Russian /u/ which already has quite a backward place of lingual articulation; rather, it affects mainly the remaining, more fronted configurations of /u/, and similarly the fronted /i/ of Fant (1960). It is interesting to note that some of these area-functions are amongst those found earlier (in Section 5.4.2.2) to have a distinctively *symmetric* shape.

Our qualitative observations of the superimposed area-functions are confirmed in Figure 5.12, where the rms errors obtained after alignment are shown by the plus symbols (joined with dashed lines), superimposed with those obtained earlier. Indeed, the largest increases in rms error compared with the previous, model-matched results, occur for /u/ of Story et al. (1996) and of Baer et al.'s (1991) speaker PN. Not unexpectedly, rms errors generally increase as a result of the model mismatch. However, in a number of cases the rms error is *only slightly worse* than before (Fant's (1960) /u/, Yang and Kasuya's (1994) /u/ and /e/, and Story et al.'s (1996) /ʌ/, /ɑ/, and /æ/); and most intriguingly, the rms error obtained for Yang and Kasuya's /ɑ/ is actually *lower* than that obtained under model-matched conditions.

The results discussed above can be summarised in terms of the mean of the rms errors for all 33 area-functions (0.37) which, although greater than the mean rms error

(0.22) found in *representing* those area-functions with the same number of parameters, might still be regarded favourably, considering the large number of fairly accurately re-estimated vocal-tract shapes shown in Appendix F. Indeed, not only do those superimposed area-functions show little evidence of *gross* errors, they also demonstrate that the inversion method is likely to succeed in capturing the main articulatory features of the majority of the vocalic area-functions considered, to within the spatial resolution afforded by the first  $M = 8$  shape parameters, and using only the first four formant frequencies and bandwidths.

### 5.5.2.3 Model-based Formant Correction

Clearly, re-estimation of directly-measured area-functions using formants synthesised under more realistic conditions than afforded by the LP vocal-tract model, generally yields worse results. In particular, we have found (in Section 5.5.2.2) that the mean rms error between original and re-estimated vocal-tract shapes under model-mismatched conditions, is about three times larger than that obtained (in Section 5.5.2.1) under model-matched conditions. Nearly two decades ago, Wakita (1979) proposed to combat this problem by using a so-called “formant conversion chart” to compensate for the effects on the formant frequencies, of the LP model’s incompleteness in regard to vocal-tract losses. However, his conversion chart apparently does not include the effects of using different numbers of vocal-tract sections; nor has there appeared a conversion chart for the formant bandwidths.

A more complete version of Wakita’s conversion chart is afforded by the formants which we have already synthesised from the 33 directly-measured area-functions, first under model-matched, then under more realistic conditions. In particular, the 4 formant frequencies listed earlier in Tables 5.3 and 5.4 are plotted along the ordinate and the abscissa, respectively, in Figure 5.13(a); similarly the 4 formant bandwidths, which are plotted in Figure 5.13(b). The solid curve shown superimposed in Figure 5.13(a), is a cubic-polynomial fit to the 132 pairs of data points in the logarithmic-frequency domain, and is described as follows:

$$\ln(F_n^{(LP)}) = 0.0494[\ln(F_n^{(FL)})]^3 - 1.14[\ln(F_n^{(FL)})]^2 + 9.71\ln(F_n^{(FL)}) - 21.8 \quad (5.21)$$

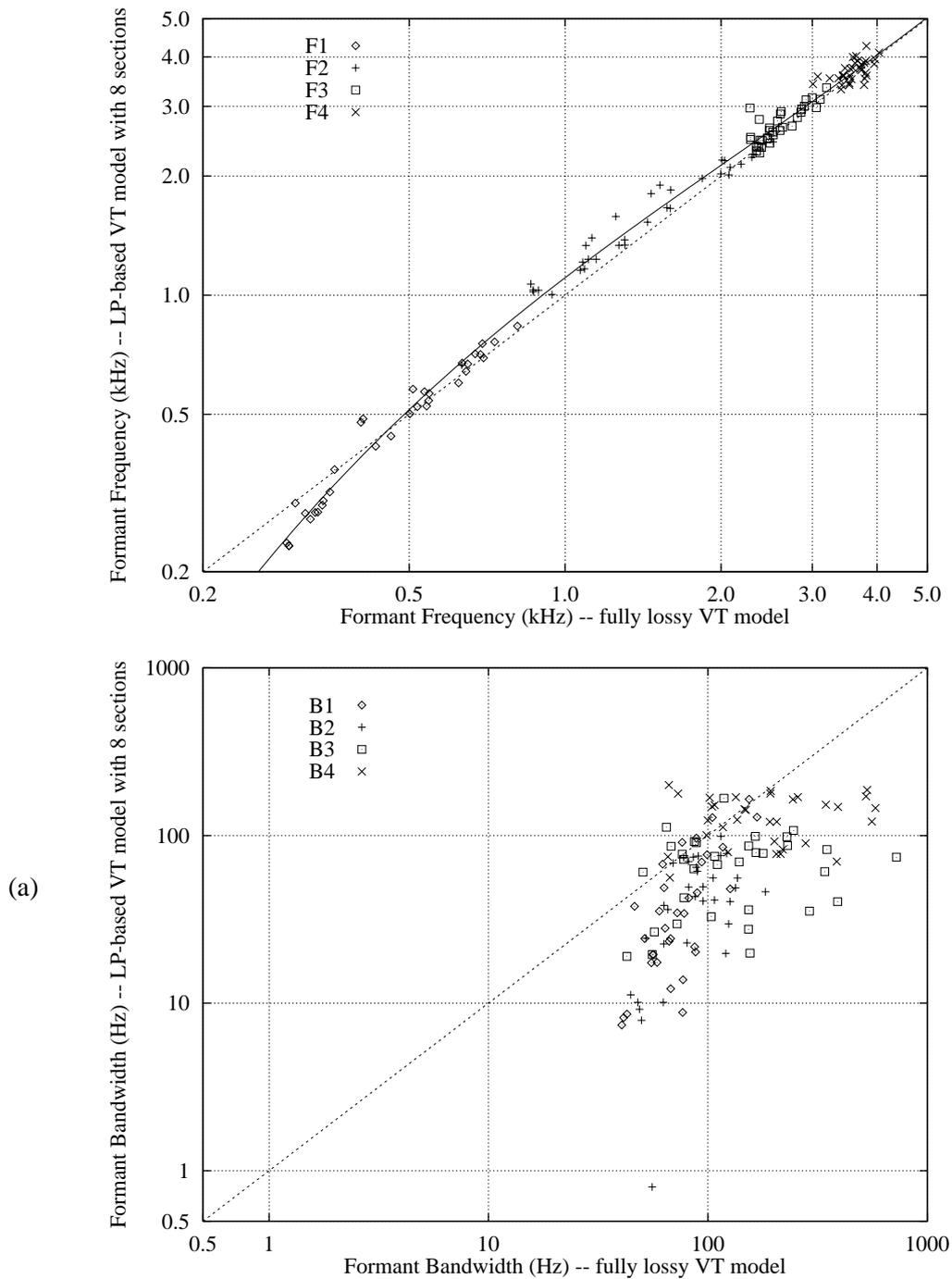


Figure 5.13: Formant frequency (a) and bandwidth (b) conversion charts (more complete version of the frequency chart first proposed by Wakita, 1979). The first four formants of each of the 33 directly-measured area-functions obtained from the literature, are synthesised using first the LP vocal-tract model with the glottal reflection coefficient fixed at a nominal value  $\mu_{\text{glott}} = 0.8$  (*ordinate*, see Table 5.3), then a lossy transmission-line analog of the vocal-tract (*abscissa*, see Table 5.4). The solid curve in (a) is the best cubic fit to the set of logarithmic-frequency data, and is described by Equation 5.21 in the text.

As pointed out by Wakita (1979), the wall-impedance effects in the low-frequency,  $F_1$  range appear to cause the most substantial deviation from the 45° straight (dashed) line. In addition, our more complete, formant frequency conversion chart highlights the extent of model mismatch caused by the combination of vocal-tract losses *and* number of sections, which appear mainly to affect the  $F_2$  range.

In contrast with the formant frequencies, our formant bandwidth conversion chart shown in Figure 5.13(b) does not suggest any regularity which might be exploited by fitting the data with a function. On the contrary, the data appear to be spread across regions of the chart which lie far away from the ideal, 45° line. Taken one bandwidth at a time, the  $B_1$  and the  $B_2$  data points (diamond and plus symbols, respectively) do appear to form an elongated cluster which may be fit with a straight line of large, positive slope; the  $B_4$  data points (cross symbols) might be considered to form a cluster which is elongated horizontally, and which might therefore be fit with a straight line having very little variation in the LP-based values along the ordinate. On the other hand, the  $B_3$  data points (square symbols) appear to be scattered across a large region of the chart, with no apparent regularity. Clearly, formant bandwidth correction is not as amenable a task as formant frequency correction.

Before becoming too despondent over these problematic results, it is worth noting that they are purely *model-based* (i.e., they are based entirely on formants synthesised using either the LP or a more lossy vocal-tract model). In this vein, we have already questioned (in Section 5.2.2) the viability of using model-based correction procedures, when it remains unknown to what degree those models can be trusted to be accurate or realistic. Indeed, the fidelity of even a fully-lossy vocal-tract model which does include many of the acoustic and aerodynamic properties missing from the LP model, has not been firmly established. Perhaps we should therefore not expect model-based formant correction procedures, even if they could be formulated, to significantly and consistently improve the plausibility of LP-derived vocal-tract shapes by rendering them somehow more “realistic” than they already are.

In that context, it is highly instructive to enquire whether those five studies from which we have taken the 33 directly-measured area-functions, have also reported results concerning the discrepancy between measured and re-synthesised formant data. Those

studies have indeed published two sets of formant frequency values based, respectively, on acoustic measurements (from recordings made either at the time of the articulatory measurements or, more commonly, on a separate occasion) and vocal-tract simulations using the measured area-functions themselves. Fant (1960, Table 2.31-1, p.109) used a digital computer (BESK) implementation of a 20-section vocal-tract model, and found that the “average deviation of calculated data from spoken data is of the order of 5 per cent in  $F_2$  and  $F_3$  and 10 per cent in  $F_1$ ”; Baer et al. (1991, Table III, p.813) synthesised waveforms from their measured area-functions using a variant of Kelly and Lochbaum’s (1963) model which includes non-ideal terminations at the lips and at the glottis, and found that the “formant frequencies of the subjects’ utterances differ significantly from the computed resonances of the area functions”; Yang and Kasuya (1994, Table 4, p.626) used Sondhi and Schroeter’s (1987) hybrid time-frequency domain synthesiser which incorporates source-tract interactions and a number of vocal-tract losses, and found differences of up to 7.2% in the formant frequencies of their male subject; Beautemps et al. (1995, Table 2, p.35) used a lossless vocal-tract model with the wall-vibration and lip-radiation taken into account by a correction to the formant frequencies and a correction to the length of the lip-section, respectively, and despite explicit optimisation of their directly-measured area-functions to match the measured and model-synthesised formants, they still obtained errors of up to 10% in the formant frequencies; Story et al. (1996, Table IV, p.549) used a “wave-reflection analog vocal tract model” which includes viscosity, wall-vibration, and lip-radiation losses, in addition to an acoustic side-branch to model the piriform sinuses, and still found deviations from measured formant frequencies up to nearly 15% in  $F_1$ , up to 16% in  $F_2$ , and up to nearly 20% in  $F_3$ .

In row (a) of Table 5.5 we list the rms difference (in Hz) between the measured and synthetic values for each of the first three formant frequencies published in those five studies, and the rms difference in the fourth formant frequencies published by Yang and Kasuya (1994) and by Beautemps et al. (1995). By comparison, row (b) lists the rms differences (in Hz) between the formants synthesised first under model-matched, then (presumed) more realistic conditions (those listed earlier in Tables 5.3 and 5.4, respectively). Both sets of results suggest rms errors on the order of 10% — clearly

	$F_1$	$F_2$	$F_3$	$F_4$	$B_1$	$B_2$	$B_3$	$B_4$
(a)	53.5	183.7	251.2	145.6	-	-	-	-
(b)	40.0	160.6	185.5	228.0	40.4	54.5	166.4	173.4

Table 5.5: Root-mean-square (rms) difference (in Hz) in each of the first four formant frequencies and bandwidths, either synthesised or measured, and pertaining to the 33 directly-measured area-functions obtained from the literature. (a): rms difference between *measured* formant frequencies and those *synthesised* from the directly-measured area-functions, as published in the five studies referred to in the text (cf. Section 5.5.2.3). (b): rms difference between formants first LP-synthesised with an 8-section representation (cf. Table 5.3), then synthesised using the original section-lengths and areas, with a lossy transmission-line analog (cf. Table 5.4).

greater than the 3–5% perceptual difference limen suggested by Flanagan (1955). However, it is even more interesting to note that for the first three formant frequencies, the rms errors listed in row (a) of Table 5.5 (those inferred from the studies cited above) are in fact *larger* than those obtained (and listed in row (b) of Table 5.5) in our evaluation of the discrepancies between the LP and a more realistic vocal-tract model!

That result alone is indicative of the futility of formant correction procedures based on model data. This assertion is indeed confirmed by an experiment in which we attempted to correct the formants listed in Table 5.4, according to the correction charts of Figure 5.13. In particular, we used the cubic-polynomial of Equation 5.21 to correct the formant frequencies, and four separate, linear functions to correct the formant bandwidths. Perhaps not surprisingly, the mean rms error yielded for the re-estimated area-functions (0.33), was found to be only slightly lower than that obtained originally under model-mismatched conditions (0.37). In addition to confirming the futility of model-based formant correction, these results also suggest that the mismatched conditions simulated in the previous section may indeed have been exaggerated, by virtue of our over-reliance on *model*-generated (or synthetic) formant data. In the next section we therefore consider the inversion problem using *real* (or measured) formants.

### 5.5.3 Estimation of Area-Functions from Measured Formants

If we are to use our hybrid LP-SM method of inversion to estimate area-functions from real, measured acoustic data, then we are compelled to alleviate the following problems which would otherwise potentially undermine the interpretability of our results. The

first problem (considered in Section 5.5.3.1) concerns the sheer *variability* which is known to plague measurements of formant bandwidths in particular. The second problem (considered in Section 5.5.3.2) concerns the potential *mismatch* between measured formant bandwidths, and those more appropriate to the LP vocal-tract model. Although both of these problems have long been acknowledged, we still lack a quantitative assessment of, and an accepted method of dealing with, their potentially harmful influences on estimated area-functions.

### 5.5.3.1 Formant-induced Variability of LP Area-Functions

Whilst formant-tracking has traditionally been concerned with reliably measuring the formant frequencies, we have shown in this chapter that the formant bandwidths do play a vitally supporting role in ensuring uniqueness in the LP-based method of inversion. Indeed, the distinctive relation which we have shown to exist between each formant bandwidth and the corresponding, symmetric component of the LP-derived, logarithmic area-function, is sufficient to suspect that large amounts of inter-frame or inter-repetition variation in bandwidth measurements might induce significant variability in the estimated vocal-tract shapes. As this would clearly be an undesirable input to our forthcoming, articulatory interpretation of the phenomenon of dichotomy, it is of utmost importance to quantify the extent to which measurement variability in either formant frequencies or bandwidths are likely to influence the estimated vocal-tract shapes.

To that end, a preliminary step is taken by computing the relative *sensitivity* of our shape parameters to variations in the formant frequencies and bandwidths. This is achieved by fitting the distribution of each of the four formant frequencies and the four bandwidths generated (in Section 5.4.2.1) from the 6561 area-functions perturbed about the neutral-tube, using a linear combination of all eight, re-estimated shape-parameters, as shown by the following, multiple-linear regression formulae:

$$F_n^{(\text{rel})} = \sum_{m=1}^{M/2} \alpha_{n,m}^{(F)} a_{2m-1} + \sum_{m=1}^{M/2} \beta_{n,m}^{(F)} b_{2m-1}, \quad n = 1, \dots, M/2, \quad (5.22)$$

$$B_n^{(\text{rel})} = \sum_{m=1}^{M/2} \alpha_{n,m}^{(B)} a_{2m-1} + \sum_{m=1}^{M/2} \beta_{n,m}^{(B)} b_{2m-1}, \quad n = 1, \dots, M/2, \quad (5.23)$$

	$a_1$	$a_3$	$a_5$	$a_7$	$b_1$	$b_3$	$b_5$	$b_7$
$F_1^{(rel)}$	<b>-0.535</b>	-0.040	-0.040	-0.044	-0.015	-0.017	-0.006	0.018
$F_2^{(rel)}$	-0.041	<b>-0.516</b>	-0.050	-0.053	-0.022	-0.009	-0.011	0.019
$F_3^{(rel)}$	-0.048	-0.053	<b>-0.479</b>	-0.071	-0.029	-0.014	-0.007	0.020
$F_4^{(rel)}$	-0.019	-0.021	-0.026	<b>-0.380</b>	-0.011	-0.006	-0.002	0.007
$B_1^{(rel)}$	0.228	0.954	0.590	0.197	<b>-0.776</b>	0.032	0.064	0.082
$B_2^{(rel)}$	-0.164	0.073	0.968	0.255	0.106	<b>-1.867</b>	0.493	0.618
$B_3^{(rel)}$	-0.340	-0.954	-0.494	0.273	0.089	0.643	<b>-2.184</b>	1.469
$B_4^{(rel)}$	0.276	-0.073	-1.064	-0.726	0.581	1.191	1.627	<b>-2.170</b>

Table 5.6: Coefficients of hyperplanar model, obtained from the 6561 pairs of acoustic/vocal-tract shape data generated in Section 5.4.2.1, for each of the first four, relative formant frequencies and bandwidths, in terms of the first eight shape parameters (cf. Equations 5.22 and 5.23). Boldface figures correspond to the so-called first-order parameter relations which were earlier shown (in Table 5.2) to have high coefficients of correlation ( $|r| > 0.9$ ).

where the superscript “(rel)” refers to the *relative* formant parameters, as defined earlier in Section 5.4.2.1. The resulting set of  $M = 8$  hyperplanes (each of which passes through the origin), are fully described by the following direction-cosines with respect to each shape parameter:  $\alpha_{n,m}^{(F)}$ ,  $\beta_{n,m}^{(F)}$ ,  $\alpha_{n,m}^{(B)}$ ,  $\beta_{n,m}^{(B)}$ ,  $n = 1, \dots, M$ ,  $m = 1, \dots, M/2$ . The coefficients themselves are found such that each hyperplane fits the 6561 data points in a least-mean-squares sense, using a generalised linear least-squares algorithm (Press et al., 1988, p.537) based on singular value decomposition (SVD).

Table 5.6 lists the coefficients of sensitivity thus obtained. In particular, we draw attention to those coefficients along the main diagonal, which quantify the sensitivity in each of the so-called first-order parameter relations, shown earlier to have the most significant coefficients of *correlation* (see Table 5.2, in Section 5.4.2.1). Indeed, similarly to our earlier results, the coefficients of sensitivity exhibit dominant and negative values along the main diagonal, corresponding to the first-order relations between  $F_n^{(rel)} \leftrightarrow a_{2n-1}$ , and between  $B_n^{(rel)} \leftrightarrow b_{2n-1}$ . The factor  $-\frac{1}{2}$  which is predicted theoretically by the SM model (cf. Equation 5.2), is determined empirically by the multiple-linear regression analysis as  $\alpha_{1,1}^{(F)} = -0.535$ ,  $\alpha_{2,2}^{(F)} = -0.516$ ,  $\alpha_{3,3}^{(F)} = -0.479$ , and  $\alpha_{4,4}^{(F)} = -0.380$ , respectively, for each of the first four formant frequencies. The analogous coefficients of sensitivity for each of the first four formant bandwidths, are  $\beta_{1,1}^{(B)} = -0.776$ ,  $\beta_{2,2}^{(B)} = -1.867$ ,  $\beta_{3,3}^{(B)} = -2.184$ , and  $\beta_{4,4}^{(B)} = -2.170$ , respectively. These

increasingly larger-valued, negative coefficients, were already foreshadowed by the slopes of the relevant nomograms shown earlier in Figure 5.5(b).

In order to obtain a first approximation of the sensitivity of each shape parameter with respect to a unit perturbation in the corresponding, relative formant parameter, we assume that the matrix found above is diagonally dominant, and simply compute the reciprocal of each of those hyperplane coefficients listed along the main diagonal. The *inverse-sensitivities* thus obtained, are  $-1.87$ ,  $-1.94$ ,  $-2.09$ , and  $-2.63$ , respectively, for each of the first four, asymmetric shape parameters  $a_{2n-1}$ , and  $-1.29$ ,  $-0.54$ ,  $-0.46$ , and  $-0.46$ , respectively, for each of the first four, symmetric shape parameters  $b_{2n-1}$ . The results of this simple calculation imply that vocal-tract shapes obtained by our hybrid LP-SM method of inversion, are *less sensitive* to unit variations in the relative *bandwidths*, than they are to unit variations in the relative formant *frequencies*. In view of the notorious difficulties of measuring formant bandwidths robustly, this preliminary analysis provides some hope that the bandwidth-induced variations in vocal-tract shapes may not be so disastrous as to obscure important information pertaining, for example, to the phonetic identity of the vowel being considered.

We now turn to the formant data of our four adult, male speakers of Australian English, in order to better assess the detrimental influences of formant variations on estimated vocal-tract shapes. As described in Chapter 3, we used both our phonetic knowledge and the formant-enhancement properties of the NDPS, in order to ensure consistent and relatively noise-free estimates of the formant frequencies in particular. However, the formant bandwidths can be expected to have relatively larger amounts of both *inter-frame* and *inter-repetition* variability, which may therefore induce a greater amount of variability in LP-derived vocal-tract shapes, *despite* the lower sensitivity of those shapes to the bandwidths than to the formant frequencies (as found above). In order to assess the relative influence of formant frequency and bandwidth variations on estimated vocal-tract shapes, the mean and standard-deviation ( $\sigma$ ) of each formant frequency and bandwidth were first computed on a per-vowel, per-speaker basis, across the 7 steady-state frames and 5 repetitions. The following two sets of formant data were then prepared for each vowel of each speaker: (i) with the 4 bandwidths fixed at their mean values, the 4 formant frequencies are set equal to  $\pm 1\sigma$  about their respective

mean values, in all possible ( $2^4 = 16$ ) combinations; and (ii) with the 4 formant frequencies fixed at their mean values, the 4 bandwidths are set equal to  $\pm 1\sigma$  about their respective mean values, in all possible ( $2^4 = 16$ ) combinations. For each of those two sets of 16 formant-patterns, our hybrid method of inversion was then used to estimate the corresponding groups of 16 area-functions, which in turn were subjected to the alignment procedure described in Section 5.5.1.

Three representative examples of the area-functions thus obtained are shown in Figure 5.14, where the two groups of 16, aligned vocal-tract shapes are superimposed in each graph. In particular, it is fairly consistently observed that the formant frequency-perturbed vocal-tract shapes (the green area-functions) cluster together more closely than the bandwidth-perturbed shapes (the red area-functions). This immediately confirms our suspicions raised in the previous paragraph, that despite the smaller *sensitivities* of vocal-tract shape parameters to variations in bandwidths, the *measured*, inter-frame and inter-repetition variations in the bandwidths are sufficiently large (compared with the measured variations in the formant frequencies) as to induce a greater amount of variability in the estimated shapes.

The results are summarised in Figure 5.15, which shows the per-vowel, mean rms dispersion (averaged over the 4 speakers) in the groups of estimated vocal-tract shapes. These results first confirm that the dispersion in vocal-tract shapes caused by  $\pm 1\sigma$  variations in measured formant bandwidths (shown by the plus symbols joined by dashed lines), is indeed consistently larger than the shape-dispersions induced by  $\pm 1\sigma$  variations in the formant frequencies (diamond symbols joined by solid lines). Nevertheless, this quantitative evidence confirms our observations in Figure 5.14, that at least for the formant data at hand, the bandwidth-induced articulatory dispersions are not so significantly larger than the formant frequency-induced dispersions, that they should cause catastrophic distortions of the estimated vocal-tract shapes. Indeed, the mean, formant frequency-induced dispersions for the two back vowels in “hod” and “hoard”, are numerically within the same range as the mean, bandwidth-induced dispersions for some of the front vowels. This finding is all the more remarkable since, as discussed earlier, the measured formant bandwidths were not subject to the rigorous constraints imposed by our methodology for extracting the steady-state formant

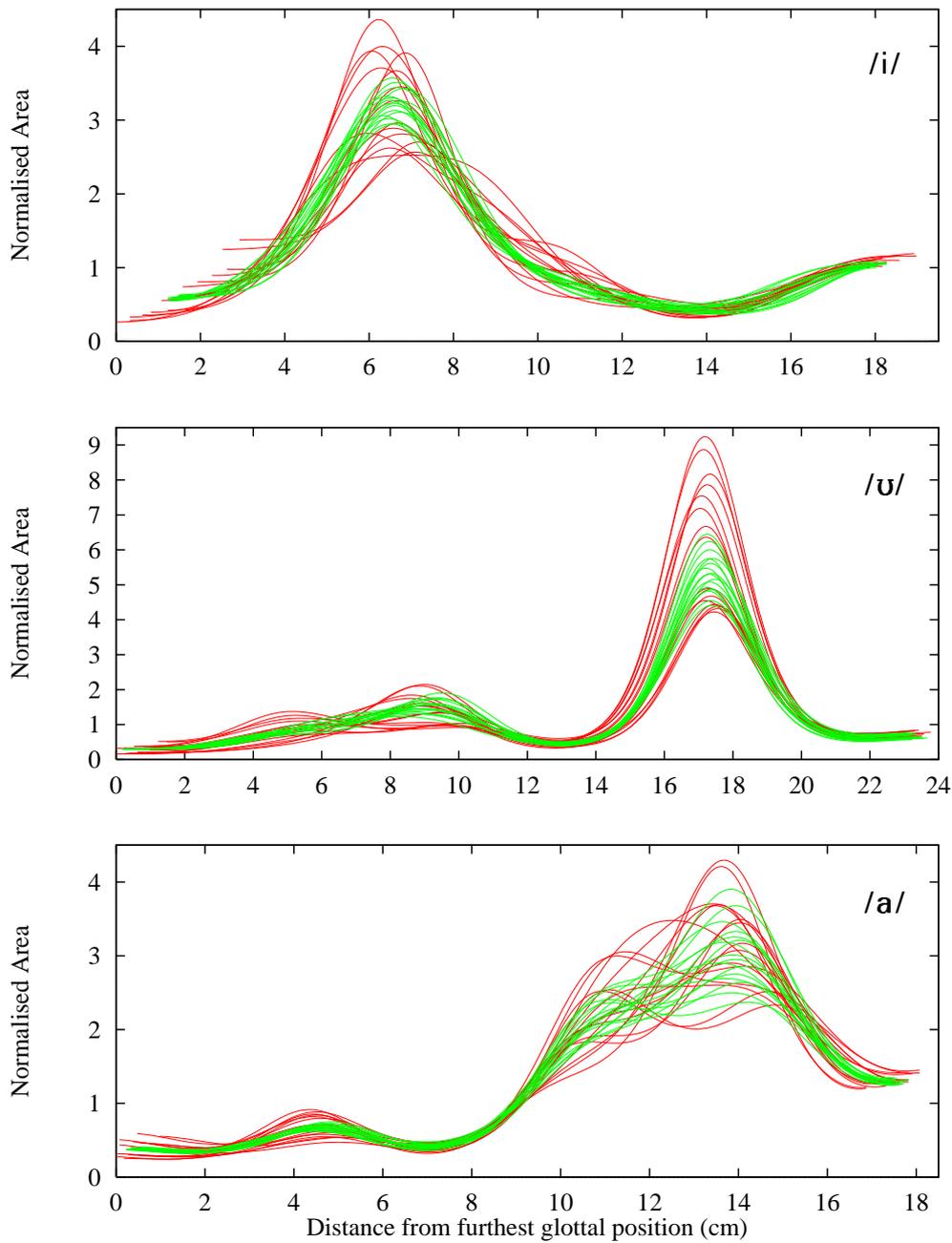
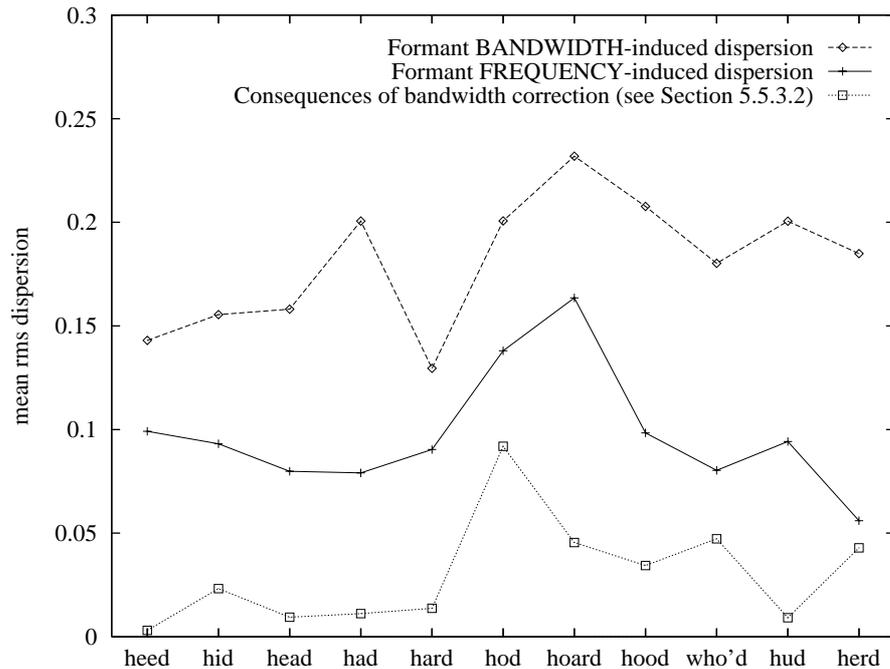


Figure 5.14: Illustration of the vocal-tract *shape dispersions* induced by all combinations of  $\pm 1\sigma$  perturbations in each of the first four, measured formant frequencies (*green area-functions*) and bandwidths (*red area-functions*) of the FC dataset. *Top panel*: vowel /i/ of speaker C. *Middle panel*: vowel /u/ of Speaker D. *Bottom panel*: vowel /a/ of Speaker B. All area-functions are estimated using our hybrid LP-SM method of inversion. As exemplified in these three graphs, it is consistently observed in all 11 vowels of all 4 speakers of the FC dataset, that the bandwidth-induced shape dispersions are larger than the formant-induced shape dispersions.



Figures 5.15: Dispersions about the prototype vocal-tract shape for each of the eleven vowels, found using the alignment procedure described in Section 5.5.1, and averaged over the four speakers. *Diamond symbols (joined by solid lines)*: average dispersions induced by all combinations of  $\pm 1\sigma$  perturbations in each of the first four formant frequencies. *Plus symbols (joined by dashed lines)*: average dispersions induced by all combinations of  $\pm 1\sigma$  perturbations in each of the first four formant bandwidths. The mean and standard-deviation for each formant parameter was computed separately for each vowel of each speaker, across the 7 steady-state frames and 5 repetitions. *Square symbols (joined by dotted lines)*: rms differences between pairs of vocal-tract shapes estimated from the mean formants of the FC dataset, before and after bandwidth correction (see Section 5.5.3.2).

frequencies; nor did we use a pitch-synchronous acoustic analysis which might have reduced the bandwidth measurement variability caused by the varying position of the analysis window relative to the peaks of glottal excitation in the speech waveform.

In sum, our results indicate that *bandwidth*-induced dispersions in vocal-tract shape (mean rms error of 0.181) are larger than formant *frequency*-induced dispersions (mean rms error of 0.097) by approximately a factor of two. It is interesting to note that by comparison, the normalised standard-deviation of the bandwidths themselves (mean value of 0.163) is larger than that of the formant frequencies (mean value of 0.048) by a factor greater than three<sup>6</sup>.

Although studies concerned with the expected measurement variability in formant bandwidths are quite rare, Markel and Gray (1976, p.178) do offer the following

<sup>6</sup> The smaller factor obtained for the dispersions in vocal-tract shapes than for the formants themselves, is partly explained by the coefficients of inverse-sensitivity found earlier.

generalisation:

“...linear prediction bandwidth estimates may be in error by as much as a factor of 2.5.”

In light of this and the evidence summarised in Figure 5.15, we must acknowledge the potentially harmful influence of the presumably non-phonetic variations in formant bandwidths, and prescribe a method of dealing with those variations. As the bandwidths do clearly play a role in determining the so-called second-half of shape components which are otherwise missing in the completely lossless model, their *phonetic* variation (i.e., from vowel to vowel) should perhaps be retained. Towards this end, we have determined to use only the mean of each bandwidth on a per-vowel basis, computed across frames, repetitions, and speakers. This prescription for controlling the non-phonetic variability in the formant bandwidths is all the more pertinent that our vowel classification experiments using simplified cepstra have already shown (in Chapter 4) that the acoustic-phonetic phenomenon of speech-speaker dichotomy is independent of these acoustic parameters.

### 5.5.3.2 Closed-glottis Correction of Formant Bandwidths

Whilst the sheer variability in measured formant bandwidths can be controlled by using only their per-vowel mean values as recommended above, a serious question remains as to the relevance of those mean bandwidths to the LP vocal-tract model. The potential mismatch stems from the fact that while “true” or measured formant bandwidths reflect all types of losses in the human vocal tract, such as those caused by the glottal source, acoustic viscosity, heat-conduction, wall-vibration, and lip radiation, the LP vocal-tract model has only a single, frequency-independent source of acoustic energy loss located at the glottal end. Hence, as stated by Fant (1980, p.85):

“The bandwidths we need for the inverse LPC-based transforms are the bandwidths of a production model which has losses at the glottis only and lacks the cavity wall shunt.”

However, as we have earlier argued (in Section 5.5.2.3), model-based formant correction procedures are not easily formulated, nor do they appear to be particularly effective. In this vein, Fant’s (1980, p.85) very next, more general statement is perhaps a more appropriate guide to pursuing the problem:

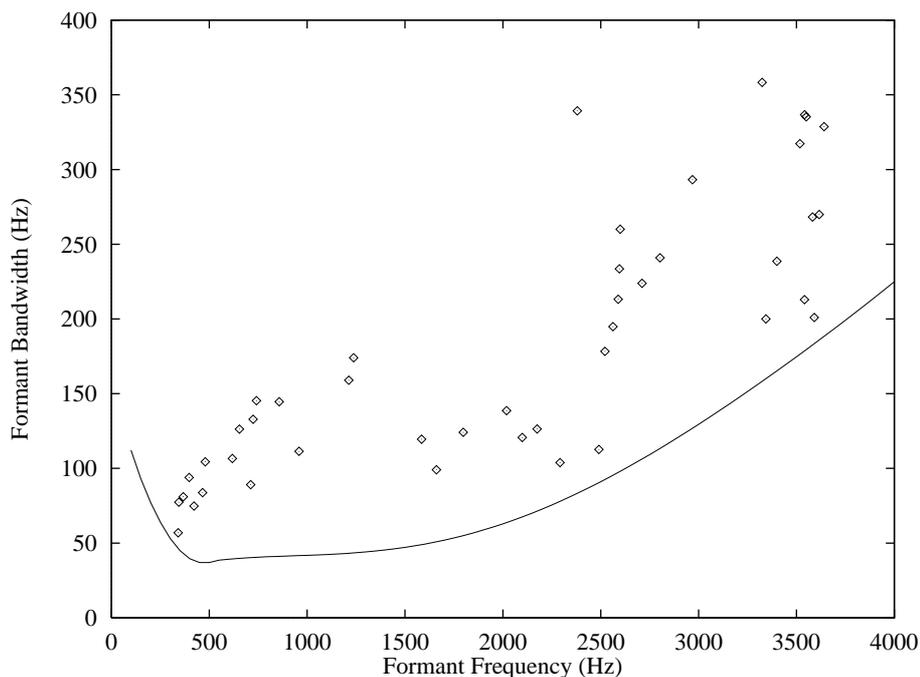
“From the true formant frequencies and bandwidths we thus have to make a best guess of what bandwidths the LPC model would generate.”

If model-based data are to be avoided, then we are naturally drawn towards measured bandwidth data, in search of a clue as to the necessary “best guess”.

In that regard, it is relevant to note that any measured bandwidth is known to be approximately an *additive* combination of its individual components (in Hz), which themselves are determined by the configuration-dependent contribution from each of the separate sources of loss in the vocal-tract (e.g., Fant, 1960; Flanagan, 1972). Bandwidths measured from the acoustic speech signal are thus the total sum of their individual components arising from glottal, vocal-tract internal, lip-radiation and other losses; by contrast, the LP vocal-tract model “would generate” bandwidths from only the glottal resistance. The question then arises whether one can approximately compensate for the losses which are missing in the LP model, by *subtracting* those components of each bandwidth which might have been measured had there been no source of loss at the speaker’s glottis.

Such measurements of so-called “*closed-glottis bandwidths*” have indeed been made on human speakers (Fant, 1962; Fujimura and Lindqvist, 1971), by applying a sweep-tone excitation signal transcutaneously at the throat while the speaker maintains a static articulatory configuration with closed glottis, and by measuring the acoustic response near the speaker’s lips. Using those empirical data, Fant (1972) then brought acoustic theory to bear on the problem of deriving a formula for each of the first three bandwidths, in terms of the first three formant frequencies and an average, fourth formant frequency. More recently, Hawks and Miller (1995) used those empirical data to derive a more general equation which is able to predict the bandwidth of any of the formant frequencies up to nearly 5kHz, given only the centre-frequency of that formant.

If we can assume that the per-vowel, mean values of the measured bandwidths embody all of the natural sources of loss which arise for the given vocalic configuration, and that Hawks and Miller’s (1995) equation embodies the relation between frequencies and bandwidths of closed-glottis formants, then we might expect that by subtracting thus estimated closed-glottis bandwidths from the measured bandwidths, we would obtain an approximation of the per-vowel, mean *glottis-only* bandwidths. It is indeed



Figures 5.16: *Illustration of the viability of closed-glottis bandwidth correction.* Solid curve: Hawks and Miller’s (1995) empirically-derived equation relating the closed-glottis bandwidth of a formant with its centre-frequency (assuming their default conditions for an adult male speaker). Diamond symbols: the per-vowel mean of each of the first four formants computed over the 7 steady-state frames, 5 repetitions, and 4 adult, male speakers of the FC dataset.

the glottis-only bandwidths which the LP vocal-tract model ideally would require in order to yield area-functions which (presumably) more closely resemble those actually produced by a given speaker. Clearly, it is difficult to test this particular hypothesis without simultaneously-measured acoustic and vocal-tract shape data; however, we may at least test the viability of the procedure by determining whether the subtraction would yield admissible (i.e., positive and non-zero) bandwidths.

To that end, in Figure 5.16 is shown Hawks and Miller’s (1995) bandwidth-versus-frequency curve (assuming their default conditions for adult male speakers), together with the mean frequencies and bandwidths of each of the first four formants of the 11 vowels of our adult, male speakers of the FC dataset (diamond symbols). This Figure shows that our measured bandwidth data do follow the expected trend of higher values for higher formant frequencies. More importantly, it clearly shows that our mean data do consistently lie *above* the curve, and that the difference (or the vertical distance) between each mean bandwidth (or data point) and the corresponding, closed-glottis bandwidth (along the curve), would indeed yield a positive and non-zero, “glottis-only”

bandwidth<sup>7</sup>.

Whilst the corrected bandwidths (listed in Appendix G, together with the measured, mean formant frequencies and bandwidths for each vowel) do confirm the viability of the procedure, it would also be interesting to know how, and to what extent, these bandwidth corrections influence the estimated vocal-tract shapes themselves. Shown in each of the 11 panels in Appendix G, are superimposed the aligned area-functions estimated from our mean formant data before (dashed line) and after (solid line) bandwidth correction. Those graphs show that the consequences of our proposed method of bandwidth correction are not only harmless, but may even be considered beneficial on the whole. Differences between the pairs of vocal-tract shapes appear to be the largest for the quasi-neutral vowel /ɜ/ (which is also somewhat shorter after bandwidth correction), the fronted /ɥ:/ (which develops a more well-defined constriction), and the back vowel /ɔ/ (whose front cavity expansion reduces in volume to a somewhat more realistic level). It is also interesting to note that after bandwidth correction, the two, high-back vowels /ɔ/ and /ʊ/ are no longer found to require the addition of a fifth, synthetic formant in order for the inversion method to determine an optimum vocal-tract length (cf. Section 5.4.1).

Our qualitative observations are quantitatively confirmed in terms of the rms differences between the pairs of (logarithmic) area-functions, shown by the square symbols (joined by dotted lines) in Figure 5.15 (see the end of the previous section). That graph clearly indicates that the shape-related consequences of our proposed method of bandwidth correction, are less severe than either the formant bandwidth- or frequency-induced dispersions found in the previous section. Indeed, the overall mean rms error (0.03) is even less than that found earlier for the 33 directly-measured area-functions re-estimated under model-matched conditions (0.04, see Section 5.5.2.1).

It is highly encouraging that the consequences of formant bandwidth correction do not appear to be as drastic as the literature would generally seem to suggest. However, we must concede that our proposed method of determining the mean, “glottis-only”

---

<sup>7</sup> It is important to note that such, “glottis-only” bandwidths would represent only *an average* of the quasi-periodic bandwidth values, owing to our speech analysis frame-lengths typically extending over more than one glottal cycle.

bandwidths which *ought* to be better matched with the LP vocal-tract model, must be more conclusively put to the test, perhaps with the help of simultaneously-measured acoustic and vocal-tract shape data. Until such time, we may rest assured that at least for the formant data at hand (the FC dataset), the proposed correction applied to the per-vowel mean bandwidths, should not endanger the phonetic interpretability of the estimated area-functions, nor their articulatory manifestations of the speech-speaker dichotomy, which we wish to investigate in the forthcoming chapter.

## 5.6 Concluding Summary

Our aim in this chapter was to find a method of mapping the geometry of the human vocal-tract from the acoustic speech signal, in order to facilitate an extension of our acoustic-phonetic investigations of the speech-speaker dichotomy (in Chapter 4) into the domain of speech production. As outlined in Section 5.1, we sought a method of inversion which would satisfy the following two criteria: *uniqueness* and *formant-based parameterisation*. These criteria served as guideposts in our bipartite rationale (in Section 5.2), where we reviewed the strengths and the limitations of the Schroeder-Mermelstein (SM) model and the linear-prediction (LP) vocal-tract model.

The main strength of the SM model is its description of vocal-tract area-functions in terms of orthogonal parameters which relate quasi-linearly and uniquely to the resonance frequencies of the acoustic-tube. Its main limitation in the context of area-function estimation, is that only one-half of the required information is available in the acoustic speech signal (i.e., only the formant frequencies). Hence, the estimated vocal-tract shapes are ambiguous in their *symmetric* components, unless further constraints are introduced.

The LP model, on the other hand, is well-known to yield a unique vocal-tract shape (for a given vocal-tract length), and is notable for its mathematical tractability and low computational complexity in comparison with other inversion methods which have been proposed. However, while the formant *bandwidths* are implicated as the necessary second-half of acoustic information required to obtain a unique LP area-function, it has never been explained *how* they contribute to the *shape*-related uniqueness of those area-functions. Furthermore, the LP-based method of inversion suffers from a number of

limitations, including the analysis artefact of an upper limit  $L_{\max}$  on allowed vocal-tract length, and the coarse, step-wise representation which inhibits comparison of area-functions of different length.

With the aim of elucidating the uniqueness property of the LP inversion method, and perhaps overcoming some of its inherent limitations, we then used the SM model as a probe to examine the formant-dependence of LP-derived vocal-tract shapes. In Section 5.3 we first validated the SM model within the LP modelling framework, and thereby showed that those two models are indeed isomorphic in regard to the quasi-linear relation between the formant *frequencies* and the *antisymmetric* (cosinusoidal) perturbations of the uniform (logarithmic) area-function. We then derived a theoretical proof of the dependence of the mean formant bandwidth on the glottal reflection coefficient, and thereby provided the motivation for regarding the *relative* bandwidths more pertinent to *shape* components of LP-derived area-functions. Remarkably, our empirical investigations then revealed those *bandwidth*-related shape components to be the *symmetric* (sinusoidal) perturbations of the uniform (logarithmic) area-function.

Given the acknowledged (but to date largely unexplained) importance of formant bandwidths in securing uniqueness in LP-derived area-functions, our proposed method of vocal-tract shape parameterisation (given in Section 5.4) is therefore an important contribution to our understanding of the uniqueness property of the LP inverse method. More importantly, it overcomes the inherent discreteness of LP-derived vocal-tract shapes, and provides a smooth representation by way of a compact and orthogonal set of parameters which also lend themselves to an acoustic-phonetic interpretation. Furthermore, the artificial limitation of  $L_{\max}$  is easily overcome by invoking the very property of the SM model which explicitly relates higher-resolution shape components with the higher formants. In particular, the vocal-tract shape is estimated to a specified degree of spatial resolution, while using a sufficient number of vocal-tract sections to ensure that the highest available formant still lies within the spectral range of the vocal-tract model.

In Section 5.4.2 we first used our hybrid LP-SM method of inversion to evaluate our proposed method of area-function parameterisation. In particular, we obtained high coefficients of correlation between each of the formant frequencies and bandwidths

synthesised from 6561 different area-functions, and the corresponding shape-parameters re-estimated from those formant data. The *acoustic* relevance of our shape parameters thus confirmed, their *physical* relevance was assessed by representing 33 directly measured, human vocal-tract area-functions obtained from the literature. On average, the most important shape-components of those area-functions were found to be the first two pairs of both antisymmetric and symmetric components. This evidence confirmed the phonetic importance of the first two formant frequencies of vowels; however, it also confirmed the importance of the bandwidths in describing unique vocal-tract shapes.

The hybrid method of inversion itself was finally evaluated in Section 5.5, with the help of a new method for aligning and quantifying the difference between similarly-shaped area-functions. We first re-estimated the 33 directly measured area-functions obtained from the literature, using the first four formant frequencies and bandwidths synthesised under model-matched conditions (i.e., a vocal-tract model with a glottal resistance as the only source of loss, and the number of vocal-tract sections equal to twice the number of formants). Whilst the phonetic range of variation in vocal-tract length was found to be exaggerated upon re-estimation, the vocal-tract shapes were fairly accurately re-estimated, with a mean rms error of 0.12 . More realistic conditions were then imposed, by re-estimating the area-functions from formants synthesised using a more lossy vocal-tract model, and by retaining the original section-lengths and areas. The mean rms error between original and re-estimated shapes under model-mismatched conditions, was found to be 0.37 .

In the hope of finding a reasonably accurate method of compensating for the LP model's simplified assumptions in regard to vocal-tract losses and number of sections, we presented a more comprehensive version of the "formant frequency conversion chart" first proposed by Wakita (1979). Whilst we were able to fit the model-generated formant data with a third-order polynomial function in logarithmic frequencies, the analogous conversion chart for formant *bandwidths* did not suggest the possibility of finding a single functional description. Indeed, the futility of approaching the formant correction problem from a purely *model*-based point of view, was further reinforced by our finding that on average, errors between formant frequencies measured from the speech waveform and those re-synthesised from the directly-measured area-functions

are in fact larger than the corresponding errors between the formant frequencies synthesised, respectively, under model-matched and model-mismatched conditions.

In Section 5.5.3 we then turned to our measured vowel formant data (of the FC dataset). As a prelude to examining formant-induced variability using those real data, we first established that LP-derived vocal-tract shapes are actually *less* sensitive to unit perturbations in the relative formant bandwidths, than to unit perturbations in the relative formant frequencies. Despite this lower sensitivity, the inter-frame and inter-repetition variability in our measured bandwidth data was large enough to induce dispersions in estimated vocal-tract shapes, approximately *twice* that induced by variability in the formant frequencies. In view of our forthcoming, articulatory investigation of the speech-speaker dichotomy which we had earlier explained in terms of the formant frequencies (in Chapter 4), our prescription was therefore to control the inter-frame, inter-repetition, and the inter-speaker variations in each formant bandwidth (while retaining their phonetic variation, the importance of which was indirectly suggested by our results in Section 5.4.2.2), by reducing the bandwidths to their mean values computed on a per-vowel basis.

Whilst the measurement variability in formant bandwidths was thus controlled, their relevance to the LP vocal-tract model was still questionable. Our approach to dealing with this problem was then to recognise that the bandwidths required by the LP model are those which would arise from a vocal-tract having only a resistive glottal termination. We therefore used Hawks and Miller's (1995) empirically-derived equation to determine the so-called "closed-glottis" bandwidths of our mean formant data, then subtracted those from the original values in order to approximate the required, glottis-only bandwidths. Although a more conclusive proof of this procedure would perhaps require simultaneously-measured acoustic and area-function data, our results do support its viability, and furthermore indicate only minor differences in the vocal-tract shapes re-estimated after bandwidth correction.

Finally, the detailed evaluation presented in Section 5.5 raises our confidence in LP-derived and parameterised vocal-tract shapes obtained by the hybrid LP-SM method described in Section 5.4, which itself secures both *uniqueness* and *resonance-based parameterisation* of estimated vocal-tract area-functions. In the following chapter we

therefore proceed to use our method of inversion and area-function parameterisation, with a view toward an investigation of the speech-speaker dichotomy in the speech production domain.

## **Chapter 6**

### **An Articulatory Explanation of Speech-Speaker Dichotomy**

#### **6.1 Introduction**

From the very start, the central theme of our thesis has been the long-standing problem of speech-speaker dichotomy. In Chapter 4 we introduced a new methodology for examining the interactions between phonetic and speaker-specific attributes in spoken vowel sounds, and thereby unfolded and explained the vowel-speaker dichotomy in the acoustic-phonetic domain. The remarkably systematic and orderly manifestations of vowel-speaker interactions across the spectral continuum, and the readily interpretable, acoustic-phonetic relevance of those spectral regions of either phonetic or speaker influence, then heightened our initial resolve (stated in Chapter 1) to gain an even more fundamental interpretation of the dichotomy, from a speech production point of view. Towards that end, in Chapter 5 we addressed the problem of mapping the geometry of the human vocal-tract from the acoustic speech signal. The formant-based parameterisation of LP-derived area-functions yielded by our extended version of the Schroeder-Mermelstein (SM) model, then provided the necessary tools to proceed with an articulatory investigation of the vowel-speaker dichotomy, with sufficient confidence in the uniqueness and plausibility of our estimated vocal-tract area-functions.

In this chapter we therefore seek an articulatory explanation of the vowel-speaker dichotomy, and for this purpose we shall use the FC dataset of /hVd/ monosyllables recorded by our four, adult male speakers of Australian English. We begin in the following section with a discussion of our overall philosophy on describing speaker differences in the articulatory domain. In particular, we shall define three potential, articulatory sources of speaker variability associated with the supralaryngeal vocal-tract, which will then serve as a descriptive framework for explaining the vowel-speaker

interactions, whose manifestations we had previously observed only in the acoustic-phonetic domain.

As we shall see, a significant component of our articulatory approach will involve speaker normalisation of the estimated vocal-tract area-functions. In Section 6.3 we therefore describe our methodology for speaker normalisation of each of the three articulatory sources of variability. In Section 6.4 we then apply our methodology to provide a three-part articulatory explanation of the dichotomy. The final proof of our articulatory explanation will then be revealed in Section 6.5, where we examine both the articulatory and acoustic-phonetic consequences of complete speaker normalisation. We conclude in Section 6.6 with a summary of our contributions.

## **6.2 Articulatory Approach**

The vowel-speaker dichotomy which we first unfolded and explained in the acoustic-phonetic domain (in Chapter 4) is, by definition, a consequence of the interactions between phonetic and speaker-specific attributes of the spoken vowel sounds considered. Insofar as the vocal-tract area-functions corresponding to those vocalic data were estimated (in Chapter 5) from the formant parameters which earlier afforded an acoustic-phonetic explanation of the dichotomy, we should therefore expect them to carry a similarly potent blend of phonetic and speaker-specific attributes. Indeed, to the extent that those estimated area-functions represent articulatory manifestations of both the phonetic and inter-speaker variabilities which exist in the original, acoustic data, they embody the very elements required to provide a physical explanation of the dichotomy.

Perhaps the simplest approach to an articulatory explanation, is to observe and to quantify the differences between our four speakers' area-functions — particularly those which correspond to the pairs of vowels, confusions amongst which were found (in Chapter 4) to be the main contributors to the dichotomy. However, the apparent simplicity of this direct approach belies the methodological complexities in comparing area-functions of different vowels and different speakers, and in formulating plausible articulatory interpretations of speaker differences, without the benefit of an adequate framework for describing those differences. Alas, nor does the literature seem to have

embraced any of the pioneering efforts (e.g., Garvin and Ladefoged, 1963; Stevens, 1971) or the more substantial attempts (e.g., Laver, 1980; Nolan, 1983) to provide such a descriptive framework, which might then facilitate our investigation of the articulatory correlates of the vowel-speaker dichotomy.

In this vein, our review of the literature (in Chapter 2) has already shown that differences in overall vocal-tract length, particularly between speakers of different genders, have received by far the greatest amount of attention, sadly to the almost complete exclusion of all other, potential articulatory sources of speaker variability. Whilst the size of the vocal-tract (and hence its length) is perhaps the most important, physical source of variation between male and female speakers, or indeed between adult and child speakers, rarely has there been an attempt to explain more subtle differences in articulatory behaviour, as might be expected amongst our four, adult male speakers of the FC dataset used to unfold the dichotomy. Those differences, whether of an idiolectal or more intrinsic nature, arise from what has been broadly termed *acquired* or *learned* articulatory behaviour, as distinct from the so-called *organic* or *anatomical* differences which may be accounted for mainly in terms of vocal-tract size (Garvin and Ladefoged, 1963; Stevens, 1971).

In that context, we have also reviewed (in Chapter 2) a number of studies which have attempted to provide a more complete account of the potential sources of so-called learned articulatory differences between speakers, from which emerges a distinction between *long-term* and so-called *realisational* differences (Nolan, 1983). Although the term “articulatory setting” was coined by Honikman (1964) mainly in regard to cross-linguistic differences in the long-term utilisation of the articulatory organs, Laver’s (1980) more comprehensive, descriptive framework of *settings* is sufficiently general to allow for speaker differences at the two broad levels which Garvin and Ladefoged (1963) have called *group* (e.g., across languages, dialects, or genders) and *individual* (e.g., within genders and idiolects). By contrast, differences in realisation or articulatory *strategy* have received attention mainly at the level of groups of speakers, such as dialectal or socio-linguistic variations in the realisation of phonemes, as accounted for by phoneticians and linguists.

From the foregoing discussion emerges an integrated perspective on the potential

articulatory (supralaryngeal) sources of speaker differences, three of which stand out as essential components in any, more extensive or elaborate, descriptive framework. Those basic components, or articulatory *features* as hereafter referred to, describe speaker differences in the overall size of the fixed vocal-tract *structure*, in the long-term bias or articulatory *setting*, and in the residual, phoneme- (or vowel-) specific articulatory *strategy*. These three features comprise what might be regarded as a minimal, yet sufficiently descriptive framework, which indeed will form the basis of our articulatory explanation of the vowel-speaker dichotomy.

If our explanation is to be tripartite as suggested, it will require isolation of the speaker differences carried by each of the three articulatory features in turn, followed by an assessment of their individual contributions to the overall, acoustic-domain manifestation of the dichotomy. Naturally, the process of speaker normalisation which is thereby called for, is best applied directly to the estimated area-functions themselves, and the normalised acoustic data then obtained by an appropriate method of articulatory synthesis. For example, the formant parameters synthesised from all the area-functions after speaker normalisation of vocal-tract structure and articulatory setting, will define an acoustic-phonetic space where only vowel-specific, articulatory strategy-related speaker differences are present. Simplified cepstra generated from those formants, can then be used in inter-speaker vowel classification experiments of the type conducted in Chapter 4, and the contributions of speaker differences in articulatory strategy to the dichotomy thus determined.

In order to maintain consistency with our LP-based method of acoustic-to-articulatory mapping, the LP model is also used for synthesising the formant parameters after each stage of speaker normalisation of the vocal-tract area-functions. Furthermore, the number of vocal-tract sections for each area-function (and hence the order  $M$  of the LP synthesis model), is the same as the number of sections used to estimate that area-function (i.e., equal to either 8 or 10, depending on whether an artificial, fifth formant was required in order to overcome the  $L_{\max}$  limitation, as described in Chapter 5). This method guarantees that the original formants (those which were used as inputs to the inverse mapping) are perfectly resynthesised from the unmodified area-functions. The formants synthesised after normalisation of pairwise articulatory sources of inter-

speaker variability, can therefore be compared with the original formants, and the acoustic-phonetic consequences of the area-function modifications thus quantified within a consistent methodological framework.

Our approach to an articulatory explanation of the vowel-speaker dichotomy is therefore based on the concept of systematically depriving the phenomenon of certain, articulatorily-defined sources of inter-speaker variability, thereby forcing a reduction in the degree of spectral manifestations of vowel-speaker interactions — a reduction, however, with pre-defined physical correlates. This approach involves the successive steps of articulatory speaker normalisation, synthesis of normalised acoustic data, inter-speaker vowel classification as a function of an increasing upper spectral limit, and an acoustic-phonetic decomposition of the behaviour of the accuracy curve thus generated. Starting with the estimated area-functions of our four speakers of Australian English (FC dataset), those steps will be carried out three times in Section 6.4, each time normalising inter-speaker variations in two of the three articulatory features described above, thus providing a threefold articulatory decomposition of the vowel-speaker dichotomy. Towards that end, we first describe (in Section 6.3) our methodology for speaker normalisation of vocal-tract structure, articulatory setting, and vowel-specific articulatory strategy.

## **6.3 Methodology for Articulatory Speaker Normalisation**

A prerequisite to speaker normalisation of the three articulatory features described in the preceding section, is a definition of each of those features in terms of quantifiable aspects of estimated vocal-tract area-functions. In the following three sections, we shall therefore motivate and justify our assumptions in regard to the definition of each of the articulatory features in turn, and thus describe our collective methodology for speaker normalisation of vocal-tract fixed structure, articulatory setting, and vowel-specific articulatory strategy.

### **6.3.1 Speaker Normalisation of VT Fixed Structure**

In our search for a definition of the overall size of each speaker’s vocal-tract anatomy or fixed structure, we are naturally drawn away from shape-related aspects of area-

functions, and focus, almost by default, on their length. Indeed, as mentioned in the previous section and discussed in greater detail in our review of the literature (in Chapter 2), vocal-tract length is the physical parameter most widely studied in regard to speaker differences, with a clear emphasis on the more radical differences between men, women, and child speakers, and an equally clear (although often implicit) connection with the overall size of the vocal-tract structure. However, as the vocal-tract length of any given speaker changes from one vowel configuration to another, the question arises whether the lengths pertaining to certain articulatory configurations are likely to be more relevant than others, to the size of that speaker's fixed vocal-tract anatomy.

Although the length along the path of the vocal-tract airway depends to a certain extent on the posture and position of the tongue, the most significant contributors to the phonetic variation in vocal-tract length are the position of the lips and of the larynx, which vary typically over a range of up to 1 or 2cm each (e.g., Perkell, 1969). The clearly lip-rounded vowels included in our (FC) dataset of Australian English, are the back vowels /ɔ/ and /ʊ/, and the fronted vowel /ɜ:/, all of which therefore have characteristically longer overall lengths (as confirmed in Chapter 5); the back vowel /ɒ/ is also known to be accompanied by lip rounding (e.g., Singh and Singh, 1976; Ladefoged, 1993). At the same time, it is well-documented (e.g., Lindblom and Sundberg, 1971; Ewan and Krones, 1974; Riordan, 1977; Högberg, 1995) that the larynx is usually lowered in all those vowels, concomitant with lip-rounding. Indeed, in an attempt to gain a more faithful measure of the "articulatory organ size" in terms of vocal-tract lengths estimated from vocalic speech data recorded by speakers of Japanese, Ishizaki (1978b, p.1050) argues thus: "Since the pronunciations of /u/ and /o/ are usually accompanied with lip protrusions and glottal movement downward, vowels /i/, /e/ and /a/ are chosen ...". On the other hand, there is additional evidence in the literature (e.g., Ewan and Krones, 1974; Riordan, 1977; Högberg, 1995) where it has been observed that the larynx is raised above its average position, in the mid- to high front vowels such as /i/, /ɪ/, and /ɛ/, for which the lips are usually either spread or neutral.

By virtue of the significant potential for variations in lip protrusion/spreading and larynx lowering/raising in those seven vowels, inter-speaker variability in the overall

lengths of their area-functions are potentially compounded by differences both in anatomical size and articulatory behaviour. It would seem, therefore, that a more accurate (or less biased) estimate of the size of a speaker’s fixed vocal-tract anatomy is provided by the lengths of only the non-rounded, mid- to low vowels, which presumably are least susceptible to extreme positions of the lips and of the larynx. In this connection, it is interesting to note Lindblom and Sundberg’s (1971) suggestion, based on vocal-tract simulation experiments, that a good estimate of the vocal-tract length is provided by  $F_3$  of *open* vowels, which would include our /æ/, /a/, and /ʌ/. In addition to being articulated more centrally and therefore being least constricted, the two vowels /ʌ/ and /ɜ/ are also known to be “neutral for the feature rounding” (Singh and Singh, 1976, p.57). Our measure of the speaker-dependent vocal-tract fixed structure is therefore defined as the mean length of the four vowels /æ/, /a/, /ʌ/, and /ɜ/. Thus for the  $s^{\text{th}}$  speaker, the mean, structure-related vocal-tract length is obtained as follows:

$$\bar{L}_s = \frac{1}{4} (\bar{L}_s^{/æ/} + \bar{L}_s^{/a/} + \bar{L}_s^{/ʌ/} + \bar{L}_s^{/ɜ/}), \quad (6.1)$$

where the mean length  $\bar{L}_s^{/v/}$  for each of the four vowels is computed over all 5 repetitions and 7 frames.

Having thus defined the first of our three articulatory features, we come now to a description of its normalisation across speakers. In that context, recall from our review of the literature (in Chapter 2) that the differences in vocal-tract lengths of male and female speakers have been shown (Chiba and Kajiyama, 1958; Högberg, 1995) to be unevenly distributed across the length of the vocal-tract — differences in the length of the pharyngeal region are generally greater than differences in the length of the oral region. However, as all four of our speakers of Australian English are male, we may assume that their pharynx-to-oral length ratios are not as disparate as might be expected of inter-gender comparisons. Consequently, the vocal-tract fixed structure is speaker normalised by *uniform* scaling of the vocal-tract lengths of all area-functions, as follows:

$$L'_{svrf} = k_s L_{svrf}, \quad (6.2)$$

where the prime superscript denotes the normalised length, and the indices refer to the

$v^{\text{th}}$  vowel, the  $r^{\text{th}}$  repetition, and the  $f^{\text{th}}$  frame. The speaker-dependent length-scaling factor itself is defined by the following expression:

$$k_s = \frac{\bar{L}^{(\text{ref})}}{\bar{L}_s}, \quad (6.3)$$

where the reference length is simply the mean, structure-related length of all  $N_s = 4$  speakers, as follows:

$$\bar{L}^{(\text{ref})} = \frac{1}{N_s} \sum_{s=1}^{N_s} \bar{L}_s. \quad (6.4)$$

It is conceivable that if the area-functions of both male and female speakers were included, then a *non-uniform* length normalisation might have to be considered. It is not at all clear, however, where the boundary between the oral and pharyngeal parts of estimated area-functions would be identified. On the other hand, the studies of Nordström (1977) and Yang and Kasuya (1995, 1996) have cast considerable doubt on the significance of non-uniform length scaling, and suggest that uniform scaling is indeed relatively more important, both from acoustic and perceptual points of view.

After speaker normalisation as described in Equation 6.2, each speaker’s mean vocal-tract length computed over the selected subset of four vowels will, by definition, be equal to the average, reference length  $\bar{L}^{(\text{ref})}$ , while inter-vowel, inter-repetition and inter-frame variations in length are naturally retained. Those length-normalised area-functions can then be assumed to have been cast in a common, reference vocal-tract anatomy or fixed structure, and the remaining speaker differences therefore related only to articulatory behaviour, which we consider next in terms of setting and vowel-specific strategy.

### 6.3.2 Speaker Normalisation of Articulatory Setting

Articulatory setting is the term preferred by Laver (1980) in his descriptive framework for “voice quality”, which has been described as “a quasi-permanent quality running through all the sound that issues from [a speaker’s] mouth” (Abercrombie, 1967, p.91). Although the term was first used in reference to “the gross oral posture ... requisite as a framework for the ... merging and integrating of the isolated sounds ... of a language”

(Honikman, 1964, p.73), it has also been described to be useful “for extralinguistic purposes, as a phonetic component of voice quality identifying the individual speaker” (Laver, 1980, p.3). It is in the latter context that we adopt the concept of articulatory setting as a useful component or feature of speaker idiosyncrasy in vowel production.

To the extent that an articulatory setting “does not imply simply the particular articulations of the individual speech sounds ..., but ... their common, rather than their distinguishing components” (Honikman, 1964, p.73), or indeed that it is “poly-segmental, ... a property of a stretch greater than a single segment” (Laver, 1980, p.3), it may therefore be computationally defined as an articulatory feature averaged over all the available, segmental data of a speaker. Analogously to our definition (in the previous section) of a speaker’s fixed vocal-tract structure in terms of the mean length of their estimated area-functions, a speaker’s articulatory setting can be defined as the mean *shape* of their area-functions. However, in contrast to our definition of structure in terms of the mean length computed over only a subset of the vowels, setting is more appropriately defined as the mean vocal-tract shape computed over all the vowels.

This working definition of articulatory setting is by no means complete — in Laver’s (1980) descriptive terminology, it can take into account only *latitudinal*, *supralaryngeal* settings, and therefore ignores the so-called longitudinal, supralaryngeal settings which comprise long-term idiosyncrasies in labial protrusion/labiodentalisation and larynx height (precisely those aspects of vowel articulation which we sought to exclude in our speaker normalisation of vocal-tract structure). However, to the extent that our definition of setting includes the shape-related (or *latitudinal*) aspects of labial idiosyncrasies, speaker-specific behaviour in the longitudinal positions of the lips and of the larynx will later be absorbed into the vocal-tract length component of per-vowel articulatory *strategy* (in Section 6.3.3). Whilst the so-called laryngeal, velopharyngeal, and overall muscular tension settings are also ignored, they are in any case irrelevant to our study of the vocal-tract area-functions of non-nasalised vowels. Our present definition of articulatory setting in terms of mean vocal-tract shape thus retains the essential components of Laver’s detailed descriptive framework, while adhering to a reasonable degree of both conceptual and computational simplicity.

That simplicity, however, is threatened by a serious question which arises in

regard to the required averaging of area-functions across all the vowels of a given speaker. As the area-functions are generally of different lengths, and do not possess any fixed articulatory landmarks except the glottal and lip ends which mark the acoustically-relevant extreme end-points, there appears to be no obvious way in which they might be aligned along the length-axis prior to averaging or any other type of comparison. In this context, recall that our method of alignment described in Chapter 5 can only account for variations in length across frames and repetitions, and is therefore applicable only on a per-vowel basis.

Our discussion in the previous section regarding the expected phonetic variations in larynx height, clearly rules out the highly simplified assumption (which is so often implicit in the literature) that vocal-tract area-functions can be aligned at the glottal end. At the same time, the obvious contributions of labial posture to the overall length of the vocal-tract cast some doubts on the robustness of aligning area-functions at the lip end. The problem of comparing area-functions of different vowels without the benefit of a fixed articulatory frame of reference, would thus appear to be insurmountable, and may indeed have partially discouraged the pursuit of articulatory explanations of basic speech phenomena using inverse methods based on area-functions.

One clue to a possible resolution of this problem lies in the observation (in the literature) that there exists a correlation between lip protrusion and larynx lowering. The significance of the correlation is such that, for example, Shirai and Honda (1977, p.284) assume as part of the control of their articulatory model, “that the position of the glottis varies with the lip parameter” in a simple, linear relationship (with a coefficient of 0.3, which implies a smaller range of variation in glottal height compared with lip protrusion). There has also been put forward an acoustic-theoretic explanation of the observed correlation: in his model-based study of the acoustical significance of articulatory maneuvers of [y], Wood (1986, p.391) concludes that “larynx depression is complementary to lip rounding and restores spectral sensitivity to palatal and pharyngeal tongue movements otherwise disturbed by the labial activity.”

Acoustic theory also provides a second clue as to the possible location of a fixed area-function landmark. In particular, the Schroeder-Mermelstein (SM) model, which formed the basis of our parameterisation (in Chapter 5) of vocal-tract shapes estimated

by the linear-prediction (LP) model, clearly asserts that for variations in the formant frequencies only, there exists a single location at the *centre* of an area-function (i.e., at  $x = \frac{1}{2}L$  cm from either end) where the cross-sectional area is invariant (see Figure 5.1). This node, which also marks a point of antisymmetry in the behaviour of formant frequency-sensitivity functions of a uniform tube, might therefore be considered a theoretically-motivated landmark (as aptly pointed out by Fuchi and Ohta, 1979) with considerable, acoustic-phonetic significance.

Both the physiological and acoustic-theoretic evidence considered, support the notion of aligning estimated vocal-tract area-functions at approximately their mid-length. Indeed, the emergent concept of *centre-alignment* is perhaps the least biased with respect to both the glottal and labial ends of area-functions which, by virtue of their amounts of overhang being determined by the vocal-tract length, are thus allowed to characterise phonetically-relevant variations in larynx lowering and lip protrusion. Whilst a somewhat off-centre alignment may be more appropriate in order to enforce a smaller range of positions in larynx height compared with that of lip protrusion (as preferred by Shirai and Honda, 1977), there is also evidence of some speakers who adopt quite comparable ranges of phonetic variation in the positions of their larynx and lips (e.g., Perkell, 1969; Högberg, 1995). Insofar as the comparative range of positional variation may well be speaker-dependent, it may be impossible to precisely define a point of reference which can then be applied with certainty to all estimated area-functions; or indeed to determine speaker-specific points of reference on the basis of estimated area-functions alone. In the face of these difficulties, we therefore retain the more tractable concept of centre-alignment, which may adequately serve our purpose without introducing undue computational complexity and methodological uncertainty.

Our method of computation and speaker normalisation of articulatory setting is therefore described as follows:

1. Speaker-normalise the vocal-tract fixed structure by scaling the lengths of all area-functions (as described in Section 6.3.1). All area-functions are then assumed to have originated from a common (or reference) vocal-tract anatomy.
2. Align the structure-normalised, per-vowel area-functions of each speaker across

the 7 frames and 5 repetitions (as described in Section 5.5.1.2), thus minimising the root-mean-square (rms) difference in vocal-tract shapes amongst each group of 35 area-functions. The resolution along the length-axis for these and all subsequent computations, is  $\Delta x = 0.05$  cm per section.

3. Compute the length of the mutually-overlapping (MOL) region for the 35 aligned area-functions of each vowel of each speaker, defined as the length over which all 35 area-functions exist (i.e., excluding the glottal and labial ends where some area-functions may extend beyond others).
4. Retaining the alignment performed in step 2, align the groups of 35 area-functions across vowels and speakers, at the mid-length of their respective MOL regions found in step 3.
5. Compute the mean vocal-tract shape per speaker, by averaging the area-functions (in the logarithmic-area domain) across all frames, repetitions, and vowels. At the glottal and labial ends where longer area-functions may extend beyond the shorter ones, averaging is still performed (at each  $\Delta x$ ) albeit over the reduced number of area-functions.
6. Compute the overall mean vocal-tract shape, by averaging the per-speaker mean area-functions (in the logarithmic-area domain) across all speakers. Overhanging area-functions at the glottal and labial ends are dealt with as described in step 5.
7. At each vocal-tract section (of length  $\Delta x$ ) and for each speaker separately, compute the ratio of overall mean area and per-speaker mean area (provided the latter is defined at that section). Analogously to Equation 6.3, these ratios are the speaker-specific area-scaling factors required to normalise the area-functions of each speaker with respect to articulatory setting as defined earlier.
8. For each individual area-function of each speaker (aligned as in step 4), multiply the area at each vocal-tract section (of length  $\Delta x$ ) by the area-scaling factor found for that speaker in step 7.

To summarise the computational procedure described above, all the area-functions are speaker-normalised with respect to vocal-tract structure, aligned to minimise inter-frame and inter-repetition variations in shape, centre-aligned across vowels and

speakers, and speaker-normalised with respect to articulatory setting. By definition, the average vocal-tract shape per speaker will then be identical to the overall mean shape computed in step 6. To the extent that the alignment of each area-function as computed in step 4 is retained in all subsequent steps, any discontinuities in the mean areas found in steps 5 and 6 (as a consequence of averaging over a reduced number of area-functions towards the glottal and labial ends) are also retained in subsequent steps, without adverse effect.

### 6.3.3 Speaker Normalisation of Vowel-Specific Articulatory Strategy

Thus far we have described our procedure for speaker normalisation of two articulatory features, both of which relate to long-term aspects of vowel production. First the vocal-tract structural differences between speakers are identified and normalised, then the vowel-averaged bias in vocal-tract shape or articulatory setting. The speaker differences which remain after normalisation of structure and setting, can therefore be attributed to idiosyncratic behaviour in articulatory strategy on a per-vowel basis.

As with the long-term features, vowel-specific articulatory strategy comprises two components — one which relates to longitudinal variations in articulatory behaviour, such as the amount of lip protrusion, larynx lowering or raising, and the posture of the tongue and other articulators as they influence the *length* of the path along the vocal-tract airway; and one which relates to latitudinal variations, such as the degree of labial constriction, and the posture of the tongue and other articulators as they determine the *shape* of (or cross-sectional areas along) the vocal-tract airway. We therefore define the former in terms of speaker differences in mean vocal-tract length per vowel, and the latter in terms of differences in mean vocal-tract shape per vowel. Assuming structural and setting-related differences to have been speaker-normalised, the residual differences in mean vocal-tract length and shape for each vowel separately then make up what we here define as articulatory strategy.

Our algorithmic procedure for the computation and speaker normalisation of the *shape*-related component of vowel-specific articulatory strategy is described as follows:

1. Starting with the structure- and setting-normalised, completely aligned area-

functions obtained after step 8 in Section 6.3.2, compute the per-vowel mean vocal-tract shape for each speaker, by averaging (in the logarithmic-area domain, and with  $\Delta x = 0.05$  cm per section) over the 7 frames and 5 repetitions.

2. Compute the mean vocal-tract shape for each vowel, by averaging the per-speaker mean area-functions obtained in step 1 above (in the logarithmic-area domain) across all speakers.
3. For each vowel separately, at each vocal-tract section (of length  $\Delta x$ ) and for each speaker, compute the ratio of overall mean area and per-speaker mean area (provided the latter is defined at that section). Similarly to step 7 in Section 6.3.2, these ratios are the speaker-specific area-scaling factors required to normalise the area-functions of each speaker with respect to articulatory strategy, on a per-vowel basis.
4. For each individual area-function of each speaker, multiply the area at each vocal-tract section (of length  $\Delta x$ ) by the area-scaling factor found for that speaker and for that vowel in step 3 above.

Following the steps described above, the per-vowel average vocal-tract shape of each speaker will, by definition, be identical to the overall mean shape computed for that vowel in step 2.

The *length*-related component of articulatory strategy is then speaker normalised by uniform length-scaling of all area-functions on a per-vowel basis, as follows:

$$L''_{svrf} = k_{sv} L'_{svrf}, \quad (6.5)$$

where the prime superscript denotes the structure-normalised length obtained earlier in Section 6.3.1, and the double-prime superscript denotes the structure- and strategy-normalised length. The length-scaling factor in Equation 6.5 is both speaker- and vowel-dependent, as defined by the following expression:

$$k_{sv} = \frac{\overline{L}_v^{(\text{ref})}}{\overline{L}_{sv}}, \quad (6.6)$$

where the denominator is the mean, structure-normalised length for each vowel of each speaker, and the numerator is simply the mean, structure-normalised length of all  $N_s = 4$  speakers for the  $v^{\text{th}}$  vowel, as follows:

$$\overline{L}_v^{(\text{ref})} = \frac{1}{N_s} \sum_{s=1}^{N_s} \overline{L}_{sv} . \quad (6.7)$$

It is important to note that the speaker normalisation of length-related differences in articulatory strategy as described above, is carried out without altering the relative alignment of area-functions along the length-axis. Indeed, the alignment secured in Section 6.3.2 is retained, by scaling (i.e., either lengthening or shortening) each area-function in equal proportions about the fixed point of reference which is defined at the centre (or mid-length) of the MOL regions found previously.

## 6.4 Articulatory Explanation of the Dichotomy

As outlined in Section 6.2, our approach to an articulatory explanation of the vowel-speaker dichotomy consists of a tripartite decomposition of the speaker differences contained in the estimated vocal-tract area-functions, followed by resynthesis of the formant parameters, and a return to the acoustic-phonetic methods of analysis and explanatory framework developed in Chapter 4. In this way, speaker differences in the three articulatory features (defined in the previous Section) may be assessed both in terms of their acoustic-phonetic correlates, and the spectral manifestations of vowel-speaker interactions which they each induce.

If the algorithmic procedures outlined in the previous Section are performed in the order that we presented them, the resulting area-functions will have been completely speaker normalised with respect to vocal-tract structure, articulatory setting, and vowel-specific articulatory strategy. However, as argued in the exposition of our articulatory approach (in Section 6.2), we require to isolate the speaker differences carried in each of those three articulatory features in turn, in order to assess their individual contributions to the overall phenomenon of vowel-speaker dichotomy. In principle, this can be achieved after having performed complete speaker normalisation, by methodically re-introducing the speaker differences found in each of the three features in turn. Of course, this process of speaker *un*-normalisation is not necessary for isolating the differences in vowel-specific articulatory strategy (as we shall require in Section 6.4.3), because the procedures described in Sections 6.3.1 and 6.3.2 already yield the desired (structure- and setting-normalised) area-functions. On the other hand,

speaker differences in vocal-tract structure (as required in Section 6.4.1) must be re-introduced in the completely-normalised data, by dividing (rather than multiplying) the length of each area-function by the appropriate, per-speaker scaling factor  $k_s$  defined in Section 6.3.1. Similarly, speaker un-normalisation of articulatory setting (as required in Section 6.4.2) is achieved by dividing (rather than multiplying) the area of each section (of length  $\Delta x$ ) of each area-function by the appropriate, per-speaker area-scaling factor for that vocal-tract section, as found in step 7 of the procedure described in Section 6.3.2. In order to ensure the section-by-section relevance of those area-scaling factors, care must be taken to perform this latter operation prior to speaker normalisation of the length component of vowel-specific articulatory strategy, as described in Section 6.3.3.

In the following three Sections we thereby examine the contributions of speaker differences in vocal-tract *structure*, articulatory *setting*, and vowel-specific articulatory *strategy*, to the vowel-speaker dichotomy.

#### 6.4.1 Contributions of VT Structure to the Dichotomy

Table 6.1 lists the mean of our speakers' estimated vocal-tract lengths, averaged over the 7 frames and 5 repetitions per vowel. In the last column are listed the speaker-averaged lengths for each vowel, which confirm our earlier observations (in Chapter 5) regarding the general validity of the phonetic variations in estimated vocal-tract length. Indeed, the four vowels with the longest mean lengths (/ʊ/, /ɔ/, /ɜ:/, and /ɒ/, in order of decreasing length) are precisely those which are known to be articulated with lip rounding and larynx lowering, and the vowels with the shortest mean lengths (/æ/, /ɛ/, /ɪ/, and /i/, in order of increasing length) are the four front vowels which are known to be articulated either with an open oral tract or with spread lips and raised larynx.

The third row from the bottom in Table 6.1 lists the mean lengths per speaker, computed over all eleven vowels. The overall mean of 17.51cm agrees well with the commonly accepted value for the typical vocal-tract length of an adult male. However, in line with our earlier discussions regarding the potential contributions of lip rounding and larynx height to variations in vocal-tract length, the range of 0.87 cm in the overall mean lengths across the four speakers presumably bears the influence of both anatomical and behavioural differences.

Vowel	Speaker				Mean
	A	B	C	D	
/i/	15.76	16.14	16.61	16.26	16.19
/ɪ/	15.39	16.09	16.33	16.37	16.05
/ɛ/	14.74	16.20	16.00	16.21	15.79
/æ/	<b>14.94</b>	<b>15.30</b>	<b>16.08</b>	<b>15.27</b>	<b>15.40</b>
/a/	<b>16.38</b>	<b>17.00</b>	<b>17.39</b>	<b>16.65</b>	<b>16.85</b>
/ʌ/	<b>16.89</b>	<b>16.82</b>	<b>17.65</b>	<b>17.14</b>	<b>17.12</b>
/ɒ/	17.26	18.28	17.59	17.86	17.74
/ɔ/	19.37	19.62	20.08	21.02	20.02
/ʊ/	20.88	19.81	21.22	22.33	21.06
/ʌɪ/	19.06	18.69	19.81	20.35	19.47
/ɜ/	<b>16.84</b>	<b>16.22</b>	<b>16.76</b>	<b>17.63</b>	<b>16.86</b>
Overall Mean	17.05	17.29	17.77	17.92	17.51
<b>Sub-Mean</b>	<b>16.26</b>	<b>16.33</b>	<b>16.97</b>	<b>16.67</b>	<b>16.56</b>
$k_s$	<b>1.018</b>	<b>1.014</b>	<b>0.976</b>	<b>0.993</b>	

Table 6.1: Mean estimated VT-lengths (cm), and length-scaling factors required to speaker normalise VT *structure*. Shown for each of the 11 vowels, are the mean VT-lengths computed over all 7 frames and 5 repetitions per speaker. The per-vowel mean VT-lengths are listed in the last column on the right, and the per-speaker mean VT-lengths are listed in the row labelled “Overall Mean”. By contrast, the row labelled “Sub-Mean” lists the mean VT-lengths computed only over the subset of four vowels deemed to be least affected by lip protrusion/retraction and larynx raising/lowering (/æ/, /a/, /ʌ/, /ɜ/, as argued in Section 6.3.1), and which therefore yield the per-speaker length-normalisation factors ( $k_s$ , listed in the bottom row) required for speaker normalisation of VT fixed *structure*.

Indeed, as argued in Section 6.3.1, there is ample evidence in the literature to suggest that speaker differences in vocal-tract fixed structure are best described in terms of the mean length of only the non-rounded, mid- to low vowels /æ/, /a/, /ʌ/, and /ɜ/, owing to their relatively low susceptibility to extreme and variable degrees of lip protrusion and larynx lowering. The mean vocal-tract lengths computed over that subset of four vowels are therefore shown in the second row from the bottom (labelled “Sub-Mean” and highlighted in bold-font) in Table 6.1. Whilst our arguments for selecting those four vowels were motivated purely on the basis of direct physiological observations reported in the literature, it is encouraging to note that the computed sub-mean lengths are consistently about 1 cm shorter than the overall mean lengths, thus lending further credibility to our assumption of having largely excluded the influences of

lip rounding and larynx lowering, and thereby arriving at a more effective measure of the speaker differences in vocal-tract fixed structure.

Listed in the bottom row of Table 6.1 are the uniform length-scaling factors  $k_s$ , which were computed according to Equation 6.3, assuming normalisation of each speaker's sub-mean length to the mean, reference value of  $\bar{L}^{(\text{ref})} = 16.56$  cm. These length-scaling factors and the sub-mean lengths from which they were derived, together suggest that differences in the size of our four speakers' vocal-tract fixed structures are indeed fairly small, amounting to a range of only 0.71 cm from the shortest (16.26 cm for speaker A) to the longest (16.97 cm for speaker C). By comparison, the estimated vocal-tract lengths averaged over the vowels /i/, /e/, and /a/ spoken by eight adult, male speakers of Japanese was found by Ishizaki (1978a) to range from 16.7 cm to 18.4 cm.

As indicated by the  $k_s$  values, normalisation of the speaker differences in vocal-tract structure entails uniform lengthening of all the area-functions of speakers A and B by 1.8 % and 1.4 %, respectively, and uniform shortening of all the area-functions of speakers C and D by 2.4 % and 0.7 %, respectively. Whilst those length-scaling factors are admittedly quite small compared with the commonly accepted average difference of about 15-20% between the vocal-tract lengths of adult males and females (as found, e.g., in the X-ray studies of Chiba and Kajiyama (1941) and Högberg (1995), or in Fant's (1966) physiological interpretation of inter-gender differences in vowel formant distributions), nor is there any reason to expect larger differences in the vocal-tract anatomical sizes of our four speakers, as they are all adult males. However, a more relevant issue in regard to our articulatory explanation of the vowel-speaker dichotomy, is the extent to which those vocal-tract structural differences play a part in the spectral manifestation of the dichotomy as unfolded in Chapter 4. In particular, we seek the contributions of speaker differences in vocal-tract structure (as revealed in Table 6.1) to the overall drop in inter-speaker vowel classification accuracy observed across the higher spectral regions.

In order to perform the relevant vowel classification experiments, we require acoustic data which contain the speaker differences in vocal-tract fixed structure alone. Towards that end, we first carry out speaker normalisation of all three articulatory

features (structure, setting, and strategy) as described in Section 6.3; then re-introduce the speaker differences in vocal-tract structure alone, by uniform un-normalisation of the area-function lengths of each speaker, according to the scaling factors listed in Table 6.1. The formants of each of the resulting area-functions are then obtained by LP synthesis, using either 8 or 10 vocal-tract sections depending on the number used originally to estimate that area-function (as explained earlier, in Section 6.2). In order to maintain consistency with our earlier assumptions regarding simplified cepstra generated for the FC dataset and used in vowel classification experiments (cf. Section 4.4.3), only the first three, synthesised formant frequencies are therefore used to compute  $NCC = 14$  simplified LP cepstral coefficients (as described in Section 3.4.2), assuming the formant bandwidths to be fixed at their respective, original mean values computed over all eleven vowels and four speakers ( $\bar{B}_1 = 99$  Hz,  $\bar{B}_2 = 128$  Hz, and  $\bar{B}_3 = 218$  Hz).

The resulting behaviour of inter-speaker vowel classification accuracy across the available spectral range is shown by the solid curve in Figure 6.1, superimposed with both the inter- and intra-speaker accuracy curves obtained earlier using the original, simplified cepstra (cf. Figure 4.13). As expected, the inter-speaker accuracies obtained using the simplified cepstra generated from the partially speaker-normalised articulatory data are consistently higher than those obtained using the original, simplified cepstra. For example, at full spectral range (5000Hz), the acoustic data which bear the influence of only the vocal-tract structural differences between the speakers yield 85.6% — clearly higher than the original 77.5%, and understandably lower than the 90.8% attained by the intra-speaker curve.

Surprisingly, as the upper spectral limit is extended across the region from 1380Hz to 1900Hz, the accuracies slightly exceed those obtained along the intra-speaker curve. On face value, this might be taken to imply that speaker differences in vocal-tract fixed structure are slightly beneficial to vowel classification across that frequency sub-band. However, the method of data-partitioning used in the inter-speaker experiment yields a larger number of training samples per vowel, which itself may lead to an improved performance in classification (as found in connection with the quadratic classifier in Chapter 4), even more so if the vocal-tract structural differences amongst

our four speakers are insufficient to counteract the presumed improvement in accuracy. In this vein, it is interesting to note that the higher accuracies occur across a spectral range where vowel-speaker interactions are yet to manifest their full detrimental influence.

Indeed, the solid curve of vowel classification accuracy shown in Figure 6.1 clearly indicates that even those small differences in vocal-tract structure (as identified in Table 6.1) are sufficient to induce vowel-speaker interactions across the higher spectral regions, which then cause the drop in classification accuracy from a peak of 89.4% at 1580Hz, to 85.3% as the spectral range is extended further to 3140Hz. As the detrimental influence of inter-speaker variability is caused only by vocal-tract structural differences between the speakers, the drop of 4.1% in accuracy is naturally smaller than the drop of 7.2% observed in the original, inter-speaker curve. Nevertheless, the contrast between the dichotomous behaviour on the one hand, and the nearly-asymptotic, intra-speaker curve on the other, does clearly underscore the contributions of vocal-tract fixed structure to the vowel-speaker dichotomy.

Analogously to our acoustic-phonetic explanation of the dichotomy in Chapter 4, a more informative perspective on the contributions of vocal-tract structure to the dichotomy is gained by plotting the relevant portions of the inter-speaker accuracy curve (the solid curve in Figure 6.1) adjacent to the abscissa and ordinate of the  $F_1F_2$  and the  $F_2F_3$  vowel formant distributions used to generate the simplified cepstra. Indeed, the acoustic-phonetic relevance of the spectral regions of primary phonetic influence is clearly indicated in Figure 6.2, by the continuous rise in vowel classification accuracy as the upper spectral limit is extended across the entire  $F_1$  range, and across the  $F_2$  range of the back vowels. The peak in classification accuracy at 1580Hz (marked by the horizontal, dashed line in Figure 6.2, and by the horizontal and vertical, dashed lines in Figure 6.3) coincides almost perfectly with the mid- $F_2$  of our four speakers' vowel space, as defined earlier in terms of the mean  $F_2$  of the quasi-neutral vowel /ɜ/. The acoustic-phonetic relevance of the vowel-speaker interactions induced by speaker differences in vocal-tract structure is then revealed in both Figures 6.2 and 6.3, by the gradual drop in classification accuracy as the spectral range is extended further to include first the  $F_2$  of the front vowels, and later the  $F_3$  of both the back and

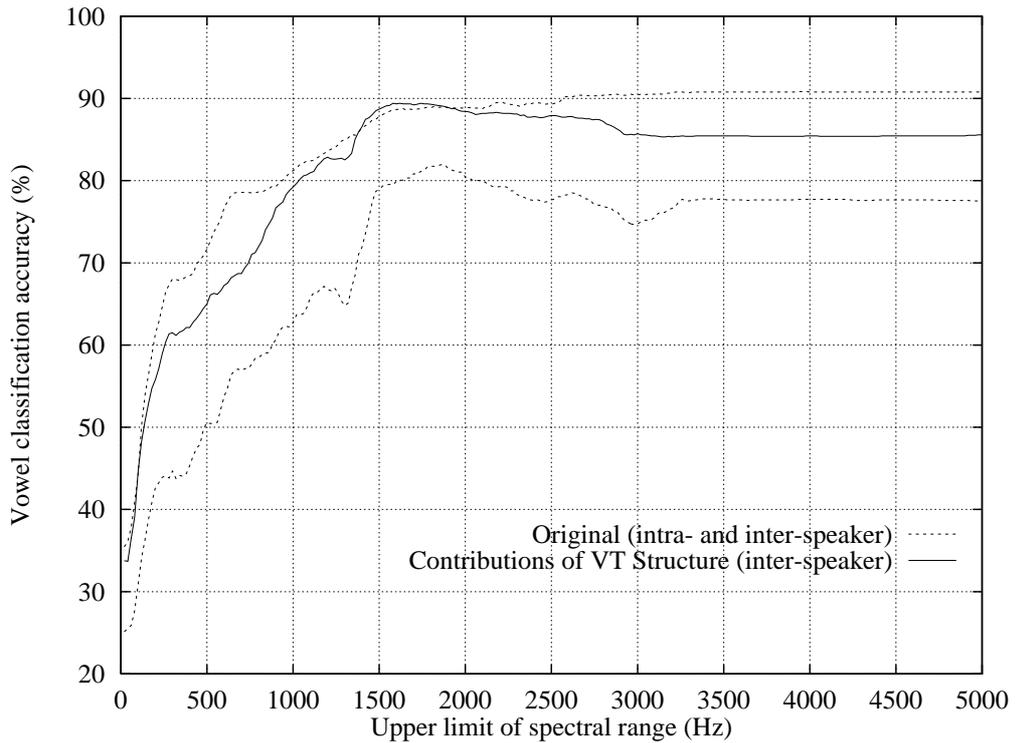


Figure 6.1: Contributions of speaker differences in vocal-tract fixed *structure*, to the dichotomous behaviour in inter-speaker vowel classification accuracy (FC dataset). The upper and lower, dashed curves are the intra- and inter-speaker accuracy curves, respectively, obtained in Chapter 4 (see Figure 4.13) using simplified cepstra generated from the first three, measured formant frequencies, with bandwidths fixed to mean values ( $B_1=99\text{Hz}$ ,  $B_2=128\text{Hz}$ ,  $B_3=218\text{Hz}$ ), sampling frequency  $F_s=10\text{kHz}$ , and  $NCC=14$ . The solid curve shows the behaviour of inter-speaker vowel classification accuracy obtained using simplified cepstra generated from the first three formant frequencies synthesised from the vocal-tract area-functions, after speaker normalisation of articulatory *setting* and vowel-specific articulatory *strategy* (with formant bandwidths fixed at the mean values listed above).

the front vowels.

Having thus confirmed the persistence of the dichotomy and its broad acoustic-phonetic explanation in the presence of only vocal-tract anatomical differences between the speakers, it seems natural to enquire about the more detailed, acoustic-phonetic manifestations of those differences. In view of our definition of vocal-tract structure in terms of mean length, the acoustic-phonetic correlates of the related speaker differences are easily predicted. In particular, uniform scaling of the vocal-tract length of any area-function induces a uniform (but inverse) scaling of the formant frequencies. Moreover, equal percentage changes in the formant frequencies give rise to larger absolute changes (in Hz) in the higher than in the lower formants. Consequently, the individual vowel formant clusters are expected to exhibit relatively larger, inter-speaker variations along

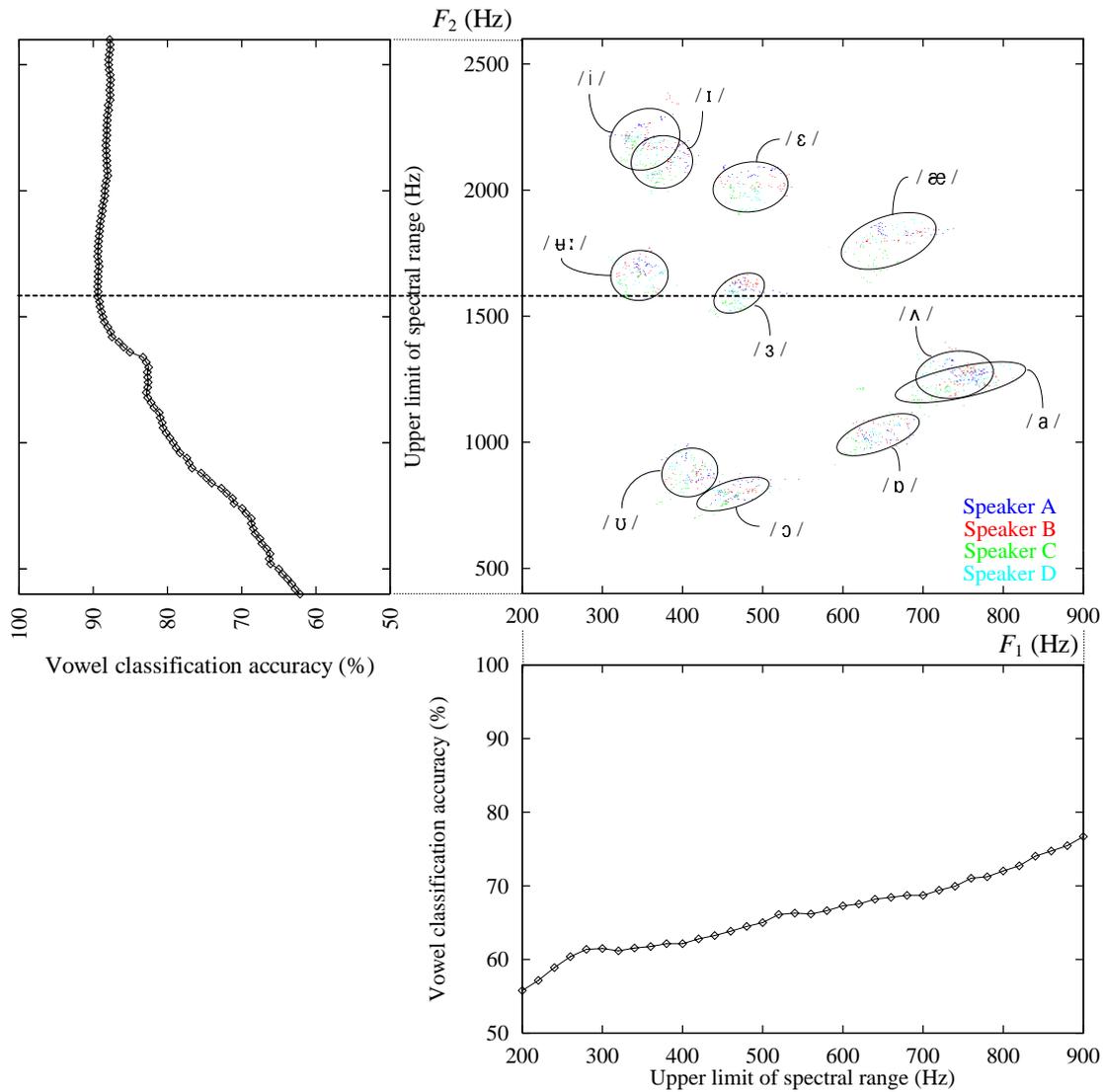


Figure 6.2:  $F_1F_2$  vowel space of all 4 male speakers (FC dataset), synthesised from their estimated area-functions after speaker normalisation, retaining differences in vocal-tract *structure* only. Each vowel cluster is shown with a  $2\sigma$  ellipse. Adjacent to the abscissa and ordinate are plotted the portions of the *inter*-speaker accuracy curve (solid line in Figure 6.1) which span the  $F_1$  and the  $F_2$  ranges, respectively. The horizontal (dashed) line intersects the accuracy curve at its peak (at 1580Hz), and cuts across the formant plane in order to emphasise the acoustic-phonetic relevance of the spectral regions of primary phonetic or speaker (VT *structure*) influence.

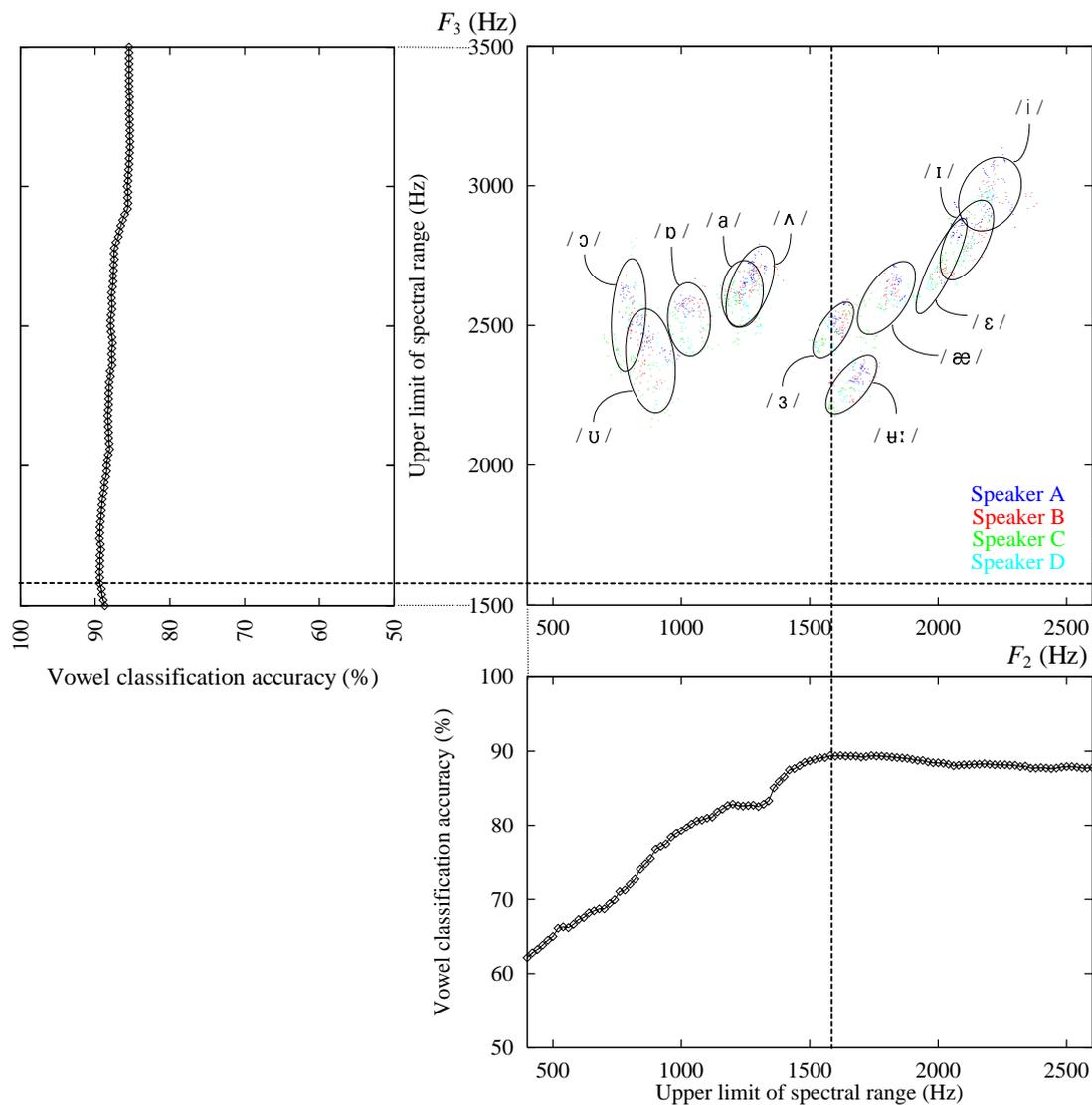


Figure 6.3:  $F_2F_3$  vowel space of all 4 male speakers (FC dataset), synthesised from their estimated area-functions after speaker normalisation, retaining differences in vocal-tract *structure* only. Each vowel cluster is shown with a  $2\sigma$  ellipse. Adjacent to the abscissa and ordinate are plotted the portions of the *inter-speaker* accuracy curve (solid line in Figure 6.1) which span the  $F_2$  and the  $F_3$  ranges, respectively. The vertical and horizontal (dashed) lines intersect the accuracy curve at its peak (at 1580Hz), and cut across the formant plane in order to emphasise the acoustic-phonetic relevance of the spectral regions of primary phonetic or speaker (VT *structure*) influence.

imaginary lines running through the origin, and the elongation along those radial lines is expected to be more pronounced in the  $F_2F_3$  than in the  $F_1F_2$  plane.

These predictions are indeed borne out by the vowel formant distributions shown in Figures 6.2 and 6.3. For example, although the vocal-tract structural differences amongst our four speakers were earlier shown to be relatively small, speaker C (who was found to have the longest mean length) clearly occupies the lower-left region of the ellipse drawn around each vowel cluster. Also, in comparison with the original formant distribution (cf. Figures 4.9 and 4.10), the ellipses are narrower in the direction perpendicular to imaginary, radial lines, and their elongation parallel to those radial lines is most clearly evident for the group of front vowels in the  $F_2F_3$  plane. Indeed, speaker normalisation of articulatory setting and vowel-specific articulatory strategy has not only reduced the overall size of each ellipse (while retaining a small degree of radial elongation), it has almost completely transformed the original, fan-shaped distribution of the front vowels in the  $F_2F_3$  plane, and replaced it with a more compact distribution which is elongated in the direction of increasing  $F_2$  and  $F_3$ . As the direction of elongation of the entire group of front vowels in that plane nearly coincides with the principal axis of each individual ellipse, confusions amongst the  $F_2$  and the  $F_3$  of neighbouring front vowels might be expected to contribute significantly to the drop in classification accuracy observed across the higher spectral regions.

Indeed, an acoustic-phonetic decomposition of the vowel misclassifications which contribute to the drop in accuracy across those higher spectral regions, reveals that the contributions arising from confusions amongst the  $F_2$  and amongst the  $F_3$  of the front vowels, are more significant than those arising from confusions amongst the  $F_3$  of the back vowels. As shown in Table 6.2, the two largest contributions are caused by the drop of 40% in the  $F_2$  range and 60% in the  $F_3$  range of the front vowel / $\epsilon$ / of speaker A due to confusions with / $\text{I}$ /, and the drop of 25% in the  $F_2$  range and 30% in the  $F_3$  range of the front vowel / $i$ / of speaker C, also due to confusions with the neighbouring vowel / $\text{I}$ /. Notwithstanding some minor differences in the exact nature of the vowel confusions, our acoustic-phonetic decomposition of the drop in accuracy in the original inter-speaker curve (see Table 4.2) showed that amongst the front vowels, / $\epsilon$ / of speaker A and / $i$ / of speaker C were indeed the main contributors to the

Vowel	Speaker			
	A	B	C	D
/i/			/ɪ/ (F <sub>2</sub> , 25%; F <sub>3</sub> , 30%)	/ɪ/ (F <sub>2</sub> , 5%)
/ɪ/	/i/ (F <sub>2</sub> , 15%)	/i/ (F <sub>2</sub> , 20%)	/ɛ/ (F <sub>2</sub> , 10%; F <sub>3</sub> , 15%)	/ɛ/ (F <sub>3</sub> , 5%)
/ɛ/	/ɪ/ (F <sub>2</sub> , 40%; F <sub>3</sub> , 60%)			
/æ/			/ɜ/ (F <sub>3</sub> , 15%)	
/a/	/ʌ/ (F <sub>3</sub> , 25%)	/ʌ/ (F <sub>3</sub> , 5%)		/ʌ/ (F <sub>3</sub> , 10%)
/ʌ/			/a/ (F <sub>3</sub> , 30%)	/a/ (F <sub>3</sub> , 5%)
/ɔ/				
/ɔ/			/ʊ/ (F <sub>3</sub> , 20%)	
/ʊ/	/ɔ/ (F <sub>3</sub> , 5%)			/ɔ/ (F <sub>3</sub> , 25%)
/ɘ/				
/ɜ/				

Table 6.2: Acoustic-phonetic decomposition of the dichotomy in inter-speaker vowel classification behaviour (solid curve in Figure 6.1), showing the contributions of speaker differences in vocal-tract *structure*, in terms of the vowel misclassifications that contribute to the drop in accuracy across the higher spectral regions which encompass the high- $F_2$  and the  $F_3$  of the speakers’ vowel formant distribution.

dichotomy, even when the data had not yet been subjected to speaker normalisation. Our present results therefore suggest that the main contributions of vocal-tract structure to the overall dichotomy are the confusions amongst the  $F_2$  and amongst the  $F_3$  of the front vowels of speakers A and C, whom we have earlier identified, respectively, with the shortest and the longest, structure-related vocal-tract lengths amongst the four adult, male speakers of the FC dataset.

#### 6.4.2 Contributions of Articulatory Setting to the Dichotomy

In Section 6.3.2 we described the steps required to compute each speaker’s overall mean vocal-tract shape, or articulatory setting. Briefly, the structure-normalised area-functions of each speaker are first aligned to minimise inter-frame and inter-repetition variations in shape, then centre-aligned across vowels (and speakers) at the mid-length of the MOL region determined for each group of frame- and repetition-aligned area-functions. The mean vocal-tract shape is then computed for each speaker, by averaging (in the logarithmic-area domain) across all frames, repetitions, and vowels, at every vocal-tract section (of length  $\Delta x = 0.05$  cm) where there exists an aligned area-function. As also pointed out in Section 6.3.2, any discontinuities at the glottal and labial ends of

the mean area-functions thus found are retained in subsequent computations, without adverse effect. However, for purely aesthetic reasons, those raw, mean vocal-tract shapes are first re-parameterised across their entire lengths, and only the resulting, smoothed area-functions are here displayed.

Figure 6.4 shows superimposed the four, mean vocal-tract shapes thus obtained. The origin on the abscissa is defined as the overall mean position of the glottis, which lies a distance equal to half the overall mean length (cf. Table 6.1) to the left of the central landmark at which all groups of area-functions are aligned; the vertical, dotted line on the right marks the overall mean position of the radiating plane at the lips, which itself is 17.51 cm from the mean position of the glottis. Each speaker's mean vocal-tract shape extends beyond those average glottis and lip markers, as far as there exists any individual, aligned area-function of that speaker.

The mean vocal-tract shapes shown superimposed in Figure 6.4 do reveal long-term idiosyncracies which may be interpreted in terms of articulatory settings adopted by each of our four speakers. First, it appears that the bulk of speaker differences in articulatory setting are manifest in approximately the front half of the vocal-tract. Indeed, the mean vocal-tract shapes are quite similar up to about 6 cm from the average glottal position, which corresponds perhaps to a large part of the pharynx. Although the root of the tongue, the lower part of the tongue body, and even the muscles of the pharynx walls are capable of either constricting or expanding the cross-sectional areas in that part of the vocal-tract on a long-term basis, thereby giving rise to so-called "pharyngeal settings" (Laver, 1980), Stevens and House's (1955, p.485) observations of X-ray data presented in the literature "suggest ... that the first few centimeters of the tract above the glottis change through only a restricted range of cross-sectional areas in comparison with other portions of the tract." Consequently, that part of the vocal-tract (which includes the so-called larynx tube) is often modelled with a fixed cross-sectional area profile and a fixed length of about 2 cm (e.g., Stevens and House, 1955; Fant, 1960; Flanagan et al., 1980; Carré and Mrayati, 1991). To the extent that any significance may be given to an anatomical (as opposed to a behavioural) interpretation of that part of the vocal-tract, the apparent similarity in our speakers' mean area-functions up to about 6 cm from the glottis is therefore consistent with our findings in

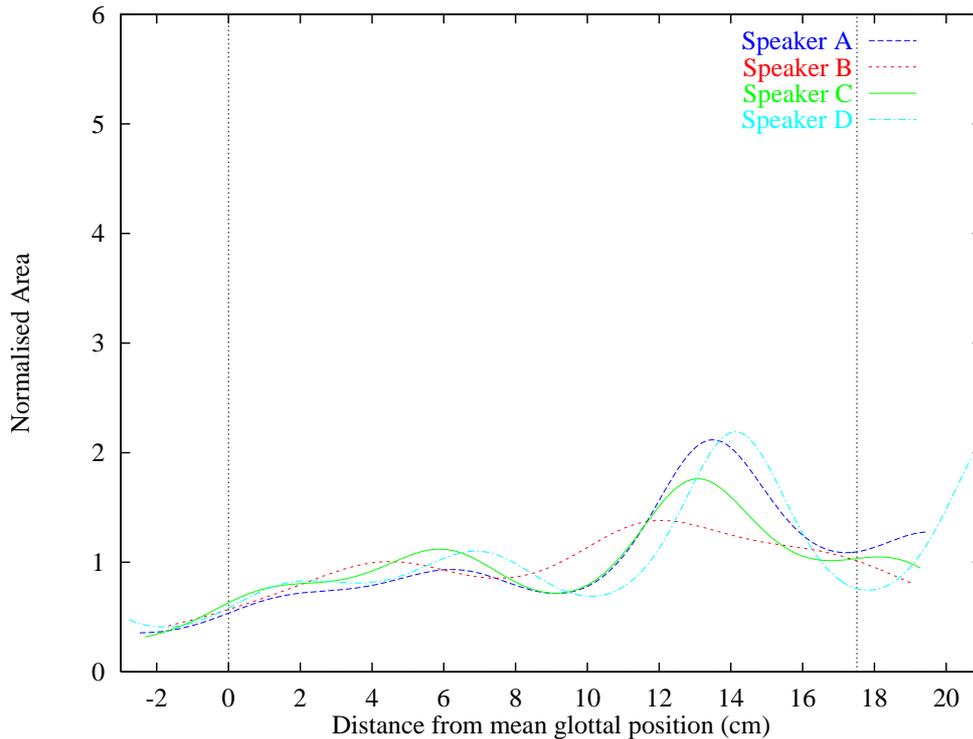


Figure 6.4: Speaker differences in articulatory *setting*. Shown superimposed are the 4 speakers’ mean VT-shapes computed over all their vowels after speaker normalisation of VT *structure*, cross-frame and -repetition alignment, and centre-alignment of all area-functions across vowels and speakers. The vertical, dotted lines on the left and the right indicate, respectively, the overall mean position of the glottis (which defines the origin on the abscissa) and of the lips (at 17.51 cm). The *setting* area-functions extend beyond those lines as far as there exists any single, aligned area-function of that speaker.

the previous Section, of the fairly similar vocal-tract anatomical sizes of these four speakers.

By contrast, the middle and front parts of the mean vocal-tract shapes shown in Figure 6.4 suggest a wider range of speaker differences in articulatory setting in the upper pharynx and oral parts of the vocal-tract. As the shapes embody, by definition, articulatory properties which are shared by all the vowels of each speaker, they are naturally less eccentric (or less constricted) compared with most of the individual area-functions of which they are the average. Nevertheless, if we consider the acoustically more salient, local minima in the cross-sectional area profiles, there appear two main articulators responsible for the locations of long-term constriction — the tongue itself, which determines the shape of the settings near the mid-length of the vocal-tract; and the lips, which determine the (latitudinal) settings towards the very front of the vocal-tract.

The location of the mean lingual constriction alone suggests a separation of the four speakers into three categories: speaker D exhibits the most anterior location at about 10.1cm from the mean position of the glottis; speaker B exhibits the most posterior location at about 7.5 cm from the mean glottal position; and speakers A and C share a location intermediate between those two extremes, at about 9.1cm from the glottis. Whilst the hard palate, the soft palate, and other landmarks along the length of the vocal-tract cannot be precisely located on our area-functions, we may nevertheless venture to label the observed lingual settings in terms of a standard vocal-tract anatomy of an adult male speaker, as depicted for example by Perkell (1969, p.6, Figure 2.3). Indeed, the same reference vocal-tract outline was also used by Mrayati and Carré (1988, p.224, Figure 9) to illustrate the physical relevance of the vocal-tract sections of their DRM model. If the total length along the vocal-tract airway from the glottis to the lips is assumed to be 17.5 cm (as found in Table 6.1), then the three distinct locations of constriction observed in Figure 6.4 might be interpreted as long-term tendencies towards *uvularised* (speaker B), *velarised* (speakers A and C), and *palato-velarised* (speaker D) lingual articulations, respectively. Whilst these terms are consistent both with the physiological theory of phonetics put forward by Peterson and Shoup (1966) and with Laver's (1980) framework for describing latitudinal, supralaryngeal settings which involve the tongue body, we should nevertheless proceed with caution, and avoid placing undue emphasis on such finely graded articulatory labels in an absolute sense. Our estimated articulatory data should rather be interpreted in terms of the overall trends of speaker differences in vocal-tract shape.

Apart from his relatively more fronted, average place of lingual constriction, Speaker D is further separated from the other speakers on the basis of his long-term (latitudinal) lip posture, which clearly appears more constricted. Although Laver (1980) provides quite a detailed ensemble of descriptors for latitudinal, labial settings, many of those detailed, physiological specifications are inapplicable to vocal-tract area-functions, which cannot by themselves discriminate, for example, between the vertical and horizontal dimensions of constriction or expansion. Speaker D might therefore be regarded simply as having a labial setting which involves a more pronounced degree of *lip rounding* in comparison with the other speakers.

In order to ascertain whether the speaker differences in articulatory setting thus far described are likely to have any contributions at all to the vowel-speaker dichotomy, we require acoustic data which contain only those setting-related differences. To this end, the original area-functions are again completely speaker-normalised with respect to all three articulatory features (structure, setting, and strategy) as described in Section 6.3, and each section of the resulting area-functions of each speaker is then divided by the appropriate area-scaling factor for that speaker (as found in step 7 in Section 6.3.2), thereby re-introducing the speaker differences in articulatory setting. The first three, LP-synthesised formant frequencies are then used with our fixed formant bandwidths to generate  $NCC = 14$  cepstral coefficients by the same methods described in the previous Section; and inter-speaker vowel classification experiments are repeated to yield another accuracy profile as a function of increasing spectral range.

The resulting behaviour of classification accuracy is shown by the solid curve in Figure 6.5, superimposed with the original intra- and inter-speaker accuracy curves first obtained in Chapter 4 (using simplified cepstra) and shown again in the previous Section (cf. Figure 6.1). At nearly every value of the upper spectral limit, the accuracies attained after retaining only speaker differences in articulatory setting are intermediate to those obtained in the original intra- and inter-speaker experiments. Indeed, the accuracy of 85.5 % attained at full spectral range (5000Hz) is nearly identical to the full-range accuracy obtained in the previous Section when only the speaker differences in vocal-tract structure were retained. The detrimental influence of speaker differences in vocal-tract structure and articulatory setting (as we have defined them) are therefore equal in magnitude when we consider the entire spectral range.

More importantly in terms of the contributions of articulatory setting to the dichotomy, the solid curve in Figure 6.5 does exhibit a quasi-dichotomous behaviour. Classification accuracy rises to a peak of 84.9 % at 1600Hz, then drops to 81.5 % as the spectral range is extended further to 2500Hz. Whilst this drop of 3.4 % is slightly less than the drop of 4.1 % found (in the previous Section) to be precipitated by speaker differences in vocal-tract structure, it does occur across a narrower spectral range. Classification accuracy then begins to rise as the spectral range is extended beyond 2500Hz, and attains a full-range value which is slightly higher than its earlier

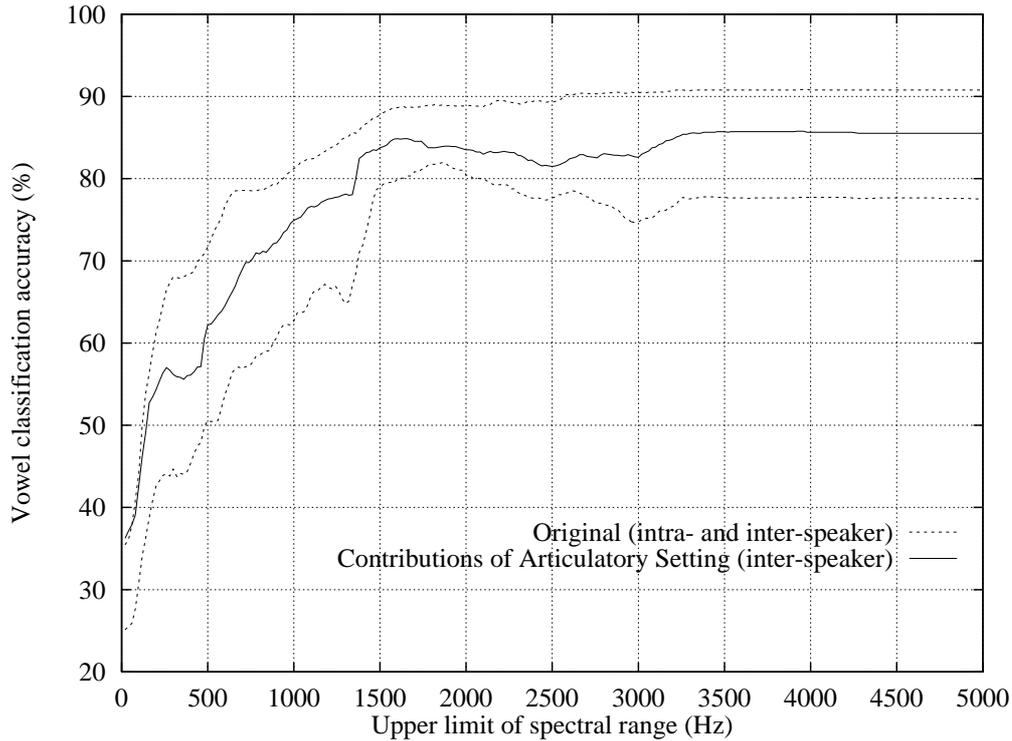


Figure 6.5: Contributions of speaker differences in articulatory *setting*, to the dichotomous behaviour in inter-speaker vowel classification accuracy (FC dataset). The upper and lower, dashed curves are the intra- and inter-speaker accuracy curves, respectively, obtained in Chapter 4 (see Figure 4.13) using simplified cepstra generated from the first three, measured formant frequencies, with bandwidths fixed to mean values ( $B_1=99\text{Hz}$ ,  $B_2=128\text{Hz}$ ,  $B_3=218\text{Hz}$ ), sampling frequency  $F_s=10\text{kHz}$ , and  $NCC=14$ . The solid curve shows the behaviour of inter-speaker vowel classification accuracy obtained using simplified cepstra generated from the first three formant frequencies synthesised from the vocal-tract area-functions, after speaker normalisation of vocal-tract fixed *structure* and vowel-specific articulatory *strategy* (with formant bandwidths fixed at the mean values listed above).

peak at 1600Hz.

An acoustic-phonetic interpretation of the quasi-dichotomous behaviour of the accuracy curve is again gained by plotting the relevant portions of that curve adjacent to the abscissa and ordinate of the  $F_1F_2$  and the  $F_2F_3$  vowel formant distributions from which the simplified cepstra were generated. As shown in Figure 6.6, classification accuracy rises across the entire  $F_1$  range, and continues to rise as the spectral range is extended across the  $F_2$  of the speakers' back vowels. The peak in accuracy occurs at very nearly the mid- $F_2$  of the formant distribution, as defined earlier in terms of the mean  $F_2$  of the quasi-neutral vowel /ɜ/. Classification accuracy then drops as the spectral range is extended across the entire  $F_2$  of the speakers' front vowels. Whilst the local minimum along the accuracy curve at 2500Hz indeed occurs beyond the highest

$F_2$ , in Figure 6.7 it is shown to occur near the middle of the speakers'  $F_3$  range. The subsequent improvement in classification accuracy is therefore attributed partly to the high- $F_3$  range of the speakers' back vowels, but perhaps more strongly to the entire  $F_3$  range of their four front vowels /i/, /ɪ/, /ɛ/, and /æ/, which are admittedly somewhat better separated than the back vowels along the  $F_3$  dimension.

Before we decompose the speaker-induced vowel misclassifications which have caused the drop in accuracy across the high  $F_2$  and the low  $F_3$  ranges, we should perhaps attempt to gain a better understanding of the acoustic-phonetic consequences of the speaker differences in setting, which we have described thus far only in articulatory terms. Some of the acoustic correlates of the lingual and labial settings described earlier in connection with Figure 6.4, were indeed discussed by Laver (1980). For example, a slightly retracted tongue body (as in the uvularised setting of speaker B) ought to lower the  $F_2$ , and the implied contraction in the uvular region should tend to raise the  $F_3$  of back vowels (Fant, 1975b). On the other hand, a relatively more raised lingual bias (as in the palato-velarised setting of speaker D) is generally expected to raise the  $F_2$ , particularly in the front vowels where the acoustic sensitivity in the palatal region is enhanced by the local constriction. By contrast, lip rounding (as also exhibited by speaker D) generally lowers all formant frequencies, with a more pronounced effect on higher formants (Lindblom and Sundberg, 1971); it should therefore lower the  $F_3$  in particular, and tend to counteract any rise in  $F_2$  of back vowels caused by concomitant palato-velarisation.

The formant distributions shown in Figures 6.6 and 6.7 allow us to verify these predictions of the acoustic correlates, as the formants were synthesised from the area-functions following the speaker normalisation procedure described earlier, which retains speaker differences in articulatory setting alone. The  $F_2F_3$  plane (Figure 6.7) in particular highlights the setting-related idiosyncrasies of speakers B and D as discussed above. For example, the formant distribution of speaker B appears to form an upper layer in  $F_3$ , especially in the back vowels (as predicted by the uvular contraction). Whilst it is tempting to attribute the lower  $F_2$  of that speaker's high front vowels to the presumed tongue body retraction, recall from Figure 6.4 that speaker B has a more constricted setting just anterior to the vocal-tract region where those vowels are likely

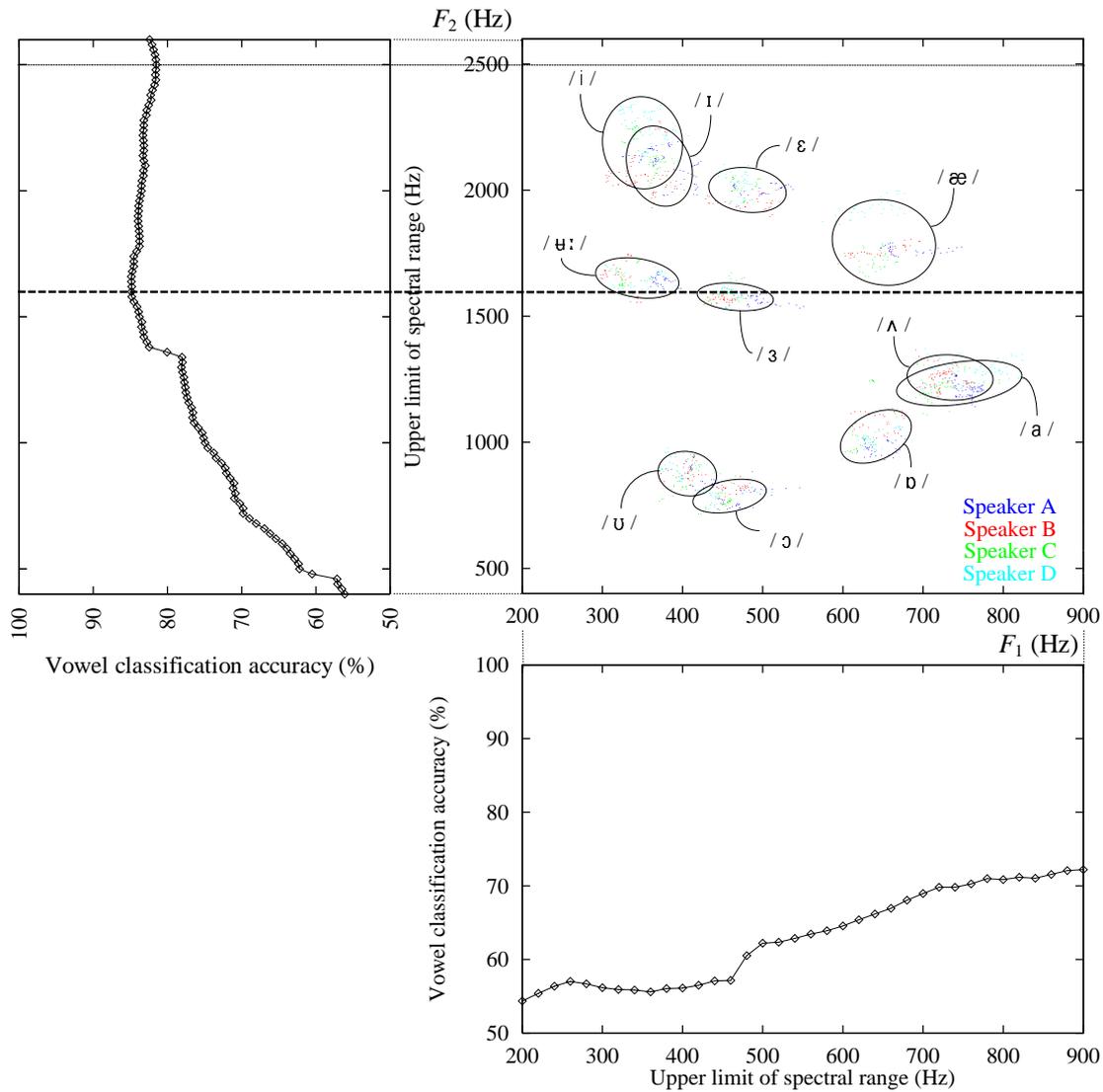


Figure 6.6:  $F_1F_2$  vowel space of all 4 male speakers (FC dataset), synthesised from their estimated area-functions after speaker normalisation, retaining differences in articulatory *setting* only. Each vowel cluster is shown with a  $2\sigma$  ellipse. Adjacent to the abscissa and ordinate are plotted the portions of the *inter*-speaker accuracy curve (solid line in Figure 6.5) which span the  $F_1$  and the  $F_2$  ranges, respectively. The horizontal, heavy-dashed line intersects the accuracy curve at its peak (at 1600Hz); the horizontal, light-dashed line intersects the curve at its local minimum (at 2500Hz). Those lines cut across the formant plane in order to emphasise the acoustic-phonetic relevance of the spectral regions of primary phonetic or speaker (articulatory *setting*) influence.

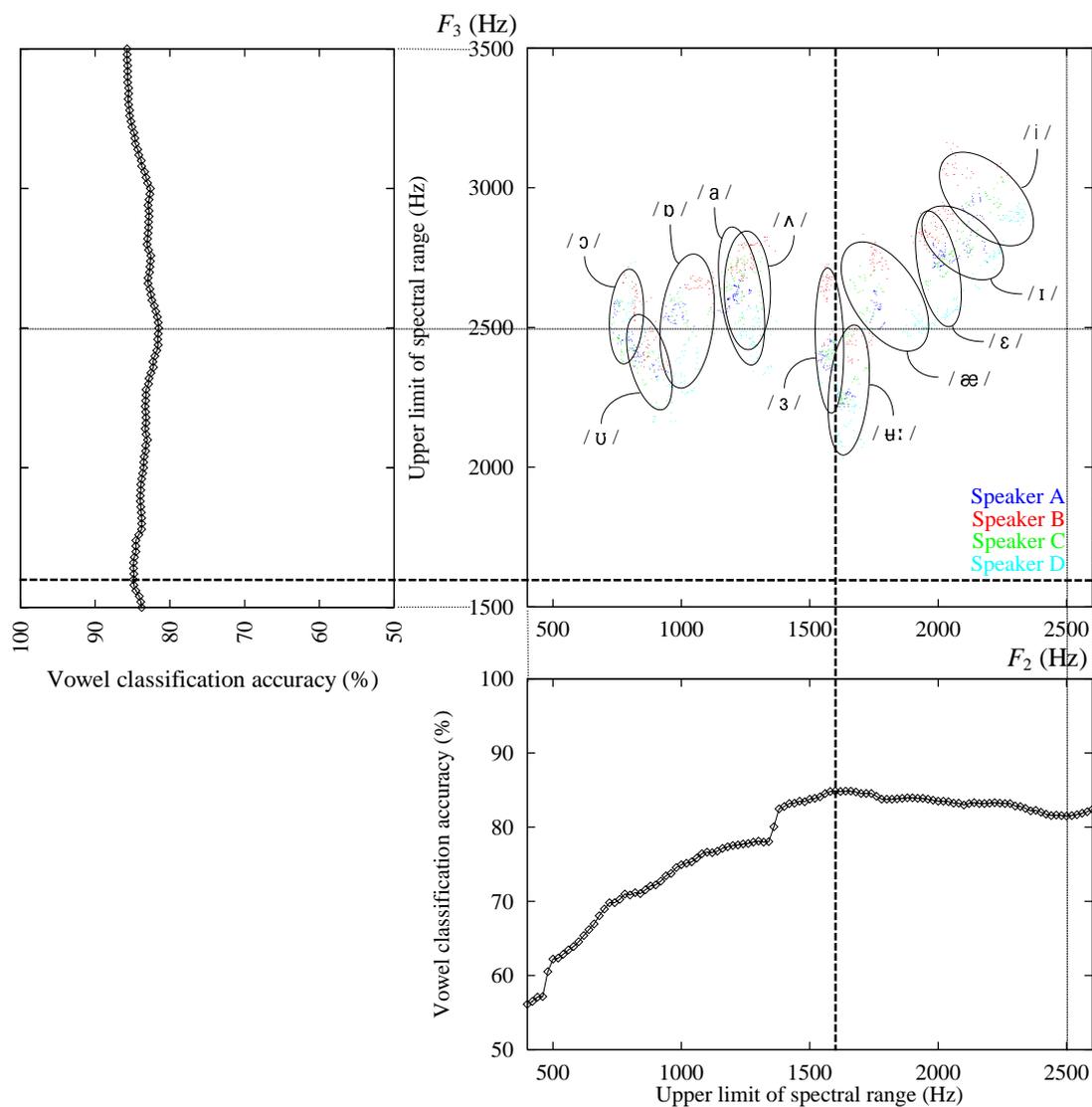


Figure 6.7:  $F_2F_3$  vowel space of all 4 male speakers (FC dataset), synthesised from their estimated area-functions after speaker normalisation, retaining differences in articulatory *setting* only. Each vowel cluster is shown with a  $2\sigma$  ellipse. Adjacent to the abscissa and ordinate are plotted the portions of the *inter-speaker* accuracy curve (solid line in Figure 6.5) which span the  $F_2$  and the  $F_3$  ranges, respectively. The vertical and horizontal, heavy-dashed lines intersect the accuracy curve at its peak (at 1600Hz); the vertical and horizontal, light-dashed lines intersect the curve at its local minimum (at 2500Hz). Those lines cut across the formant plane in order to emphasise the acoustic-phonetic relevance of the spectral regions of primary phonetic or speaker (articulatory *setting*) influence.

to have their main place of linguo-palatal constriction (from about 12cm to 15cm from the glottis). The lowering of  $F_2$  and the concurrent raising of  $F_3$  observed in Figure 6.7 is therefore explained by the effective lengthening of the back cavity and shortening of the front cavity, respectively, which would arise from the resulting, more anterior place of constriction in those high front vowels.

By contrast, the formant distribution of speaker D appears to form a lower layer in  $F_3$ , as predicted by the greater degree of lip rounding observed in Figure 6.4. At the same time, the presumed tendency to raise the tongue body indeed causes a higher  $F_2$  in the front vowels, especially in /æ/ (see also the  $F_1F_2$  distribution in Figure 6.6), whose constriction near the palato-velar region of the vocal-tract (Peterson and Shoup, 1966, p.45, Figure 2) would tend to be further constricted by the palato-velarisation suggested in Figure 6.4 for speaker D.

In comparison with the  $F_2F_3$  formant distribution synthesised from the area-functions containing speaker differences only in vocal-tract structure (cf. Figure 6.3 in the previous Section), each of the front vowel clusters shown in Figure 6.7 is elongated in a direction perpendicular to radial lines, thus clearly reflecting the speaker differences in long-term vocal-tract shape rather than length. In view of these clearly identifiable acoustic-phonetic differences which are also justified articulatorily, the question then arises to what extent the speaker differences described above induce the drop in classification accuracy observed across the high  $F_2$  and the low  $F_3$  ranges.

An acoustic-phonetic decomposition of the vowel misclassifications which occur in those higher spectral regions is shown in Table 6.3. The largest contributions amongst the front vowels are the confusions of the vowel /i/ of speaker B with /ɪ/ (owing to the low  $F_2$  of the former, which we explained in terms of a more fronted place of lingual constriction), and of the vowels /ɪ/ and /æ/ of speaker D with /i/ and /ɛ/, respectively (owing to the higher  $F_2$ , which we explained in terms of a palato-velarised setting, or a more raised tongue body). The largest contribution amongst the back vowels arises from confusions of the vowel /ɑ/ of speaker C in the  $F_3$  range with the highly overlapping vowel /ʌ/. Amongst the more central vowels, the consistently high  $F_3$  values of speaker B cause confusions of his vowel /ɥ:/ with the quasi-neutral /ɜ/ — confusions which had also occurred when the original simplified cepstra were

Vowel	Speaker			
	A	B	C	D
/i/	/ɪ/ (F <sub>2</sub> , 15%)	/ɪ/ (F <sub>2</sub> , 40%)	/ɪ/ (F <sub>2</sub> , 10%)	
/ɪ/	/ɛ/ (F <sub>2</sub> , 25%)	/ɛ/ (F <sub>2</sub> , 15%)	/i/ (F <sub>2</sub> , 10%)	/i/ (F <sub>2</sub> , 30%)
/ɛ/				/ɪ/ (F <sub>2</sub> , 20%) /i/ (F <sub>2</sub> , 5%)
/æ/				/ɛ/ (F <sub>2</sub> , 25%)
/a/	/ɒ/ (F <sub>3</sub> , 5%)		/ʌ/ (F <sub>3</sub> , 35%)	/ʌ/ (F <sub>3</sub> , 5%)
/ʌ/		/a/ (F <sub>3</sub> , 5%)	/a/ (F <sub>3</sub> , 10%)	/a/ (F <sub>3</sub> , 5%)
/ɒ/		/a/ (F <sub>3</sub> , 15%)		
/ɔ/			/ʊ/ (F <sub>3</sub> , 5%)	
/ʊ/				
/ʉ/		/ɜ/ (F <sub>3</sub> , 20%)		
/ɜ/				

Table 6.3: Acoustic-phonetic decomposition of the dichotomy in inter-speaker vowel classification behaviour (solid curve in Figure 6.5), showing the contributions of speaker differences in articulatory *setting*, in terms of the vowel misclassifications that contribute to the drop in accuracy across the higher spectral regions which encompass the high- $F_2$  and the low- $F_3$  of the speakers' vowel formant distribution (from the peak at 1600Hz to the local minimum at 2500Hz).

used (as listed in Table 4.2).

Clearly, this acoustic-phonetic decomposition confirms our earlier observations that only the  $F_2$  of the front vowels and the  $F_3$  of the back (and central) vowels are responsible for the drop in accuracy which is induced by speaker differences in articulatory setting, and which was observed to occur between 1600Hz and 2500Hz. An acoustic-phonetic decomposition of the rise in accuracy beyond 2500Hz also confirms our earlier predictions concerning the role of setting-related differences in the  $F_3$  of the speakers' front vowels — improvements in classification accuracy occur mainly in the  $F_3$  range of the vowel /i/ of speaker B and the vowels /ɪ/, /ɛ/, and /æ/ of speaker D. This is in clear contrast with the results of the inter-speaker experiments performed in the previous Section, where it was found that both the  $F_2$  and the  $F_3$  of front vowels were the main contributors to the dichotomy induced by speaker differences in vocal-tract structure.

### 6.4.3 Contributions of Articulatory Strategy to the Dichotomy

Thus far we have accounted for the two, long-term articulatory sources of inter-speaker variability in vowel production, namely *structure*, which we defined in terms of the

mean vocal-tract length of the mid- to low vowels, and *setting*, which we defined in terms of the mean vocal-tract shape computed over all the vowels following two stages of area-function alignment. The contributions of vocal-tract structure to the dichotomy were found to arise mainly in the  $F_2$  and the  $F_3$  of the front vowels of speakers A and C, who themselves were found to have the smallest and the largest vocal-tract anatomical sizes, respectively, amongst our four speakers. By contrast, the contributions of articulatory setting to the dichotomy were found to arise mainly in the  $F_2$  of the front vowels and the  $F_3$  of the back and central vowels of the two speakers B and D, who had earlier been found to have vocal-tract anatomical sizes close to the average of our four speakers, and whose articulatory settings were then described in terms of long-term tendencies, respectively, towards uvularisation concomitant with a more forward place of constriction in front vowels, and towards palato-velarisation (or raising of the tongue body) concomitant with lip rounding.

Upon speaker normalisation of both structure and setting, the area-functions are assumed to be associated with a common, or reference vocal-tract fixed anatomy, and to have a common, or reference long-term articulatory bias. The remaining, or residual speaker differences can then be described in terms of both the lengths and shapes of the normalised area-functions on a per-vowel basis, and thus collectively regarded as idiosyncratic behaviour in vowel-specific articulatory *strategy*. Insofar as the articulators are capable of altering the effective length along the vocal-tract airway from the glottis to the radiating plane at the lips, residual differences in length can arise from idiosyncratic articulatory behaviour in the amount of lip protrusion and larynx lowering or raising, and perhaps to a lesser extent in the posture and position of the tongue. On the other hand, residual differences in shape are determined by the so-called latitudinal components of inter-speaker articulatory variability, which are measured perpendicular to the direction of airflow, and which involve the degrees of lingual and labial constriction, and the areas of the vocal-tract cavities and constrictions formed by the tongue and the opposing tract walls.

These residual differences in articulatory strategy are shown in Figure 6.8, where each of the eleven panels displays superimposed the per-vowel mean (or prototype) area-functions of our four speakers, as computed in step 1 of the algorithmic procedure

described in Section 6.3.3. The lengths of the displayed area-functions have thus been subjected to speaker normalisation of vocal-tract structure as described in Section 6.3.1 and later quantified in Table 6.1. Their shapes have likewise been subjected to speaker normalisation of articulatory setting, as described in Section 6.3.2 and discussed in the previous Section in connection with Figure 6.4. As in the latter, the left and right vertical (dotted) lines in each panel of Figure 6.8 indicate, respectively, the overall average position of the glottis (which defines the origin on the abscissa) and of the lip termination (at 17.51 cm).

Whilst the length-related speaker differences in articulatory strategy are conveyed graphically in Figure 6.8, they are also quantified in Table 6.4, which lists the mean, structure-normalised vocal-tract length of each speaker, on a per-vowel basis. In the last column are listed the speaker-averaged mean lengths for each vowel (cf. Equation 6.7 in Section 6.3.3), which then allow computation of the length-scaling factors  $k_{sv}$  (according to Equation 6.6) required for speaker normalisation of the length component of vowel-specific articulatory strategy (cf. Equation 6.5).

A brief examination of each group of superimposed area-functions in Figure 6.8 reveals relatively smaller amounts of inter-speaker variation in the lower parts of the vocal-tract which include the first 4 to 6 cm above the glottis. A similar observation was made in connection with the speakers' vowel-averaged vocal-tract shapes or settings (in the previous Section), in which context we offered an explanation in terms of the relative inflexibility of the lower part of the pharynx and the so-called larynx tube. It is therefore remarkable that the structure- and setting-normalised area-functions (in Figure 6.8) should also have relatively invariant shapes across speakers on a per-vowel basis, in that part of the vocal-tract which is more likely than other parts to be associated with the speakers' fixed anatomy. In this regard, it could well be argued that the plausibility of our estimated area-functions is partially supported by the absence of unrealistically large areas or expanded cavities in that lowest part of the vocal-tract which, as we have noted, is often modelled explicitly with a fixed length and cross-sectional area profile.

Rather, the bulk of speaker differences in articulatory strategy appear to be manifest in the vocal-tract regions which correspond to the entire front or oral cavity, and approximately the upper half of the pharynx. Speaker variations occur in the size of

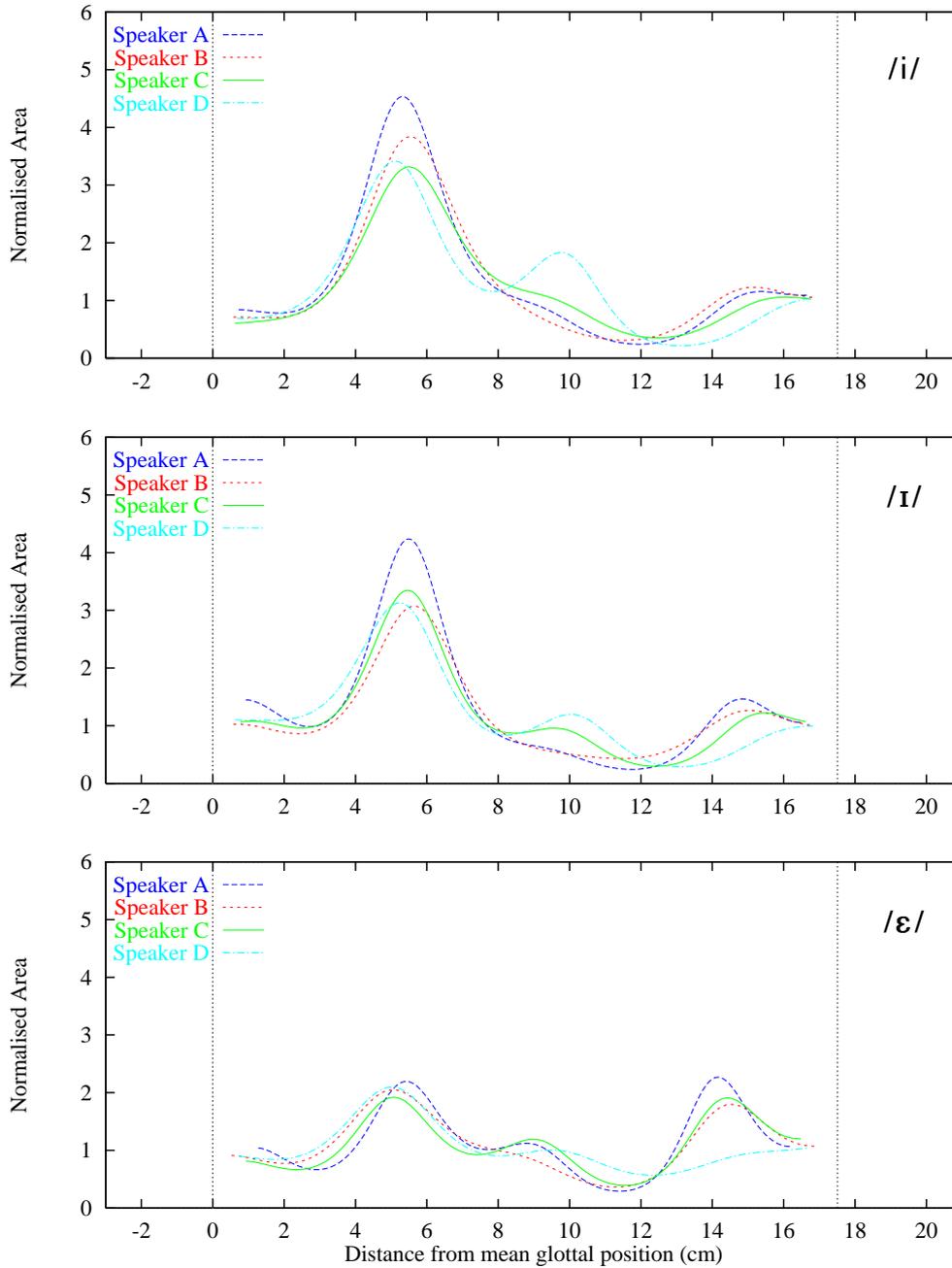


Figure 6.8: Speaker differences in vowel-specific articulatory *strategy* (4 pages of graphs). Shown superimposed for each of the 11 vowels, are the 4 speakers’ prototype area-functions obtained after speaker normalisation of VT *structure* and articulatory *setting*. The vertical, dotted lines shown in each graph on the left and the right indicate, respectively, the overall mean position of the glottis (which defines the origin on the abscissa) and of the lips (at 17.51 cm). Area-functions are aligned at the centre of the mutually overlapping region per vowel per speaker (as described in Section 6.3.2).

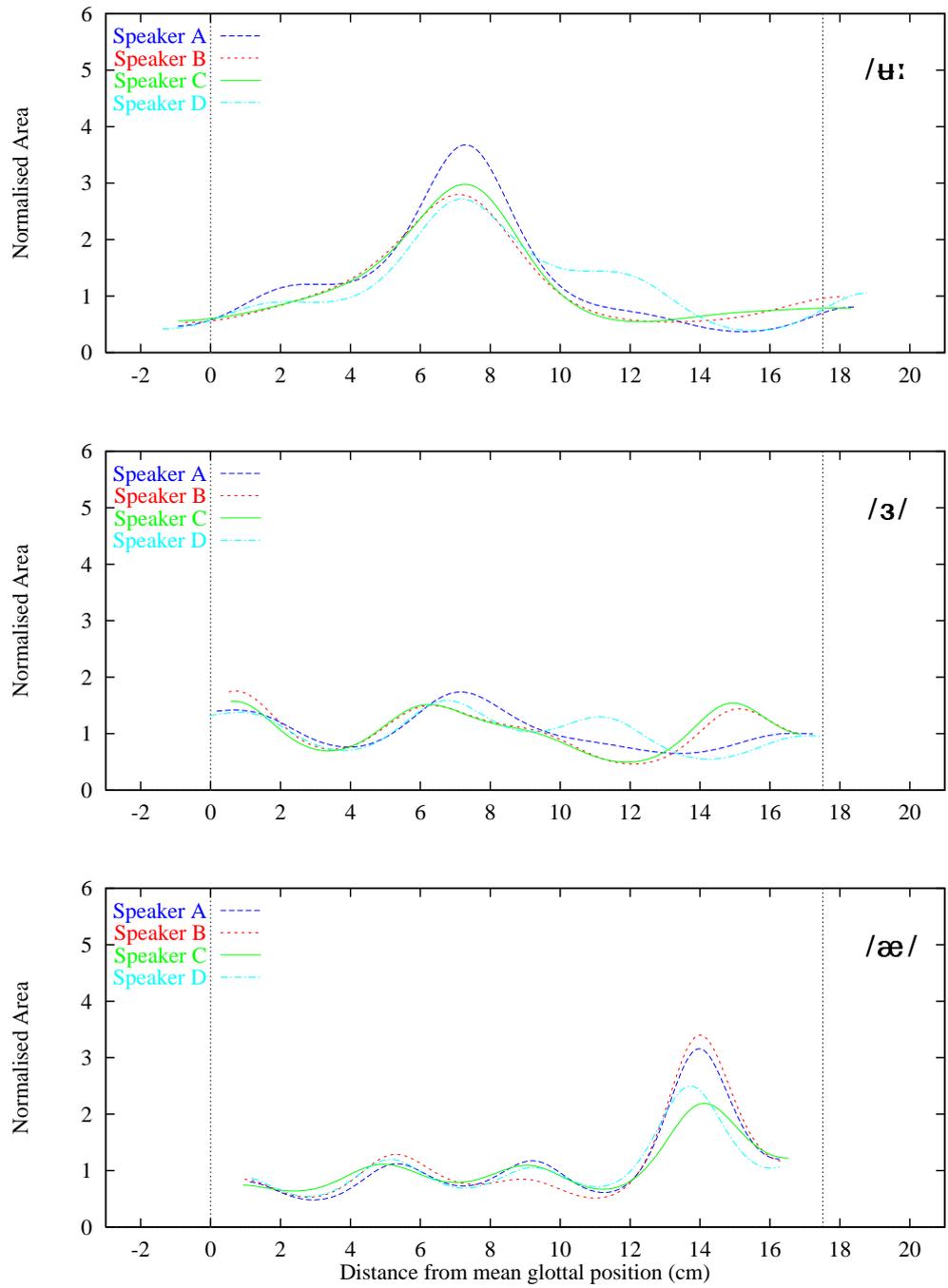


Figure 6.8: (continued from previous page).

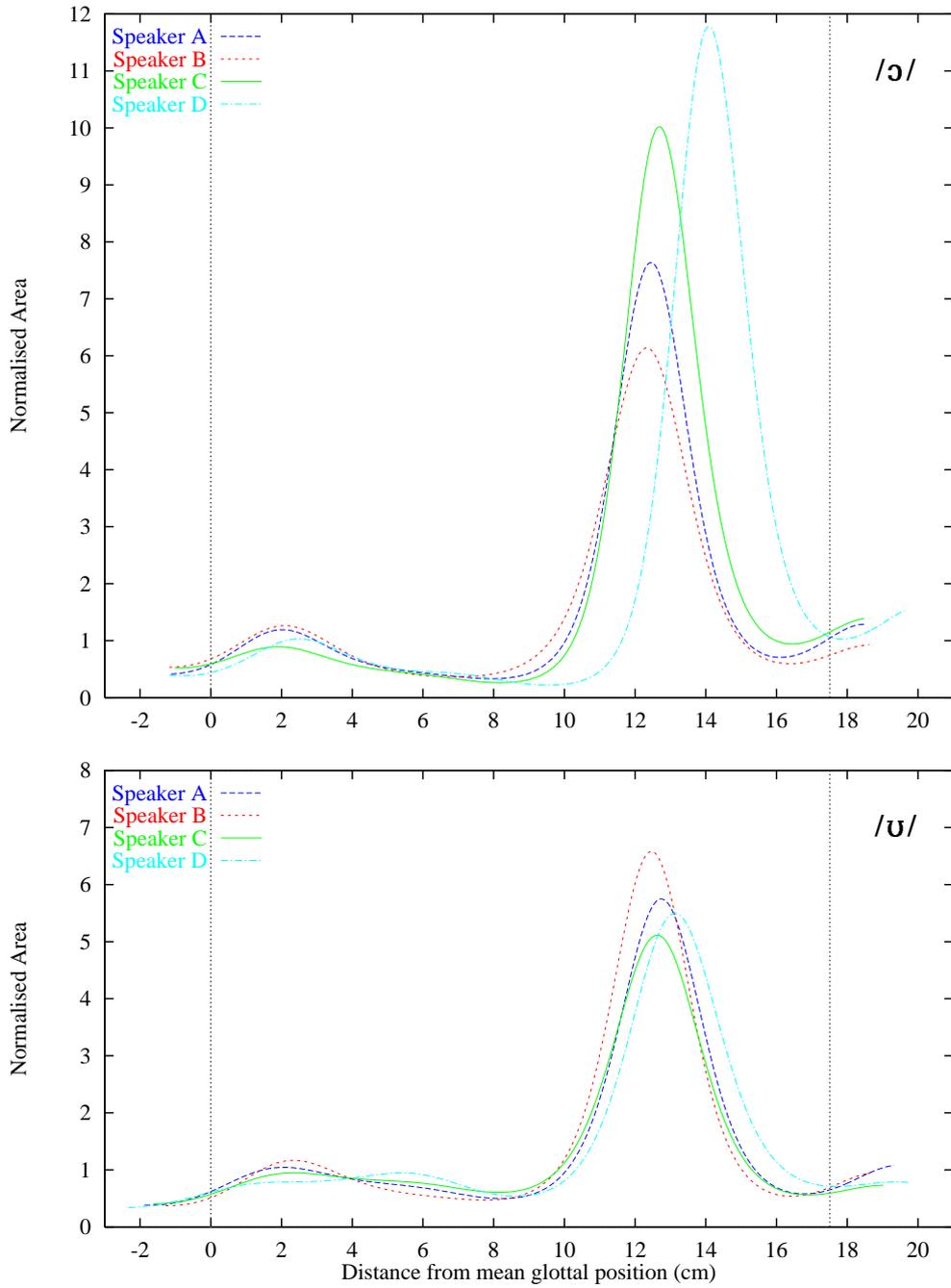


Figure 6.8: (continued from previous page).

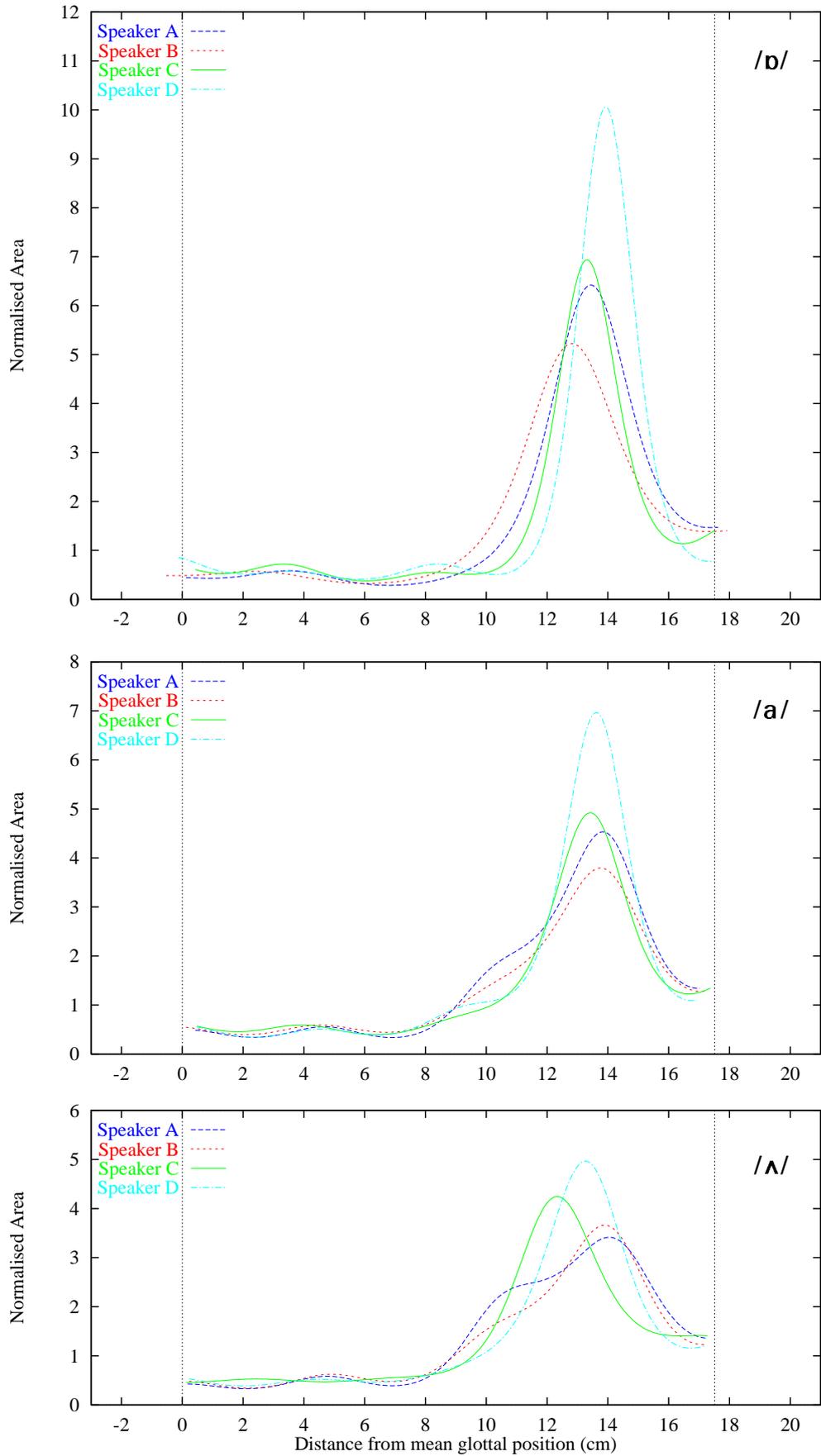


Figure 6.8: (continued from previous page).

expanded regions, in the degrees and locations of primary (or secondary) constriction, and in the amount of lip rounding and protrusion. For example, in the two, high front vowels /i/ and /ɪ/, speaker D appears to have a distinctly more forward place of primary lingual constriction, while that of speaker B is both longer and more retracted. In those vowels and also in the fronted /ɛ:/, speaker A appears to have a greater volume in the back cavity expansion compared with the other speakers. In the mid-front vowel /ɛ/, speaker D has quite a distinct vocal-tract configuration in the oral cavity, with a more fronted place but reduced degree of constriction; on the other hand, speaker A exhibits a more voluminous front cavity, a greater degree of constriction, and an overall shorter vocal-tract length (as quantified in Table 6.4). In the lip-rounded back vowels /ɔ/, /ʊ/, and /ɒ/, speaker D exhibits a more anterior place of lingual constriction, a greater amount of lip rounding, and larger areas in the front-cavity expansion; on the other hand, speaker B exhibits a more retracted place of constriction which also appears to extend over a longer part of the vocal-tract. Whilst the area-functions of all the back vowels suggest a fairly large amount of speaker variability in the shape and volume of the expanded, front cavity, it is doubtful that these types of physical differences will have a proportional impact in the acoustic domain, as the formant sensitivity to changes in cross-sectional area generally reduces with increasing area.

Indeed, it remains to be determined whether any of the speaker differences in articulatory strategy shown in Figure 6.8 and listed in Table 6.4, will contribute to the vowel-speaker dichotomy which is spectrally manifest in the behaviour of inter-speaker vowel classification accuracy. In order to perform the necessary vowel classification experiments, we require acoustic data which contain speaker differences in articulatory strategy alone. Consistently with the methodology adopted in Chapter 4 and in the previous two Sections, the first three formant frequencies LP-synthesised from the structure- and setting-normalised area-functions are therefore used with fixed formant bandwidths, to generate  $NCC = 14$  simplified cepstral coefficients. Those cepstra are then used to perform inter-speaker vowel classification experiments with the same increasing spectral range, as described in Section 4.2.3.1.

The resulting accuracy curve is shown in Figure 6.9 (solid curve), superimposed

Vowel		Speaker				Mean
		A	B	C	D	
/i/	VTL	16.04	16.36	16.21	16.15	16.19
	$k_{sv}$	<b>1.009</b>	<b>0.990</b>	<b>0.999</b>	<b>1.003</b>	
/ɪ/	VTL	15.67	16.32	15.94	16.27	16.05
	$k_{sv}$	<b>1.024</b>	<b>0.984</b>	<b>1.007</b>	<b>0.987</b>	
/ɛ/	VTL	15.00	16.42	15.61	16.10	15.79
	$k_{sv}$	<b>1.052</b>	<b>0.961</b>	<b>1.011</b>	<b>0.981</b>	
/æ/	VTL	15.22	15.51	15.69	15.17	15.40
	$k_{sv}$	<b>1.012</b>	<b>0.993</b>	<b>0.981</b>	<b>1.015</b>	
/a/	VTL	16.68	17.23	16.97	16.54	16.85
	$k_{sv}$	<b>1.010</b>	<b>0.978</b>	<b>0.993</b>	<b>1.019</b>	
/ʌ/	VTL	17.20	17.05	17.22	17.03	17.12
	$k_{sv}$	<b>0.996</b>	<b>1.004</b>	<b>0.994</b>	<b>1.006</b>	
/ɒ/	VTL	17.57	18.53	17.16	17.74	17.74
	$k_{sv}$	<b>1.010</b>	<b>0.958</b>	<b>1.034</b>	<b>1.001</b>	
/ɔ/	VTL	19.73	19.89	19.60	20.88	20.02
	$k_{sv}$	<b>1.015</b>	<b>1.007</b>	<b>1.022</b>	<b>0.959</b>	
/ʊ/	VTL	21.26	20.08	20.71	22.18	21.06
	$k_{sv}$	<b>0.991</b>	<b>1.048</b>	<b>1.017</b>	<b>0.950</b>	
/ʌɪ/	VTL	19.40	18.94	19.33	20.21	19.47
	$k_{sv}$	<b>1.004</b>	<b>1.028</b>	<b>1.007</b>	<b>0.963</b>	
/ɜ/	VTL	17.15	16.45	16.36	17.51	16.86
	$k_{sv}$	<b>0.984</b>	<b>1.025</b>	<b>1.031</b>	<b>0.963</b>	

Table 6.4: Mean VT-lengths (cm) obtained after speaker normalisation of VT *structure*, and the factors required to speaker normalise length-related differences in vowel-specific articulatory *strategy*. Shown for each of the 11 vowels, are the mean VT-lengths computed over all 7 frames and 5 repetitions per speaker, after normalisation of VT *structure* (cf. Table 6.1). The per-vowel mean VT-lengths are listed in the right-most column, and the per-speaker length-normalisation factors ( $k_{sv}$ ) required to speaker-normalise length-related differences in articulatory *strategy*, are listed in the rows directly below the lengths for each vowel. Shading indicates a significant contribution of the length-related differences in strategy to the vowel-speaker dichotomy (see text for detailed discussion).

with both the inter- and intra-speaker accuracy curves obtained using the original, simplified cepstra. Evidently, the speaker differences in vowel-specific articulatory strategy do significantly contribute to the vowel-speaker dichotomy. Indeed, of the three accuracy curves generated thus far in this Chapter and shown by the solid curves in Figures 6.1, 6.5, and 6.9, this latter curve is clearly the closest to the original, inter-speaker accuracy curve at every step along the spectral continuum.

At full spectral range (5000Hz), the accuracy of 78.1% is only marginally better than that of 77.5% obtained using the original, simplified cepstra. This indicates that

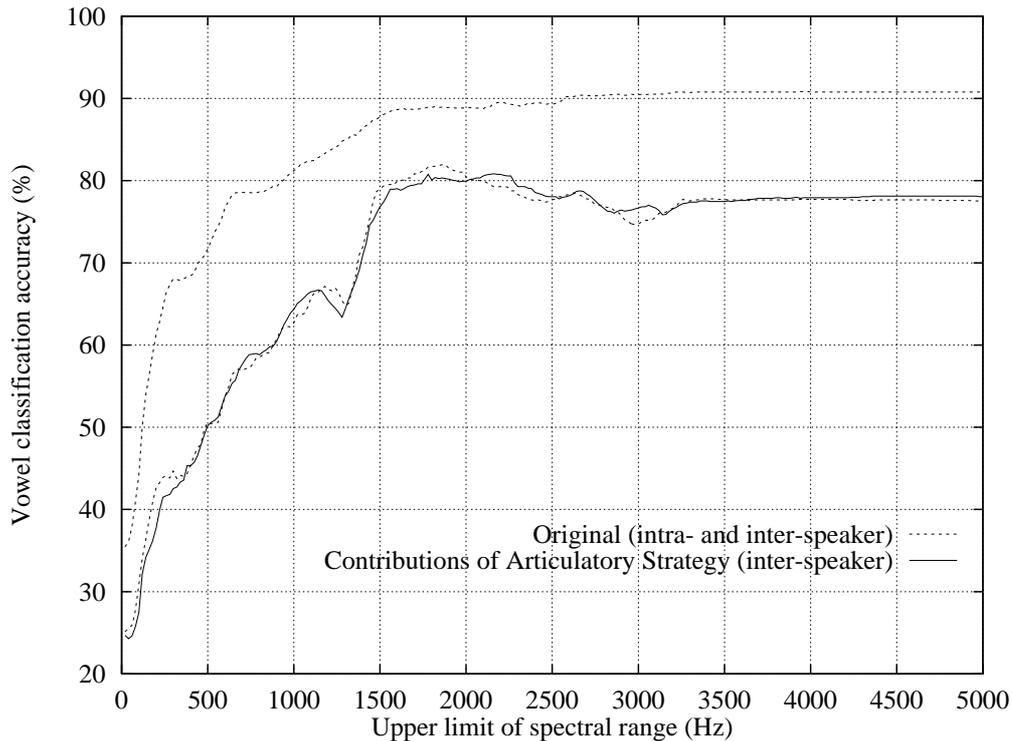


Figure 6.9: Contributions of speaker differences in vowel-specific articulatory *strategy*, to the dichotomous behaviour in inter-speaker vowel classification accuracy (FC dataset). The upper and lower, dashed curves are the intra- and inter-speaker accuracy curves, respectively, obtained in Chapter 4 (see Figure 4.13) using simplified cepstra generated from the first three, measured formant frequencies, with bandwidths fixed to mean values ( $B_1=99\text{Hz}$ ,  $B_2=128\text{Hz}$ ,  $B_3=218\text{Hz}$ ), sampling frequency  $F_s=10\text{kHz}$ , and  $NCC=14$ . The solid curve shows the behaviour of inter-speaker vowel classification accuracy obtained using simplified cepstra generated from the first three formant frequencies synthesised from the vocal-tract area-functions, after speaker normalisation of vocal-tract fixed *structure* and articulatory *setting* (with formant bandwidths fixed at the mean values listed above).

the detrimental influence of speaker differences in articulatory strategy (as defined in our modelling methodology) is, on the whole, nearly identical in magnitude to that of the three articulatory sources of speaker variability combined. By comparison, the accuracies attained at full spectral range in Figures 6.1 and 6.5 were about 8% higher, indicating that the overall detrimental influences of speaker differences in either structure or setting alone, are relatively lower in magnitude.

As far as vowel-speaker interactions are concerned in the higher spectral regions, a closer examination of Figure 6.9 reveals a slightly lower peak in accuracy, and a less pronounced drop in accuracy across the higher spectral regions. In particular, classification accuracy rises to a maximum of 80.8% as the spectral range is extended to 1780Hz, dips slightly to 79.9% at 1960Hz and rises again to 80.8% at 2160Hz,

then drops to 75.8% as the spectral range is extended further to 3140Hz. The most extensive drop in accuracy therefore amounts to 5.0% which, although less than the original drop of 7.2%, is still larger than the drop in accuracy attributed to structure- and setting-related speaker differences (4.1% and 3.4%, respectively). Amongst the three, articulatory sources of inter-speaker variability, vowel-specific articulatory strategy therefore appears to claim the largest contribution to the vowel-speaker dichotomy.

Whilst the sheer extent of the drop in accuracy across the higher spectral regions is a good indication of the magnitude of the contribution to the dichotomy, the appearance of a double-peak in accuracy which spans nearly 400Hz (from 1780Hz to 2160Hz) raises an important question regarding the acoustic-phonetic relevance of the respective spectral regions over which the presumed influences of vowel-speaker interactions play either a small, or a more substantial role. Naturally, that question is best addressed by plotting, as we are now accustomed to do, the relevant portions of the inter-speaker accuracy curve adjacent to the abscissa and ordinate of the synthesised  $F_1F_2$  (in Figure 6.10) and  $F_2F_3$  (in Figure 6.11) vowel formant distributions which were earlier used to generate the simplified cepstra.

In Figure 6.10, classification accuracy is shown to steadily rise across the four speakers' entire  $F_1$  range. That steady rise in accuracy is briefly interrupted in the low- $F_2$  range, by a local peak and a small drop in accuracy which is caused by confusions amongst the  $F_2$  of the highly overlapping vowels /a/ and /ʌ/ — interestingly, a similar dip in accuracy was observed in the original inter-speaker curve, and also in the two accuracy curves shown earlier in this Chapter. However, as the spectral range is extended further, classification accuracy rises substantially to the first of the two global peaks, at approximately the low  $F_2$  of the three, high front vowels. Indeed, as shown in both Figures 6.10 and 6.11, the spectral region between 1780Hz and 2160Hz where the small, speaker-induced drop in accuracy is matched by an equal rise back to the peak of 80.8%, encompasses the  $F_2$  of mainly the three front vowels /i/, /ɪ/, and /ε/. By comparison, a more substantial and persistent deterioration in classification performance occurs as the spectral range is extended beyond the second peak (at 2160Hz), and across nearly the entire  $F_3$  distribution.

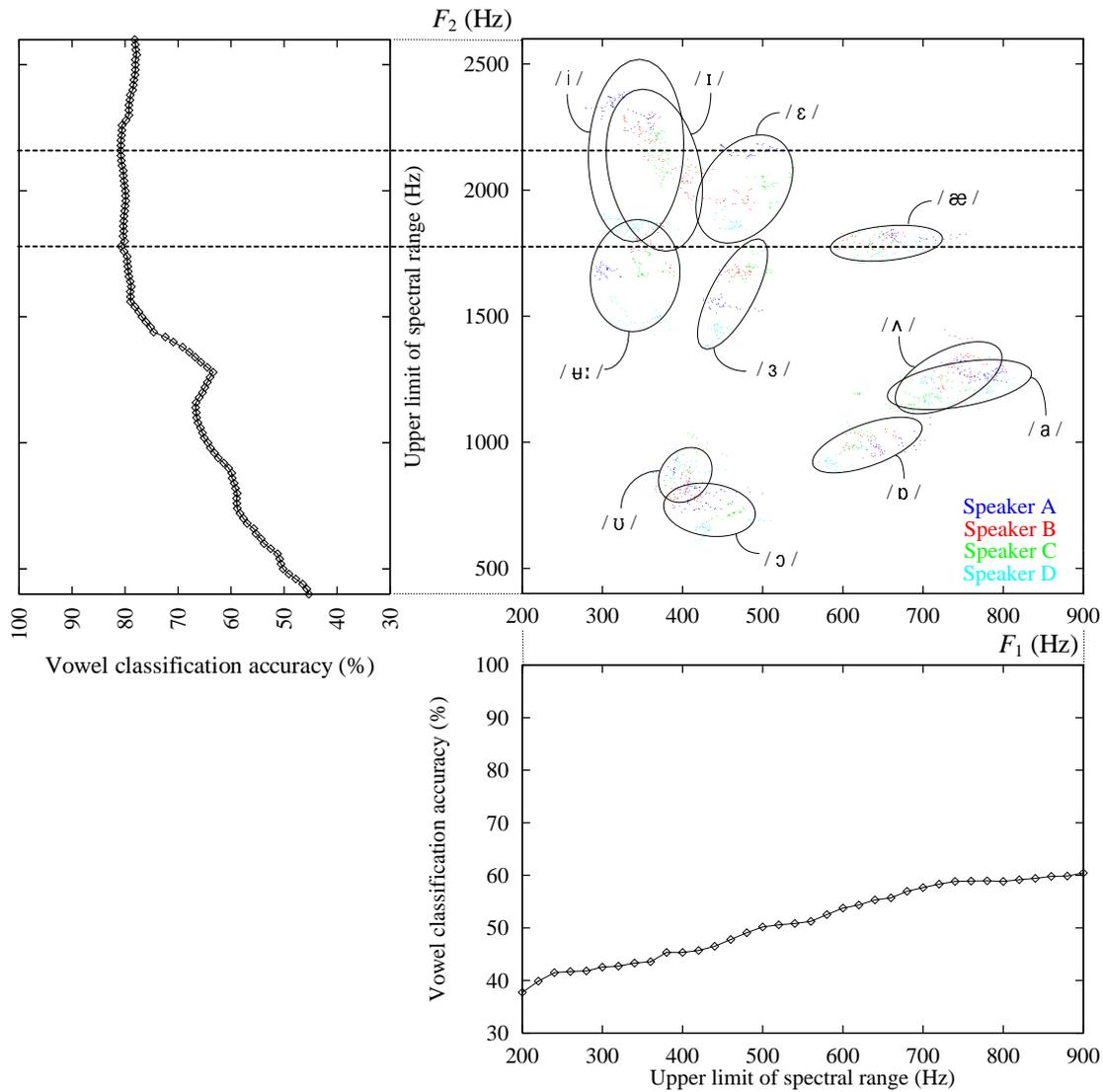


Figure 6.10:  $F_1F_2$  vowel space of all 4 male speakers (FC dataset), synthesised from their estimated area-functions after speaker normalisation, retaining differences in vowel-specific articulatory *strategy* only. Each vowel cluster is shown with a  $2\sigma$  ellipse. Adjacent to the abscissa and ordinate are plotted the portions of the *inter-speaker* accuracy curve (solid line in Figure 6.9) which span the  $F_1$  and the  $F_2$  ranges, respectively. The horizontal, dashed lines intersect the accuracy curve at its two peaks (at 1780Hz and 2160Hz, respectively), and cut across the formant plane in order to emphasise the acoustic-phonetic relevance of the spectral regions of primary phonetic or speaker (vowel-specific articulatory *strategy*) influence.

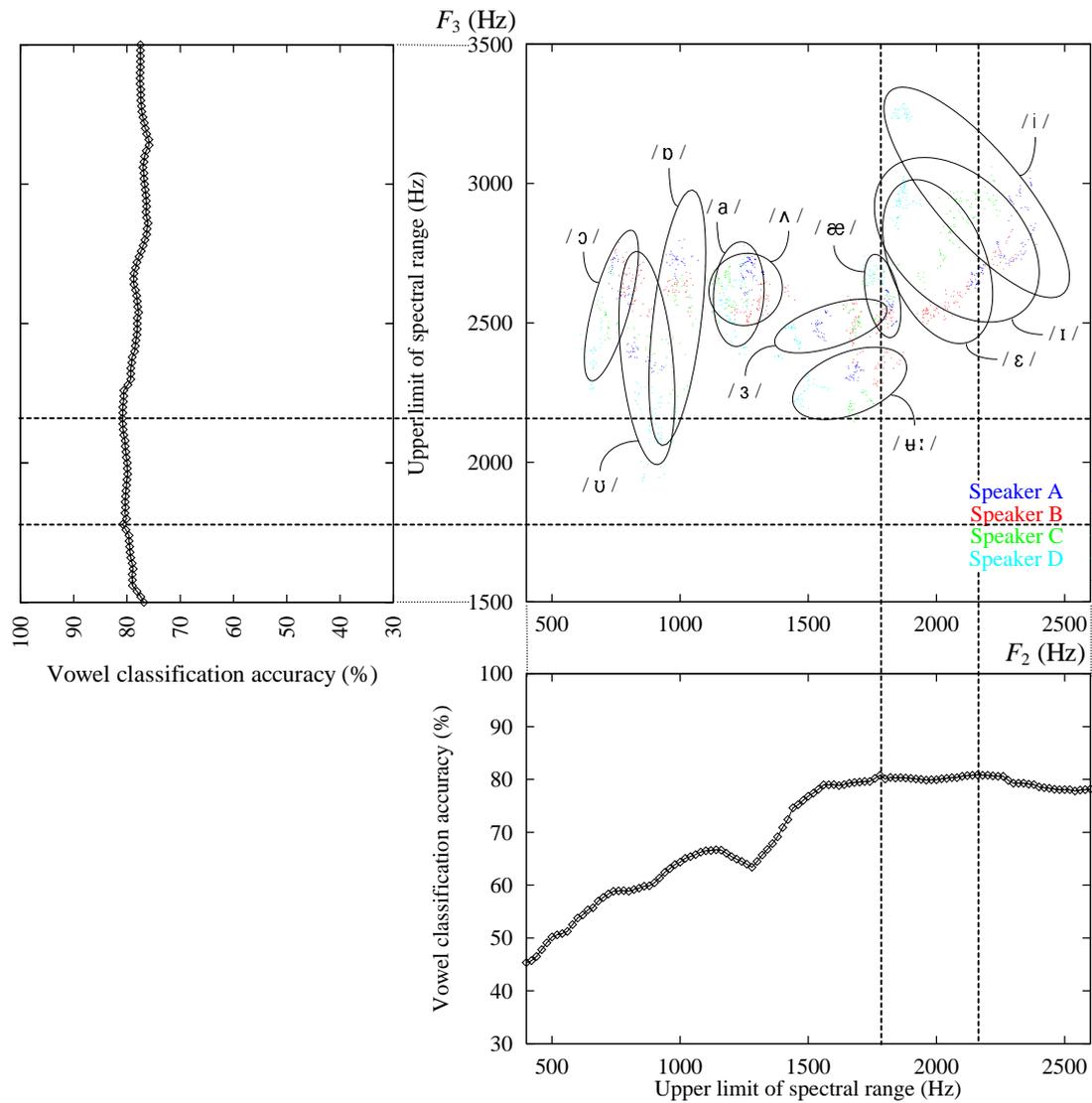


Figure 6.11:  $F_2F_3$  vowel space of all 4 male speakers (FC dataset), synthesised from their estimated area-functions after speaker normalisation, retaining differences in vowel-specific articulatory *strategy* only. Each vowel cluster is shown with a  $2\sigma$  ellipse. Adjacent to the abscissa and ordinate are plotted the portions of the *inter*-speaker accuracy curve (solid line in Figure 6.9) which span the  $F_2$  and the  $F_3$  ranges, respectively. The vertical and horizontal, dashed lines intersect the accuracy curve at its two peaks (at 1780Hz and 2160Hz, respectively), and cut across the formant plane in order to emphasise the acoustic-phonetic relevance of the spectral regions of primary phonetic or speaker (vowel-specific articulatory *strategy*) influence.

Vowel	Speaker			
	A	B	C	D
/i/		/ɪ/ (F <sub>3</sub> , 100%)	/ɪ/ (F <sub>2</sub> , 20%) /ɛ/ (F <sub>2</sub> , 5%)	
/ɪ/		/i/ (F <sub>2</sub> , 20%) /ɛ/ (F <sub>3</sub> , 60%)	/i/ (F <sub>2</sub> , 25%) /ɛ/ (F <sub>3</sub> , 10%)	/ʉ:/ (F <sub>2</sub> , 20%)
/ɛ/	/i/ (F <sub>2</sub> , 100%)	/ɪ/ (F <sub>3</sub> , 5%)		
/æ/				
/a/	/ʌ/ (F <sub>3</sub> , 80%)			/ʌ/ (F <sub>3</sub> , 10%)
/ʌ/		/a/ (F <sub>3</sub> , 15%)	/a/ (F <sub>3</sub> , 5%)	
/ɔ/				/ʊ/ (F <sub>3</sub> , 80%)
/ʊ/	/ɔ/ (F <sub>3</sub> , 15%)	/ɔ/ (F <sub>3</sub> , 75%)		
/ʉ:/				/ɜ/ (F <sub>2</sub> , 35%)
/ɜ/			/ɛ/ (F <sub>2</sub> , 5%)	

Table 6.5: Acoustic-phonetic decomposition of the dichotomy in inter-speaker vowel classification behaviour (solid curve in Figure 6.9), showing the contributions of speaker differences in vowel-specific articulatory *strategy*, in terms of the vowel misclassifications that contribute to the drop in accuracy across the higher spectral regions which encompass the high- $F_2$  and the  $F_3$  of the speakers' vowel formant distribution.

These observations are confirmed by an acoustic-phonetic decomposition (in Table 6.5) of the vowel misclassifications which contribute to the drop in accuracy across the higher spectral regions (including the region between the two global peaks). Significant contributions to the drop in accuracy are found to occur as a result of confusions mainly amongst the  $F_3$  of the back vowels and amongst the  $F_3$  of the front vowels. However, contributing confusions are also found to occur amongst the  $F_2$  of the front vowels, despite the concurrent improvement in classification performance across that spectral range as observed earlier. The most significant of the dichotomous vowel confusions listed in Table 6.5 are next discussed with reference to the formant distributions of Figures 6.10 and 6.11, and with reference to the articulatory manifestations of speaker differences in vowel-specific strategy shown earlier in the area-functions of Figure 6.8 and the vocal-tract lengths listed in Table 6.4.

Confusions amongst the  $F_2$  of the front vowels are indeed listed in Table 6.5 for all four speakers, but none more significant than the 100% drop in accuracy caused by confusions of the vowel /ɛ/ of speaker A with the vowel /i/. Precisely the same vowel

confusion was reported earlier using the original acoustic data (both the measured and the simplified cepstra, cf. Tables 4.1 and 4.2, respectively), but not in Tables 6.2 and 6.3 where the contributions of structure and setting were identified. We may therefore assume that this major contribution to the dichotomy is attributed entirely to speaker differences in articulatory strategy. As shown in Figure 6.8, those differences in strategy are nevertheless rather subtle. In particular, it would appear that the articulatory strategy adopted by speaker A to produce the vowel / $\epsilon$ / involves a shorter vocal-tract length (see also the shaded entry in Table 6.4) and a slightly greater degree of linguo-palatal constriction, compared with the other speakers. The higher  $F_2$  values observed in Figures 6.10 and 6.11 for the vowel / $\epsilon$ / of speaker A (higher values which are also clearly evident in the original formant distributions shown in Chapter 4, cf. Figure 4.9), are therefore explained by a combination of potentially three articulatory strategies: a more raised larynx, a greater amount of lip spreading, and a closer approximation of the tongue to the hard palate.

A less dramatic contribution to the dichotomy which also occurs in the  $F_2$  range of the front vowels, is caused by confusions of the vowel / $\text{I}$ / of speaker D with the fronted / $\text{u}$ :/ . An acoustic-phonetic explanation is provided in Figure 6.10, which shows clearly lower values of  $F_2$  for / $\text{I}$ / of speaker D — so low that they encroach upon the ellipse drawn around the formant cluster of the neighbouring, lip-rounded vowel (as also observed in the original formant distribution, in Figure 4.9). Although not itself implicated in the observed misclassifications, that speaker's  $F_3$  distribution for / $\text{I}$ / is shown in Figure 6.11 (and indeed in Figure 4.10) to be relatively high. In this vein, it is well known that the second and third formants of high front vowels are strongly affiliated, respectively, with the back and the front vocal-tract cavities which are separated by the linguo-palatal constriction (e.g., Fant, 1960; Lindblom and Sundberg, 1971). The lower  $F_2$  and higher  $F_3$  observed in the high front vowels of speaker D might therefore be brought about by an anterior shift in the place of constriction, which would lengthen the back cavity (i.e., lower  $F_2$ ) and simultaneously shorten the front cavity (i.e., raise  $F_3$ ). This is indeed the articulatory explanation emergent from Figure 6.8, which clearly suggests a relatively more forward place of constriction in the high front vowels of that speaker. As a consequence of this idiosyncratic articulatory

behaviour, vowel-speaker interactions are induced in the acoustic domain which cause the observed confusions in the  $F_2$  range with the neighbouring, lip-rounded vowel.

Table 6.5 also reveals confusions in the  $F_2$  range of the fronted vowel / $\mathfrak{u}$ :/ of speaker D, with the more central vowel / $\mathfrak{3}$ /. A lower distribution of  $F_2$  in the former is again implicated in the formant plots of Figures 6.10 and 6.11, and the articulatory explanation suggested in Figure 6.8 is similar to that discussed above — speaker D appears to prefer a relatively more anterior place of constriction. However, there is also evidence of a longer vocal-tract length (as quantified in the shaded entry in Table 6.4). The implied lowering of the larynx would further lengthen the back cavity and thus help to lower the  $F_2$ ; by contrast, the equally implied lip protrusion would help to maintain a nearly constant front-cavity length, thereby counteracting the potential rise in  $F_3$  caused by the more fronted place of constriction. Indeed, this is confirmed in the formant plot of Figure 6.11, where it is shown that the  $F_3$  distribution for / $\mathfrak{u}$ :/ of speaker D is close to the average for that vowel.

This latter explanation is of particular significance, as it also partly explains the rise in accuracy as the spectral range is extended to the second, global peak at 2160Hz. The vowel / $\mathfrak{u}$ :/ of speaker D is indeed one of the main contributors to that rise in accuracy, owing to a gradual correction of the misclassifications discussed above, as the upper spectral limit is extended across the low- $F_3$  range. However, a comparison of Figure 6.11 with the  $F_2F_3$  vowel formant distributions shown in the previous two Sections and in Chapter 4, indicates that while the originally lowered  $F_2$  of that speaker's / $\mathfrak{u}$ :/ is entirely explained by his idiosyncratic behaviour in articulatory strategy, there is also originally a lowered  $F_3$ , which we have attributed to his articulatory setting (the greater degree of lip rounding on a vowel-averaged or long-term basis, as discussed in the previous Section). Whilst that lowered  $F_3$  was not found by itself to cause a drop in classification accuracy, its nullification by speaker normalisation of articulatory setting (as observed in Figure 6.11) has caused a *rise* in accuracy, which occurs only after the strategy-induced drop in the  $F_2$  range.

This example highlights an important aspect of our methodology: when potential articulatory sources of speaker variability are considered in isolation, they may induce certain vowel-speaker interactions (either to the detriment or to the benefit of vowel

classification accuracy) which had not transpired and were therefore latent in the original data, where the sources of variability are combined and themselves allowed to interact. Another example listed in Table 6.5 is a contribution to the strategy-induced dichotomy in the  $F_3$  range of the vowel /a/ of speaker A, which is confused with /ʌ/. Admittedly, those two vowels are nearly indistinguishable in both the  $F_1F_2$  and the  $F_2F_3$  planes. However, this particular confusion which contributes an 80% drop in accuracy across the higher spectral regions, had not occurred until the speaker differences in articulatory strategy were explicitly isolated.

The two significant contributions to the dichotomy in the  $F_3$  range of the back vowels, which *had* also occurred using the original, simplified cepstra (cf. Table 4.2), are listed in Table 6.5 as confusions of the vowel /ɒ/ of speaker D with /ʊ/, and of the vowel /ʊ/ of speaker B with /ɔ/. The formant distributions in Figures 6.10 and 6.11 offer an acoustic-phonetic explanation of the former in terms of a lower  $F_1$ ,  $F_2$ , and especially  $F_3$  in /ɒ/ of speaker D, and of the latter in terms of a slightly lower  $F_2$  but much higher  $F_3$  in /ʊ/ of speaker B.

The superimposed area-functions for /ɒ/ in Figure 6.8 certainly indicate a range of speaker differences not only in the length and volume of the oral cavity which is formed by the retracted and lowered tongue body, but also in the place of lingual constriction. In particular, speaker D appears to have the most expanded oral cavity, a main place of lingual constriction which is displaced further down the pharynx compared with that of the other speakers (at about 5.8cm from the glottis), and a more fronted, secondary lingual constriction (at about 10.3cm from the glottis). Attesting to the articulatory plausibility of this double-constriction, are directly measured area-functions of similar, low back vowels reported by Wood (1979, p.26, Figure 1: /ɑ/) for a speaker of British English, and by Story et al. (1996, p.545, Figure 6: /ɔ/) for a speaker of American English. According to the SM model (cf. Figure 5.1), the  $F_3$  of a uniform acoustic tube can be lowered by alternately expanding and constricting the areas at multiples of  $\frac{1}{5}L$  cm from the lips. Thus, given the overall mean length  $L = 17.74$  cm for /ɒ/ (as listed earlier in Table 6.1), the half-wavelength of the  $F_3$  shape-component is  $\frac{1}{5}L \approx 3.5$  cm. Indeed, as shown in Figure 6.8, the more expanded front cavity and the secondary lingual constriction exhibited by the area-function for /ɒ/ of speaker D do occur,

respectively, at about 3.5 cm and 7 cm from the mean position of the lips. Those shape-related idiosyncrasies in articulatory strategy therefore explain the low values of  $F_3$  observed for /ɒ/ of speaker D.

The superimposed area-functions for /ʊ/ in Figure 6.8 suggest that speaker B has a more retracted place of constriction, and one which extends over a longer part of the vocal-tract. In addition, his vocal-tract length for that vowel is markedly shorter (see also the shaded entry in Table 6.4). As discussed previously in connection with this speaker's articulatory setting (which we cautiously labelled uvularisation), a contraction of the vocal-tract areas in the uvular region will tend to raise the  $F_3$  of back vowels. The idiosyncratic behaviour in articulatory strategy adopted by speaker B in his production of /ʊ/, which indeed appears to involve more constricted areas in the upper pharynx in addition to an overall shorter vocal-tract length (implying a less lowered larynx and less protruded lips, as shown in Figure 6.8), therefore explains the higher values of  $F_3$  which have caused confusions with /ɔ/, and hence the contribution to the drop in accuracy observed in the higher spectral regions.

Finally, significant contributions to the drop in classification accuracy across the mid- to high range of our speakers'  $F_3$  distribution, are listed in Table 6.5 as arising from confusions amongst the  $F_3$  of the high front vowels of speaker B. In particular, a drop of 100% in accuracy occurs as that speaker's /i/ is confused with /ɪ/, and a drop of 60% occurs as his vowel /ɪ/ is confused with /ɛ/. Whilst both of these vowel confusions were also found to occur in Chapter 4 (using either the original or the simplified cepstra, cf. Tables 4.2 and 4.3, respectively), their contributions to the dichotomy as shown in Table 6.5 are considerably greater. This implication of the speaker differences in articulatory strategy playing a major role in the observed vowel confusions, is further underscored by the complete absence of those confusions in our earlier acoustic-phonetic decompositions in Tables 6.2 and 6.3. The acoustic-phonetic manifestations of those differences in strategy as shown in Figure 6.11, indicate a lower distribution of  $F_3$  in the high front vowels of speaker B, which clearly explain the confusions observed in the  $F_3$  range with the neighbouring vowels. The articulatory explanation, as noted earlier in connection with Figure 6.8, is a more retracted place of linguo-palatal constriction in those high front vowels of speaker B — as the third

resonance is mainly affiliated with the vocal-tract cavity in front of the constriction, a more retracted place of constriction would lengthen that front cavity, and hence lower the  $F_3$  as observed in the vowel formant distributions.

In sum, our acoustic-phonetic decomposition of the *strategy*-induced drop in inter-speaker vowel classification accuracy observed across the higher spectral regions, has revealed that the largest contributions arise from confusions amongst the  $F_3$  of the back vowels of speakers B and D, amongst the  $F_3$  of the high front vowels of speaker B, and to a lesser extent amongst the  $F_2$  of the mid-front vowel of speaker A. These vowel confusions are caused by idiosyncrasies in vowel-specific articulatory strategy, which are summarised (together with their observed acoustic-phonetic consequences) as follows: a less lowered larynx in the back vowels and a generally more retracted place of lingual constriction for speaker B (causing a higher  $F_3$  in the high back vowels, and a lower  $F_3$  in the high front vowels); a larger front-cavity expansion in the back vowels and a generally more forward (even if secondary) place of constriction for speaker D (causing a lower  $F_2$  in the front and central vowels, and a lower  $F_3$  in the back vowels); a more raised larynx and a greater degree of linguo-palatal constriction in the mid- to high front vowels of speaker A (causing a higher  $F_2$  in those front vowels). By comparison, as shown by the vowel formant distributions, the area-functions, and the list of vowel misclassifications, speaker C is the most representative (or the least extreme) of the four speakers of the FC dataset.

## 6.5 The Dichotomy Undone

The methodology used in the previous Section to obtain a tripartite articulatory explanation of the vowel-speaker dichotomy, relied on speaker normalisation of only two of the three articulatory features in turn, thereby retaining the speaker differences in the third or remaining feature. If our explanations of the dichotomy have indeed revealed the influences of three separate and complementary, articulatory sources of inter-speaker variability, then it might be expected that speaker normalisation of all three components would completely remove the inter-speaker variations in our dataset, and retain the influences of only the inherent, phonetic variability contained therein. Thus completely deprived of the inter-speaker component of variability, the vowel-

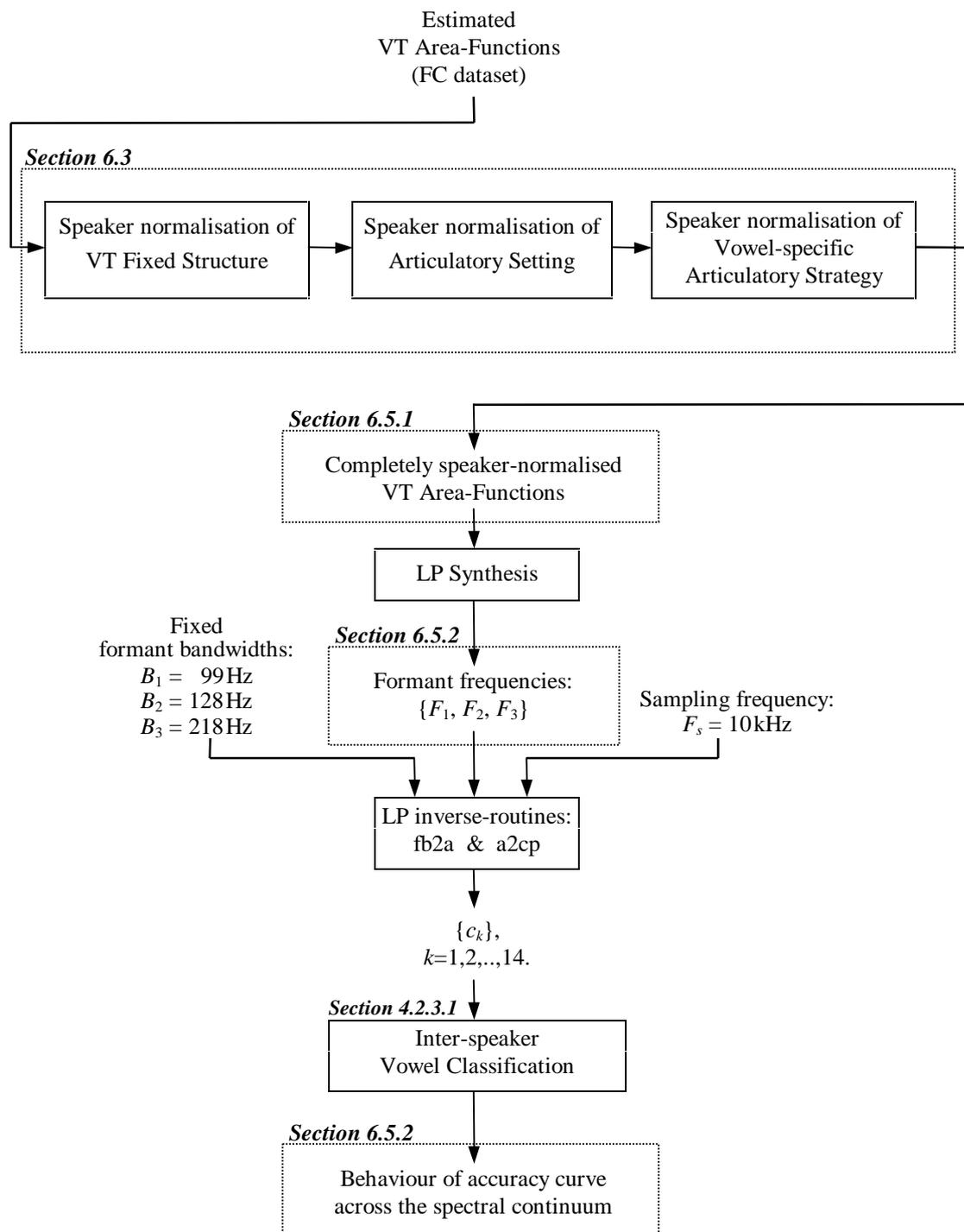


Figure 6.12: *Undoing the vowel-speaker dichotomy by complete articulatory speaker normalisation.* Flow-chart of the computational procedure adopted to speaker-normalise the estimated vocal-tract area-functions, LP-synthesise formant frequencies, generate simplified cepstra, and perform inter-speaker vowel classification as a function of an increasing spectral range. Enumerated at the top-left of computational modules or output data modules, are the Sections in this or in previous Chapters where the respective algorithmic procedures or generated data are discussed.

speaker interactions which we have observed and explained in both the acoustic-phonetic and articulatory domains should cease to be manifest, and the remaining, phonetic component might then be regarded as prototypical of the speakers contained in the given dataset.

In Figure 6.12 is shown a flow-chart of the computational procedure adopted to reveal the articulatory and acoustic-phonetic consequences of undoing the vowel-speaker dichotomy as suggested above. First, our four speakers' estimated vocal-tract area-functions are subjected to all three stages of articulatory speaker normalisation (*structure*, *setting*, and *strategy*) as described in Section 6.3. In Section 6.5.1 we briefly examine those completely speaker-normalised area-functions, and discuss their merit in representing prototypical vocal-tract shapes of the Australian English monophthongal vowels which comprise the phonetic dimension of our FC dataset. Those speaker-normalised area-functions are then input to the subsequent stages of LP synthesis of formants, of generation of simplified cepstra, and of inter-speaker vowel classification as a function of an increasing spectral range. Both the synthesised formant distribution and the resulting behaviour of classification accuracy across the spectral continuum are therefore examined in Section 6.5.2, where we finally bring to light the acoustic-phonetic consequences of our articulatory method of complete speaker normalisation.

### 6.5.1 Articulatory Consequences

As described above, the area-functions obtained after complete speaker normalisation of vocal-tract structure, articulatory setting, and vowel-specific articulatory strategy, contain only the phonetic component of the original variability, and are therefore expected to be immune to vowel-speaker interactions. However, they are also expected to contain the original, *intra*-speaker variations which are retained throughout the process of speaker normalisation. The per-vowel *mean* of the completely speaker-normalised area-functions (computed across frames, repetitions, and speakers, as in step 2 of the strategy-normalisation procedure described in Section 6.3.3), may therefore be regarded as a *prototypical* vocal-tract configuration for the steady-state of that vowel recorded in /hVd/ context by our four, adult male speakers of Australian English.

In Figure 6.13 are shown the per-vowel *prototype* area-functions thus computed,

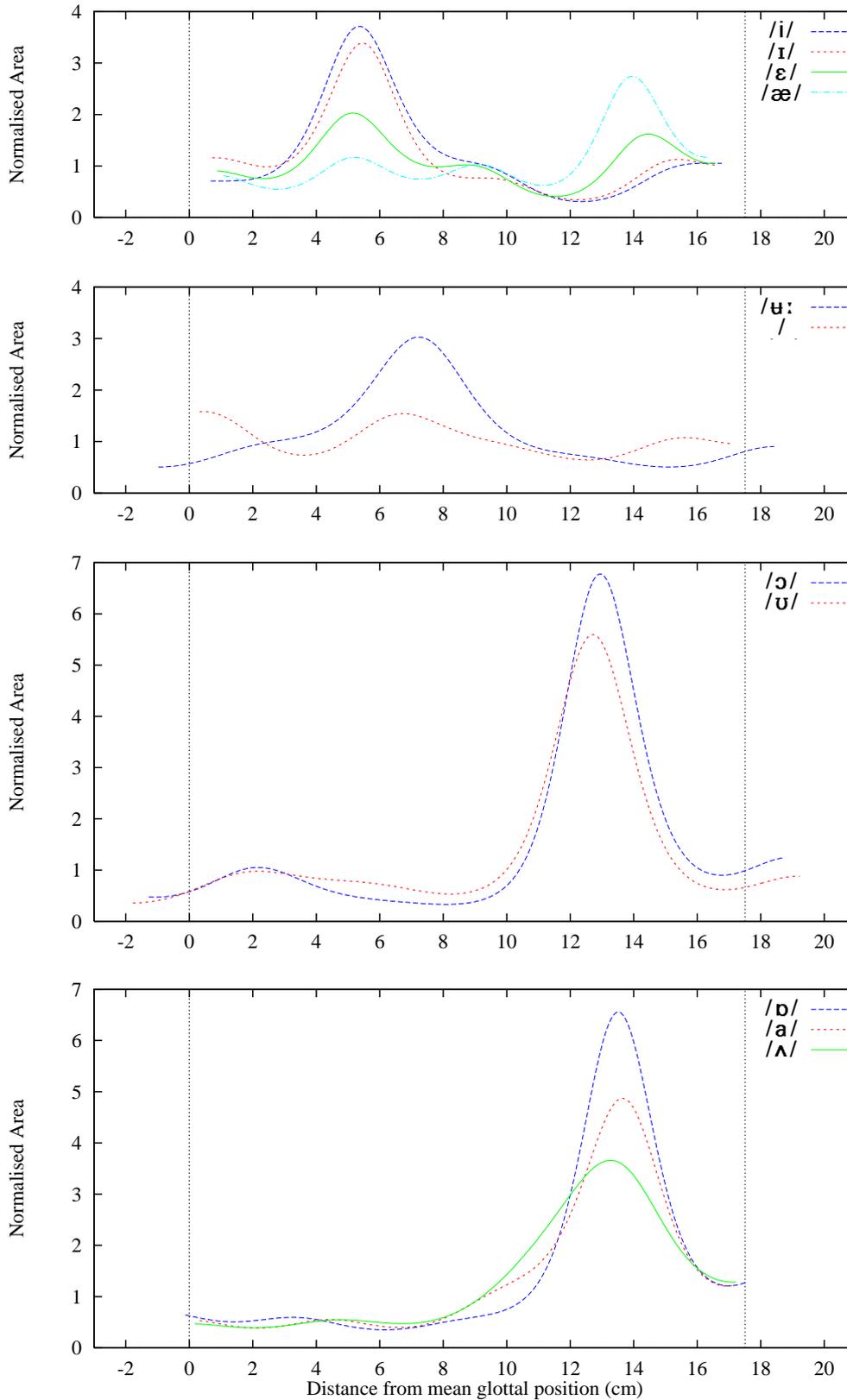


Figure 6.13: *Prototype vocal-tract area-functions of the 11 Australian English vowels (FC dataset), as obtained in step 2 of the procedure described in Section 6.3.3. The vertical, dotted lines on the left and the right indicate, respectively, the overall mean position of the glottis (which defines the origin on the abscissa) and of the lips (at 17.51 cm). Area-functions are aligned at the centre of the mutually overlapping (MOL) region per vowel per speaker (as described in Section 6.3.2).*

grouped approximately according to their main place of constriction. If these area-functions are indeed to be trusted as prototypical of the Australian English vowels included in our FC dataset, then they should embody at least certain basic properties which are normally expected of vocal-tract configurations for those or similar vowels. For example, the individual lengths of the prototype area-functions are expected to confirm what we know about the phonetic relevance of lip rounding and larynx lowering. The prototype vocal-tract lengths which appear in Figure 6.13 and which were also listed earlier in the last column of Table 6.4, indeed confirm that the lip-rounded and larynx-lowered vowels are, in order of decreasing length, the back vowels /ʊ/ and /ɔ/, the fronted /ɜ:/, and to a lesser extent the back vowel /ɒ/. By comparison, the lengths of the central to low vowels /ɜ/, /ʌ/, and /ɑ/ are fairly close to the average, and thus confirm our arguments (in Section 6.3.1) which led to the selection of a vowel subset more likely to convey speaker differences in vocal-tract fixed structure. In that context, whilst the low front vowel /æ/ (which we also included in that vowel subset) appears in Figure 6.13 to have the most inward-displaced glottal end (albeit marginally), it is difficult to separate the influences of our methods of inter-repetition alignment and centre-alignment in that regard. However, in regard to vocal-tract length, our prototype area-functions of the four front vowels are not surprisingly the shortest, and differ only by a slight progression of increasing length with vowel height.

As mentioned above, the four groups of area-functions in Figure 6.13 are arranged approximately according to their main place of constriction along the length of the vocal-tract. In the top graph are shown superimposed the four front vowels, with their main places of lingual constriction clearly located in the palatal region, between 10cm and 13cm from the overall mean position of the glottis. Consistent with our knowledge of the constrictive properties of these vowels, their place and degree of linguo-palatal constriction appear to be more retracted and more expanded, respectively, in the order from the high-front /i/ to the low-front /æ/. In addition, there appear alternative or secondary places of constriction in the pharyngeal part of the vocal-tract for the lower of the front vowels, as often remarked or portrayed in the articulatory literature (cf. for example the MRI-measured area-functions for /æ/ shown in Baer et al., 1991, and in Story et al., 1996).

In the second graph from the top are shown superimposed the prototype area-functions of the very fronted /**ɥ**:/ and the quasi-neutral /**ɜ**/. In accordance with its relative quasi-neutrality, the latter area-function is indeed the least constricted along its entire length. However, it does appear to have a tendency towards a fronted lingual posture, which indeed confirms the slightly fronted position of its formant distribution in the  $F_1F_2$  plane (e.g., Figure 4.9), with a lower  $F_1$  and a higher  $F_2$  compared with the standard reference of  $F_n = (2n - 1)500$  Hz.

Of all the area-functions shown in Figure 6.13, the prototype for the vowel /**ɥ**:/ appears to be the most fronted, with its main place of constriction located in perhaps the most anterior part of the hard palate. It is encouraging to note that Liljencrants' (1971) projections of a Swedish speaker's X-ray measured mid-sagittal tongue contours into a Cartesian coordinate system, also suggest a more fronted place of constriction for the vowel /**ɥ**/ than for /**i**/. Whilst that vowel may indeed be articulated with extreme frontness in Australian English (cf. its characteristically low  $F_3$  in, e.g., Figure 4.10), there does appear to be a tendency for the raised tongue body to continue its approximation to the roof of the mouth along nearly the entire length of the palate. In this vein, recall from Figure 6.8 that two of our speakers (B and C) appeared to have their main place of constriction for /**ɥ**:/ at about the same location as for /**i**/ along the hard palate (at about 12cm from the overall mean position of the glottis); consistent with Wood's (1986) articulatory and acoustic findings on the /**i**/-/**ɥ**/ contrast, speakers B and C thus appeared to distinguish /**i**/ from /**ɥ**:/ by their degrees of lip rounding and larynx lowering. By contrast, speakers A and D were shown in Figure 6.8 to have a more fronted place of constriction for the lip-rounded vowel. Our prototype area-function for /**ɥ**:/ portrayed in Figure 6.13 with (the shape-inferred articulatory features) rounded lips and lowered larynx, thus clearly bears the influence of the articulatory strategies of all four speakers.

In the third graph from the top (in Figure 6.13) are shown superimposed the prototype area-functions for the two, high back vowels /**ɔ**/ and /**ʊ**/. The velarised articulations of these two vowels are confirmed by the location of the lingual constriction at about 8cm to 9cm from the overall mean position of the glottis. Consistent with the more extreme position of the formant cluster for /**ɔ**/ in the  $F_1F_2$

plane (e.g., Figure 4.9), it is shown in Figure 6.13 with a greater degree and a longer place of constriction. Both vowels are shown correctly with large amounts of lip rounding and larynx lowering.

In the bottom graph are shown superimposed the remaining three, prototype area-functions for /ɒ/, /ɑ/, and /ʌ/, which are appropriately portrayed as having the most open vocal-tract configurations. Moreover, their places of constriction are correctly located in the pharyngeal part of the vocal-tract, at about 5 to 7cm from the mean position of the glottis. By comparison with the front vowel /æ/ which was shown earlier to involve a secondary, low place of constriction, these three back vowels have the lowest places of constriction in the pharynx, with an upward progression in place (and a more subtle progression towards a lesser degree of constriction) from /ɒ/ to /ʌ/. In accordance with the principle of conservation of the tongue body mass, the same order of progression is also observed in the area of the most expanded part of the oral cavity.

Our presentation in Figure 6.13 of the purely phonetic range of variation embodied in prototype area-functions obtained after complete speaker normalisation, was partly inspired by a similar presentation of X-ray measured area-functions by Wood (1979), who thereby demonstrated the existence of four, so-called preferred locations of constriction for spoken vowels. Analogously to his main contention in that work, our prototype vocal-tract shapes are broadly categorised into three (rather than four) main places of constriction, as follows<sup>1</sup>: (i) in the palatal region for the fronted vowels such as /i/, (ii) in the velar region for the high back vowels such as /ɔ/, and (iii) in the pharyngeal region for the low vowels such as /ɑ/. The distinction drawn by Wood (1979) between the places of constriction for /ɔ/ (further back) and /ʊ/ (more anterior) does not appear in our prototype area-functions — nor is it suggested, however, in the formant distributions of those two vowels, which appear (e.g., in Figure 4.9) partially overlapping in the extreme high-back region of the  $F_1F_2$  plane. Nevertheless, the lingual constriction of our prototype for the more extreme (or tense) vowel /ɔ/ does

---

<sup>1</sup> These three categories are represented in Figure 6.13 by the area-functions in the top and in the bottom two panels; /ɜ:/ is grouped together with the fronted configurations in category (i), and /ɜ/ is disregarded in this context owing to its quasi-neutrality.

appear to extend further back into the upper pharynx compared with the area-function for /ʊ/, thus partly agreeing with Wood (1979), despite the differences in the phonetic quality of the vowels spoken by his subjects (among them a speaker of Southern British English) and our speakers of Australian English.

In the final analysis, we shall perhaps never be justified in claiming that area-functions estimated by acoustic-to-articulatory mapping are exactly the vocal-tract area-functions actually produced by the given speaker. In this vein, however, it is well known that direct methods of area-function measurement such as X-radiography or MRI are also prone to various, non-negligible sources of error. Although in Chapter 5 we evaluated our hybrid method of area-function parameterisation and estimation mainly from a theoretical point of view, a more lasting proof of viability is that which emerges from articulatory interpretations of the variability contained in those estimated area-functions. In this Section, we have provided further evidence of the adequacy of our inverse method in yielding sufficiently realistic area-functions, in terms of the more widely documented and certainly better understood, *phonetic* dimension of variability in our spoken vowel data.

### **6.5.2 Acoustic-Phonetic Consequences**

As argued earlier, normalisation of the speaker differences in vocal-tract structure, articulatory setting, and vowel-specific articulatory strategy, is expected to yield area-functions which retain variability only along the phonetic and intra-speaker dimensions. Consequently, the vowel-speaker interactions which have been the focus of our study, should cease to exist, and their manifestations in the acoustic-phonetic domain effaced. In the previous Section we discussed the articulatory consequences of complete speaker normalisation, in terms of the phonetic relevance of the per-vowel prototype area-functions thus obtained. We now examine its acoustic-phonetic consequences, particularly in terms of the spectral manifestations of the vowel-speaker dichotomy.

As outlined earlier in Figure 6.12, the formants of the completely speaker-normalised area-functions are first obtained by LP synthesis.  $NCC = 14$  simplified cepstral coefficients are then generated from the first three formant frequencies together with the fixed formant bandwidths which we have adopted throughout this chapter.

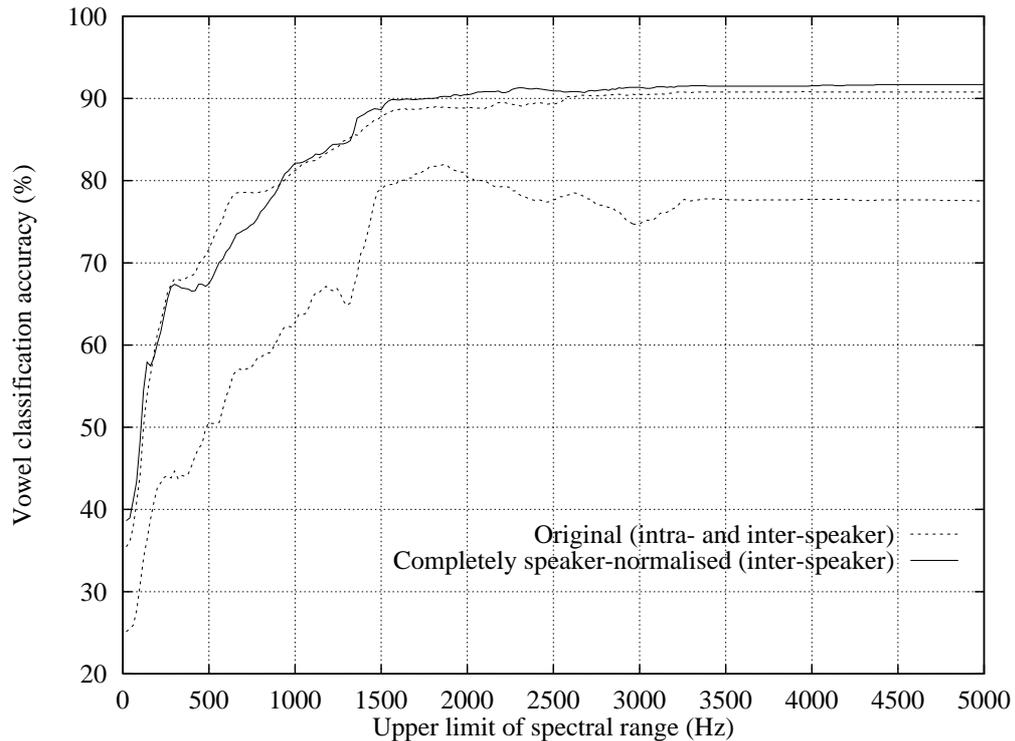


Figure 6.14: *The dichotomy undone by complete articulatory speaker normalisation (FC dataset).* The upper and lower, dashed curves are the intra- and inter-speaker accuracy curves, respectively, obtained in Chapter 4 (see Figure 4.13) using simplified cepstra generated from the first three, measured formant frequencies, with bandwidths fixed to mean values ( $B_1=99\text{Hz}$ ,  $B_2=128\text{Hz}$ ,  $B_3=218\text{Hz}$ ), sampling frequency  $F_s=10\text{kHz}$ , and  $NCC=14$ . The solid curve shows the behaviour of inter-speaker vowel classification accuracy obtained using simplified cepstra generated from the first three formant frequencies synthesised from the vocal-tract area-functions, after speaker normalisation of vocal-tract fixed *structure*, articulatory *setting*, and vowel-specific articulatory *strategy* (with formant bandwidths fixed at the mean values listed above).

Applying the same methodology as described in Section 4.2.3.1, those cepstra are then used to obtain our final profile of inter-speaker, vowel classification accuracy as a function of increasing spectral range.

The behaviour of the resulting accuracy curve is shown in Figure 6.14 (solid line), superimposed with the intra- and inter-speaker accuracy curves (dashed lines) which were obtained in Chapter 4 using the original, simplified cepstra. The solid curve rises to 89.9% as the spectral range is increased to 1580Hz, beyond which it changes very little; the best performance of 91.7% is achieved at full spectral range (5000Hz). Indeed, the nearly-asymptotic behaviour of the solid curve is quite comparable to that of the original, *intra*-speaker curve, thus confirming that the inter-speaker variations in our spoken vowel data have been effectively normalised, and only within-speaker variations retained. Unquestionably, the vowel-speaker dichotomy has been undone!

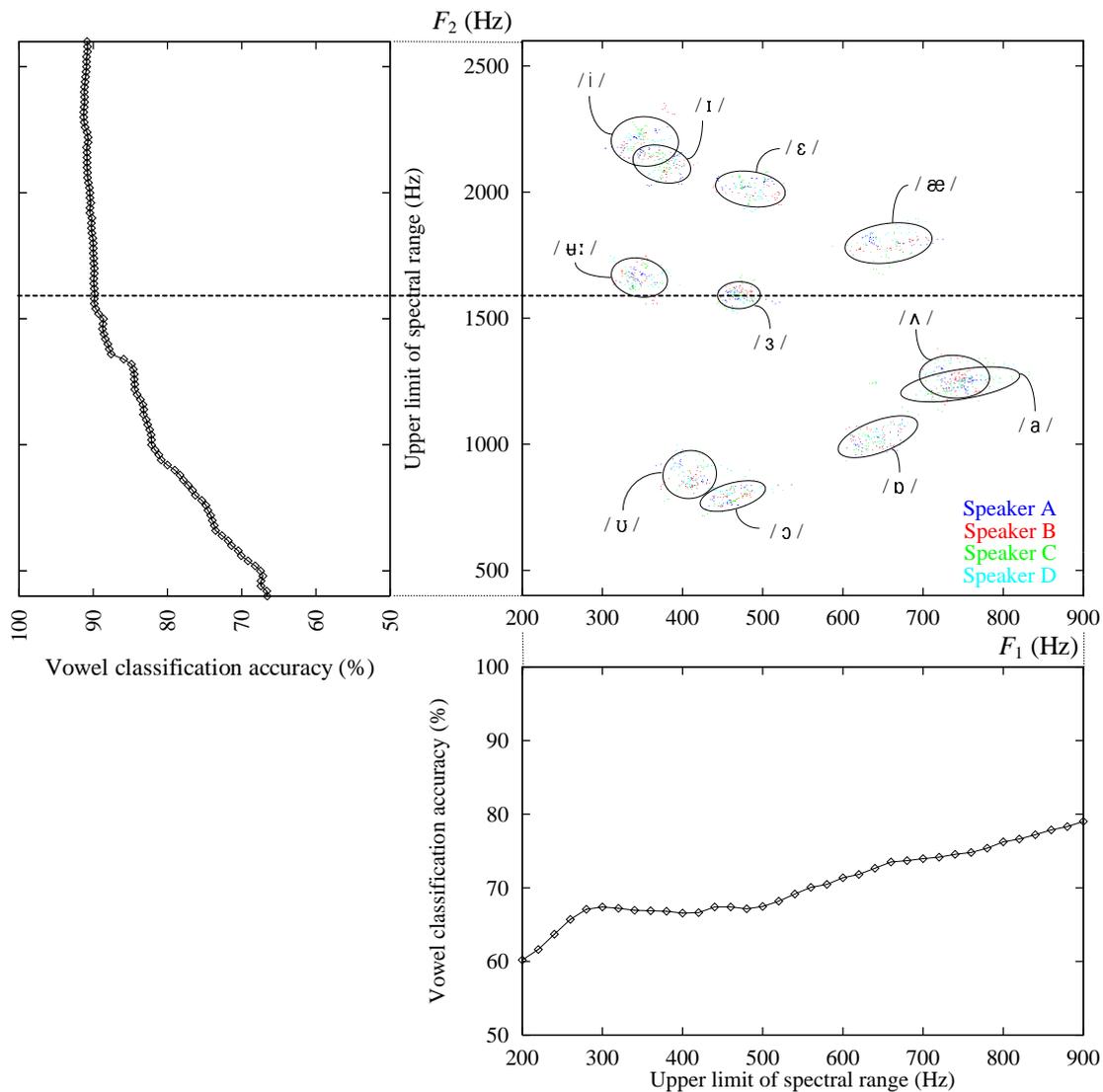


Figure 6.15:  $F_1F_2$  vowel space of all 4 male speakers (FC dataset), synthesised from their estimated area-functions after complete speaker normalisation of vocal-tract *structure*, articulatory *setting*, and vowel-specific articulatory *strategy*. Each vowel cluster is shown with a  $2\sigma$  ellipse. Adjacent to the abscissa and ordinate are plotted the portions of the *inter-speaker* accuracy curve (solid line in Figure 6.14) which span the  $F_1$  and the  $F_2$  ranges, respectively. The horizontal (dashed) line cuts through the centre of the ellipse for the quasi-neutral vowel /ɜ:/; the mid- $F_2$  thus defined (at about 1590Hz), also intersects the accuracy curve at what might be regarded as its “knee”.

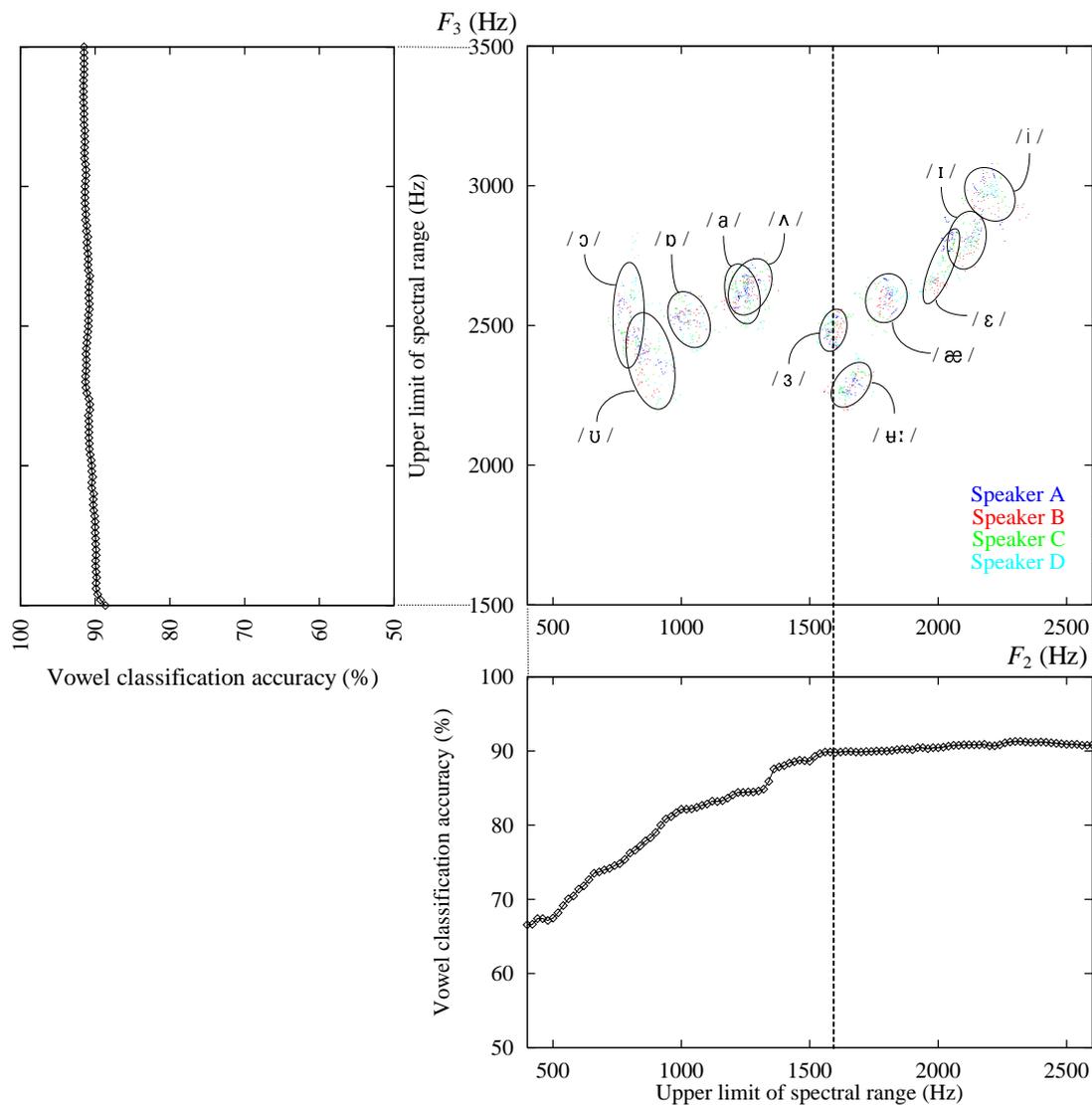


Figure 6.16:  $F_2F_3$  vowel space of all 4 male speakers (FC dataset), synthesised from their estimated area-functions after complete speaker normalisation of vocal-tract *structure*, articulatory *setting*, and vowel-specific articulatory *strategy*. Each vowel cluster is shown with a  $2\sigma$  ellipse. Adjacent to the abscissa and ordinate are plotted the portions of the *inter-speaker* accuracy curve (solid line in Figure 6.14) which span the  $F_2$  and the  $F_3$  ranges, respectively. The vertical (dashed) line cuts through the centre of the ellipse for the quasi-neutral vowel /ɜ:/; the mid- $F_2$  thus defined (at about 1590Hz), also intersects the accuracy curve at what might be regarded as its “knee”.

Articulatorily *and* acoustically, we have therefore succeeded in normalising the data of our four speakers to those of a single, prototype speaker, whose vocal-tract organic and articulatory behavioural characteristics are, by definition, the average of our four speakers' characteristics. Owing to the four-fold increase in the amount of (effectively) *single*-speaker data, it is likely that the resulting, within-speaker variations represent an even greater degree of homogeneity compared with the original, non-normalised data of each speaker separately. In this context, we note that as the spectral range is extended beyond 1340Hz, the accuracies along the solid curve in Figure 6.14 are consistently higher than those obtained along the original, intra-speaker curve; at full spectral range (5000Hz), the normalised data yield an improvement of 0.9%. Indeed, this small but consistent improvement is explained by the following two factors: (i) as noted above, the speaker-normalised data afford a greater richness of (effectively) within-speaker variability; (ii) in addition, our method of data-partitioning yields a greater number of per-vowel training samples in inter-speaker experiments.

The reduction of inter-speaker variability and the consequent increase in the *density* of per-vowel variations, are perhaps best portrayed by the  $F_1F_2$  and  $F_2F_3$  vowel formant distributions which were earlier used to generate the simplified cepstra. As shown in Figures 6.15 and 6.16, the vowel formant clusters are indeed much smaller than those depicted in the original formant distributions (cf. Figures 4.9 and 4.10). As the means of the clusters are hardly affected by the speaker normalisation procedure, the overall vowel separability is therefore enhanced, both in the formant space, and in the cepstral domain as indicated earlier by the higher accuracies obtained. Moreover, the density of each vowel formant cluster is clearly much higher, owing to the four-fold increase in the sheer number of data points which represent effectively the frames and repetitions of our single, prototype speaker.

Adjacent to the abscissa and the ordinate of the vowel formant planes in Figures 6.15 and 6.16 are plotted those portions of the inter-speaker accuracy curve (solid line in Figure 6.14) which span the respective formant ranges. Similarly to the behaviour of the original *intra*-speaker curve (cf. Figure 4.9), classification accuracy rises across the entire  $F_1$  range, and continues to rise as the spectral range is extended further across the low  $F_2$  of our four speakers' articulatorily-normalised formant distribution. At the

mid- $F_2$  (as defined by the mean  $F_2$  of the quasi-neutral vowel /ɜ/, which is equal to 1591Hz in the present, resynthesised data), classification accuracy nearly reaches its asymptotic value of a little over 90%, and improves only marginally as the spectral range is extended across the high  $F_2$  and the entire  $F_3$  range.

In further support of our conclusions in Chapter 4 regarding the intra-speaker vowel classification results, the behaviour of our final, inter-speaker accuracy curve (shown in Figures 6.14, 6.15 and 6.16) first confirms the phonetic importance of the low spectral regions which extend up to the mid- $F_2$  of the vowel formant distribution. It also confirms that the spectral regions higher than the mid- $F_2$ , where we have observed the manifestations of speaker differences to be the most potent, contribute relatively little to vowel discriminability even in the absence of the detrimental influence of inter-speaker variations. Perhaps more importantly, the behaviour of our final, inter-speaker accuracy curve attests to the effectiveness of our articulatory method of speaker normalisation, which itself yielded (in Section 6.4) a three-part, physical explanation of the vowel-speaker dichotomy. This undoing of the dichotomy by complete speaker normalisation of the estimated vocal-tract area-functions, is indeed our final proof both of the speaker-related potency of the higher spectral regions, and of the viability of our tripartite definition and normalisation of the articulatory sources of speaker differences in spoken vowels.

## 6.6 Concluding Summary

In this Chapter we presented a new methodology for speaker normalisation of vocal-tract area-functions, which first facilitated a physical explanation of the vowel-speaker dichotomy in terms of three articulatory (supralaryngeal) sources of inter-speaker variability. Those sources of speaker differences were defined as follows: (i) vocal-tract fixed *structure*, determined by the mean vocal-tract length computed over the mid- to low vowels; (ii) articulatory *setting*, determined by the mean vocal-tract shape computed over all the vowels; and (iii) vowel-specific articulatory *strategy*, determined by the mean vocal-tract length and shape computed on a per-vowel basis. Our articulatory approach to explain the vowel-speaker dichotomy then consisted of: normalising the speaker differences in two of the three articulatory features in turn;

	Vowel formant range			Contribution to dichotomy (%)
	$F_2$ of front	$F_3$ of back	$F_3$ of front	
<i>Structure</i>	✓		✓	4.1
<i>Setting</i>	✓	✓		3.4
<i>Strategy</i>		✓	✓	5.0
Original	✓	✓	✓	7.2

Table 6.6: Contributions of vocal-tract *structure*, articulatory *setting*, and vowel-specific articulatory *strategy*, to the vowel-speaker dichotomy, in terms of: (i) the vowel formant ranges of our four adult, male speakers of Australian English (FC dataset) across which were observed significant contributions to the drop in inter-speaker vowel classification accuracy, and (ii) the largest extent of that drop in accuracy across the higher spectral regions. Also shown for comparison are the formant ranges and the total drop in accuracy obtained (in Chapter 4) using the *original*, simplified cepstra.

synthesising the first three formant frequencies of the resulting, normalised area-functions; generating simplified cepstra from those synthetic formants; performing inter-speaker vowel classification experiments as in Chapter 4; and finally carrying out an acoustic-phonetic decomposition of the drop in classification accuracy observed across the higher spectral regions.

Indeed, all three classification accuracy curves thus obtained were found to exhibit a drop in accuracy across the higher spectral regions. However, systematic differences were found not only in the magnitude of the largest drop in each accuracy curve, but also in the spectral range across which the most significant vowel confusions occurred. Table 6.6 summarises the contributions of each of the three articulatory features to the dichotomy, as found in this Chapter for the FC dataset. The baseline for comparison is the behaviour of the inter-speaker vowel classification accuracy curve obtained in Chapter 4 using the original, simplified cepstra. Our acoustic-phonetic decomposition of that accuracy curve revealed significant vowel confusions across the high  $F_2$  and the entire  $F_3$  range, yielding an overall drop in accuracy of 7.2 %. By contrast, each of the three articulatory sources of speaker variability were found (in Section 6.4) to have contributions to the dichotomy mainly across two of the three, vowel formant ranges. For example, the strategy-induced vowel-speaker interactions which yielded the largest overall drop of 5.0% in accuracy, were manifest mainly across the entire  $F_3$  range. The

					Speaker					
					A	B	C	D		
<i>Structure</i>	<b>Smallest (16.26 cm)</b> ↳ higher formant distribution, especially in $F_2$ and $F_3$ of front vowels			<b>Largest (16.97 cm)</b> ↳ lower formant distribution, especially in $F_2$ and $F_3$ of front vowels						
<i>Setting</i>	<b>Velarised</b>	<b>Uvularised</b> ↳ high $F_3$ in back vowels <b>More anterior constriction along hard-palate</b> ↳ low $F_2$ and high $F_3$ in high-front vowels	<b>Velarised</b>	<b>Palato-velarised</b> ↳ high $F_2$ in low-front vowels <b>Lip-rounded</b> ↳ low $F_3$ in back vowels						
<i>Strategy</i>	<b>More raised larynx &amp; greater degree of constriction in mid-to high-front vowels</b> ↳ high $F_2$ in mid-to high-front vowels	<b>Less lowered larynx in back vowels, &amp; generally more retracted place of constriction</b> ↳ high $F_3$ in high-back vowels ↳ low $F_3$ in high-front vowels		<b>Larger front-cavity expansion in back vowels, &amp; generally more anterior place of constriction</b> ↳ low $F_2$ in front & central vowels ↳ low $F_3$ in back vowels						

Table 6.7: Articulatory interpretations and acoustic-phonetic consequences of the speaker differences in vocal-tract *structure*, articulatory *setting*, and vowel-specific articulatory *strategy*, which were found in Section 6.4 to contribute to the vowel-speaker dichotomy in the FC dataset.

next largest contribution of 4.1% was caused by structure-induced, inter-speaker vowel confusions mainly across the  $F_2$  and the  $F_3$  range of the front vowels. By contrast, the smallest contribution of 3.4% was caused by setting-related speaker differences mainly in the  $F_2$  range of the front and the  $F_3$  range of the back vowels.

Nor did we find the articulatory sources of dichotomous vowel confusions to be uniformly distributed across all four speakers. Table 6.7 summarises our interpretations of the speaker idiosyncracies in each of the three articulatory features, together with their most pronounced, acoustic-phonetic consequences which were found to contribute to the dichotomy. For example, the main contribution of speaker C to the dichotomy was found to occur in the  $F_2$  and the  $F_3$  range of the front vowels, and was explained almost exclusively in terms of his relatively larger vocal-tract anatomy. Although the vowel confusions across the same formant ranges for speaker A were similarly explained by his relatively smaller vocal-tract anatomy amongst our four speakers, his

contribution to the dichotomy in the  $F_2$  range of the mid- to high-front vowels was, in addition, explained by idiosyncratic behaviour in articulatory strategy in those vowels. By contrast, speakers B and D, who were found to have vocal-tract anatomical sizes close to the average of our four speakers, were then found to contribute almost exclusively in their setting- and strategy-related idiosyncracies in vowel production.

As stated repeatedly, we cannot simply assume (nor do we claim) that our estimated vocal-tract area-functions are those which the speakers actually produced during their recording sessions. On the contrary, numerous sources of error are known to exist along the path from the recorded acoustic speech signal, through formant measurements and the many simplifying assumptions of the inverse method and its underlying vocal-tract model, to the estimated area-functions themselves. Whilst those sources of inaccuracy were considered in Chapter 5 mainly from a vocal-tract modelling point of view, we reserved a more conclusive judgement on their significance, pending further analyses of the area-functions, aimed at examining and interpreting those data in the context of an actual speech phenomenon. In this Chapter we have indeed achieved precisely what we set out to achieve in Chapter 1, namely to gain some articulatory insights into the phenomenon of vowel-speaker dichotomy, at least as it is manifest in the given dataset of spoken vowels. The credibility of our articulatory explanation can only uphold the plausibility of our estimated area-functions, and lead us to advocate further use of the hybrid LP-SM method of area-function parameterisation and estimation, in order to better understand some basic speech phenomena which still await exploration or elucidation from a speech production point of view.

# Chapter 7

## Concluding Discussion

As stated in the opening paragraph of this thesis, one of the most eminent and long-standing problems in understanding and describing the processes of human speech communication, whether in production, acoustics, or perception, is that of *variability*. As also explained in our Introduction (Chapter 1), the kernel of the variability problem is the *speech-speaker dichotomy*, which has indeed been the central theme of the present work. In our interpretive review of the literature (in Chapter 2), we assembled the evidence presented and the approaches used in previous studies which bear, either directly or indirectly, on the dichotomy problem. In the four intervening, core chapters, we then described our own methods and investigative results concerning the dichotomy manifest in the steady-state, segmental properties of spoken vowels. In the following, concluding Sections 7.1 and 7.2 of this thesis, we endeavour to consolidate our main contributions, whilst discussing implications and raising questions which point to future work, from both an *acoustic-phonetic* and an *acoustic-articulatory* perspective.

### 7.1 Acoustic-Phonetic Perspective

Our acoustic-phonetic investigations of the speech-speaker dichotomy described in Chapter 4, contribute both a more complete ensemble of experimental methods (Section 7.1.1) and a more unified description of the spectral manifestations of the dichotomy (Section 7.1.2) than have been offered to date.

#### 7.1.1 Experimental Methods

##### 7.1.1.1 *A Posteriori* Selection of Frequency Sub-bands

Perhaps the most important concept which has permeated throughout the present work, is that the phonetic and the speaker-specific components of acoustic variability are not

equally distributed along the entire frequency range. It was not without some surprise, therefore, that we previously noted the absence from the literature (despite its long history of support for that concept), of a convenient and computationally efficient method of accessing spectral information within any available frequency sub-band. Indeed, as reviewed in Chapter 2, previous studies which have at all shed light on the frequency-band dependence of phonetic and speaker variability, have either been stuck with the inflexibility of low-pass, high-pass, or banks of band-pass filters; or relied on re-partitioning of FFT spectra; or resorted to more computationally intensive, re-analyses of the acoustic speech signal at each, new frequency sub-band.

In that regard, our parametric cepstral distance measure (PCD) and partial, quefrequency-weighted cepstrum (P-QCEP), both derived in Chapter 4, present a significant advance towards greater flexibility and computational simplicity in the powerful and popular, cepstral comparison of speech sounds. Indeed, our parametric formulations of cepstral distance computation and cepstrum derivation, respectively, have greatly facilitated our investigations of the frequency-band dependence of phonetic-speaker interactions. In particular, it is the *a posteriori* selection of any frequency sub-band within the available spectral range, using only the low-order cepstral coefficients, which renders those two methods potentially useful in a wide range of applications including the recognition framework which we have adopted throughout our investigation of the dichotomy.

### **7.1.1.2 Recognition Framework for Diagnosing the Dichotomy**

One important aspect of our literature review exposed in Chapter 2, has been to identify experimental probes into the dichotomy and its acoustic-phonetic manifestations. In this regard, we therein suggested and, in subsequent chapters, demonstrated that *vowel-speaker interactions* are suitably implicated in machine recognition of vowels on a speaker-independent basis, where both phonetic and inter-speaker variabilities (and their interactions) have a direct influence on the level of accuracy attained within each selected frequency sub-band. Our experimental approach to the problem of speech-speaker dichotomy manifest in spoken vowels, was accordingly founded in a vowel

recognition framework.

However, we have also shown in Chapter 4 that the recognition methodology therein adopted is able to yield a comprehensive diagnosis of the dichotomy, which is unprecedented. First, the PCD and the P-QCEP (used with the hyper-planar and the hyper-quadratic classifiers, respectively) afford a more complete profile of the frequency-band dependence of vowel-speaker interactions. In particular, our whole-spectrum approach avoids the inherent problems of the discrete, formant-based spectral representation (as discussed in Section 2.3.2.2), and yields arbitrarily-detailed profiles of vowel classification accuracy as a function of an increasing upper spectral limit. Second, our method of contrasting the accuracy profiles obtained on an intra- and an inter-speaker basis, leaves very little doubt that the observed differences are caused by interactions between phonetic and *inter-speaker* variabilities.

No less important are our post-recognition interpretation and decomposition of the generated accuracy curves. First, an acoustic-phonetic explanation of the gross features of the accuracy profiles was afforded by plotting the relevant portions of those curves adjacent to the speakers'  $F_1F_2$  and  $F_2F_3$  vowel formant distributions. Second, the overall mean, inter-speaker accuracy curve was decomposed in terms of its constituent, per-vowel and per-speaker accuracy curves, thus facilitating a rank ordering of the vowels and speakers according to their individual contributions to the dichotomy.

It would be tempting at this point to advance that the necessity of building up an entire, acoustic-phonetic methodology, is itself indicative of the relative immaturity of the current outlook on the long-standing problem of speech-speaker dichotomy — even for the most well-known class of speech sounds (vowels) recorded by the most widely studied group of speakers (adult males). For the least, however, we should like to express the firm view that the ensemble of methods summarised above, indeed, yield a long-awaited procedure for diagnosing the speech-speaker dichotomy manifest in any segmental, acoustic dataset of spoken language (with the caveat that if non-vocalic sounds are involved, the formant-based interpretation would need to be relinquished for a more suitable one).

### 7.1.2 Spectral Manifestations of the Dichotomy and Implications

The methodology recapitulated in the previous section, was used in Chapter 4 first to unfold, then to provide a detailed, acoustic-phonetic explanation of the dichotomy manifest in the steady-state vowels recorded in /hVd/ context by four, adult male speakers of Australian English. The nearly asymptotic behaviour of the *intra*-speaker accuracy curve, confirmed the well-known phonetic importance of the low spectral range encompassing the first two formants. In particular, *intra*-speaker vowel classification accuracy was found to be relatively independent of the spectral regions higher than the mid- $F_2$  of the speakers' vowel formant distribution. By contrast, the behaviour of the *inter*-speaker accuracy curve exposed the speaker-related potency of those spectral regions higher than the mid- $F_2$ . Indeed, the detrimental influences of *inter*-speaker variability were clearly manifest by the drop in accuracy, caused by misclassifications mainly amongst the  $F_2$  of front vowels, amongst the  $F_3$  of back vowels, and amongst the  $F_3$  of front vowels, in this decreasing order of detrimental effects. It is interesting to note that this bipartite, acoustic-phonetic interpretation of the manifestations of vowel-speaker interactions in terms of the *front-back* vowel distinction, meshes quite well with Broad's (1981) insightful, multi-speaker extension of Broad and Wakita's (1977) bi-planar model of the first three formant frequencies of spoken vowels.

Our confidence in the intrinsic nature of the observed phenomenon was then reinforced, by obtaining essentially the same behaviour of vowel classification accuracy using first a more powerful classifier; then so-called simplified cepstra; and finally, two different datasets containing larger numbers of speakers of American and Australian English, respectively.

First, the more statistically-reliant, hyper-quadratic classifier accentuated the importance of using sufficient amounts of training data in order to avoid the so-called "curse of dimensionality". Indeed, it is well known that the severity of that problem increases both with the number of parameters used, and with the statistical complexity of the classifier; hence, the highly data-driven nature of state-of-the-art ASR systems. Our approach to avoiding this problem was to acknowledge the intrinsically smaller

dimensionality of the *partial* cepstrum, and therefore retain proportionally fewer P-QCEP coefficients in classification experiments performed in frequency sub-bands. The drop in inter-speaker vowel classification accuracy thus obtained across the higher spectral regions was found to persist, thereby lending further support for the intrinsic nature of the dichotomy.

The dichotomy also persisted upon using so-called simplified cepstra generated from only the first three formant frequencies (with fixed bandwidths). Whilst this result certainly paved the way for using the formant-based, but more populous, PB and JB datasets, at a deeper level it partially addressed the important question of the influence of spectral representation on the manifestations of vowel-speaker interactions. In particular, it provided strong evidence that, consistently with findings reported in the auditory-perception literature, the frequency locations of the formant peaks are indeed the most important features of the whole-spectrum representation of spoken vowels. In addition, it proved the pertinence of our formant-based interpretations of the gross features of the accuracy profiles.

In that context, it is relevant to recall that the NDPS representation consistently used in all our classification experiments, does already enhance the formant peaks and suppress non-phonetic influences such as the overall spectral tilt. However, an intriguing question now arises as to the dependence of the dichotomy on the number of frequency-weighted cepstral coefficients retained, or equivalently, the degree of smoothing applied to that formant-enhanced spectral representation. On the one hand, given the inconsistency of the (discrete) formant representation itself in revealing the detrimental influences of phonetic-speaker interactions in speaker-independent vowel recognition experiments reported in the literature (as discussed at length in Section 2.3.2.2), there is good reason to suspect that the clearly dichotomous behaviour of our accuracy curves obtained using 14th-order cepstra might begin to blur if the cepstral distance measure is rendered more sensitive by increasing the number of cepstral coefficients, and thereby sharpening the formant peaks of the smoothed NDPS. On the other hand, given the reported, phonetic relevance of the low-order LP and PLP models (e.g., Strube, 1980; Hermansky, 1990), further smoothing of the NDPS by truncation of the cepstrum to less than 14 coefficients may be expected, even without any

perceptually-motivated pre-processing, to yield levels of accuracy comparable to those attained at the peak of our inter-speaker accuracy curve near the mid- $F_2$  of the speakers' vowel formant distribution. Further work is warranted, however, in order to gain more detailed insights into the dependence of the dichotomy on spectral smoothing.

We also sought to evaluate the dependence of the dichotomy on speaker homogeneity, using the PB dataset of vowels recorded by 32 adult, male speakers of American English. In sum, we found that a spectrally more homogeneous group of speakers is more likely to exhibit a blurred dichotomy; and conversely, that a spectrally more heterogeneous group is more likely to exhibit a clearly dichotomous accuracy curve (where spectral homogeneity can be quantified, for example, by per-vowel, inter-speaker variation in each formant frequency). Whilst a quantitative measure of each speaker's degree of "sheepiness" or "goatiness", respectively, was therefore defined on the basis of the gross features of the speakers' individual accuracy curves, a more informative index or rank-ordering criterion might also take into account the more detailed, *phonetic* decomposition of each speaker's accuracy curve.

Finally, the JB dataset of 14 Broad, 11 General, and 11 Cultivated, adult male speakers of Australian English was used to investigate the dependence of the dichotomy on idiolectal speaker differences. Briefly, the dichotomous behaviour of the inter-speaker accuracy curve was found to be independent of the speakers' (perceived) idiolectal homogeneity — regardless of the presence or absence of idiolectal differences, we were still able to explain the drop in accuracy across the higher spectral regions in terms of speaker-induced confusions amongst the  $F_2$  of front (and mid-) vowels, and amongst the  $F_3$  of back vowels. This consistency in the acoustic-phonetic decomposition of the dichotomous vowel confusions across three different datasets of spoken vowels, does attest to the basic nature of the observed phenomenon. It also encourages further investigations using our acoustic-phonetic methodology, applied to datasets of vowels recorded by speakers of different dialects, and of different languages.

Admittedly, the implications of the dichotomy for ASR would be more completely assessed by extending our investigations to include other classes of speech sounds than the steady-state vowels of English. However, at least for the spoken vowels considered

(the relative importance of which is well established), our results provide a far more solid basis than has been offered to date, for the very old concept of phonetic and speaker-specific influences being more strongly manifest, respectively, in the low and in the higher ranges along the spectral continuum. Moreover, our recognition results obtained before and after speaker normalisation leave little doubt that speaker-independent ASR systems stand to gain by focusing on the more phonetically relevant, lower spectral regions of spoken vowels, and by exploiting the higher spectral regions of those sounds for speaker adaptation. Conversely, for automatic *speaker* recognition, our results raise the strong possibility that classification accuracy would be improved by focusing on the higher spectral regions of vocalic speech sounds. Whilst the benefits of auditorily-motivated frequency scales already attest to the importance of emphasising the lower and de-emphasising the higher spectral regions, the PCD and the P-QCEP arguably afford more direct control of frequency sub-bands and, *a fortiori*, of the spectral influences of phonetic and speaker variabilities.

## **7.2 Acoustic-Articulatory Perspective**

Analogously to the bipartite structure of the previous section, our acoustic-articulatory investigations of the speech-speaker dichotomy in Chapters 5 and 6 contribute both a methodology for acquiring and analysing vocal-tract shapes (Section 7.2.1), and a first attempt at using those estimated shapes for a more complete description than has been offered to date, of the likely physical correlates of the speaker differences implicated in the acoustic-phonetic phenomenon of dichotomy (Section 7.2.2).

### **7.2.1 Experimental Methods**

#### **7.2.1.1 Vocal-tract Shape Parameterisation**

As discussed at length in Section 2.4.3.2 and later in Chapter 5, acoustic-to-articulatory mapping is in general plagued by the problem of non-uniqueness, which itself depends crucially on the articulatory parameterisation and the acoustic vocal-tract model used. In regard to the latter, it has long been known (Atal, 1970; Wakita, 1973) that the linear prediction (LP) vocal-tract model is the only one for which we know that uniqueness (within certain limits) is an inherent property. However, the LP method of inversion is

generally regarded with disdain (e.g., Schroeter and Sondhi, 1994), partly owing to its presumed incapability of yielding sufficiently plausible or realistic vocal-tract shapes; but also partly because of its highly discretised representation of the vocal-tract area-function, which prohibits physiologically meaningful comparisons of vocal-tract shapes of different lengths, across different vowels and different speakers.

In Chapter 5 we addressed these issues at a fundamental level, by invoking the completely lossless, Schroeder-Mermelstein (SM) model to provide a new explanation of the uniqueness property of the LP model, and ultimately to parameterise the latter in terms of our extended version of the SM model. In particular, we first showed that the LP model shares with the SM model, the fundamental, quasi-unique relation between the formant *frequencies* and the odd-indexed coefficients of the *cosine* series of the logarithmic area-function. Our extension of the SM model then emerged from our discovery of an analogous, quasi-unique relation in the LP model, between the formant *bandwidths* and the odd-indexed coefficients of the *sine* series of the logarithmic area-function. Our resulting parameterisation of the discrete-sectioned LP area-function thus provides the first ever, fundamentally shape-related explanation of the LP model's inherent uniqueness; at the same time, it affords an acoustically meaningful, smooth representation of LP-derived vocal-tract shapes, which are then more amenable to cross-vowel and cross-speaker comparisons.

An important pending question regarding the hybrid LP-SM method of inversion, is the generality of the MAD criterion used to estimate the vocal-tract length (VTL). In that regard, one of the advantages of more physiologically-oriented articulatory models (e.g., Mermelstein, 1973; Coker, 1976) is that they obviate the necessity of VTL as an explicit parameter. However, apart from the prohibitive, speaker-dependence of those models, they have also been found to require both static and dynamic constraints to overcome their inherent non-uniqueness. In that context, the principle of *minimal articulatory effort* which was first applied by Zue (1969) as a static constraint in area-function estimation, has since been widely used (in various formulations, and under different pseudonyms) with reasonable effect. Indeed, our results also suggest that reasonably plausible lengths (and concomitantly, reasonably plausible vocal-tract shapes) are yielded by that constraint, provided a sufficient number of acoustic

parameters (e.g., 4 formants) are used. Further work along the lines pursued recently (e.g., Sorokin's (1992) more physiologically-oriented definition of minimum "muscle work", or Yehia and Itakura's (1994) use of the speaker's average vocal-tract shape as a reference area-function in place of the uniform tube), may unveil even more realistic alternatives to, or interpretations of the MAD criterion.

Another outstanding issue is the choice of formant bandwidths which, as we have been able to establish theoretically and empirically, play a vitally supporting role in ensuring uniqueness of LP-derived vocal-tract shapes. Indeed, those acoustic parameters provide the crucial second-half of shape components, which otherwise remain ambiguous in the completely lossless model on which the original SM parameterisation is based. Owing to the notorious unreliability of bandwidths measured from the acoustic speech signal, we first averaged the values of each bandwidth across repetitions and speakers, while retaining their phonetic variation. We then acknowledged the fact that the LP vocal-tract model has only a single source of acoustic energy loss lumped at the glottal end, by subtracting from those averaged bandwidths, the so-called closed-glottis bandwidths predicted by Hawks and Miller's (1995) equations derived from empirical data. The plausibility of our thus estimated area-functions encourages us to speculate that a more sophisticated procedure for determining the most appropriate formant bandwidths may not be necessary. However, further work is called for if our method is to be applied (at least in spirit) to datasets such as PB and JB, which comprise only formant frequencies.

Another promising avenue for further research, would be to derive an articulatory distance formula which yields the rms distance between area-functions, using the parameters of our extended SM model. As the relation between our vocal-tract shape parameters and the logarithmic area-function (i.e., the cosinusoidal series expansion) is quite similar to the relation between the cepstral coefficients and the log-magnitude spectrum (or similarly, the relation between the QCEP and the NDPS), the derivation of such a distance measure should follow closely our derivation (in Chapter 4) of the PCD. Although there is the additional complexity that area-functions will potentially have different lengths, a *parametric articulatory distance* measure (PAD) may afford a reduced computational load and a greater flexibility in quantifying shape-related

differences within physiologically-motivated, sub-regions of the vocal-tract.

### 7.2.1.2 Articulatory Speaker Normalisation

Our review (in Chapter 2) of previous methods proposed to overcome the speech-speaker dichotomy by speaker normalisation, showed that on the one hand, purely acoustic-phonetic approaches have rarely afforded insights into the underlying causes of the dichotomy; and that on the other hand, articulatorily-motivated methods have rarely transcended the persistent trend of dealing with only the gross differences in vocal-tract length. One of the outstanding deficiencies in previous works has clearly been the unwillingness to adopt a sufficiently complete, theoretical framework for describing speaker differences in physical terms. For example, it has long been suggested that the physical sources of inter-speaker variability are both organic (or anatomical) and learned (or behavioural) in nature, yet by far the majority of previous studies concerned with importing articulatory principles in speaker normalisation, have cared to account for only the former. In a similar vein, whilst the concept of long-term average acoustic parameters has been vociferously embraced in automatic speaker recognition and other applications, it has rarely been connected with the old and neglected concept of long-term articulatory setting.

Our articulatory method of speaker normalisation expounded in Chapter 6, does transcend the incompleteness of prevalent approaches, by acknowledging three basic sources of speaker differences — namely, vocal-tract fixed *structure*, long-term articulatory *setting*, and vowel-specific articulatory *strategy*. In order to quantify and to speaker-normalise any combination of these articulatory sources of speaker variation, they were each given an operational definition in terms of the estimated area-functions of any given speaker. Thus, *structure* was defined as the mean vocal-tract length, computed over only the non-rounded, mid- to low vowels, which are presumably least influenced by speaker differences in extreme laryngeal and labial postures; *setting* was defined as the mean vocal-tract shape, computed over all the vowels after speaker normalisation of structure; and *strategy* was defined as the mean vocal-tract shape and length computed on a per-vowel basis, after speaker normalisation of both structure and setting. Each one of these three articulatory features was then retained for cross-

speaker comparison, by normalisation of the remaining two, while normalisation of all three features completely removed the corresponding articulatory sources of inter-speaker variation.

However, a methodological obstacle in comparing any pair of area-functions, or indeed in computing an average vocal-tract shape as required in the normalisation procedure, is that of lengthwise alignment. As area-functions are generally of different lengths, and as there appears to be no reliable way of identifying on them fixed anatomical landmarks, it is not at all surprising that the literature fails to provide a commonly accepted method of alignment. In Chapter 6 we therefore proposed to conform with the acoustic theory on which the SM model is founded, and *centre-align* our estimated area-functions, thereby imposing the least amount of bias to the labial and glottal ends. Although the mean vocal-tract shapes and the comparisons thereof were found to be quite plausible, our method of alignment may be rendered even more realistic by acknowledging that phonetic variations in the inferred, articulatory features of larynx height and lip protrusion, are in general of *different* magnitude; and that the difference in magnitude itself is potentially speaker-dependent.

### **7.2.2 Articulatory Explanation of the Dichotomy**

An articulatory explanation of the speech-speaker dichotomy was offered in Chapter 6, the completeness of which was confirmed in the penultimate section therein. Indeed, whilst the *articulatory* completeness of our normalisation method is intuitively clear from the description of the method itself, its *acoustic-phonetic* completeness was only confirmed by the nearly asymptotic, *inter-speaker* accuracy profile obtained after normalisation of all three articulatory components of inter-speaker variability. That result provided the final proof of the effectiveness of our articulatory method of normalisation, and at the same time, it confirmed the validity of our recognition framework as a powerful diagnostic of the speech-speaker dichotomy.

However, the articulatory explanation itself was obtained *en route* to that ultimate proof, whereby the individual contributions to the drop in accuracy across the higher spectral regions were identified for the FC dataset of four adult, male speakers of Australian English. Indeed, we were able to identify contributions to the dichotomy

from two of the speakers' differences in vocal-tract structure; the other two speakers' idiosyncratic articulatory settings; and three of the speakers' idiosyncratic articulatory strategies. Speaker differences in vocal-tract fixed structure were found to induce misclassifications mainly amongst the  $F_2$  and amongst the  $F_3$  of front vowels. Not surprisingly, these were caused by the two speakers who were found to have the shortest and the longest, structure-related vocal-tract lengths. Perhaps less anticipated, is the fact that the admittedly marginal difference between those speakers' mean, structural lengths of 16.26 cm and 16.97 cm, respectively, is apparently sufficient to contribute to the observed drop in vowel classification accuracy across the higher spectral regions. The other two speakers were then found to have dichotomy-inducing idiosyncrasies in long-term articulatory setting, with one speaker preferring a slightly uvularised setting together with a more anterior place of constriction for high-front vowels, and the other speaker preferring a palato-velarised and slightly lip-rounded setting. These long-term behavioural idiosyncrasies were found to induce misclassifications mainly amongst the  $F_2$  of front vowels and amongst the  $F_3$  of back vowels. However, perhaps not unexpectedly for our dataset of four, adult male speakers, the largest contributions to the dichotomy were found to be caused by differences in vowel-specific articulatory strategy, including vocal-tract shape-inferred variations (i) in larynx height of both front and back vowels, (ii) in the maximum area of the front cavity expansion of back vowels, and (iii) in the place of constriction along the length of the vocal-tract. These dimensions of speaker variation in strategy were found to induce misclassifications mainly amongst the  $F_3$  of both front and back vowels.

In view of the clearly marked differences in both articulatory setting and strategy observed in Chapter 6 for speaker D, it is interesting to consider whether those articulatory-behavioural differences can be regarded as embodying at least some of the distinguishing characteristics of spoken Australian English. Informal auditory impressions (Clermont, 1991) indeed suggest that, of the four speakers in the FC dataset, speaker D tends mostly towards the so-called Cultivated variety, while the other three speakers tend towards the General variety of Australian English. Therefore, a question of practical importance is whether the *perceptually*-determined idiolectal labelling of the speakers in the JB dataset (used to validate the dichotomy in Chapter 4),

can be confirmed *articulatorily*; and by extension, whether our acoustic-articulatory methodology might then be useful in idiolectal labelling of speakers based on the characteristics of their inferred vocal-tract shapes.

Naturally, a more global question arises to what extent our articulatory explanations could be generalised across different speaker datasets. Even considering only those languages which have an inventory of steady-state vowels similar to that of English, we feel compelled to suggest that in general, different groups of speakers will most probably exhibit different types of articulatory variations. For example, dialectal or idiolectal differences may be caused by different types of vowel-specific articulatory strategies, while individual differences in “voice quality” (in the sense of Laver, 1980) may be caused by various types of articulatory settings, and more diverse groups of speakers which include adult males, adult females, and children, would most certainly exhibit a much wider range of vocal-tract structural differences. However, within the limitations of our study, we have determined that if a speaker dataset is endowed with the types of physical variations described in Chapter 6 and summarised above, then a dichotomous behaviour in inter-speaker vowel classification accuracy is indeed very likely.

In this vein, we are admittedly rather fortunate in having had recourse to such an exemplary dataset of four speakers, whose physical differences in vowel production appear to be so uniformly distributed as to allow an unbiased investigation of the acoustic-phonetic consequences of variations in structure, setting, and strategy. It is indeed clearly important to acknowledge (as we have found) that vowel-speaker interactions in the higher spectral regions can be induced by speaker differences in any one of those three, articulatory sources of variability. Furthermore, our results suggest that each of those three types of physical idiosyncrasies may have a distinctive, acoustic-phonetic signature which can be decoded on the basis of the spectral range of greatest detrimental influence. Thus, for example, we might expect that a group of speakers with a particularly wide range of variation in vocal-tract size, would manifest the largest drop in inter-speaker vowel classification accuracy across the spectral ranges encompassing the  $F_2$  and the  $F_3$  of their front vowels.

Dare we now hope for a more “universal” explanation of the speech-speaker

dichotomy than we have been able to offer? In this vein, Peterson's (1959) visionary statement quoted much earlier in Section 2.1.1, concerning the contrastive influences of "gross" and more "exact" vocal-tract shapes on the lower and the higher formants, respectively (indeed, a statement which clearly foreshadows the SM model to appear more than half a decade later), is perhaps the most direct, and the most "universal" explanation yet, of the contrastive roles of phonetic and speaker influences in spoken vowel sounds. Whilst an even more "universal" interpretation will require an extension of our methods in order to cope with greater complexity in both the phonetic and speaker components of variability, our present results do contribute a far more detailed *substantiation* of that explanation than has ever been attempted. Nor do they preclude the possibility of a more general interpretation, which may lead towards the definition of the principal, articulatory components of inter-speaker variation in vowel production.

Indeed, perhaps the most distinguishing characteristic of our articulatory method of speaker normalisation, is its chief purpose not as a method of normalisation *per se*, but as an integral component of an acoustic and articulatory, *explanatory framework*! In this vein, we are reminded of Payan and Perrier's (1993, p.417) statement which we quoted earlier in Section 2.4.3.1, that a method of speaker normalisation "is efficient if it can account for the causes of variability." We have of course argued that such efficiency is not to be expected from purely acoustic-domain methods which primarily aim to increase recognition accuracy using the entire available spectral range; nor from acoustic-domain methods informed only by auditory-perceptual criteria. Rather, it is best attained by returning to the acoustic domain only after having explicitly accounted for the articulatory sources of speaker variability.

In that light, our acoustic-articulatory methodology does represent a significant departure from previous works, as we have succeeded in actually using estimated vocal-tract area-functions for a wider purpose than the mere investigation of the inverse problem itself. While we cannot claim to have rendered either "perfect" or the "most realistic" vocal-tract shapes, we were clearly able to advance understanding of the dichotomy problem in an acoustic-articulatory framework where direct physical measurements were not involved. Therefore, notwithstanding the need for further

evaluation of our method of inversion and area-function parameterisation, the research work reported in this thesis lends support to the following, emerging contention. Namely, that the imperfections of acoustic-to-articulatory mapping (in particular those ascribed to the over-neglected LP model) do not and should not impede progress towards a better understanding of the articulatory bases of phonetic and speaker variability in the acoustics of speech.



## Appendix A

### Formant-tracking Analysis Conditions and Sequence Charts

In this appendix are shown, for each of the four (adult, male) speakers of the FC dataset (cf. Chapter 3), a table which summarises the formant-tracking problems encountered and LP analysis conditions finally adopted, and a figure (“formant sequence chart”, as used by Potter and Steinberg, 1950) which shows the entire, steady-state vowel formant-frequency data finally extracted from the vocalic nuclei of the recorded, /hVd/ monosyllabic words.

	Repetition					Summary
	I	II	III	IV	V	
heed	? F3	? F4, F3 M18 FA10	? F3, F4 M18 FA13	? F3, F4 M18 FA25	? F3, F4 FA10	? F3
hid		? F4, F3 M13 FA15	? F3, F4 FA25	? F3, F4 FA25	? F3, F2, F4 FA10	
head	? F4 FA10	? F4, F3 M13 FA10	? F4 adPR FA20	? F4 FA30	? F4 PR100 FA10	? F4 FA↓
had	? F4, F3 M17 FA15	? F4, F1 adPR FA25	? F4, F1, F2 FA10	? F4 M15 FA12	? F4, F1 M15 FA10	? F4 FA↓
hard	? F4 M12, adPR	? F4	? F4 M12, PR80 FA20	? F4 M11, PR80 FA20	? F4 M11, PR80 FA20	? F4
hod	? F3, F4 FA20	? F3, F4 adPR	? F4, F3 FA20	? F4, F3, F2 M16 FA35	? F3 M15 (?F1,F2:sts)	? F3
hoard	? F4 M15	? F4 M17, adPR	? F4 M15	? F4 M12	(?F2,F4:sts)	? F4
hood	? F4 PR90 FA40	? F2, F4 M12 <b>DP2: F1, F2</b>	? F4 M12, PR96 FA30	? F4 M12 FA30	? F4 M12, PR90 FA20	? F4
who'd	? F4 (spur) <b>DP1:uwCep</b>	? F4 M16				
hud	? F4	? F4 M13	? F4, F3 M12, PR99 FA14	? F4 M15	? F4, F3 M12, PR100 FA18	? F4
herd		? F4 M15	? F4 M16, PR90 (?F2,F4:sts)	? F2 M15	? F4 M17, PR90	

Table A.1: Formant-tracking LP analysis conditions for speaker A. The first column lists the orthographic representations of the /hVd/ monosyllabic words, the vowel nuclei in which the formant measurements were taken. The second through sixth columns list, for each of the 5 repetitions of each vowel, the ill-tracked formant(s) and the LP analysis conditions finally chosen to secure “good” (see Section 3.3.2.1) formants. (? : problematic formant-track(s); M: LP analysis order; PR: coefficient of pre-emphasis multiplied by 100; adPR: frame-by-frame adaptive pre-emphasis; FA: frame advance in msec multiplied by 10; **DP1:uwCep** refers to the unweighted cepstral distance measure used in the formant tracker; **DP2**: refers to the use of a formant-trajectory smoothing function in the formant tracker; spur: refers to a spurious-pole trajectory; sts: refers to a steady-state detection problem). An empty entry implies no formant tracking problems using the default analysis conditions (see Section 3.3.2). The last column lists a summary, where applicable, of the consistently encountered (i.e., across all 5 repetitions) problem(s) and required correction(s) for each vowel.

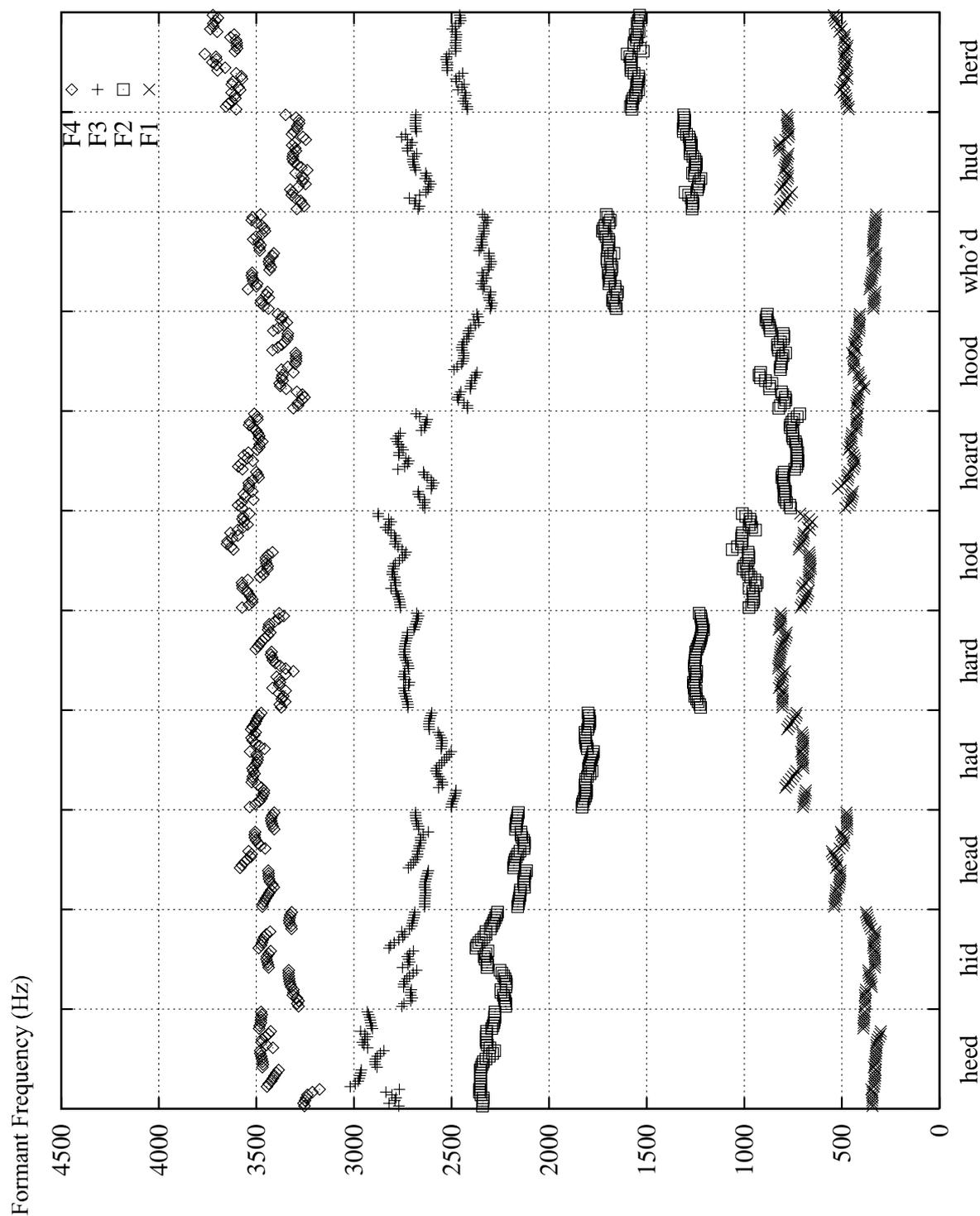


Figure A.1: *Formant-frequency sequence chart for speaker A.* Along the abscissa are listed the orthographic representations of the /hVd/ monosyllabic words, the vowel nuclei in which the formant measurements were taken. For each vowel, there are 35 data points per formant, corresponding to the 7 steady-state frames in each of the 5 repetitions.

	Repetition					Summary
	I	II	III	IV	V	
heed	? F2, F3, F4 M9, PR88	? F3, F2, F4 M10	? F3, F4, F2 M16 <b>DP2: F3</b>	? F4(spl), F2 M12, PR99	? F4(spl), F2 M11, PR99	? F4, F2
hid	? F4(spl) M10	? F4(spl), F3 M9	? F4(spl) M10	? F4(spl) M10	? F4(spl) M12	? F4(spl) M↓
head	? F4 M15, PR95	? F1, F3 M13		? F4 (..F5) M13	? F4 (..F5) M12 <b>DP2: F1..3</b>	
had	? F3 M13	? F4 M13	? F1, F2 adPR			
hard			? F3, F4 M13	? F4, F2 M13 <b>DP2: F2</b>	? F4,F2,(F1) M12 <b>DP2: F1..3</b>	
hod	? F4, F3 M12	? F3, F4 M12 (?F4:sts)	? F3,F4,(F2) M11, PR90	? F3, F4 M12	? F4(spl), F2 M11, PR80 <b>DP2: F2</b>	? F4 M↓
hoard			? F2, F4 M13		? F2, F3, F4 M13 <b>DP2: F1..3</b>	
hood	? F4(spl), F3 M9	? F4(spl) M12	? F4, F3 M12 <b>DP2: F3</b>	? F4, F2, F3 M12, PR100 <b>DP2: F2,F3</b>		
who'd	? F2 M13, PR95	? F4 M13	? F2, F4 M12, PR95	? F3 (..F4) M13	? F3, F4 M13	M↓
hud	? F1..4 M17, adPR <b>DP2: F2,F3</b>		? F2 M12 <b>DP2: F2</b>	? F3 M13	? F4, F2, F3 M15, PR95 <b>DP2: F2</b>	
herd	? F4, F2 M12 <b>DP2: F2,F3</b>	? F4 M15	? F2	? F4 M15, adPR	? F2, F3, F4 M13	

Table A.2: *Formant-tracking LP analysis conditions* for speaker B. The first column lists the orthographic representations of the /hVd/ monosyllabic words, the vowel nuclei in which the formant measurements were taken. The second through sixth columns list, for each of the 5 repetitions of each vowel, the ill-tracked formant(s) and the LP analysis conditions finally chosen to secure “good” (see Section 3.3.2.1) formants. (? : problematic formant-track(s); M: LP analysis order; PR: coefficient of pre-emphasis multiplied by 100; adPR: frame-by-frame adaptive pre-emphasis; FA: frame advance in msec multiplied by 10; **DP2**: refers to the use of a formant-trajectory smoothing function in the formant tracker; spl: refers to a “split” formant trajectory; sts: refers to a steady-state detection problem). An empty entry implies no formant tracking problems using the default analysis conditions (see Section 3.3.2). The last column lists a summary, where applicable, of the consistently encountered (i.e., across all 5 repetitions) problem(s) and required correction(s) for each vowel.

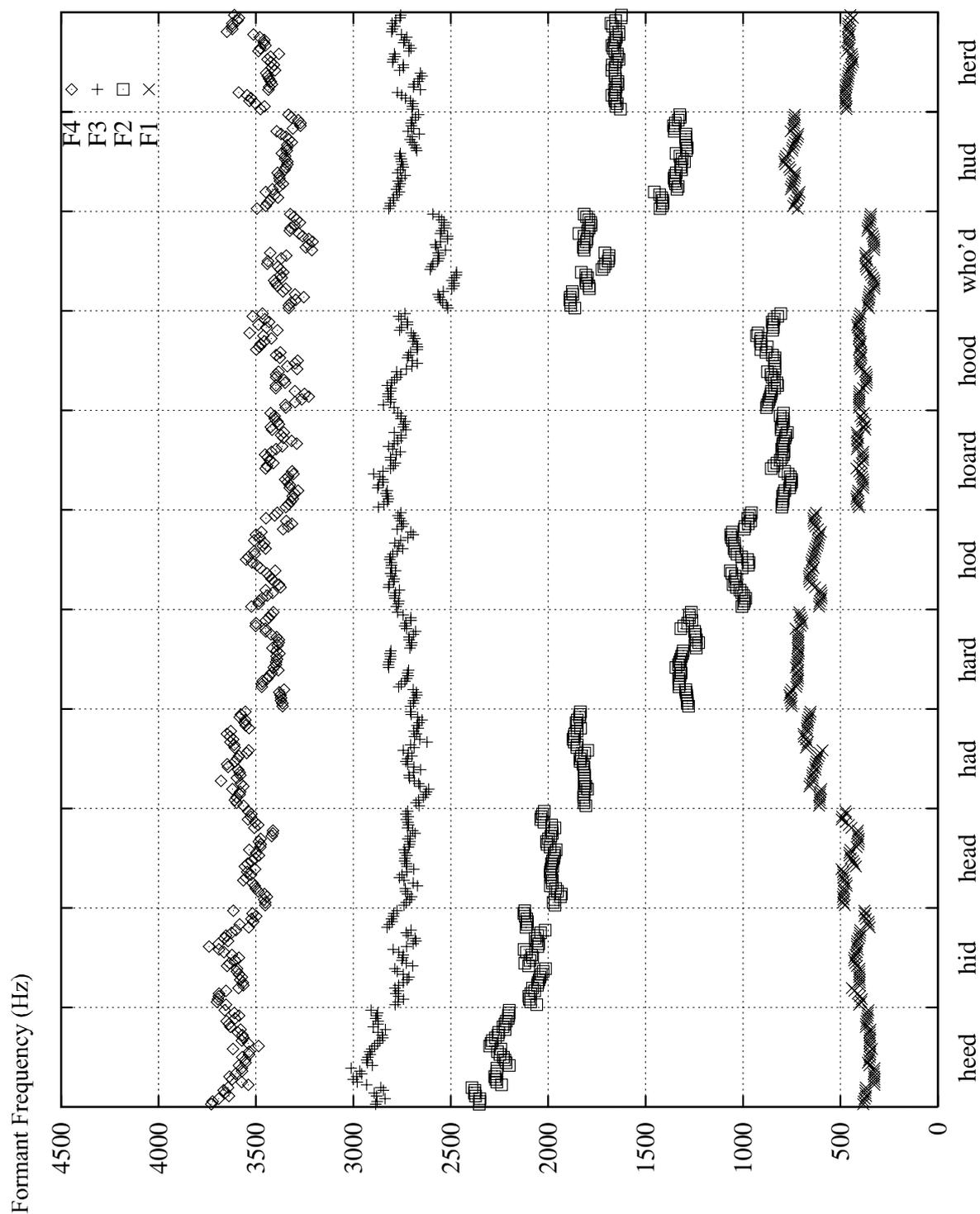


Figure A.2: *Formant-frequency sequence chart for speaker B.* Along the abscissa are listed the orthographic representations of the /hVd/ monosyllabic words, the vowel nuclei in which the formant measurements were taken. For each vowel, there are 35 data points per formant, corresponding to the 7 steady-state frames in each of the 5 repetitions.

	Repetition					Summary
	I	II	III	IV	V	
heed	? F3, F4 M16, adPR	? F4, F3, F2 M13, adPR	? F3 M13	? F3 M15, adPR	? F3 M15, adPR	? F3
hid	? F1, F3 M15	? F4 M15	? F2, F3 M15 <b>DP2: F3</b>	? F3, F4 M18 <b>DP2: F3</b>	? F4 M15	M↑
head		? F4 M15		? F4 M18, adPR		IRV: F3, F4
had		? F4 M15, adPR	? F2, F4 M12 <b>DP2: F2</b>	? F4 M12		IRV: F3
hard	? F3 M13	? F3 M13	? F2, F4 M13		? F2, F4 M13	
hod	? F1..4 M13	? F3, F4 M13	? F3, F4 M15	? F3, F4 M13	? F4, F1, F3 M17, PR90	? F3, F4
hoard	? F3 M13	? F2, F4 M13				
hood	? F1, F3 M13	? F3, F4 M16, adPR <b>DP2: F3</b>	? F1, F2, F3 M13 <b>DP2: F1..3</b>	? F3, F4 M13, adPR	? F3 adPR	? F3
who'd	? F4 M13	? F1, F2, F4 M13	? F3, F4 M13	? F2, F4 M15	? F3, F4 M13	? F4
hud	? F2, F3 M13, adPR <b>DP2: F3</b>	? F3, F1 M13, adPR <b>DP2: F3</b>	?F2(spl)F1,3 M12	? F2, F4 M13	? F2(spl), F3 M13	M↓
herd	? F2, F3 M12				? F2 M13	

Table A.3: *Formant-tracking LP analysis conditions* for speaker C. The first column lists the orthographic representations of the /hVd/ monosyllabic words, the vowel nuclei in which the formant measurements were taken. The second through sixth columns list, for each of the 5 repetitions of each vowel, the ill-tracked formant(s) and the LP analysis conditions finally chosen to secure “good” (see Section 3.3.2.1) formants. (? : problematic formant-track(s); M: LP analysis order; PR: coefficient of pre-emphasis multiplied by 100; adPR: frame-by-frame adaptive pre-emphasis; FA: frame advance in msec multiplied by 10; **DP2**: refers to the use of a formant-trajectory smoothing function in the formant tracker; spl: refers to a “split” formant trajectory). An empty entry implies no formant tracking problems using the default analysis conditions (see Section 3.3.2). The last column lists a summary, where applicable, of the consistently encountered (i.e., across all 5 repetitions) problem(s) and required correction(s) for each vowel. (IRV: refers to the apparently intrinsic presence of inter-repetition variability.)

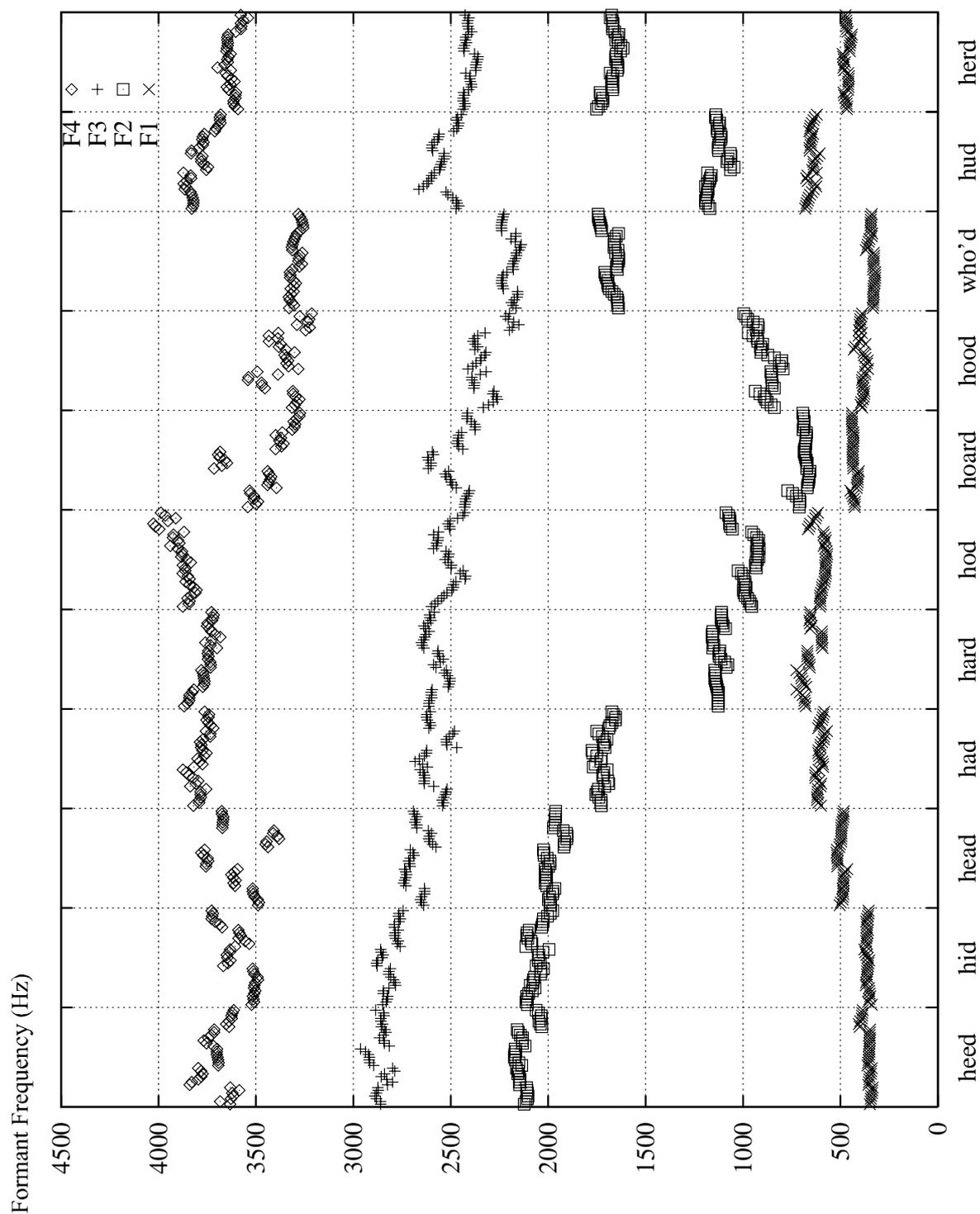


Figure A.3: *Formant-frequency sequence chart for speaker C.* Along the abscissa are listed the orthographic representations of the /hVd/ monosyllabic words, the vowel nuclei in which the formant measurements were taken. For each vowel, there are 35 data points per formant, corresponding to the 7 steady-state frames in each of the 5 repetitions.

	Repetition					Summary
	I	II	III	IV	V	
heed		? F3, F4 M13	? F3, F4 M17 <b>DP2: F3</b>			
hid	? F3 M13	? F3 M13	? F3, F4 M18, adPR	? F3 M13	? F3, F4 M15	? F3
head	? F4 M15	? F4 M15, adPR	? F4, F3 M15, adPR	? F4 M18	? F4 M18	? F4 M↑
had	? F3, F4 M15	? F4, F3 M15	? F3 M13, adPR	? F3 <b>DP2: F3</b>	? F3 adPR	? F3 IRV: F4
hard	? F3 M12, adPR	? F3 M12, adPR	? F2, F3, F1 M13, adPR <b>DP2: F1..3</b>	?F3(spl)F1,2 M13, adPR	? F3, F4 M12, adPR	? F3 M↓, adPR
hod	? F3, F1, F2 M15, adPR	? F3 M11, PR96	? F3 M15	? F4 M16, adPR		
hoard	? F4, F1,2,3 M15, adPR <b>DP2: F1..3</b>		? F4 M15, PR96	? F3, F4 M16	? F3 M15	
hood	? F3, F4 M13, adPR <b>DP2: F3</b>	? F3, F4 M13 <b>DP2: F3</b>	? F4, F3 M13 <b>DP2: F3</b>	? F3, F4 M12, PR93 <b>DP2: F3</b>	? F3, F4 M17	? F3, F4
who'd		? F2, F3 M13			? F4 M16	
hud	? F3, F4 M12, PR96 <b>DP2: F3</b>	? F2 M13, adPR <b>DP2: F2,F3</b>	? F1, F3, F4 M13 <b>DP2: F3</b>	? F1, F4 M12	? F3 M12	M↓
herd				? F4 M13		

Table A.4: *Formant-tracking LP analysis conditions* for speaker D. The first column lists the orthographic representations of the /hVd/ monosyllabic words, the vowel nuclei in which the formant measurements were taken. The second through sixth columns list, for each of the 5 repetitions of each vowel, the ill-tracked formant(s) and the LP analysis conditions finally chosen to secure “good” (see Section 3.3.2.1) formants. (? : problematic formant-track(s); M: LP analysis order; PR: coefficient of pre-emphasis multiplied by 100; adPR: frame-by-frame adaptive pre-emphasis; FA: frame advance in msec multiplied by 10; **DP2**: refers to the use of a formant-trajectory smoothing function in the formant tracker; spl: refers to a “split” formant trajectory). An empty entry implies no formant tracking problems using the default analysis conditions (see Section 3.3.2). The last column lists a summary, where applicable, of the consistently encountered (i.e., across all 5 repetitions) problem(s) and required correction(s) for each vowel. (IRV: refers to the apparently intrinsic presence of inter-repetition variability.)

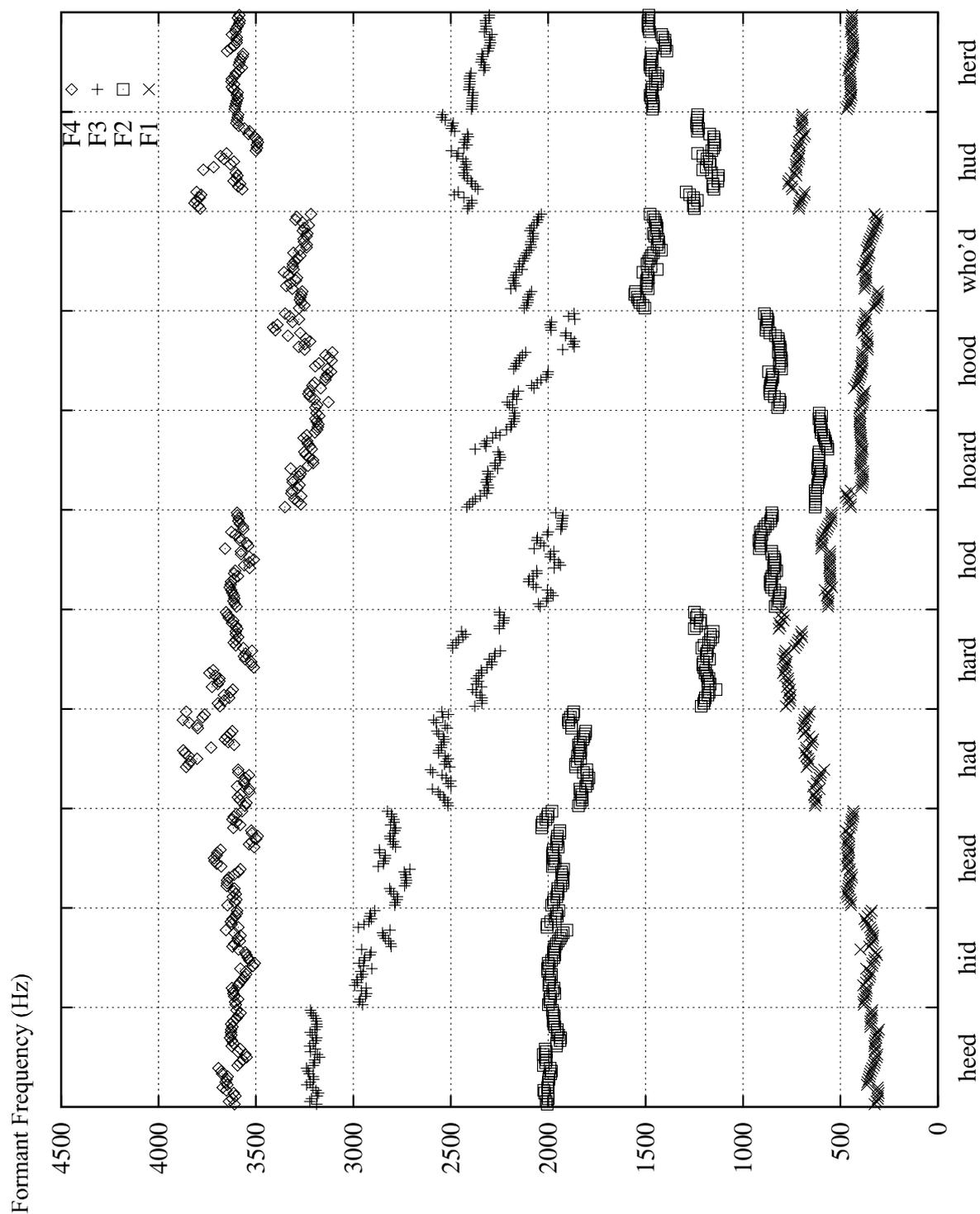


Figure A.4: *Formant-frequency sequence chart for speaker D.* Along the abscissa are listed the orthographic representations of the /hVd/ monosyllabic words, the vowel nuclei in which the formant measurements were taken. For each vowel, there are 35 data points per formant, corresponding to the 7 steady-state frames in each of the 5 repetitions.



## **Appendix B**

### **Mathematical Derivation of the SM Model**

For the sake of completeness, we herein offer a complete re-derivation of the SM vocal-tract model (cf. Chapter 5) which was first presented by Schroeder and Mermelstein (1965) and later expanded by each author separately (Schroeder, 1967; Mermelstein, 1967). Whilst we follow mainly Schroeder's (1967) mathematical derivation in the latter half of this Appendix, we start with some basic assumptions and properties of a completely lossless acoustic tube, which help to clarify certain aspects of the nonuniqueness problem referred to in Chapter 5. The final result of the SM model itself is given in Equations B.22 and B.20, which are re-stated as Equations 5.1 and 5.2, respectively, in Section 5.2.1.



A static configuration of the supralaryngeal vocal-tract is usually modelled as an acoustic tube with a cross-sectional area-function  $A(x)$  which varies along its length (the  $x$ -axis) from the glottis (at  $x = 0$ ) to the lips (at  $x = L$ ). Only planar (i.e., one-dimensional) wave propagation is considered, which is a valid assumption for frequencies below about 4 kHz, where acoustic wavelengths are larger than a typical cross-sectional dimension. If the tube is further regarded as completely lossless and rigid-walled, then the spatio-temporal behaviour of acoustic pressure  $p(x,t)$  and volume velocity  $u(x,t)$  within the vocal-tract are governed by the following two, well-known equations (e.g., Morse and Ingard, 1968, p.243) of *momentum* (Equation B.1) and *continuity of mass* (Equation B.2):

$$\frac{\partial p(x,t)}{\partial x} = -\frac{\rho}{A(x)} \frac{\partial u(x,t)}{\partial t} \quad (\text{B.1})$$

$$\frac{\partial u(x,t)}{\partial x} = -\frac{A(x)}{\rho c^2} \frac{\partial p(x,t)}{\partial t}, \quad (\text{B.2})$$

where  $\rho$  is the air density and  $c$  is the velocity of sound propagation through the air inside the acoustic tube. These differential equations are combined to yield the well-known Webster's Horn Equation, which can be expressed either in terms of acoustic pressure:

$$\frac{\partial}{\partial x} \left\{ A(x) \frac{\partial p(x,t)}{\partial x} \right\} = \frac{A(x)}{c^2} \frac{\partial^2 p(x,t)}{\partial t^2}, \quad (\text{B.3})$$

or in terms of volume velocity:

$$\frac{\partial}{\partial x} \left\{ \frac{1}{A(x)} \frac{\partial u(x,t)}{\partial x} \right\} = \frac{1}{c^2 A(x)} \frac{\partial^2 u(x,t)}{\partial t^2}. \quad (\text{B.4})$$

Without loss of generality, the time-dependence of acoustic pressure and volume velocity is taken to be complex sinusoidal, as follows:

$$p(x,t) = p(x)e^{j\omega t}, \quad (\text{B.5})$$

$$u(x,t) = u(x)e^{j\omega t}, \quad (\text{B.6})$$

where  $j$  is the imaginary unit, and  $\omega$  is the angular frequency. If the vocal-tract area-function is then assumed to be that of a uniform tube (such that  $A(x) = A_0$ , some constant), Equation B.3 reduces (by also eliminating the time-dependence using

Equation B.5) to the simple one-dimensional wave equation for acoustic pressure:

$$\frac{d^2 p(x)}{dx^2} = -\frac{\omega^2}{c^2} p(x), \quad (\text{B.7})$$

and Equation B.4 likewise reduces (using Equation B.6) to the simple one-dimensional wave equation for volume velocity:

$$\frac{d^2 u(x)}{dx^2} = -\frac{\omega^2}{c^2} u(x). \quad (\text{B.8})$$

Simplified boundary conditions are defined by first regarding the glottis as a perfect volume-velocity source with infinite impedance, such that:

$$u(0) = 0, \quad \frac{dp(0)}{dx} = 0. \quad (\text{B.9})$$

Ignoring end corrections, the open lip-end of the vocal-tract is then assumed to have zero impedance, such that:

$$p(L) = 0, \quad \frac{du(L)}{dx} = 0. \quad (\text{B.10})$$

With these boundary conditions, the solutions to Equations B.7 and B.8 are therefore characterised by the following eigenvectors for acoustic pressure and volume velocity, respectively:

$$p_n(x) = \cos\left(\frac{(2n-1)\pi x}{2L}\right), \quad n = 1, 2, 3, \dots \quad (\text{B.11})$$

$$u_n(x) = \frac{1}{Z_0} \sin\left(\frac{(2n-1)\pi x}{2L}\right), \quad n = 1, 2, 3, \dots, \quad (\text{B.12})$$

and by the eigenvalues:

$$\lambda_n = \frac{\omega_n^2}{c^2} = \left(\frac{(2n-1)\pi}{2L}\right)^2, \quad n = 1, 2, 3, \dots, \quad (\text{B.13})$$

where  $n$  denotes the index of the mode of resonance, and  $Z_0 = \rho c / A_0$  is the characteristic impedance of the uniform acoustic tube.

Mermelstein (1967) then proceeds to invoke first-order perturbation theory to yield an expression for  $\delta\lambda_n$  in terms of an assumed form of perturbation of the uniform area-function. Deeper insights are gained, however, by re-tracing Schroeder's (1967) derivation, as it continues to build on the acoustic properties of the vocal-tract model.

In particular, Schroeder's derivation is founded on the physical property that a perturbation  $\delta A(x)$  of the uniform area-function induces a change in the total energy  $E_n$  of the  $n^{\text{th}}$  mode of resonance, as follows:

$$\delta E_n = -\int_0^L P_n(x) \delta A(x) dx, \quad (\text{B.14})$$

where  $P_n(x)$  is the spatial distribution of *acoustic radiation pressure* for the  $n^{\text{th}}$  resonance mode, and is given by the difference between the potential energy per unit volume ( $p_n^2(x) / 2\rho c^2$ ) and the kinetic energy per unit volume ( $Z_0^2 u_n^2(x) / 2\rho c^2$ ). Using Equations B.11 and B.12, the resultant difference in energy densities is given by the following:

$$P_n(x) = \frac{1}{2\rho c^2} \cos\left(\frac{(2n-1)\pi x}{L}\right). \quad (\text{B.15})$$

The *total energy density* of the  $n^{\text{th}}$  resonance mode, on the other hand, is the sum of the potential and kinetic energy densities for that mode. Using Equations B.11 and B.12, the following relation is obtained:

$$E_n = \frac{L}{2cZ_0}, \quad (\text{B.16})$$

which, when combined with Equation B.15, yields the following result for acoustic radiation pressure:

$$P_n(x) = \frac{E_n}{A_0 L} \cos\left(\frac{(2n-1)\pi x}{L}\right). \quad (\text{B.17})$$

Returning to Equation B.14, Schroeder (1967) then suggests an expression for the area-function perturbation in terms of a *Fourier cosine series* which, in view of the form of  $P_n(x)$  in Equation B.17, secures the principle of orthogonality in computing the integral, and thereby eliminates all but one term, per resonance mode. Thus, assuming:

$$\delta A(x) = A_0 \sum_{m=1}^{\infty} a_m \cos\left(\frac{m\pi x}{L}\right), \quad (\text{B.18})$$

and inserting Equations B.17 and B.18 into Equation B.14, orthogonality of the components of the cosine series leads to the following, rather elegant solution:

$$\delta E_n = -\frac{1}{2} E_n a_{2n-1}, \quad (\text{B.19})$$

which states that a perturbation of the uniform area-function by the single cosine term  $a_{2n-1}$  will, to first approximation, change the total energy of only the  $n^{\text{th}}$  mode of resonance.

The final step invokes a theorem by Ehrenfest which is “*independent of the type of wave equation*” (Schroeder, 1967, p.1003), and which states that, for small, adiabatic perturbations of an undamped linear oscillator, the relative change in the *total energy* of the  $n^{\text{th}}$  resonance mode is equal to the relative shift in the *frequency* of that mode. This means that Equation B.19 can be re-written in terms of the  $n^{\text{th}}$  resonance (or formant) frequency  $F_n = \omega_n / 2\pi$ , as follows:

$$\frac{\delta F_n}{F_n} = -\frac{1}{2} a_{2n-1}. \quad (\text{B.20})$$

The essence of the model is captured in this last result, which directly relates each formant frequency to a unique vocal-tract shape parameter.

The final form of the vocal-tract shape parameterisation which led to this result, is obtained by first observing that for small perturbations, the natural-logarithm of the area-function can be approximated as follows:

$$\ln[A(x)] = \ln[A_0 + \delta A(x)] \approx \ln(A_0) + \frac{\delta A(x)}{A_0}. \quad (\text{B.21})$$

As suggested by Schroeder (1967), a logarithmic representation guarantees positive-valued areas, and in addition, it linearly separates the constant scale-factor from the perturbation term. This is clearly seen by inserting  $\delta A(x)$  from Equation B.18 into Equation B.21, to obtain the following:

$$\ln A(x) = \ln A_0 + \sum_{m=1}^M a_m \cos\left(\frac{m\pi x}{L}\right), \quad (\text{B.22})$$

where, in practice, the summation is limited to a finite number of terms  $M$ .

The complete model, which we refer to as the Schroeder-Mermelstein (SM) model, is summarised in Equation B.22 (reproduced as Equation 5.1 in Chapter 5) which defines the vocal-tract shape parameterisation, and in Equation B.20 (reproduced as Equation 5.2 in Chapter 5) which defines the relation between the acoustic- and the articulatory-domain parameters.

## **Appendix C**

### **A Lossy Vocal-Tract Acoustic Model**

This appendix describes the lossy vocal-tract acoustic model used in Chapter 5 (first in Section 5.2.1 where we disregarded all the losses, then in Section 5.5.2.2 where the losses were used) to synthesise the formants of given area-functions (in particular, the formant frequencies shown in the nomograms in Figure 5.1, and the formant frequencies and bandwidths listed in Table 5.4). The model is a transmission-line analog of the vocal-tract, suitable for the synthesis of the steady-state formants of static vocal-tract configurations corresponding to non-nasalised, vocalic speech sounds, with the vocal-tract excited at the glottis and with the radiating plane at the lips.



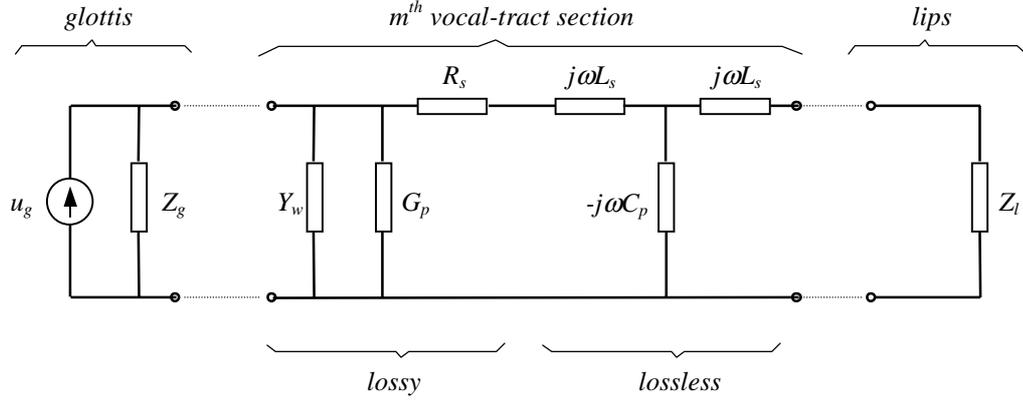


Figure C.1: Electrical transmission-line analog of the vocal-tract.

Our implementation of a lossy vocal-tract acoustic model and the associated method of formant synthesis is based primarily on the numerical method described by Atal et al. (1978, pp.1539-1542), together with other, well-known references such as Fant (1960), Flanagan (1972), Wakita and Fant (1978), Badin and Fant (1984), and Ohmura (1993). In particular, as shown in Figure C.1, an  $M$ -section vocal-tract area-function is represented as a serial concatenation of  $M$  electrical transmission-line sections, terminated at one end by a constant current source and a (time-invariant) glottal impedance, and at the other end by a lip impedance across which acoustic energy is radiated. The relation between the acoustic pressure  $p$  (or electrical voltage) and volume velocity  $u$  (or electrical current) at the input side (the glottal end, denoted by the subscript  $g$ ) and at the output side (the lip end, denoted by the subscript  $l$ ) of the entire network, is then given by the following expression:

$$\begin{bmatrix} p_g(s) \\ u_g(s) \end{bmatrix} = \begin{bmatrix} A & B \\ \Gamma & \Delta \end{bmatrix} \begin{bmatrix} p_l(s) \\ u_l(s) \end{bmatrix}, \quad (\text{C.1})$$

where  $s = \sigma + j\omega$  is the complex-frequency variable, and the so-called, overall transmission matrix is given as follows:

$$\begin{bmatrix} A & B \\ \Gamma & \Delta \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1/Z_g & 1 \end{bmatrix} \prod_{m=1}^M \begin{bmatrix} \alpha_m & \beta_m \\ \gamma_m & \delta_m \end{bmatrix}. \quad (\text{C.2})$$

In Equation C.2 above, the first matrix on the right-hand side is the transmission matrix of the glottal impedance, which itself is defined as follows:

$$Z_g = R_g + j\omega L_g = \left( \frac{12\nu d_g l_g^2}{A_g^3} + \frac{0.875}{A_g} \sqrt{2\rho P_s} \right) + j \left( \frac{\omega \rho d_g}{A_g} \right), \quad (C.3)$$

where  $\nu = 1.86 \times 10^{-4}$  dyne.sec.cm<sup>-2</sup> is the coefficient of viscosity,  $\rho = 1.14 \times 10^{-3}$  g.cm<sup>-3</sup> is the density of air,  $d_g = 0.3$  cm and  $l_g = 1.8$  cm are the assumed depth (or vertical thickness) and length of the (quasi-stationary) glottal slit,  $P_s = 8.0$  cmH<sub>2</sub>O is the assumed pressure drop across the glottis; and the glottal opening area itself is computed by first determining the equivalent glottal resistance of the given area-function (with glottal-end section-area  $A(0)$ ), glottal reflection coefficient  $\mu_{\text{glott}}$ , and the speed of sound in the vocal-tract airway  $c = 35300$  cm/sec) as follows:

$$R_g = \frac{\rho c}{A(0)} \frac{(1 + \mu_{\text{glott}})}{(1 - \mu_{\text{glott}})}, \quad (C.4)$$

then equating the result with the first term on the right-hand side of Equation C.3 and numerically solving for  $A_g$  (which is a smooth and monotonically decreasing function of  $R_g$ ).

The second matrix on the right-hand side of Equation C.2 is the transmission matrix of the  $m^{\text{th}}$  vocal-tract section, and is given as follows:

$$\begin{bmatrix} \alpha_m & \beta_m \\ \gamma_m & \delta_m \end{bmatrix} = \begin{bmatrix} 1 & R_s \\ G_p + Y_w & 1 + R_s(G_p + Y_w) \end{bmatrix}_m \begin{bmatrix} \cosh(sl_m/c) & (\rho c/S_m) \sinh(sl_m/c) \\ (S_m/\rho c) \sinh(sl_m/c) & \cosh(sl_m/c) \end{bmatrix} \quad (C.5)$$

In Equation C.5 above, the first matrix on the right-hand side is the transmission matrix of the lossy component of the  $m^{\text{th}}$  section, while the second matrix is that of its lossless component (see Figure C.1). The frequency-dependent elements of the former comprise the following, series resistance (which represents the viscosity-induced loss contributed at the  $m^{\text{th}}$  section):

$$R_s = \frac{l_m S_m}{A_m^2} \sqrt{\frac{\rho \nu \omega}{2}}, \quad (C.6)$$

the following, parallel conductance (which represents the heat-conduction loss contributed at the  $m^{\text{th}}$  section):

$$G_p = \frac{(\eta-1)l_m S_m}{\rho c^2} \sqrt{\frac{\lambda \omega}{2\xi\rho}}, \quad (\text{C.7})$$

and the following, parallel admittance (which represents the loss contributed by the reciprocal of the wall-vibration impedance at the  $m^{\text{th}}$  section):

$$Y_w = \left[ \left( \frac{c}{F_w} \right)^2 \left( \frac{\rho}{2\pi l_m A_m} \right) \left( B_{w0} + j \frac{\omega}{2\pi} \right) \right]^{-1}. \quad (\text{C.8})$$

In Equations C.5 through C.8, the variables and constants are defined as follows:  $A_m$  and  $l_m$  are the cross-sectional area ( $\text{cm}^2$ ) and the length (cm), respectively, of the  $m^{\text{th}}$  vocal-tract section;  $S_m = SF \times 2\sqrt{\pi A_m}$  cm is its circumference (where Fant (1972) suggests that the so-called “shape factor” be set to the value  $SF = 2$ , which corresponds to an elliptical cross-sectional shape);  $\eta = 1.4$  is the adiabatic gas constant;  $\lambda = 5.5 \times 10^{-5} \text{ cal.cm}^{-1}\text{sec}^{-1}\text{deg}^{-1}$  is the coefficient of heat-conduction of air;  $\xi = 2.4 \times 10^{-1} \text{ cal.g}^{-1}\text{deg}^{-1}$  is the specific heat of air; and the closed-tract formant frequency and bandwidth values of  $F_w = 190 \text{ Hz}$  and  $B_{w0} = 80 \text{ Hz}$  are approximately those used by Badin and Fant (1984).

Returning now to Equation C.1 and substituting for the acoustic pressure at the lip end  $p_l(s) = Z_l u_l(s)$  in terms of the lip impedance  $Z_l$ , the following expression is obtained for the *volume-velocity transfer function* of the entire vocal-tract model:

$$H(s) = \frac{u_l(s)}{u_g(s)} = \frac{1}{\Gamma Z_l + \Delta}. \quad (\text{C.9})$$

The lip impedance is assumed to be that of a piston in an infinite plane baffle, and is computed as follows (Flanagan, 1972; Ohmura, 1993):

$$Z_l = \frac{\rho c}{A(L)} \left( \sum_{q=1}^Q \frac{(-1)^{q+1}}{q!(q+1)!} \left( \frac{\omega}{c} \sqrt{\frac{A(L)}{\pi}} \right)^{2q} - j \frac{4}{\pi} \sum_{q=0}^Q \frac{(-1)^q}{(2q+1)!!(2q+3)!!} \left( \frac{2\omega}{c} \sqrt{\frac{A(L)}{\pi}} \right)^{2q+1} \right), \quad (\text{C.10})$$

where  $A(L)$  is the cross-sectional area at the lip end,  $q!! = 1 \times 3 \times 5 \times \dots \times (q-2) \times q$ , and the number of terms  $Q = 8$  in order to allow sufficient accuracy for frequencies up to about 5 kHz.

The resonances (or formants) of a given area-function are then simply the *poles* of the volume-velocity transfer function defined in Equation C.9 (where  $\Gamma$  and  $\Delta$  are found by performing the matrix multiplication in Equation C.2), and are located in the second quadrant of the  $s$ -plane. In practice, they are found by searching for the *zeros* of the *reciprocal* of the transfer function in that quadrant, in increasing order of frequency  $\omega$ . Atal et al. (1978) describe a search algorithm which we have found to be effective in many cases, but which still requires manual intervention in order to ensure that a formant has not been erroneously skipped. Having determined (and perhaps corrected) the first few formants  $s_n = \sigma_n + j\omega_n$ ,  $n = 1, 2, \dots$  (where  $\sigma_n \leq 0$  and  $0 < \omega_n < \omega_{n+1}$ ) in the  $s$ -plane, the  $n^{\text{th}}$  formant frequency is then given by:

$$F_n = \omega_n / 2\pi , \quad (\text{C.11})$$

and its bandwidth is given by:

$$B_n = -\sigma_n / \pi . \quad (\text{C.12})$$

The method of synthesis described above is particularly flexible, as it allows selective inclusion of vocal-tract losses. For example, if any particular loss element is not required, then its equivalent resistance is simply set equal to zero. Thus, in our illustration of the formant frequency nomograms in Figure 5.1, we used a completely lossless implementation of the above procedure, by appropriately nullifying all of the resistances in the transmission-line analog. By contrast, to obtain the formant frequencies and bandwidths listed in Table 5.4, we used the completely lossy version of the vocal-tract acoustic model.

## **Appendix D**

### **The LP Method of Inversion “Stepped Down”**

In this appendix, we describe and illustrate one way of “stepping down” the LP method of inversion, such that it yields the same vocal-tract area-function which the completely lossless, SM model would yield using the same set of formant frequencies, and assuming Mermelstein’s (1967) constraint of vocal-tract shape antisymmetry. This supplements our empirical evidence presented in Section 5.3.1.2, concerning the fundamental equivalence of the LP model and the SM model, in regard to their quasi-linear and one-to-one relation between each formant frequency and the corresponding, antisymmetric vocal-tract shape parameter.



As stated in Section 5.3.1.2, a complementary, empirical validation of the fundamental similarity of the SM and the LP vocal-tract models, is to show that the LP model can be used to derive the same vocal-tract shapes which are yielded by the completely lossless, SM model, starting from the same set of formant frequencies. To this end, we first note that the essential difference in the way in which the two models secure uniqueness in estimated area-functions, can be stated in terms of the acoustic parameters required as input to the respective method of inversion. Indeed, whilst the SM model requires the formant frequencies and the *frequencies of the poles of the lip impedance function*, the LP model requires the formant frequencies and the *formant bandwidths*. Initially, one may be tempted to try to simulate the completely lossless SM model, simply by setting the formant bandwidths to zero in the LP model. However, this is almost equivalent to underspecifying the inverse problem, and in any case results in errors when converting the LP autoregressive coefficients to reflection coefficients using the well-known recursive algorithm (Markel and Gray, 1976). A different constraint must therefore be applied to the formant bandwidths which are used as input to the LP method of inversion, while ensuring that their values are positive, finite, and non-zero.

In order to resolve the nonuniqueness problem using the SM method of inversion, Mermelstein (1967) imposed a constraint of *antisymmetry* on allowed vocal-tract shapes (by allowing only the odd-indexed cosine parameters to be non-zero). If we are to coerce the LP model to yield the same vocal-tract shapes that would be obtained by the SM model under these conditions, it is therefore necessary to assume that only the target formant *frequencies* are available, and that the values of the formant *bandwidths* are to be determined such as to obtain a purely *antisymmetrical* vocal-tract shape. The block diagram shown in Figure D.1 shows our solution to this problem. Since only the formant frequencies are assumed to be relevant, the input formant bandwidths are initially set to arbitrary values. The LP inverse routines are then used (either with a fixed vocal-tract length, or by applying Wakita's (1977) method as briefly discussed in Section 5.2.2) to obtain a discrete-sectioned, normalised LP area-function, the antisymmetric components of which are found by inverse discrete cosine transformation (IDCT) of the logarithmic area-function, while its symmetric components are ignored by setting the even-indexed parameters of the SM model to zero.

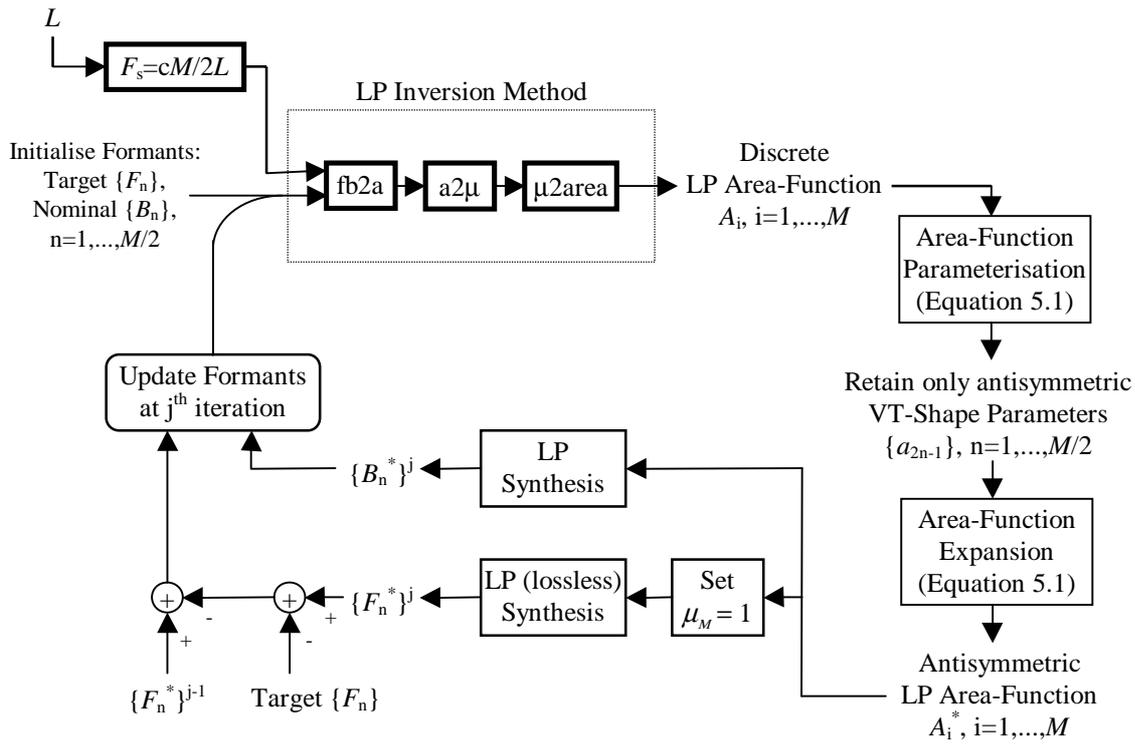


Figure D.1: Block-diagram of area-function estimation method based on the LP model *stepped down* to a completely lossless model, with the constraint of logarithmic area-function *antisymmetry*.

Having thus imposed the constraint of antisymmetry on the logarithmic area-function, the LP synthesis routines are used to obtain a new set of formant bandwidths. The arbitrary bandwidths initially used as input to the LP inverse method are then replaced with the bandwidths just synthesised from the antisymmetric area-function, and a new vocal-tract shape is estimated using the LP method of inversion. This process is repeated until the bandwidths are sufficiently similar from one iteration to the next. However, as shown in Figure D.1, the formant frequencies also need to be iteratively updated, in order to account mainly for the second-order effect of discarding the symmetric components, but also to counteract the small, systematic influence of a finite LP glottal resistance (as discussed in Chapter 5, in regard to the nomograms in Figure 5.3). This is achieved by LP-resynthesising the formant frequencies of the area-function which is forced to be antisymmetric at every iteration, with the glottal reflection coefficient  $\mu_M = 1$  to simulate complete closure of the acoustic tube at the glottal end. The difference between these formant frequencies and the original, target values then defines the incremental change at each successive iteration.

Convergence is finally declared when the magnitude of all required adjustments to formant frequencies and bandwidths does not exceed a certain limit (typically 1.0 Hz). The vocal-tract shape yielded by the LP inverse method at the last iteration is then expected to be almost perfectly antisymmetric (to within numerical accuracies determined by the condition for convergence), with a completely lossless LP resynthesis yielding the target formant frequencies. An interesting by-product of this optimisation scheme is, in answer to the question posed earlier, the identification of the formant bandwidths required for the LP model to yield an antisymmetric area-function.

Figure D.2 illustrates our results obtained by applying this method to the formant frequency data computed by Mermelstein (1967, Table II, p.1288), who used a completely lossless vocal-tract model to obtain the resonances of the six Russian vowels of Fant (1960), from the quantised area-functions (the so-called “exact areas”) measured by Fant from mid-sagittal X-ray images. As only the first three formant frequencies are given by Mermelstein, we use an LP model with  $M = 6$  vocal-tract sections, and the three formant bandwidths are initialised each to a nominal value of 30Hz. In order to avoid artifactual influences on the vocal-tract shapes that might arise from also estimating their lengths, these were fixed to Fant’s original, X-ray measured values. The optimisation process described above was then found to converge in no more than 5 iterations for each of the six Russian vowels, yielding the final, antisymmetric LP area-function shown in each panel of Figure D.2 together with the final values of the formant frequencies and bandwidths required as input to the LP inverse routine.

An interesting, preliminary observation in regard to the formant bandwidths required to finally secure an antisymmetric area-function, is that although they are individually perturbed from their arbitrary initial values, the *mean* formant bandwidth of 30Hz is preserved for each of the six vocalic configurations tested. This observation suggests that the *relative* bandwidths, rather than their absolute values, are more relevant to the LP vocal-tract *shape*. Indeed, this already foreshadows our theoretical and empirical results presented in Section 5.3.2, where we take up this important issue in greater detail.

How do the vocal-tract shapes yielded by our iterative, LP-based method,

compare with those determined by the SM model? Superimposed on each of the six, LP-derived area-functions shown in Figure D.2, is the antisymmetric vocal-tract shape obtained by applying Mermelstein’s (1967, Appendix B) approach to optimise only the odd-indexed parameters of the SM model, using the same, target formant frequencies which we used in our LP-based approach. In deciding the number of sections with which to represent each area-function, we note that Mermelstein (1967, Appendix A) himself chose section-lengths of 0.5 cm in order for the representation “to be valid up to a frequency of 4kHz”. Whilst a greater number of sections would more faithfully represent an area-function having many details and discontinuities, both the LP and the SM model imply that the *theoretical minimum* of  $M = 2LF_s / c$  sections ensures that the vocal-tract shape is acoustically valid up to a frequency of  $F_s / 2$  Hz (with an upper limit of about 4kHz imposed by the assumption of planar acoustic-wave propagation). In view of the inherent smoothness of area-functions generated by the SM model (cf. Equation 5.1), and in order to maintain consistency with the LP method, the theoretical minimum of 6 vocal-tract sections (which corresponds to a frequency range encompassing the first three formants) was therefore used to implement the SM method of inversion.

The estimated, 6-section area-functions are shown in Figure D.2 by the diamond symbols, plotted at the centre of each section; dashed curves are used to emphasise the essentially smooth and continuous representation of vocal-tract shapes afforded by the SM model. Notwithstanding minor differences in implementational details, our smooth area-functions are quite similar to those obtained by Mermelstein (1967, Figure 3). Furthermore, the close proximity of the diamond symbols to the corresponding sections of the LP-derived area-functions, indicate that the latter are essentially a discrete-sectioned representation of the smooth area-functions obtained by the SM model. These results demonstrate the success of our optimisation scheme of Figure D.1, in *stepping down* the LP method of inversion such that it yields area-functions otherwise attainable only by using a completely lossless vocal-tract model. Most importantly, our transformation of the LP model was founded on the very principle of symmetric versus antisymmetric shape perturbations, which itself is the basis of the SM model.

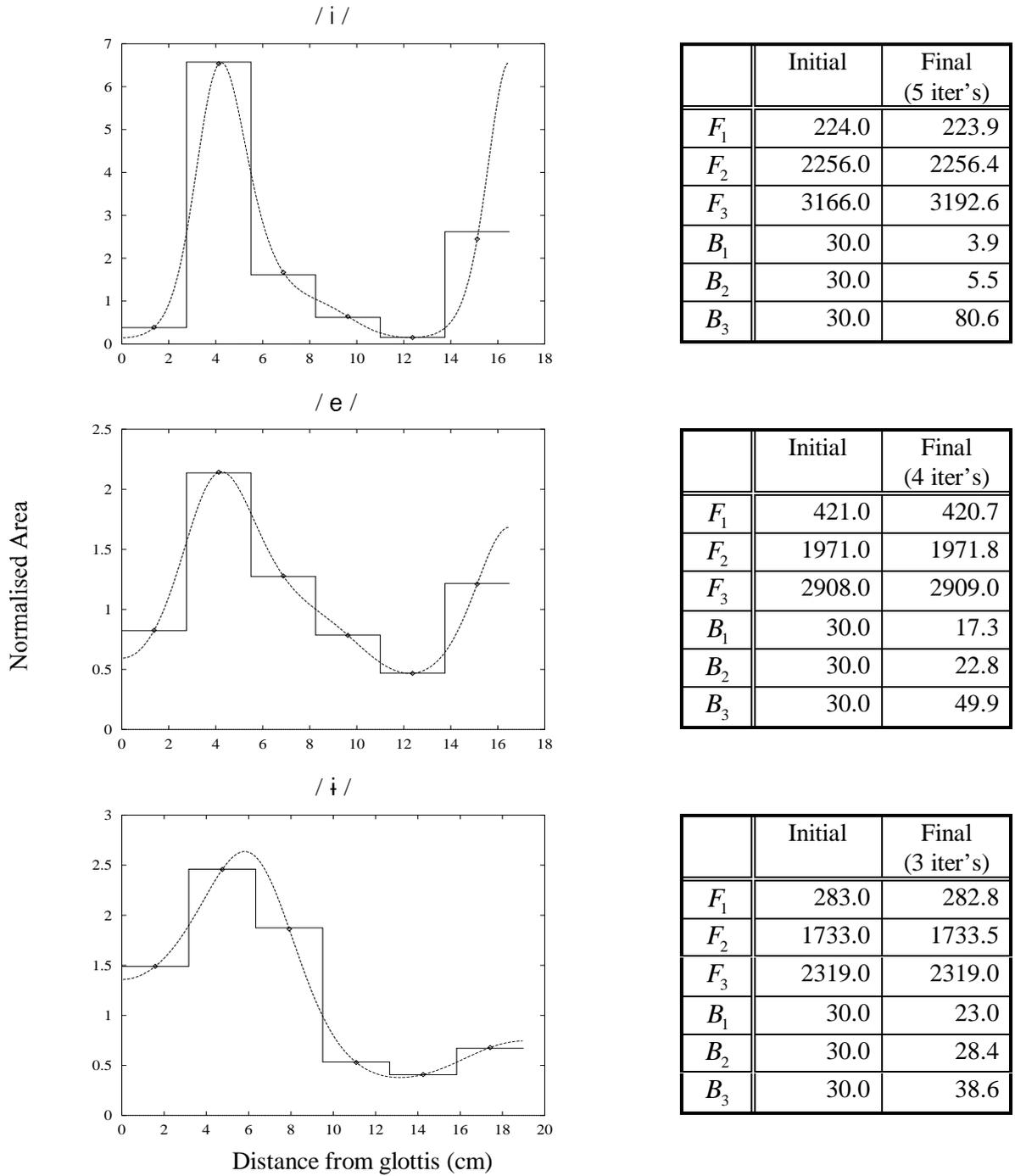
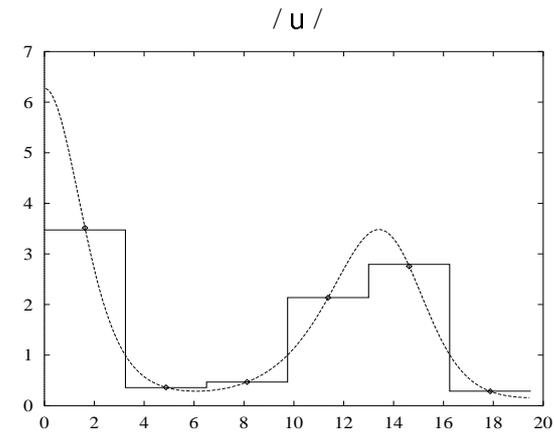
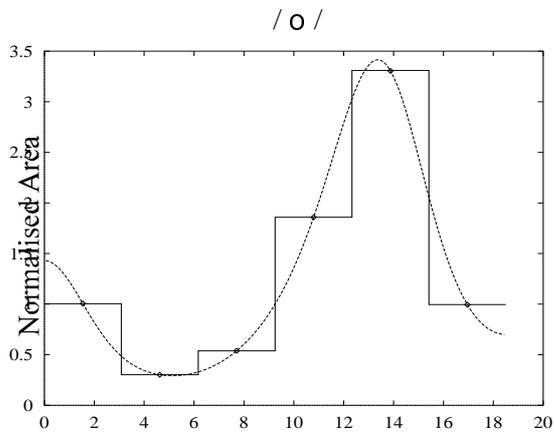


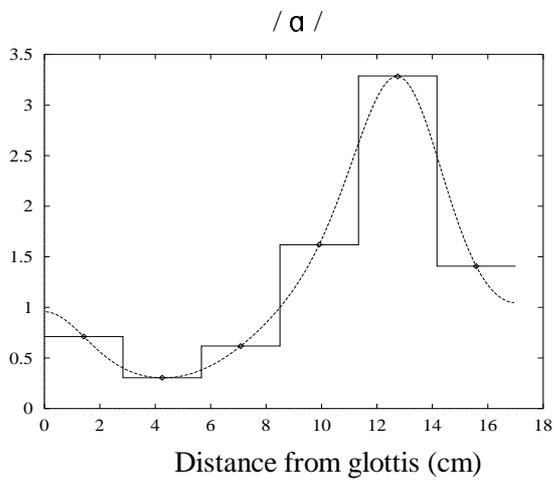
Figure D.2: Vocal-tract area-functions estimated from the first three formant frequencies given by Mermelstein (1967, Table II) for Fant's (1960) six Russian vowels. *Solid lines*: 6-section area-functions obtained by Mermelstein's (1967) optimisation method with the constraint  $a_{2n}=0$ , using a completely lossless vocal-tract model. *Diamond symbols*: 6-section area-functions represented at their section-centres, and obtained by the stepped-down LP inversion method shown in Figure D.1. *Dashed curves*: smoothed version of the 6-section, LP-derived area-functions shown by the diamond symbols, obtained by expanding the estimated vocal-tract shape parameters according to Equation 5.1. *Table to right of each graph*: first three formant frequencies and bandwidths used to initialise the stepped-down LP inversion method (frequencies taken from Mermelstein (1967), bandwidths set to nominal value 30Hz), and those yielded by our method (as input to the LP inverse routines) at the final iteration prior to convergence.



	Initial	Final (5 iter's)
$F_1$	250.0	249.1
$F_2$	601.0	599.1
$F_3$	2308.0	2308.1
$B_1$	30.0	57.7
$B_2$	30.0	28.2
$B_3$	30.0	4.1



	Initial	Final (5 iter's)
$F_1$	523.0	523.4
$F_2$	911.0	909.7
$F_3$	2350.0	2350.0
$B_1$	30.0	45.0
$B_2$	30.0	35.1
$B_3$	30.0	9.8



	Initial	Final (4 iter's)
$F_1$	669.0	669.2
$F_2$	1139.0	1137.6
$F_3$	2487.0	2487.5
$B_1$	30.0	40.7
$B_2$	30.0	36.2
$B_3$	30.0	13.0

Figure D.2: (continued from previous page).

## Appendix E

### Representation of Directly Measured Area-Functions

In this appendix are shown, for each of the 33, directly measured area-functions obtained from the literature (cf. Section 5.4.2.2), the profile of root-mean-square (rms) error as a function of the spatial resolution  $M/2$  used to represent those area-functions, using our method of parameterisation (cf. Equation 5.18). In each figure, the *diamond symbols* (joined by solid lines) show the rms errors obtained in representing that area-function with both symmetric and antisymmetric shape components; the *plus symbols* (joined by dashed lines) show the rms errors using only the antisymmetric shape components; and the *square symbols* (joined by dotted lines) show the rms errors using only the symmetric shape components. The overall mean of each of those three, rms error profiles (across all 33 area-functions), was shown in Figure 5.8.

As discussed in Section 5.4.2.2, the rms errors obtained using only antisymmetric shape components are consistently larger for the mid- to high-, back vowels which have a place of constriction near the mid-length of the vocal-tract (e.g., Fant's (1960) Russian /u/; Baer et al.'s (1991) British English /u/ and American English /u/; Yang and Kasuya's (1994) Japanese /u/ and /o/; Beutemps et al.'s (1995) French /u/; and Story et al.'s (1996) American English /o/, /u/, and /u/).

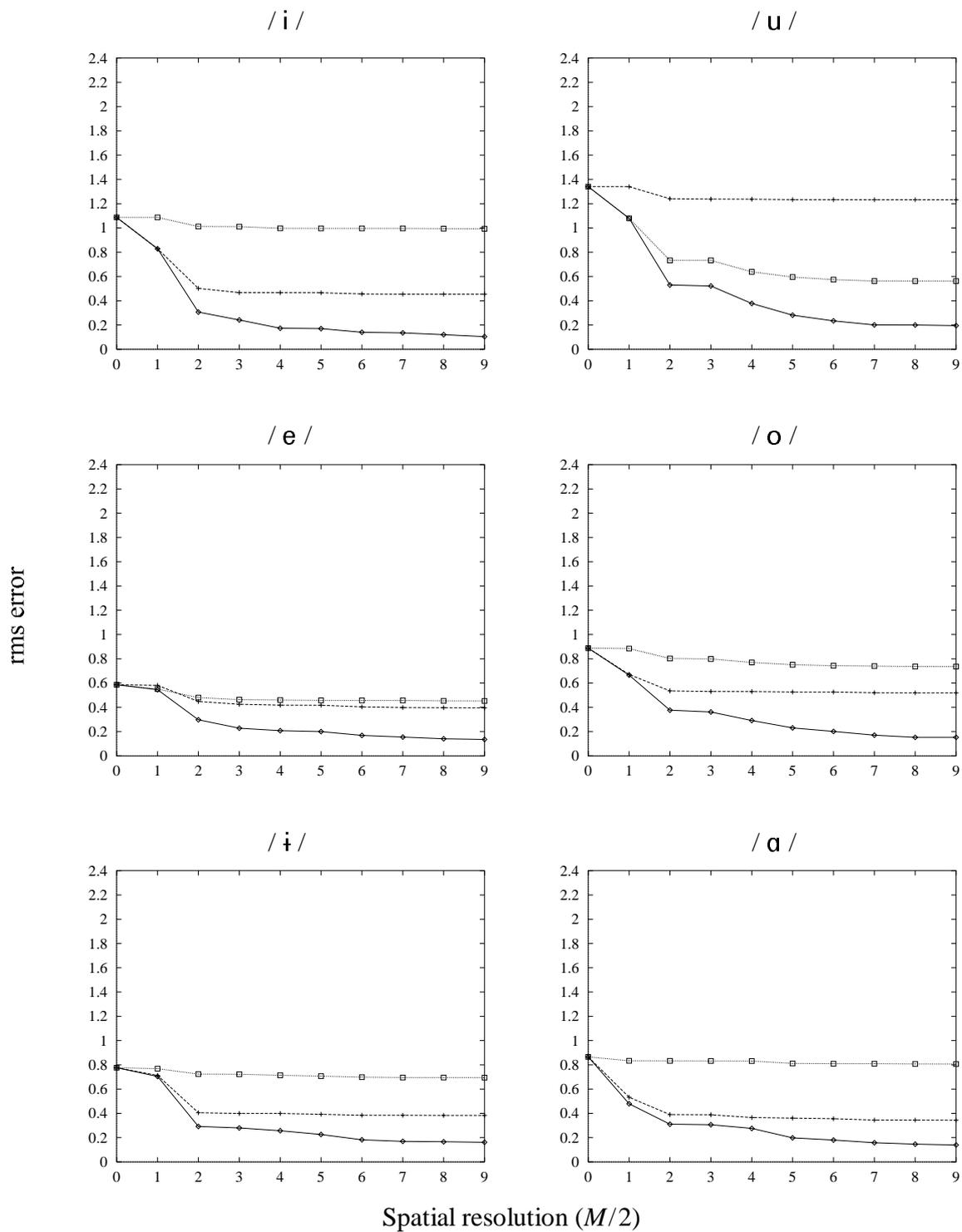


Figure E.1: Profiles of rms error in representing the area-functions of the 6 Russian vowels of Fant (1960).

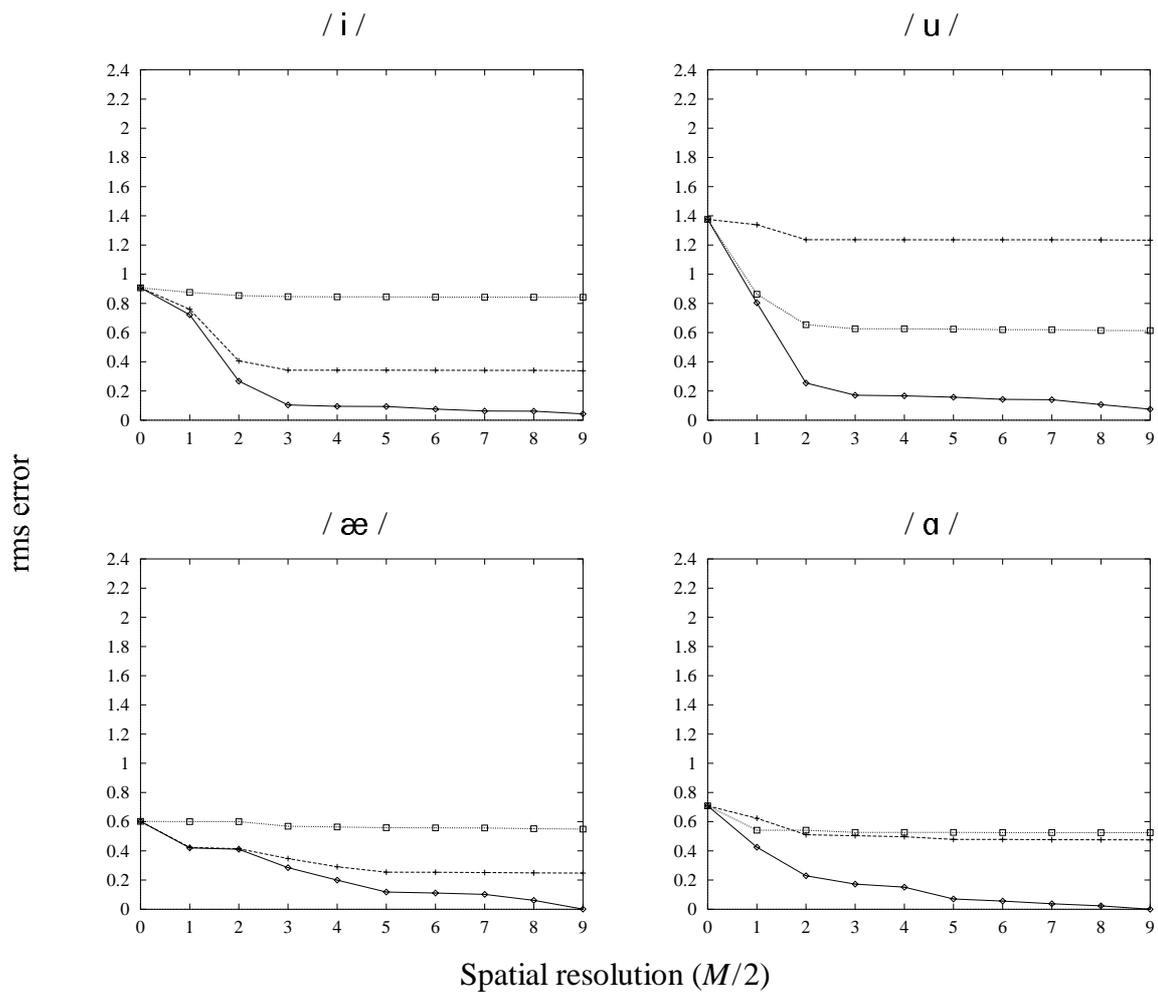


Figure E.2: Profiles of rms error in representing the area-functions of the 4 British English vowels (Speaker PN) of Baer et al. (1991).

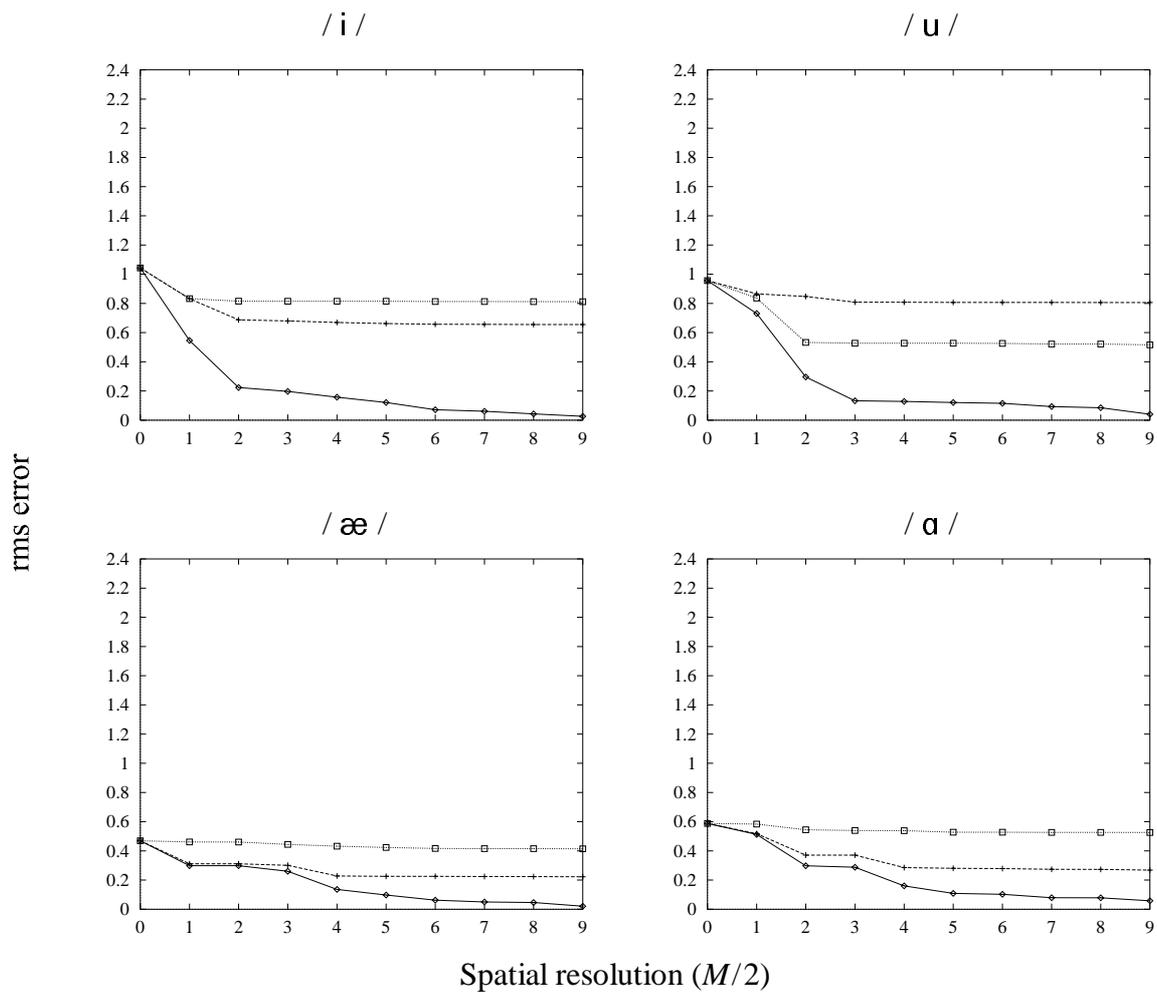


Figure E.3: Profiles of rms error in representing the area-functions of the 4 American English vowels (Speaker TB) of Baer et al. (1991).

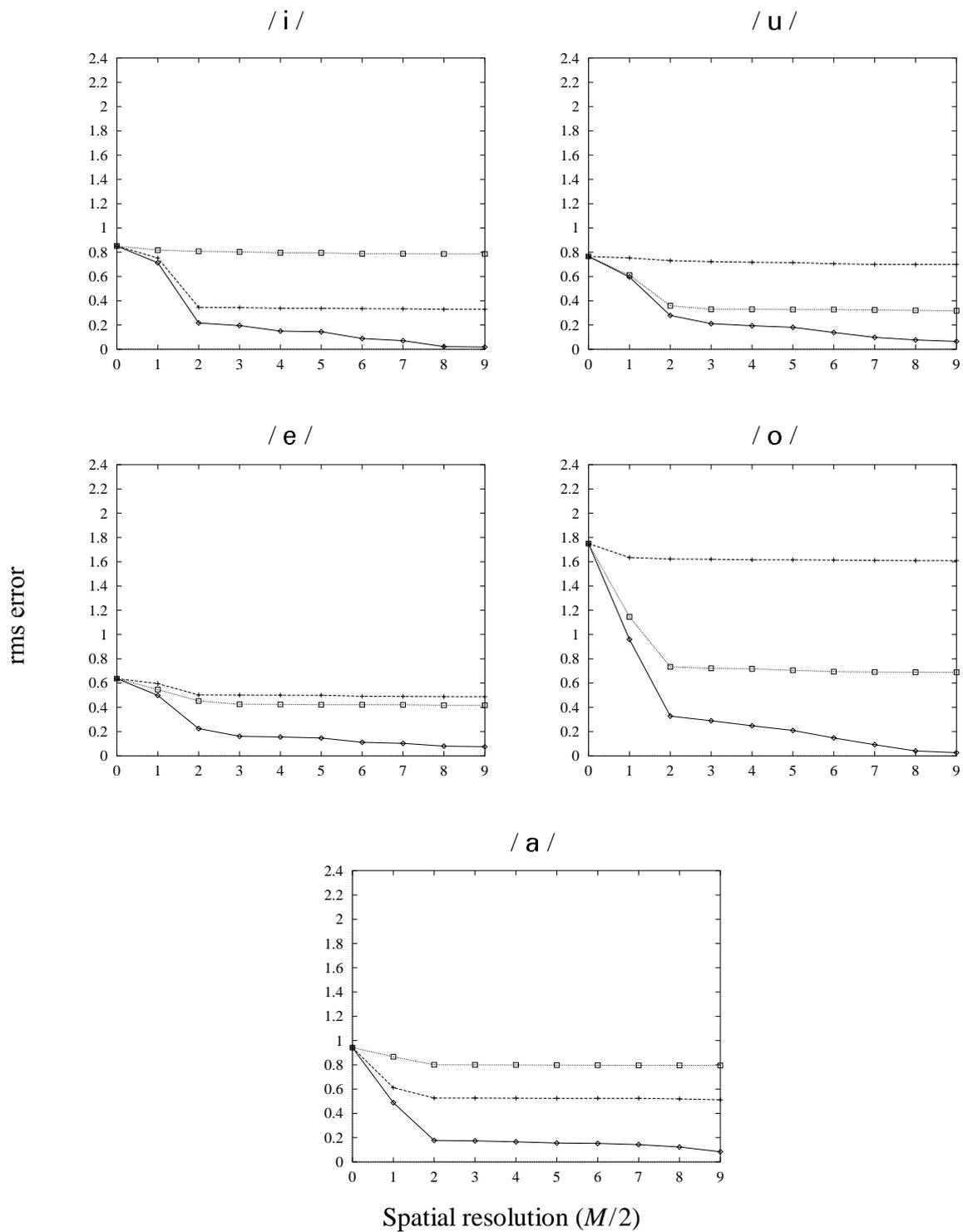


Figure E.4: Profiles of rms error in representing the area-functions of the 5 Japanese vowels (adult, male speaker) of Yang and Kasuya (1994).

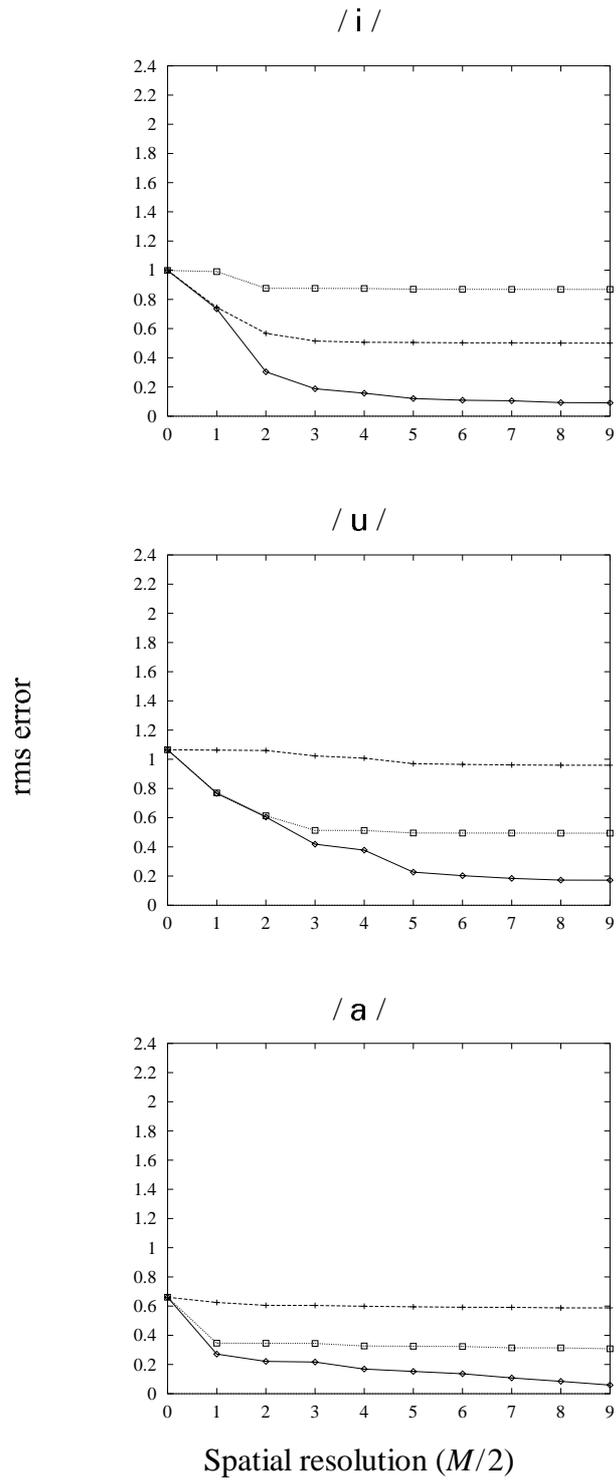


Figure E.5: Profiles of rms error in representing the area-functions of the 3 French vowels of Beutemps et al. (1995).

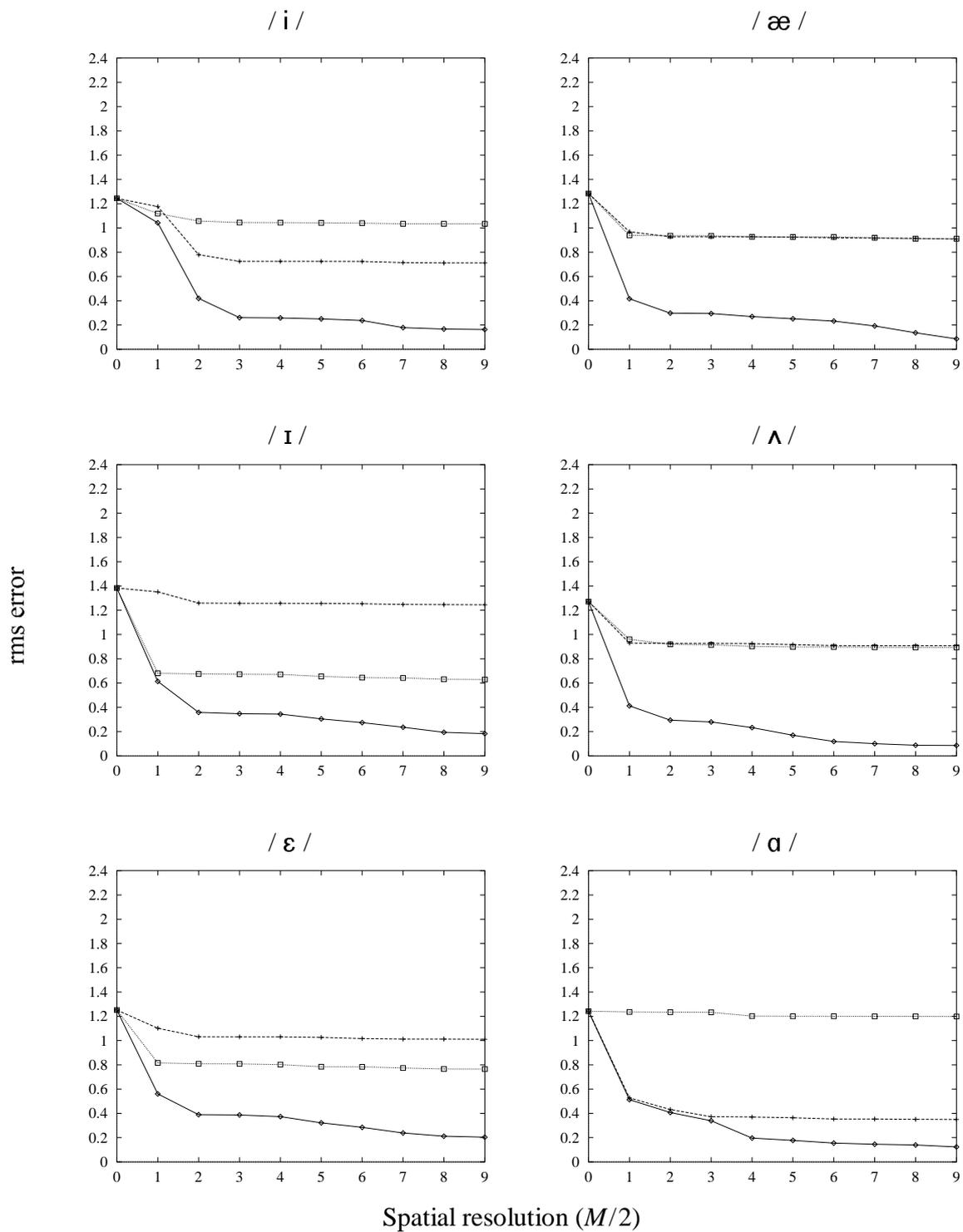


Figure E.6: Profiles of rms error in representing the area-functions of the 11 American English vowels of Story et al. (1996). (Continued on the following page.)

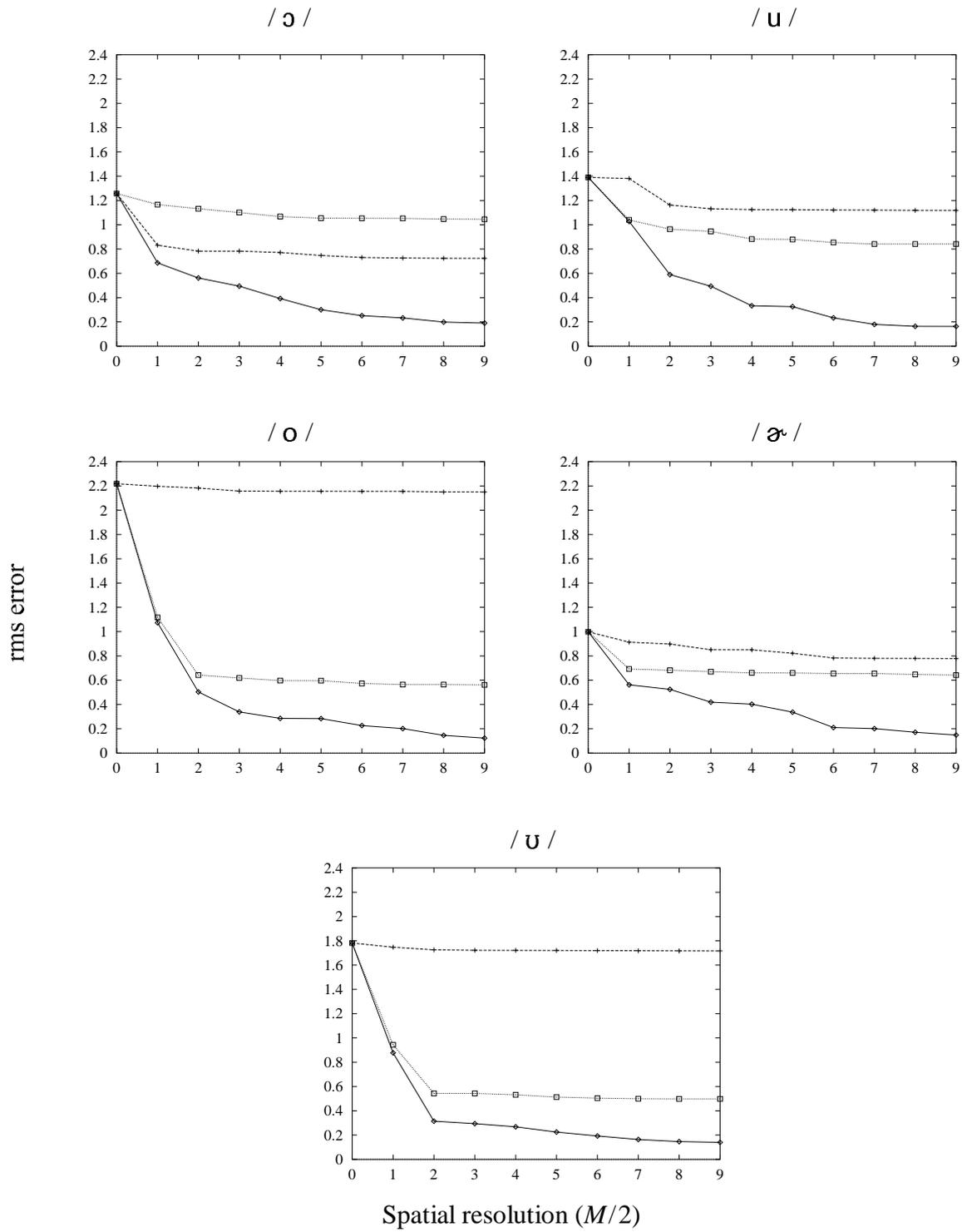


Figure E.6: (continued from previous page).

## Appendix F

### Re-estimation of Directly Measured Area-Functions

This appendix pertains to our discussions in Section 5.5.2, concerning re-estimation of the 33, directly measured area-functions (obtained from the literature, cf. Section 5.4.2.2), using our hybrid LP-SM method of inversion. In each panel is shown the original, step-wise area-function<sup>1</sup> (*solid line*), the smooth area-function re-estimated under model-matched conditions (*dashed curve*), and the smooth area-function re-estimated under model-mismatched conditions (*dotted curve*). The latter two area-functions are optimally aligned with respect to the original area-function, according to our method of alignment described in Section 5.5.1.

*Model-matched conditions* (cf. Section 5.5.2.1) imply that the first four formant frequencies and bandwidths used as inputs to the inversion (cf. Table 5.3), were synthesised using an 8-section representation of the original area-function, and an LP vocal-tract model with a nominal, glottal reflection coefficient of  $\mu_8 = 0.8$ .

*Model-mismatched conditions* (cf. Section 5.5.2.2) imply that the first four formant frequencies and bandwidths used as inputs to the inversion (cf. Table 5.4), were synthesised using the original section lengths and areas, and with a vocal-tract acoustic model (see Appendix C) which includes glottal impedance ( $\mu_{\text{glott}} = 0.8$ ), viscosity, heat-conduction, wall-vibration, and lip-radiation losses.

---

<sup>1</sup> As explained in Chapter 5, all area-functions are first “*normalised*” such that the area scaling factor in Equation 5.1 is equal to unity ( $A_0=1$ ). Hence the dimensionless ordinate on each graph, and the label “Normalised Area”, which should not be confused with the *speaker normalisation* of area-functions referred to in Chapter 6.

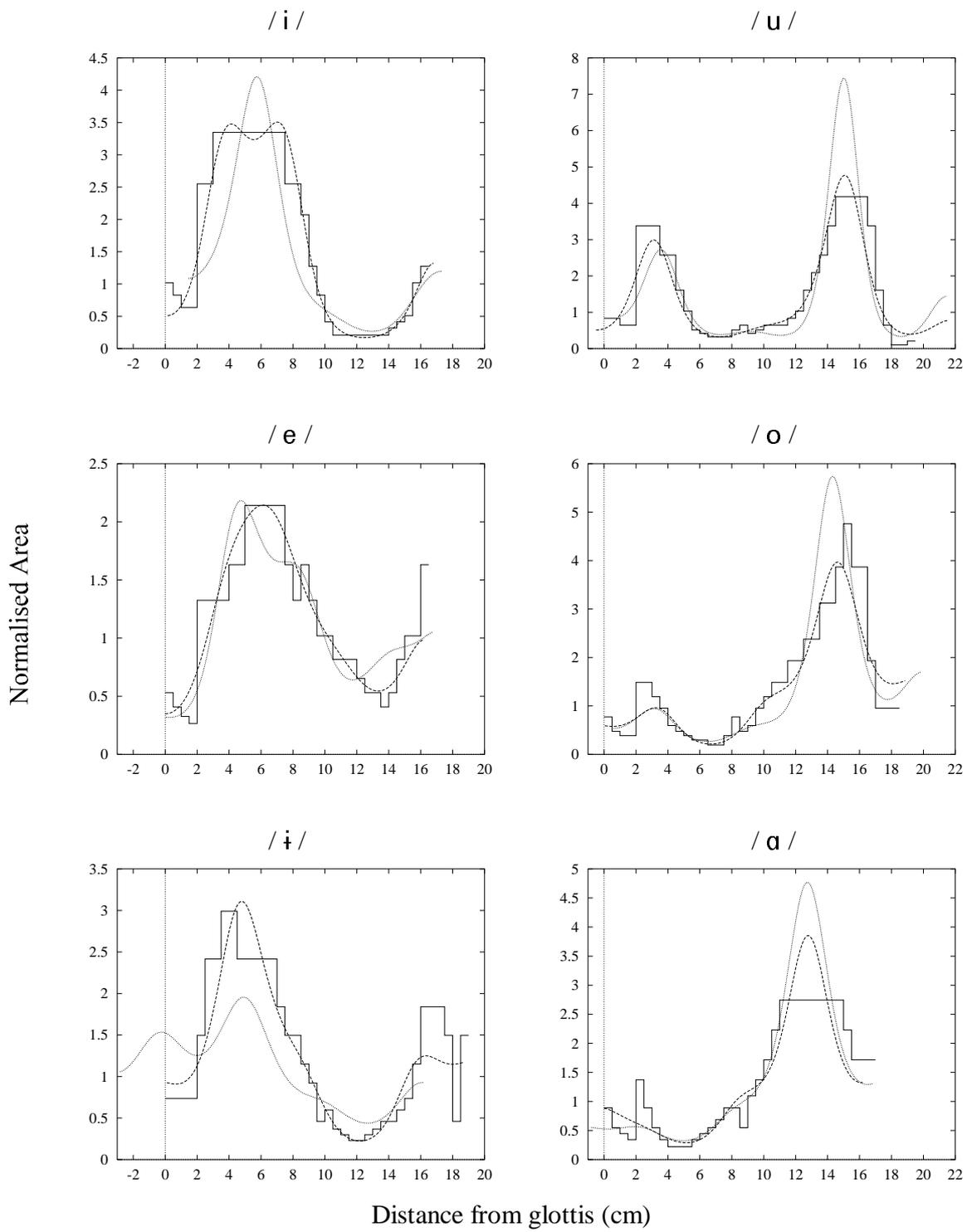


Figure F.1: Original and re-estimated area-functions of the 6 Russian vowels of Fant (1960).

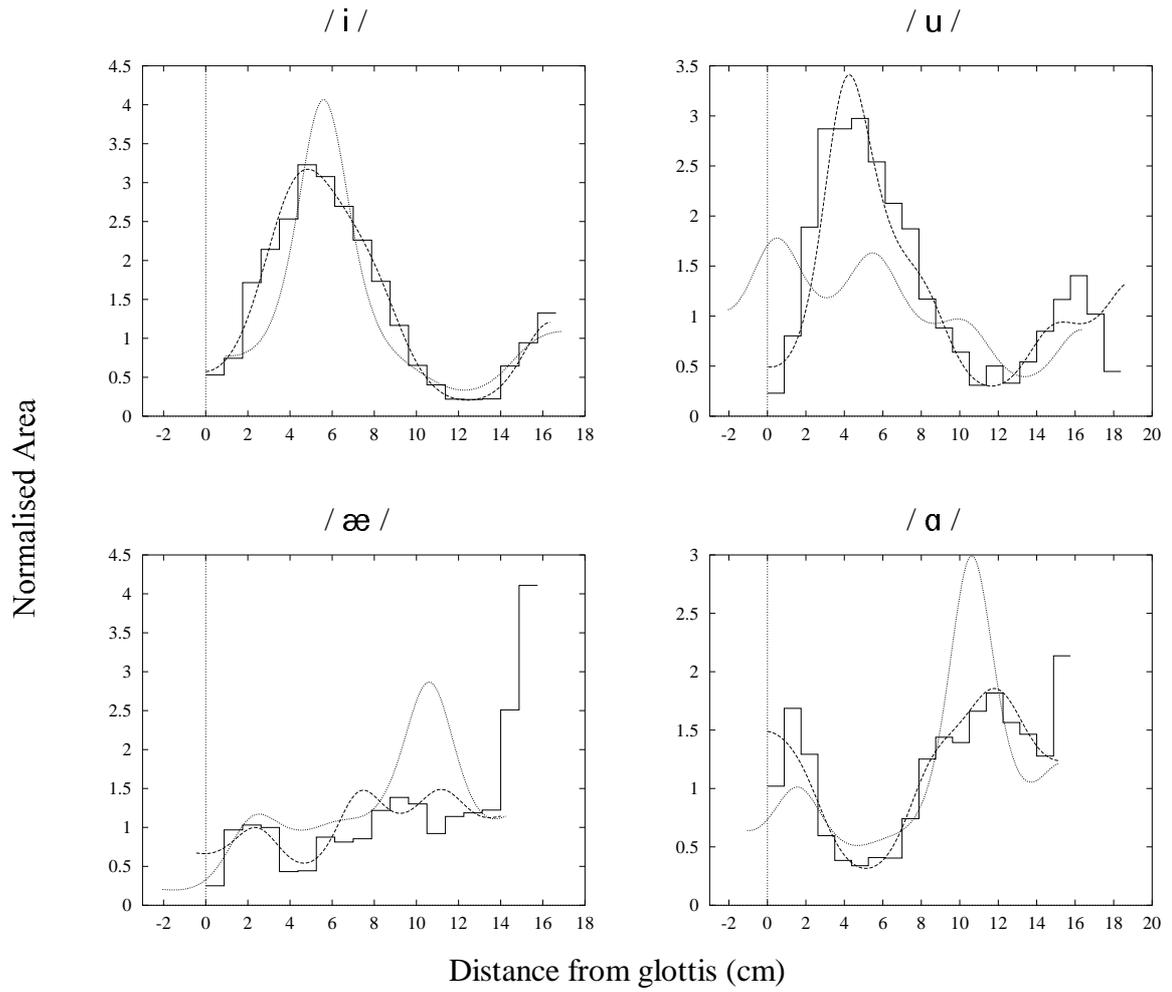


Figure F.2: Original and re-estimated area-functions of the 4 British English vowels (Speaker PN) of Baer et al. (1991).

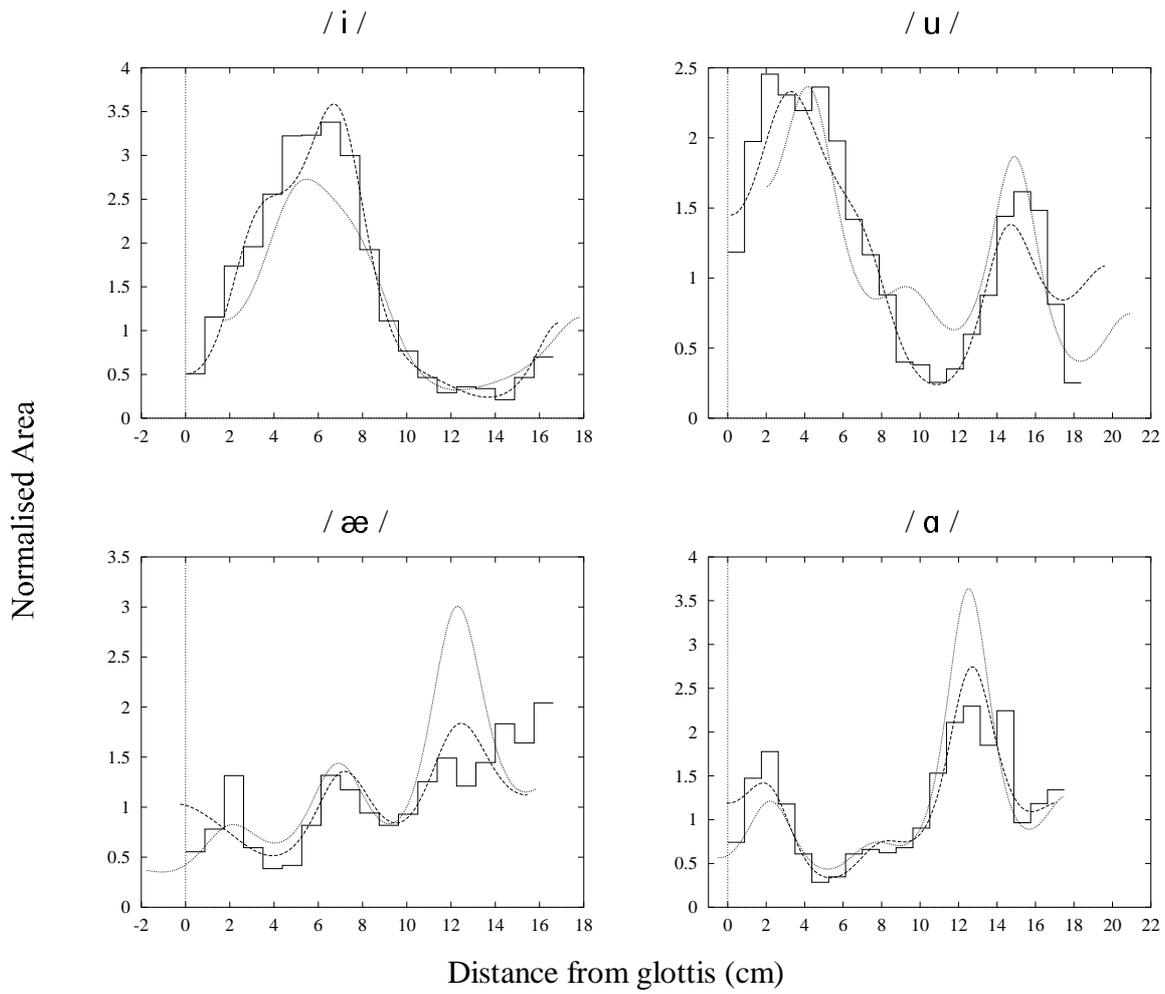


Figure F.3: Original and re-estimated area-functions of the 4 American English vowels (Speaker TB) of Baer et al. (1991).

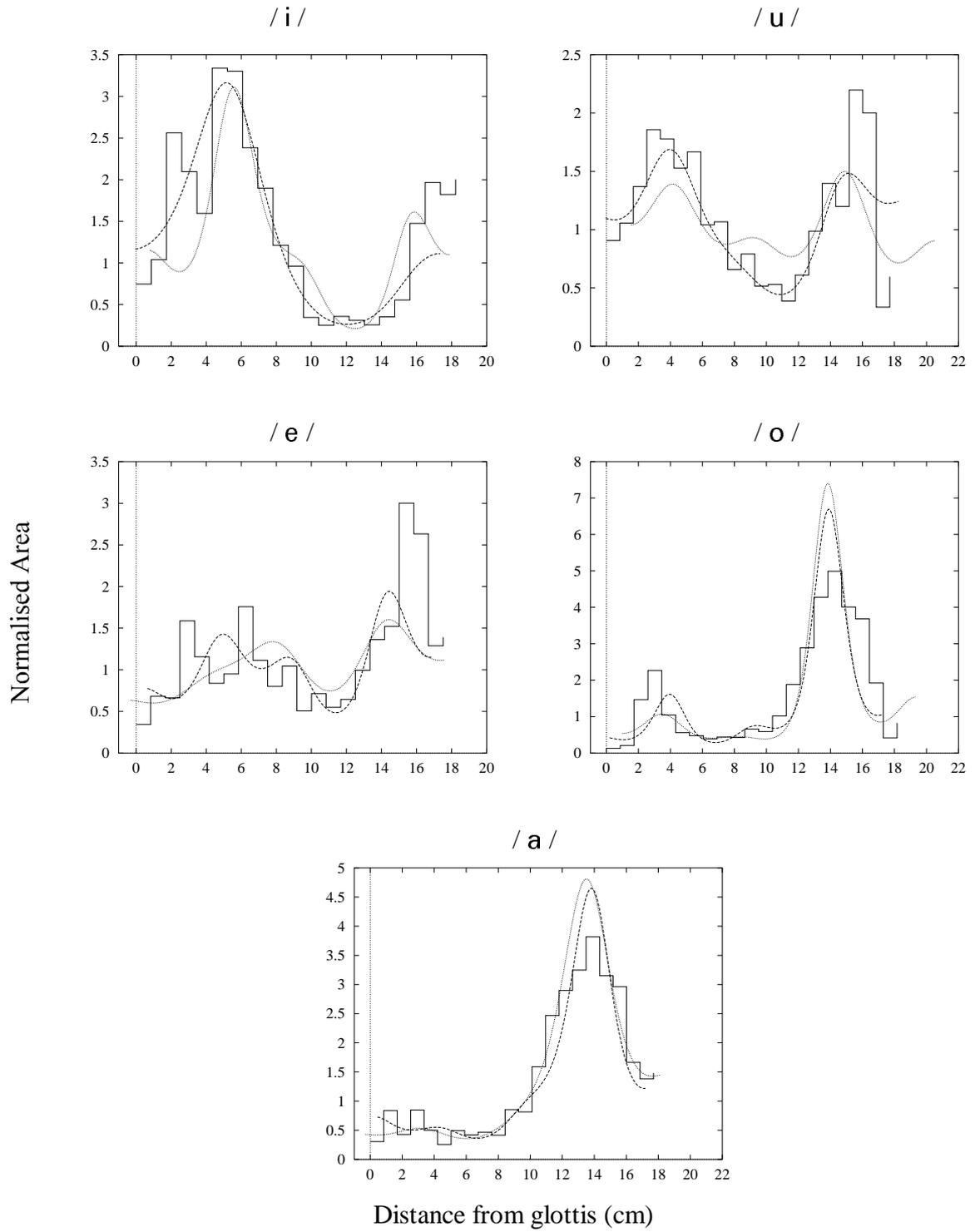


Figure F.4: Original and re-estimated area-functions of the 5 Japanese vowels (adult, male speaker) of Yang and Kasuya (1994).

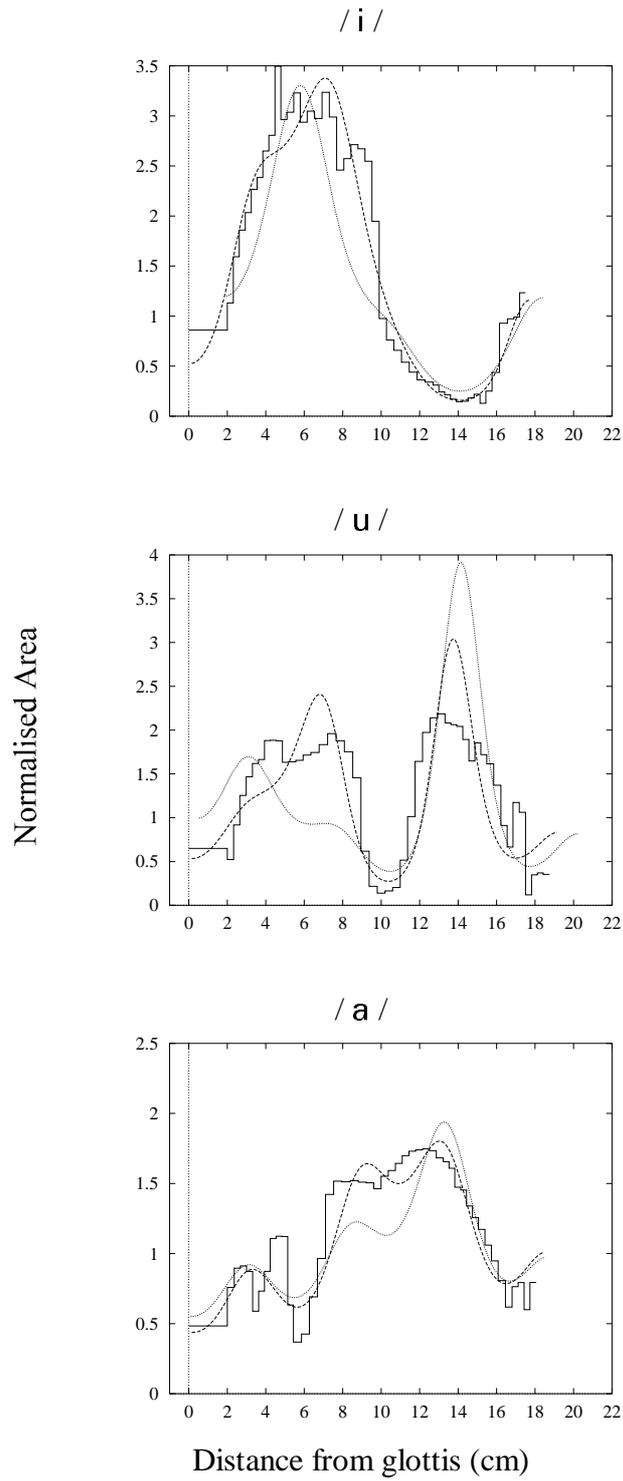


Figure F.5: Original and re-estimated area-functions of the 3 French vowels of Beautemps et al. (1995).

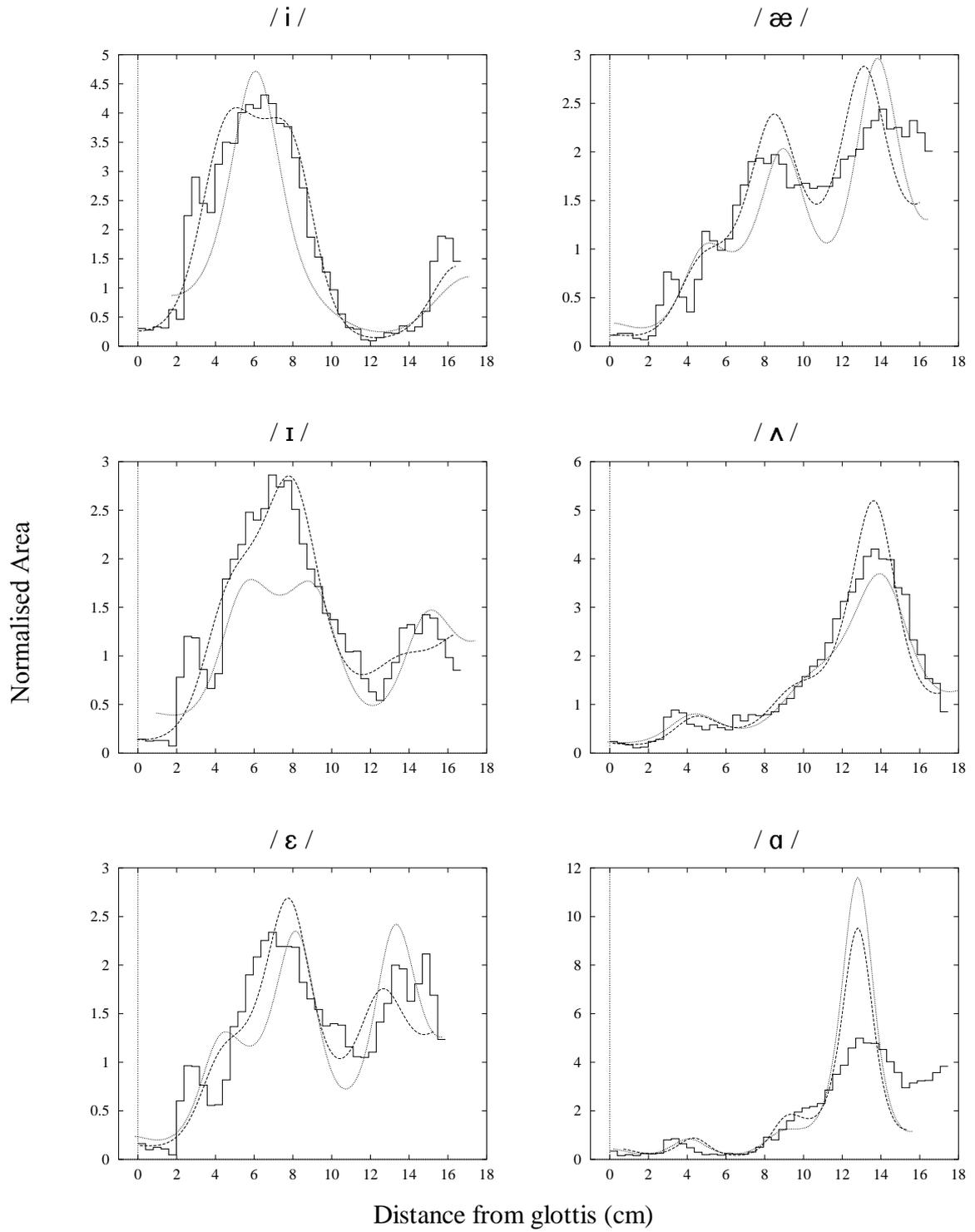


Figure F.6: Original and re-estimated area-functions of the 11 American English vowels of Story et al. (1996). (Continued on the following page.)

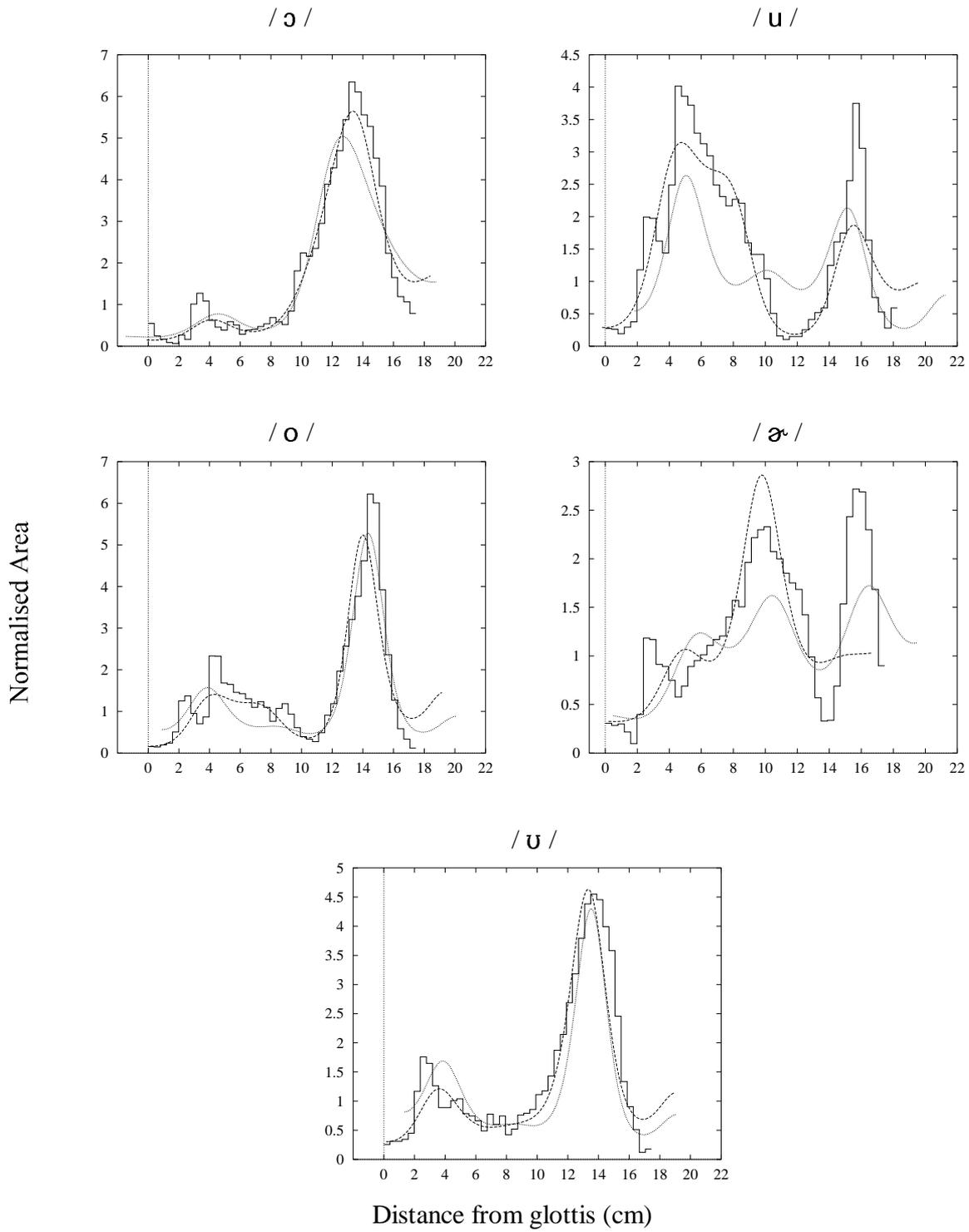


Figure F.6: (continued from previous page).

## Appendix G

### Closed-glottis Correction of Formant Bandwidths

This appendix pertains to our discussions in Section 5.5.3.2, concerning correction of measured formant bandwidths in order to obtain the “glottis-only” bandwidths ideally required by the LP vocal-tract model. [Table G.1](#) lists the per-vowel *mean* formant frequencies and bandwidths computed over all 7 steady-state frames, 5 repetitions and 4 speakers of the FC dataset; it also lists the corrected, or so-called “glottis-only” mean bandwidths, obtained by subtracting the “closed-glottis” value predicted by Hawks and Miller’s (1995) equation. [Figure G.1](#) then shows the influence of the proposed bandwidth correction on LP-derived area-functions, estimated using the mean formant data both before and after correction.

	$\bar{F}_1$ (Hz)	$\bar{F}_2$ (Hz)	$\bar{F}_3$ (Hz)	$\bar{F}_4$ (Hz)	$\bar{B}_1$ (Hz)	$\bar{B}_2$ (Hz)	$\bar{B}_3$ (Hz)	$\bar{B}_4$ (Hz)
/i/	345	2175	2968	3582	77 31	126 55	293 166	268 85
/ɪ/	367	2099	2803	3541	81 38	121 53	241 128	213 34
/ɛ/	479	2019	2711	3542	104 67	139 75	224 118	337 158
/æ/	654	1798	2589	3641	126 86	124 69	213 116	329 141
/a/	741	1213	2595	3550	145 104	159 116	234 136	335 156
/ɒ/	618	959	2521	3616	107 68	111 69	178 85	270 84
/ɔ/	422	712	2563	3400	75 37	89 49	195 100	239 74
/ʊ/	397	857	2381	3325	94 54	145 104	339 256	358 200
/ɜ:/	341	1660	2292	3344	57 11	99 48	104 26	200 40
/ʌ/	724	1237	2599	3518	133 93	174 131	260 162	317 141
/ɜ/	465	1585	2491	3591	84 47	120 71	113 23	201 18

Table G.1: *Per-vowel mean formant frequencies and bandwidths*, computed across all 7 steady-state frames, 5 repetitions, and 4 speakers of the FC dataset. The second row of figures for each vowel lists the mean formant bandwidths obtained after subtracting the so-called closed-glottis bandwidth predicted by Hawks and Miller’s (1995) equation. These mean formant values (both before and after bandwidth correction) are used to estimate the area-functions shown in Figure G.1.

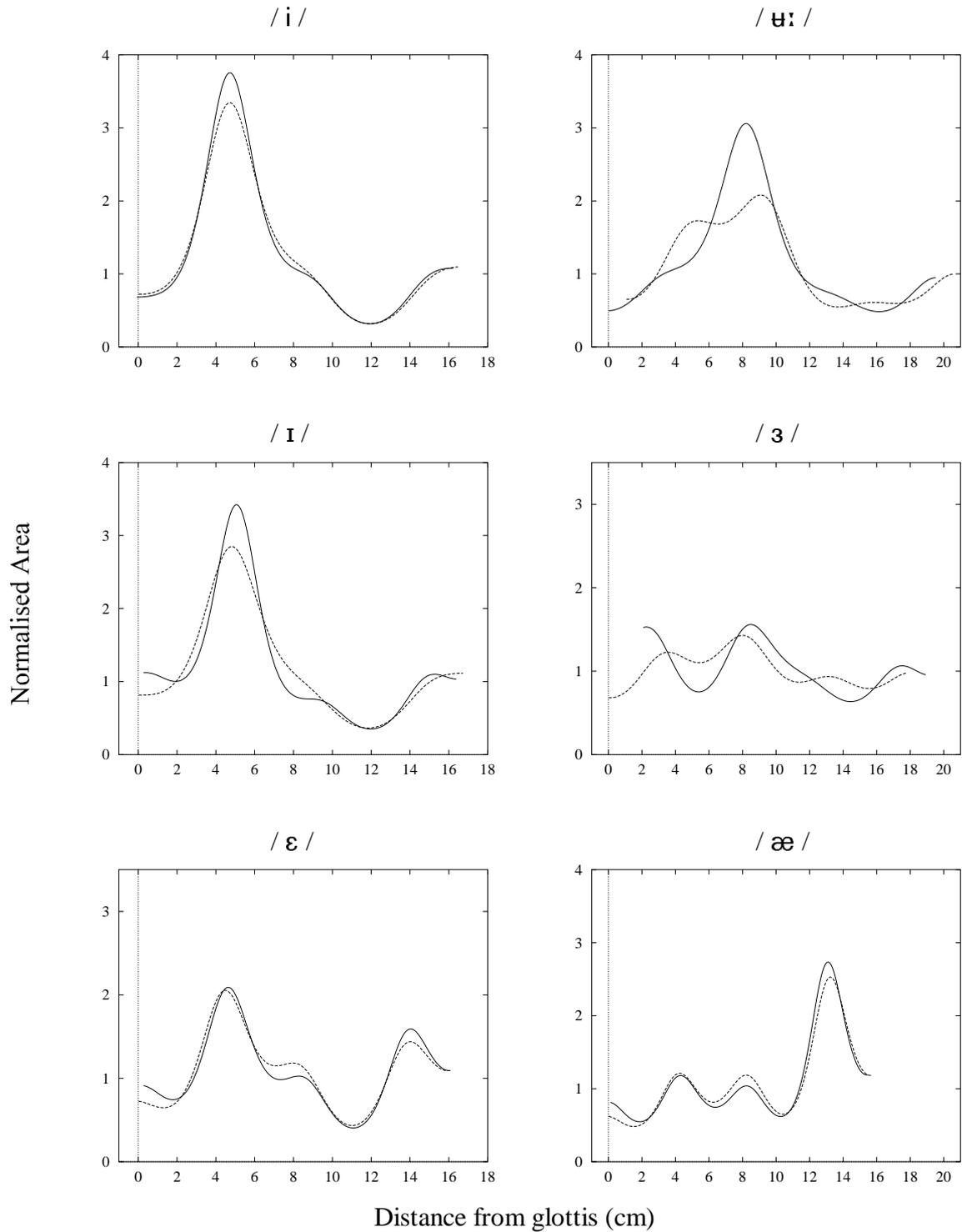


Figure G.1: *Influence of closed-glottis bandwidth correction on LP-derived area-functions.* Shown in each of the 11 panels (continued on next page) are the area-functions estimated from the per-vowel mean formant data (FC dataset) listed in Table G.1, before (*dashed line*) and after (*solid line*) correction of the bandwidths by subtraction of the closed-glottis bandwidth obtained by Hawks and Miller's (1995) equation. Each pair of area-functions thus obtained, is aligned using the method described in Section 5.5.1.

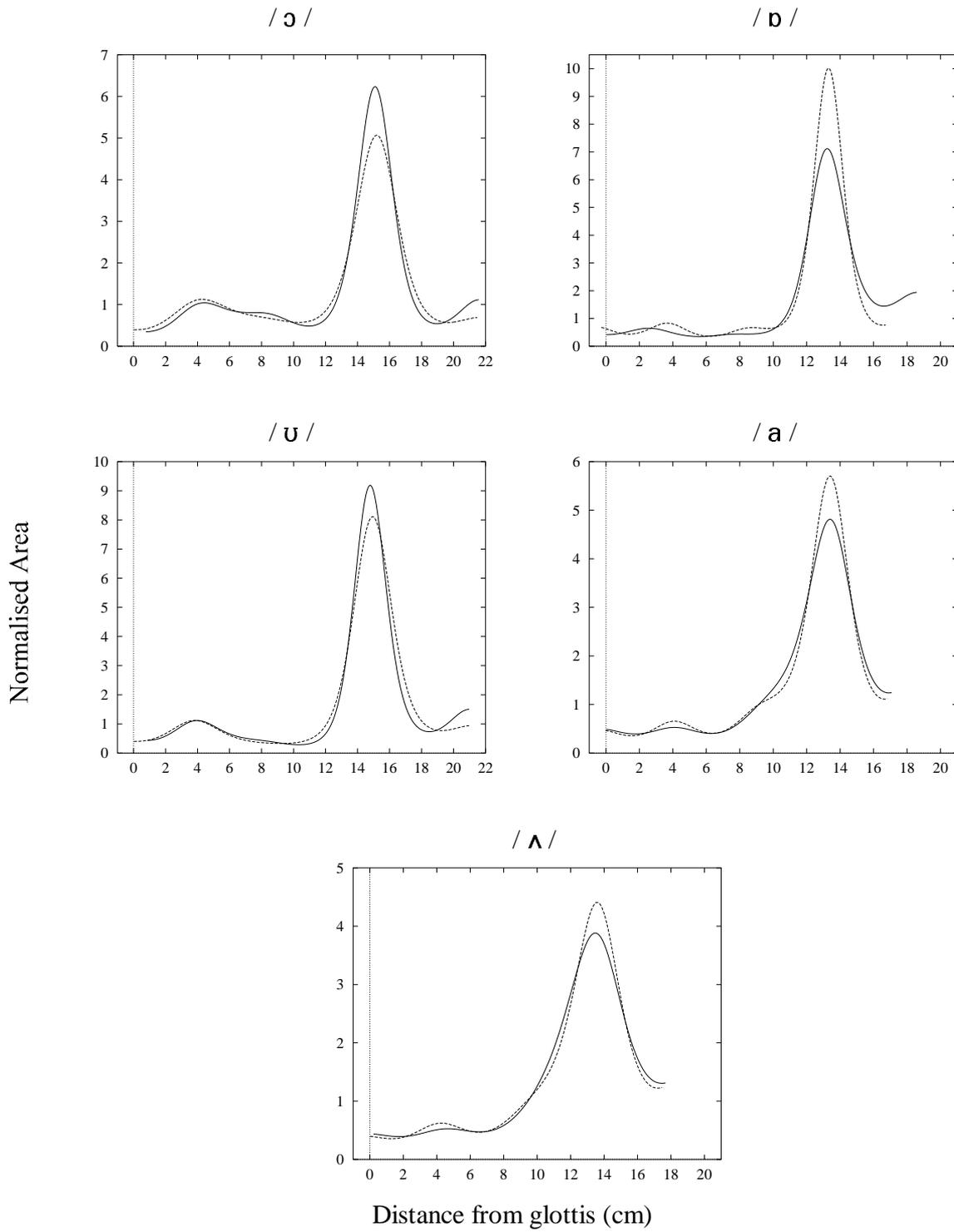


Figure G.1: (continued from previous page).

## Bibliography

- [1] Abercrombie, D. (1967), *Elements of general phonetics* (Edinburgh University Press, Edinburgh).
- [2] Ainsworth, W.A. and Foster, H.M. (1985), "The use of dynamic frequency warping in a speaker-independent vowel classifier", in De Mori, R. and Suen, C.Y. (eds.), *Proceedings of the NATO Advanced Study Institute on New Systems and Architectures for Automatic Speech Recognition and Synthesis* (Springer-Verlag, Berlin, Heidelberg): 389-403.
- [3] Arslan, L.M. and Hansen, J.H.L. (1997), "A study of temporal features and frequency characteristics in American English foreign accent", *Journal of the Acoustical Society of America* **102**: 28-40.
- [4] Assaleh, K.T. and Mammone, R.J. (1994), "New LP-Derived Features for Speaker Identification", *IEEE Transactions on Speech and Audio Processing* **2**: 630-638.
- [5] Assmann, P.F., Nearey, T.M. and Hogan, J.T. (1982), "Vowel identification: Orthographic, perceptual, and acoustic aspects", *Journal of the Acoustical Society of America* **71**: 975-989.
- [6] Atal, B.S. (1970), "Determination of the Vocal-Tract Shape Directly from the Speech Wave", *Journal of the Acoustical Society of America* **47**: S65.
- [7] Atal, B.S. (1974), "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", *Journal of the Acoustical Society of America* **55**: 1304-1312.
- [8] Atal, B.S., Chang, J.J., Mathews, M.V. and Tukey, J.W. (1978), "Inversion of articulatory-to-acoustic transformation in the vocal tract by a computer-sorting technique", *Journal of the Acoustical Society of America* **63**: 1535-1555.
- [9] Atal, B.S. and Hanauer, S.L. (1971), "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave", *Journal of the Acoustical Society of America* **50**: 637-655.
- [10] Auckenthaler, R. and Mason, J.S. (1997), "Equalizing sub-band error rates in speaker recognition", *Proceedings of the 5th European Conference on Speech Communication and Technology*: 2303-2306.

- [11] Badin, P. and Fant, G. (1984), "Notes on Vocal Tract Computation", *Speech Transmission Laboratory, Quarterly Progress and Status Report* **2-3**: 53-108.
- [12] Baer, T., Gore, J.C., Gracco, L.C. and Nye, P.W. (1991), "Analysis of vocal tract shape and dimensions using magnetic resonance imaging: Vowels", *Journal of the Acoustical Society of America* **90**: 799-828.
- [13] Bartholomew, W.T. (1934), "A Physical Definition of "Good Voice-Quality" in the Male Voice", *Journal of the Acoustical Society of America* **6**: 25-33.
- [14] Beautemps, D., Badin, P. and Laboissière, R. (1995), "Deriving vocal-tract area functions from midsagittal profiles and formant frequencies: A new model for vowels and fricative consonants based on experimental data", *Speech Communication* **16**: 27-47.
- [15] Békésy, G. von (1960), *Experiments in hearing* (McGraw-Hill, New York, Toronto, London).
- [16] Bernard, J.R.L. (1967), "Some measurements of some sounds of Australian English", unpublished Doctoral Thesis, The University of Sydney, Sydney, Australia.
- [17] Bernard, J.R.L. (1970), "Toward the acoustic specification of Australian English", *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* **23**: 113-128.
- [18] Bernard, J.R.L. (1989), "Quantitative aspects of the sounds of Australian English", in Collins, P. and Blair, D. (eds.), *Australian English: The Language of a New Society* (University of Queensland Press, Queensland, Australia): 187-204.
- [19] Besacier, L. and Bonastre, J.-F. (1997), "Independent Processing and Recombination of Partial Frequency Bands for Automatic Speaker Recognition", *Proceedings of the International Conference on Speech Processing*: 575-579.
- [20] Bladon, A. (1982), "Arguments against formants in the auditory representation of speech", in Carlson, R. and Granström, B. (eds.), *The Representation of Speech in the Peripheral Auditory System* (Elsevier Biomedical Press, Amsterdam, New York, Oxford): 95-102.
- [21] Bladon, R.A.W., Henton, C.G. and Pickering, J.B. (1983), "Outline of an Auditory Theory of Speaker Normalization", in Van den Broecke, M.P.R. and Cohen, A. (eds.), *Proceedings of the Tenth International Congress of Phonetic Sciences* (Foris Publications, Dordrecht, The Netherlands): 313-317.
- [22] Bladon, R.A.W. and Lindblom, B. (1981), "Modeling the judgment of vowel quality differences", *Journal of the Acoustical Society of America* **69**: 1414-1422.

- [23] Bogert, B.P. (1953), "On the Band Width of Vowel Formants", *Journal of the Acoustical Society of America* **25**: 791-792.
- [24] Bonder, L.J. (1983a), "The  $n$ -Tube Formula and Some of its Consequences", *Acustica* **52**: 216-226.
- [25] Bonder, L.J. (1983b), "Between Formant Space and Articulation Space", *Proceedings of the 10th International Congress of Phonetic Sciences*: 347-353.
- [26] Bonder, L.J. (1983c), "Equivalency of Lossless  $n$ -Tubes", *Acustica* **53**: 193-200.
- [27] Borg, G. (1946), "Eine Umkehrung der Sturm-Liouvilleschen Eigenwertaufgabe", ("An inversion of the Sturm-Liouville eigenvalue problem", in German), *Acta Mathematica* **78**: 1-96.
- [28] Broad, D.J. (1972), "Formants in Automatic Speech Recognition", *International Journal of Man-Machine Studies* **4**: 411-424.
- [29] Broad, D.J. (1976), "Toward Defining Acoustic Phonetic Equivalence for Vowels", *Phonetica* **33**: 401-424.
- [30] Broad, D.J. (1981), "Piecewise-planar vowel formant distributions across speakers", *Journal of the Acoustical Society of America* **69**: 1423-1429.
- [31] Broad, D.J. (1982), "Generalised Acoustic Phonetics I. Determinants of Acoustic Parameter Values", *manuscript of talk presented at Speech Technology Laboratory, Santa Barbara, California, USA*.
- [32] Broad, D.J. and Shoup, J.E. (1975), "Concepts for Acoustic Phonetic Recognition", in Reddy, D.R. (ed.), *Speech Recognition: Invited Papers Presented at the 1974 IEEE Symposium* (Academic Press, New York): 243-274.
- [33] Broad, D.J. and Wakita, H. (1977), "Piecewise-planar representation of vowel formant frequencies", *Journal of the Acoustical Society of America* **62**: 1467-1473.
- [34] Broad, D.J. and Wakita, H. (1978), "A phonetic approach to automatic vowel recognition", in Bolc, L. (ed.), *Speech Communication with Computers* (Carl Hanser Verlag, München, Wien): 55-92.
- [35] Burgess, N. (1969), "A spectrographic investigation of some diphthongal phonemes in Australian English", *Language and Speech* **12**: 238-246.
- [36] Butler, S.J. and Wakita, H. (1982), "Articulatory constraints on vocal tract area functions and their acoustic implications", *Journal of the Acoustical Society of America* **72**: S79.

- [37] Candille, L. and Meloni, H. (1995), “Automatic Speech Recognition Using Production Models”, *Proceedings of the 13th International Congress of Phonetic Sciences* **4**: 256-259.
- [38] Carré, R. (1971), “Identification des locuteurs; exploitation des données relatives aux fréquences des formants”, *Proceedings of the 7th International Congress on Acoustics*: 29-32.
- [39] Carré, R., Chennoukh, S. and Mrayati, M. (1992), “Vowel-consonant-vowel transitions: Analysis, modeling, and synthesis”, *Proceedings of the 2nd International Conference on Spoken Language Processing*: 819-822.
- [40] Carré, R. and Mrayati, M. (1991), “Vowel-vowel trajectories and region modeling”, *Journal of Phonetics* **19**: 433-443.
- [41] Charpentier, F. (1984), “Determination of the vocal tract shape from the formants by analysis of the articulatory-to-acoustic nonlinearities”, *Speech Communication* **3**: 291-308.
- [42] Chiba, T. and Kajiyama, M. (1958), *The vowel: its nature and structure* (Phonetic Society of Japan, Tokyo; first published 1941).
- [43] Chistovich, L.A. and Lublinskaya, V.V. (1979), “The ‘center of gravity’ effect in vowel spectra and critical distance between the formants: Psychoacoustical study of the perception of vowel-like stimuli”, *Hearing Research* **1**: 185-195.
- [44] Chrystal, G. (1964), *Algebra: An elementary text-book*, Part 1, Seventh Edition, (Chelsea Publishing Company, New York).
- [45] Chung, H., Makino, S. and Kido, K. (1988), “Analysis, perception and recognition of Korean isolated vowels”, Paper presented at the Second Joint Meeting of the Acoustical Societies of America and Japan, *Journal of the Acoustical Society of America* **84**: S213.
- [46] Claes, T., Dologlou, I., ten Bosch, L. and Van Compernelle, D. (1997), “New transformations of cepstral parameters for automatic vocal tract length normalization in speech recognition”, *Proceedings of the 5th European Conference on Speech Communication and Technology*: 1363-1366.
- [47] Clermont, F. (1991), “Formant-contour models of diphthongs: A study in acoustic phonetics and computer modelling of speech”, unpublished Doctoral Thesis, Computer Sciences Laboratory, Research School of Physical Sciences and Engineering, Australian National University, Canberra, Australia.
- [48] Clermont, F. (1993), “Spectro-temporal description of diphthongs in  $F_1$ - $F_2$ - $F_3$  space”, *Speech Communication* **13**: 377-390.

- [49] Clermont, F. (1996), "Multi-speaker formant data on the Australian English vowels: A tribute to J.R.L. Bernard's (1967) pioneering research", *Proceedings of the 6th Australian International Conference on Speech Science and Technology*: 145-150.
- [50] Clermont, F. and Broad, D.J. (1995), "Back-Front Classification of English Vowels using a Cepstrum-to-Formant Model", *Journal of the Acoustical Society of America* **98**: 2966.
- [51] Clermont, F. and Mokhtari, P. (1994), "Frequency-band specification in cepstral distance computation", *Proceedings of the 5th Australian International Conference on Speech Science and Technology*: 354-359.
- [52] Coker, C.H. (1976), "A Model of Articulatory Dynamics and Control", *Proceedings of the IEEE* **64**: 452-460.
- [53] Compton, A.J. (1963), "Effects of Filtering and Vocal Duration upon the Identification of Speakers, Aurally", *Journal of the Acoustical Society of America* **35**: 1748-1752.
- [54] Cooper, C. and Clermont, F. (1994), "An investigation of the speaker factor in vowel nuclei", *Proceedings of the 5th Australian International Conference on Speech Science and Technology*: 368-373.
- [55] Crichton, R.G. and Fallside, F. (1974), "Linear Prediction Model of Speech Production with Applications to Deaf Speech Training", *Proceedings of IEE* **121**: 865-873.
- [56] Dautrich, B.A., Rabiner, L.R. and Martin, T.B. (1983), "On the Effects of Varying Filter Bank Parameters on Isolated Word Recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **31**: 793-807.
- [57] Davis, S.B. and Mermelstein, P. (1980), "Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **28**: 357-366.
- [58] Delattre, P. (1951), "The physiological interpretation of sound spectrograms", *Publications of the Modern Language Association of America* **66**: 864-875.
- [59] Delattre, P., Liberman, A.M., Cooper, F.S. and Gerstman, L.J. (1952), "An experimental study of the acoustic determinants of vowel color; observations on one- and two-formant vowels synthesised from spectrographic patterns", *Word* **8**: 195-210.
- [60] Denes, P.B. (1963), "On the statistics of spoken English", *Journal of the Acoustical Society of America* **35**: 892-904.
- [61] Di Benedetto, M.-G. and Liénard, J.-S. (1992), "Extrinsic normalization of vowel formant values based on cardinal vowels mapping", *Proceedings of the 2nd International Conference on Spoken Language Processing*: 579-582.

- [62] Duda, R.O. and Hart, P.E. (1973), *Pattern Classification and Scene Analysis* (John Wiley & Sons, New York).
- [63] Dukiewicz, L. (1970), "Frequency-band dependence of speaker identification", in Jassem, W. (ed.), *Speech Analysis and Synthesis, Vol.II* (Polish Academy of Sciences, Warsaw): 41-50.
- [64] Dunn, H.K. (1961), "Methods of Measuring Vowel Formant Bandwidths", *Journal of the Acoustical Society of America* **33**: 1737-1746.
- [65] Eatock, J.P. and Mason, J.S. (1994), "A quantitative assessment of the relative speaker discriminating properties of phonemes", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing Vol. I*: 133-136.
- [66] Endres, W., Bambach, W. and Flösser, G. (1971), "Voice Spectrograms as a Function of Age, Voice Disguise, and Voice Imitation", *Journal of the Acoustical Society of America* **49**: 1842-1848.
- [67] Ewan, W.G. and Krones, R. (1974), "Measuring larynx movement using the thyroumbrometer", *Journal of Phonetics* **2**: 327-335.
- [68] Fant, G. (1960), *Acoustic Theory of Speech Production* (Mouton, The Hague, The Netherlands).
- [69] Fant, G. (1962), "Formant bandwidth data", *Speech Transmission Laboratory, Quarterly Progress and Status Report* **1**: 1-2.
- [70] Fant, G. (1966), "A note on vocal tract size factors and non-uniform F-pattern scalings", *Speech Transmission Laboratory, Quarterly Progress and Status Report* **4**: 22-30.
- [71] Fant, G. (1972), "Vocal tract wall effects, losses, and resonance bandwidths", *Speech Transmission Laboratory, Quarterly Progress and Status Report* **2-3**: 28-52.
- [72] Fant, G. (1975 a), "Non-uniform vowel normalization", *Speech Transmission Laboratory, Quarterly Progress and Status Report* **2-3**: 1-19.
- [73] Fant, G. (1975 b), "Vocal-tract area and length perturbations", *Speech Transmission Laboratory, Quarterly Progress and Status Report* **4**: 1-14.
- [74] Fant, G. (1980), "The Relations between Area Functions and the Acoustic Signal", *Phonetica* **37**: 55-86.
- [75] Fant, G. and Risberg, A. (1962), "Auditory matching of vowels with two formant synthetic sounds", *Speech Transmission Laboratory, Quarterly Progress and Status Report* **4**: 7-11.

- [76] Flanagan, J.L. (1955), "A difference limen for vowel formant frequency", *Journal of the Acoustical Society of America* **27**: 613-617.
- [77] Flanagan, J.L. (1972), *Speech Analysis, Synthesis and Perception*, Second Edition, (Springer-Verlag, Berlin, Heidelberg, New York).
- [78] Flanagan, J.L., Ishizaka, K. and Shipley, K.L. (1980), "Signal models for low bit-rate coding of speech", *Journal of the Acoustical Society of America* **68**: 780-791.
- [79] Foley, D.H. (1972), "Considerations of Sample and Feature Size", *IEEE Transactions on Information Theory* **18**: 618-626.
- [80] Fuchi, K. (1977), "Vowel Approximation by Multi-band Filtering Characteristics and Estimation of Antimetrical Vocal Tract Shapes", *Electrotechnical Laboratory Progress Report on Speech Research* **14**: 56-58.
- [81] Fuchi, K. and Ohta, K. (1978), "Observation on Group Delay Characteristics of Connected Vowels", *Electrotechnical Laboratory Progress Report on Speech Research* **18**: 44-47.
- [82] Fuchi, K. and Ohta, K. (1979), "Estimation of Symmetrical Acoustic Tubes Representing Vowel Characteristics", *Electrotechnical Laboratory Progress Report on Speech Research* **20**: 5-9.
- [83] Fujimura, O. and Lindqvist, J. (1971), "Sweep-tone measurements of vocal-tract characteristics", *Journal of the Acoustical Society of America* **49**: 541-558.
- [84] Furui, S. (1981), "Cepstral Analysis Technique for Automatic Speaker Verification", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **29**: 254-272.
- [85] Furui, S. (1989), "Unsupervised Speaker Adaptation Based on Hierarchical Spectral Clustering", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **37**: 1923-1930.
- [86] Furui, S. (1991), "Speaker-dependent-feature extraction, recognition and processing techniques", *Speech Communication* **10**: 505-520.
- [87] Furui, S. (1992), "Speaker-Independent and Speaker-Adaptive Recognition Techniques", in Furui, S. and Sondhi, M.M. (eds.), *Advances in Speech Signal Processing* (Marcel Dekker, New York, Basel, Hong Kong): 597-622.
- [88] Furui, S. and Akagi, M. (1985), "Perception of voice individuality and physical correlates", *Journal of the Acoustical Society of Japan* (English reprint), **H-85-18**: 1-8.
- [89] Garvin, P.L. and Ladefoged, P. (1963), "Speaker Identification and Message Identification in Speech Recognition", *Phonetica* **9**: 193-199.

- [90] Gath, I. and Yair, E. (1988), "Analysis of vocal tract parameters in Parkinsonian speech", *Journal of the Acoustical Society of America* **84**: 1628-1634.
- [91] Gay, T., Boë, L-J., Perrier, P., Feng, G. and Swayne, E. (1991), "The acoustic sensitivity of vocal tract constrictions: a preliminary report", *Journal of Phonetics* **19**: 445-452.
- [92] Gay, T., Lindblom, B. and Lubker, J. (1981), "Production of bite-block vowels: Acoustic equivalence by selective compensation", *Journal of the Acoustical Society of America* **69**: 802-810.
- [93] Gerstman, L.J. (1968), "Classification of Self-Normalized Vowels", *IEEE Transactions on Audio and Electroacoustics* **16**: 78-80.
- [94] Goldstein, U.G. (1979), "Modelling children's vocal tracts", *Speech Communication Papers presented at the 97th Meeting of the Acoustical Society of America*, Paper I22: 139-142.
- [95] Golibersuch, R.J. (1983), "Automatic prediction of linear frequency warp for speech recognition", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*: 769-772.
- [96] Gopinath, B. and Sondhi, M.M. (1970), "Determination of the Shape of the Human Vocal Tract from Acoustical Measurements", *The Bell System Technical Journal* **49**: 1195-1214.
- [97] Gray, A.H. and Markel, J.D. (1974), "A Spectral Flatness Measure for Studying the Autocorrelation Method of Linear Prediction of Speech Analysis", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **22**: 207-217.
- [98] Gray, A.H. and Markel, J.D. (1976), "Distance Measures for Speech Processing", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **24**: 380-391.
- [99] Gupta, S.K. and Schroeter, J. (1993), "Pitch-synchronous frame-by-frame and segment-based articulatory analysis by synthesis", *Journal of the Acoustical Society of America* **94**: 2517-2530.
- [100] Hafer, E.H. and Coker, C.H. (1975), "Determining tongue body motion from the acoustic speech wave", *Journal of the Acoustical Society of America* **57**: S3.
- [101] Hansen, J.H.L. and Womack, B.D. (1996), "Feature Analysis and Neural Network-Based Classification of Speech Under Stress", *IEEE Transactions on Speech and Audio Processing* **4**: 307-313.

- [102] Hanson, B.A. and Wakita, H. (1987), "Spectral Slope Distance Measures with Linear Prediction Analysis for Word Recognition in Noise", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **35**: 968-973.
- [103] Harshman, R., Ladefoged, P. and Goldstein, L. (1977), "Factor analysis of tongue shapes", *Journal of the Acoustical Society of America* **62**: 693-707.
- [104] Hawks, J.W. and Miller, J.D. (1995), "A formant bandwidth estimation procedure for vowel synthesis", *Journal of the Acoustical Society of America* **97**: 1343-1344.
- [105] Hayakawa, S. and Itakura, F. (1994), "Text-dependent speaker recognition using the information in the higher frequency band", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* **Vol. I**: 137-140.
- [106] Hecker, M.H.L. (1971), *Speaker Recognition: An Interpretive Survey of the Literature*, ASHA Monographs Number 16 (American Speech and Hearing Association, Washington D.C.).
- [107] Hermansky, H. (1990), "Perceptual linear predictive (PLP) analysis of speech", *Journal of the Acoustical Society of America* **87**: 1738-1752.
- [108] Hermansky, H. and Broad, D.J. (1989), "The effective second formant F2' and the vocal tract front-cavity", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*: 480-483.
- [109] Hill, D.R., Manzara, L. and Taube-Schock, C.-R. (1995), "Real-time articulatory speech-synthesis-by-rules", *Proceedings of the 14th Annual International Voice Technologies and Applications Conference of the American Voice I/O Society*: 27-44.
- [110] Hillenbrand, J. and Gayvert, R.T. (1993), "Vowel Classification Based on Fundamental Frequency and Formant Frequencies", *Journal of Speech and Hearing Research* **36**: 694-700.
- [111] Hillenbrand, J., Getty, L.A., Clark, M.J. and Wheeler, K. (1995), "Acoustic characteristics of American English vowels", *Journal of the Acoustical Society of America* **97**: 3099-3111.
- [112] Höfker, U. (1976), "Die Eignung verschiedener Sprachlaute für die automatische Sprechererkennung", *Proc. IITB Kolloquium "Akustische Mustererkennung"*.
- [113] Högberg, J. (1995), "From sagittal distance to area function and male to female scaling of the vocal tract", *Speech Transmission Laboratory, Quarterly Progress and Status Report* **4**: 11-53.

- [114] Hogden, J., Lofqvist, A., Gracco, V., Zlokarnik, I., Rubin, P. and Saltzman, E. (1996), "Accurate recovery of articulator positions from acoustics: New conclusions based on human data", *Journal of the Acoustical Society of America* **100**: 1819-1834.
- [115] Holmes, J.N., Holmes, W.J. and Garner, P.N. (1997), "Using formant frequencies in speech recognition", *Proceedings of the 5th European Conference on Speech Communication and Technology*: 2083-2086.
- [116] Honda, K., Hashi, M., Wu, C.-M. and Westbury, J.R. (1997), "Effects of the size and form of the orofacial structure on vowel production", *Journal of the Acoustical Society of America* **102**: 3133.
- [117] Honikman, B. (1964), "Articulatory Settings", in Abercrombie, D., Fry, D.B., MacCarthy, P.A.D., Scott, N.C. and Trim, J.L.M. (eds.), *In Honour of Daniel Jones* (Longmans, Green and Co., London): 73-84.
- [118] Hughes, O.M. and Abbs, J.H. (1976), "Labial-Mandibular Coordination in the Production of Speech: Implications for the Operation of Motor Equivalence", *Phonetica* **33**: 199-221.
- [119] ILS (1983), *Interactive Laboratory System V4.1: Programmer's Guide*, Signal Technology Incorporated.
- [120] Ishizaki, S. (1978a), "Interspeaker Normalization Using Vocal Tract Length and Vowel Feature Extraction", *Electrotechnical Laboratory Progress Report on Speech Research* **18**: 66-69.
- [121] Ishizaki, S. (1978b), "Vowel Discrimination by Use of Articulatory Model", *Proceedings of the 4th International Joint Conference on Pattern Recognition*: 1050-1052.
- [122] Itahashi, S. (1984), "On the relation between cepstra and formants of speech", *Preprints of the Spring Meeting of the Acoustical Society of Japan*, Paper 3-2-8: 185-186.
- [123] Itahashi, S. (1988), "On properties of speech cepstra", *Denshi Joho Tsushin Gakkai Ronbunshi (Trans. Inst. Electron. Inform. Comm. Eng. Jpn.)* **J71-D**: 1839-1842.
- [124] Itakura, F. (1975), "Minimum Prediction Residual Principle Applied to Speech Recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **23**: 67-72.
- [125] Itakura, F. and Umezaki, T. (1987), "Distance measure for speech recognition based on the smoothed group delay spectrum", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*: 1257-1260.

- [126] Jesorsky, P. (1978), "Principles of automatic speaker-recognition", in Bolc, L. (ed.), *Speech Communication with Computers* (Carl Hanser Verlag, München, Wien): 93-137.
- [127] Juang, B.-H., Rabiner, L.R. and Wilpon, J.G. (1987), "On the Use of Bandpass Liftering in Speech Recognition", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **35**: 947-953.
- [128] Kashyap, R.L. (1976), "Speaker Recognition from an Unknown Utterance and Speech-Speaker Interaction", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **24**: 481-488.
- [129] Kasuya, H. and Wakita, H. (1979), "An Approach to Segmenting Speech into Vowel- and Nonvowel-Like Intervals", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **27**: 319-327.
- [130] Kelly, J.L., Jr. and Lochbaum, C. (1963), "Speech synthesis", *Proceedings of the Speech Communication Seminar*, Stockholm, Speech Transmission Laboratory, Royal Institute of Technology, **Vol. II**: Paper F7.
- [131] Kitamura, T. and Akagi, M. (1994), "Speaker Individualities in Speech Spectral Envelopes", *Proceedings of the 3rd International Conference on Spoken Language Processing*: 1183-1186.
- [132] Kitamura, T. and Akagi, M. (1996), "Relationship between physical characteristics and speaker individualities in speech spectral envelopes", *Journal of the Acoustical Society of America* **100**: 2600.
- [133] Kiukaanniemi, H., Siponen, P. and Mattila, P. (1982), "Individual Differences in the Long-Term Speech Spectrum", *Folia Phoniatria* **34**: 21-28.
- [134] Klatt, D.H. (1982), "Prediction of perceived phonetic distance from critical-band spectra: A first step", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*: 1278-1281.
- [135] Klatt, D.H. (1986), "The problem of variability in speech recognition and in models of speech perception", in Perkell, J.S. and Klatt, D.H. (eds.), *Invariance and variability in speech processes* (Lawrence Erlbaum Associates, Hillsdale, New Jersey): 300-319.
- [136] Klein, W., Plomp, R. and Pols, L.C.W. (1970), "Vowel Spectra, Vowel Spaces, and Vowel Identification", *Journal of the Acoustical Society of America* **48**: 999-1009.
- [137] Koenig, W., Dunn, H.K. and Lacy, L.Y. (1946), "The sound spectrograph", *Journal of the Acoustical Society of America* **18**: 19-49.
- [138] Kopp, G.A. and Green, H.C. (1946), "Basic Phonetic Principles of Visible Speech", *Journal of the Acoustical Society of America* **18**: 74-89.

- [139] Kumar, K. (1996), "Computer recognition of Australian English vowels based on multi-speaker spectrographic data", unpublished M.Inf.Sc. Thesis, School of Computer Science, University of New South Wales, Australian Defence Force Academy, Canberra, Australia.
- [140] Kuwabara, H. and Takagi, T. (1991), "Acoustic parameters of voice individuality and voice-quality control by analysis-synthesis method", *Speech Communication* **10**: 491-495.
- [141] Ladefoged, P. (1993), *A Course In Phonetics*, Third Edition (Harcourt Brace Jovanovich, Florida).
- [142] Ladefoged, P. and Broadbent, D.E. (1957), "Information Conveyed by Vowels", *Journal of the Acoustical Society of America* **29**: 98-104.
- [143] Ladefoged, P., Harshman, R. and Goldstein, L. (1977), "Vowel articulations and formant frequencies", *UCLA Working Papers in Phonetics* **38**: 16-40.
- [144] Larar, J.N., Schroeter, J. and Sondhi, M.M. (1988), "Vector Quantization of the Articulatory Space", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **36**: 1812-1818.
- [145] Laver, J. (1980), *The phonetic description of voice quality* (Cambridge University Press, Cambridge).
- [146] Lea, W.A., Medress, M.F. and Skinner, T.E. (1975), "A Prosodically Guided Speech Understanding Strategy", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **23**: 30-38.
- [147] Lee, L. and Rose, R. (1998), "A Frequency Warping Approach to Speaker Normalization", *IEEE Transactions on Speech and Audio Processing* **6**: 49-60.
- [148] Lewis, D. (1936), "Vocal Resonance", *Journal of the Acoustical Society of America* **8**: 91-99.
- [149] Lewis, D. and Tuthill, C. (1940), "Resonant Frequencies and Damping Constants of Resonators Involved in the Production of Sustained Vowels "O" and "Ah"", *Journal of the Acoustical Society of America* **11**: 451-456.
- [150] Li, K.-P. and Hughes, G.W. (1974), "Talker differences as they appear in correlation matrices of continuous speech spectra", *Journal of the Acoustical Society of America* **55**: 833-837.
- [151] Liljencrants, J. (1971), "Fourier series description of the tongue profile", *Speech Transmission Laboratory, Quarterly Progress and Status Report* **4**: 9-18.

- [152] Lin, Q. and Che, C. (1995), “Normalizing The Vocal Tract Length For Speaker Independent Speech Recognition”, *IEEE Signal Processing Letters* **2**: 201-203.
- [153] Lin, Q. and Fant, G. (1989), “Vocal-tract area-function parameters from formant frequencies”, *Proceedings of the 1st European Conference on Speech Communication and Technology*: 673-676.
- [154] Lin, Q., Jan, E.-E., Che, C.W., Yuk, D.-S. and Flanagan, J. (1996), “Selective use of the speech spectrum and a VQGMM method for speaker identification”, *Proceedings of the 4th International Conference on Spoken Language Processing*: 2415-2418.
- [155] Lindblom, B., Lubker, J. and Gay, T. (1979), “Formant frequencies of some fixed-mandible vowels and a model of speech motor programming by predictive simulation”, *Journal of Phonetics* **7**: 147-161.
- [156] Lindblom, B.E.F. and Sundberg, J.E.F. (1971), “Acoustical Consequences of Lip, Tongue, Jaw, and Larynx Movement”, *Journal of the Acoustical Society of America* **50**: 1166-1179.
- [157] Linggard, R.L. (1985), *Electronic synthesis of speech* (Cambridge University Press, Cambridge).
- [158] Liou, H.-S. and Mammone, R.J. (1995), “Application of phonetic weighting to the neural tree network based speaker recognition system”, *Proceedings of the 4th European Conference on Speech Communication and Technology*: 643-646.
- [159] Lobanov, B.M. (1971), “Classification of Russian Vowels Spoken by Different Speakers”, *Journal of the Acoustical Society of America* **49**: 606-608.
- [160] Luck, J.E. (1969), “Automatic Speaker Verification Using Cepstral Measurements”, *Journal of the Acoustical Society of America* **46**: 1026-1032.
- [161] Maeda, S. (1979), “An articulatory model of the tongue based on a statistical analysis”, *Speech Communication Papers presented at the 97th Meeting of the Acoustical Society of America*, Paper I2: 67-70.
- [162] Maeda, S. (1988), “Improved articulatory model”, *Journal of the Acoustical Society of America* **81**: S146.
- [163] Maeda, S. (1991), “On articulatory and acoustic variabilities”, *Journal of Phonetics* **19**: 321-331.
- [164] Makhoul, J. (1975 a), “Linear Prediction: A Tutorial Review”, *Proceedings of the IEEE* **63**: 561-580.

- [165] Makhoul, J. (1975b), “Spectral Linear Prediction: Properties and Applications”, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **23**: 283-296.
- [166] Mammone, R.J., Zhang, X. and Ramachandran, R.P. (1996), “Robust Speaker Recognition”, *IEEE Signal Processing Magazine* **13(5)**: 58-71.
- [167] Markel, J.D. and Gray, A.H. (1976), *Linear Prediction of Speech* (Springer-Verlag, Berlin, Heidelberg, New York).
- [168] Mathieu, B. and Laprie, Y. (1997), “Adaptation of Maeda’s model for acoustic to articulatory inversion”, *Proceedings of the 5th European Conference on Speech Communication and Technology*: 2015-2018.
- [169] Matsumoto, H. and Wakita, H. (1978), “Vowel normalization by frequency warping”, *Journal of the Acoustical Society of America* **64**: S180.
- [170] Matsumoto, H. and Wakita, H. (1986), “Vowel normalization by frequency warped spectral matching”, *Speech Communication* **5**: 239-251.
- [171] McGowan, R.S. (1994), “Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests”, *Speech Communication* **14**: 19-48.
- [172] McGowan, R.S. (1997), “Normalization for articulatory recovery”, *Journal of the Acoustical Society of America* **101**: 3175.
- [173] Meisel, W.S. (1972), *Computer-oriented approaches to pattern recognition* (Academic Press, New York, London).
- [174] Mella, O. (1994), “Extraction of formants of oral vowels and critical analysis for speaker characterization”, *Proceedings of the ESCA Workshop on Automatic Speaker Recognition, Identification and Verification, Martigny, Switzerland*: 193-196.
- [175] Mermelstein, P. (1967), “Determination of the Vocal-Tract Shape from Measured Formant Frequencies”, *Journal of the Acoustical Society of America* **41**: 1283-1294.
- [176] Mermelstein, P. (1973), “Articulatory model for the study of speech production”, *Journal of the Acoustical Society of America* **53**: 1070-1082.
- [177] Morse, P.M. and Ingard, K.U. (1968), *Theoretical Acoustics* (McGraw-Hill, New York).
- [178] Mrayati, M., Carré, R. and Guérin, B. (1988), “Distinctive Regions and Modes: A new theory of speech production”, *Speech Communication* **7**: 257-286.

- [179] Nakajima, T., Ohmura, H., Tanaka, K. and Ishizaki, S. (1973), "Estimation of Vocal Tract Area Functions by Adaptive Inverse Filtering Methods", *Bulletin of the Electrotechnical Laboratory* **37**: 462-481.
- [180] Narayanan, S., Alwan, A. and Song, Y. (1997), "New results in vowel production: MRI, EPG, and acoustic data", *Proceedings of the 5th European Conference on Speech Communication and Technology*: 1007-1010.
- [181] Nearey, T.M. (1978), "Phonetic feature systems for vowels", Indiana University Linguistics Club, Bloomington, Indiana, USA.
- [182] Nolan, F. (1983), *The phonetic bases of speaker recognition* (Cambridge University Press, Cambridge).
- [183] Nordström, P-E. (1977), "Female and infant vocal tracts simulated from male area functions", *Journal of Phonetics* **5**: 81-92.
- [184] Ohmura, H. (1993), "An Algorithm for Calculating Vocal Tract Transfer Functions in terms of Reflection Coefficients", *Electrotechnical Laboratory Progress Report on Speech Research*: 1-23.
- [185] Oppenheim, A.V., Schafer, R.W. and Stockham, T.G. (1968), "Nonlinear Filtering of Multiplied and Convolved Signals", *Proceedings of the IEEE* **56**: 1264-1291.
- [186] Owren, M.J. and Bachorowski, J.-A. (1997), "Reliable cues to gender and talker identity are present in a short vowel segment recorded in running speech", *Journal of the Acoustical Society of America* **102**: 3132.
- [187] Paige, A. and Zue, V.W. (1970), "Calculation of Vocal Tract Length", *IEEE Transactions on Audio and Electroacoustics* **18**: 268-270.
- [188] Paliwal, K.K. (1982), "On the performance of the quefrency-weighted cepstral coefficients in vowel recognition", *Speech Communication* **1**: 151-154.
- [189] Paliwal, K.K. (1984a), "Effectiveness of different vowel sounds in automatic speaker identification", *Journal of Phonetics* **12**: 17-21.
- [190] Paliwal, K.K. (1984b), "Effect of preemphasis on vowel recognition performance", *Speech Communication* **3**: 101-106.
- [191] Paliwal, K.K. and Ainsworth, W.A. (1985), "Dynamic frequency warping for speaker adaptation in automatic speech recognition", *Journal of Phonetics* **13**: 123-134.
- [192] Paliwal, K.K. and Rao, P.V.S. (1982), "Evaluation of various linear prediction parametric representations in vowel recognition", *Signal Processing* **4**: 323-327.

- [193] Parthasarathy, S. and Coker, C.H. (1992), "On automatic estimation of articulatory parameters in a text-to-speech system", *Computer Speech and Language* **6**: 37-75.
- [194] Payan, Y. and Perrier, P. (1993), "Vowel normalization by articulatory normalization: First attempts for vowel transitions", *Proceedings of the 3rd European Conference on Speech Communication and Technology*: 417-420.
- [195] Perkell, J.S. (1969), *Physiology of Speech Production: Results and Implications of a Quantitative Cineradiographic Study* (Research Monograph No.53, The M.I.T. Press, Cambridge, Massachusetts).
- [196] Perkell, J.S., Matthies, M.L., Svirsky, M.A. and Jordan, M.I. (1993), "Trading relations between tongue-body raising and lip rounding in production of the vowel /u/: A pilot "motor equivalence" study", *Journal of the Acoustical Society of America* **93**: 2948-2961.
- [197] Perrier, P., Apostol, L. and Payan, Y. (1995), "Evaluation of a vowel normalisation procedure based on speech production knowledge", *Proceedings of the 4th European Conference on Speech Communication and Technology*: 1925-1928.
- [198] Peters, R.W. (1954), "Studies in extra messages: Listener identification of speakers' voices under conditions of certain restrictions imposed upon the voice signal", U.S. Naval School of Aviation Medicine, Joint Project NM 001-064-01, Rpt.30, Pensacola, Florida.
- [199] Peterson, G.E. (1952), "The Information-Bearing Elements of Speech", *Journal of the Acoustical Society of America* **24**: 629-637.
- [200] Peterson, G.E. (1959), "The Acoustics of Speech — Part II. Acoustical Properties of Speech Waves", in Travis, L.E. (ed.), *Handbook of Speech Pathology* (Peter Owen, London): 137-173.
- [201] Peterson, G.E. (1961), "Parameters of Vowel Quality", *Journal of Speech and Hearing Research* **4**: 10-29.
- [202] Peterson, G.E. and Barney, H.L. (1952), "Control Methods Used in a Study of the Vowels", *Journal of the Acoustical Society of America* **24**: 175-184.
- [203] Piper, J. (1992), "Variability and bias in experimentally measured classifier error rates", *Pattern Recognition Letters* **13**: 685-692.
- [204] Plomp, R., Pols, L.C.W. and van de Geer, J.P. (1967), "Dimensional Analysis of Vowel Spectra", *Journal of the Acoustical Society of America* **41**: 707-712.
- [205] Pols, L.C.W., van der Kamp, L.J.Th. and Plomp, R. (1969), "Perceptual and Physical Space of Vowel Sounds", *Journal of the Acoustical Society of America* **46**: 458-467.

- [206] Pols, L.C.W., Tromp, H.R.C. and Plomp, R. (1973), “Frequency analysis of Dutch vowels from 50 male speakers”, *Journal of the Acoustical Society of America* **53**: 1093-1101.
- [207] Potter, R.K. and Steinberg, J.C. (1950), “Toward the Specification of Speech”, *Journal of the Acoustical Society of America* **22**: 807-820.
- [208] Press, W.H., Flannery, B.P., Teukolsky, S.A. and Vetterling, W.T. (1988), *Numerical Recipes in C: The Art of Scientific Computing* (Cambridge University Press).
- [209] Rabiner, L. and Juang, B.-H. (1993), *Fundamentals of Speech Recognition* (Prentice Hall, New Jersey).
- [210] Richards, H.B., Mason, J.S., Hunt, M.J. and Bridle, J.S. (1995), “Deriving articulatory representations of speech”, *Proceedings of the 4th European Conference on Speech Communication and Technology*: 761-764.
- [211] Riordan, C.J. (1977), “Control of vocal-tract length in speech”, *Journal of the Acoustical Society of America* **62**: 998-1002.
- [212] Rosenberg, A.E. and Sambur, M.R. (1975), “New Techniques for Automatic Speaker Verification”, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **23**: 169-176.
- [213] Saito, S. and Itakura, F. (1983), “Frequency spectrum deviation between speakers”, *Speech Communication* **2**: 149-152.
- [214] Saito, S. and Nakata, K. (1985), *Fundamentals of Speech Signal Processing* (Academic Press, Tokyo, Florida, London).
- [215] Sakoe, H. and Chiba, S. (1978), “Dynamic programming algorithm optimization for spoken word recognition”, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **26**: 43-49.
- [216] Sambur, M.R. (1975), “Selection of Acoustic Features for Speaker Identification”, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **23**: 176-182.
- [217] Sarma, V.V.S. and Venugopal, D. (1977), “Performance Evaluation of Automatic Speaker Verification Systems”, *IEEE Transactions on Acoustics, Speech, and Signal Processing* **25**: 264-266.
- [218] Sato, S., Yokota, M. and Kasuya, H. (1982), “Statistical Relationships among the First Three Formant Frequencies in Vowel Segments in Continuous Speech”, *Phonetica* **39**: 36-46.

- [219] Savariaux, C., Perrier, P. and Orliaguet, J.P. (1995), "Compensation strategies for the perturbation of the rounded vowel [u] using a lip tube: A study of the control space in speech production", *Journal of the Acoustical Society of America* **98**: 2428-2442.
- [220] Schroeder, M.R. (1967), "Determination of the Geometry of the Human Vocal Tract by Acoustic Measurements", *Journal of the Acoustical Society of America* **41**: 1002-1010.
- [221] Schroeter, J. and Sondhi, M.M. (1989), "Dynamic Programming Search of Articulatory Codebooks", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*: 588-591.
- [222] Schroeter, J. and Sondhi, M.M. (1992), "Speech Coding Based on Physiological Models of Speech Production", in Furui, S. and Sondhi, M.M. (eds.), *Advances in Speech Signal Processing* (Marcel Dekker, New York, Basel, Hong Kong): 231-268.
- [223] Schroeter, J. and Sondhi, M.M. (1994), "Techniques for Estimating Vocal-Tract Shapes from the Speech Signal", *IEEE Transactions on Speech and Audio Processing* **2**: 133-150.
- [224] Sejnoha, V. and Mermelstein, P. (1983), "Speaker normalizing transforms for automatic recognition", *Journal of the Acoustical Society of America* **74**: S17.
- [225] Shikano, K. and Itakura, F. (1992), "Spectrum Distance Measures for Speech Recognition", in Furui, S. and Sondhi, M.M. (eds.), *Advances in Speech Signal Processing* (Marcel Dekker, New York, Basel, Hong Kong): 419-452.
- [226] Shirai, K. and Honda, M. (1977), "Estimation of articulatory motion", in Sawashima, M. and Cooper, F.S. (eds.), *Dynamic Aspects of Speech Production* (University of Tokyo Press, Tokyo): 279-304.
- [227] Singh, S. and Singh, K.S. (1976), *Phonetics: principles and practice* (University Park Press, Baltimore, London, Tokyo).
- [228] Singh, S. and Woods, D.R. (1971), "Perceptual Structure of 12 American English Vowels", *Journal of the Acoustical Society of America* **49**: 1861-1866.
- [229] Sondhi, M.M. (1979), "Estimation of Vocal-Tract Areas: The Need for Acoustical Measurements", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **27**: 268-273.
- [230] Sondhi, M.M. and Schroeter, J. (1987), "A Hybrid Time-Frequency Domain Articulatory Speech Synthesizer", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **35**: 955-967.

- [231] Soquet, A. and Saerens, M. (1994), "A comparison of different acoustic and articulatory representations for the determination of place of articulation of plosives", *Proceedings of the 3rd International Conference on Spoken Language Processing*: 1643-1646.
- [232] Sorokin, V.N. (1992), "Determination of vocal tract shape for vowels", *Speech Communication* **11**: 71-85.
- [233] Spiegel, M. (1994), "Re: Peterson and Barney's data", article 1519 of the *comp.speech* Internet newsgroup.
- [234] Stevens, K.N. (1971), "Sources of inter- and intra-speaker variability in the acoustic properties of speech sounds", *Proceedings of the 7th International Congress of Phonetic Sciences*: 206-232.
- [235] Stevens, K.N. and House, A.S. (1955), "Development of a Quantitative Description of Vowel Articulation", *Journal of the Acoustical Society of America* **27**: 484-493.
- [236] Stevens, K.N. and House, A.S. (1963), "Perturbation of Vowel Articulations By Consonantal Context: An Acoustical Study", *Journal of Speech and Hearing Research* **6**: 111-128.
- [237] Stevens, K.N., Williams, C.E., Carbonell, J.R. and Woods, B. (1968), "Speaker Authentication and Identification: A Comparison of Spectrographic and Auditory Presentations of Speech Material", *Journal of the Acoustical Society of America* **44**: 1596-1607.
- [238] Story, B.H., Titze, I.R. and Hoffman, E.A. (1996), "Vocal tract area functions from magnetic resonance imaging", *Journal of the Acoustical Society of America* **100**: 537-554.
- [239] Strube, H.W. (1977), "Can the area function of the human vocal tract be determined from the speech wave?", in Sawashima, M. and Cooper, F.S. (eds.), *Dynamic Aspects of Speech Production* (University of Tokyo Press, Tokyo): 233-250.
- [240] Strube, H.W. (1980), "Linear prediction on a warped frequency scale", *Journal of the Acoustical Society of America* **68**: 1071-1076.
- [241] Sundberg, J. (1974), "Articulatory interpretation of the 'singing formant'", *Journal of the Acoustical Society of America* **55**: 838-844.
- [242] Sundberg, J. (1995), "The singer's formant revisited", *Speech Transmission Laboratory, Quarterly Progress and Status Report* **2-3**: 83-96.
- [243] Suomi, K. (1984), "On talker and phoneme information conveyed by vowels: A whole spectrum approach to the normalization problem", *Speech Communication* **3**: 199-209.

- [244] Syrdal, A.K. and Gopal, H.S. (1986), "A perceptual model of vowel recognition based on the auditory representation of American English vowels", *Journal of the Acoustical Society of America* **79**: 1086-1100.
- [245] Tanaka, K. and Nakajima, T. (1975), "Estimation of Vocal Tract Length and Formant Frequencies by Adaptive Enhancing Filter", *Electrotechnical Laboratory Progress Report on Speech Research* **9**: 10-14.
- [246] Tohkura, Y. (1986), "A Weighted Cepstral Distance Measure for Speech Recognition", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*: 761-764.
- [247] Tou, J.T. and Gonzalez, R.C. (1974), *Pattern Recognition Principles* (Addison-Wesley, Reading, Massachusetts).
- [248] Toussaint, G.T. (1974), "Bibliography on Estimation of Misclassification", *IEEE Transactions on Information Theory* **20**: 472-479.
- [249] Umesh, S., Cohen, L., Marinovic, N. and Nelson, D. (1996), "Frequency-Warping in Speech", *Proceedings of the 4th International Conference on Spoken Language Processing*: 414-417.
- [250] van den Heuvel, H., Cranen, B. and Rietveld, A.C.M. (1993), "Speaker-variability in spectral bands of Dutch vowel segments", *Proceedings of the 3rd European Conference on Speech Communication and Technology*: 635-638.
- [251] van den Heuvel, H. and Rietveld, T. (1992), "Speaker related variability in cepstral representations of Dutch speech segments", *Proceedings of the 2nd International Conference on Spoken Language Processing*: 1581-1584.
- [252] van Nierop, D.J.P.J., Pols, L.C.W. and Plomp, R. (1973), "Frequency Analysis of Dutch Vowels from 25 Female Speakers", *Acustica* **29**: 110-118.
- [253] Wakita, H. (1973), "Direct Estimation of the Vocal Tract Shape by Inverse Filtering of Acoustic Speech Waveforms", *IEEE Transactions on Audio and Electroacoustics* **21**: 417-427.
- [254] Wakita, H. (1977), "Normalization of Vowels by Vocal-Tract Length and Its Application to Vowel Identification", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **25**: 183-192.
- [255] Wakita, H. (1979), "Estimation of Vocal-Tract Shapes from Acoustical Analysis of the Speech Wave: The State of the Art", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **27**: 281-285.

- [256] Wakita, H. and Fant, G. (1978), "Toward a better vocal tract model", *Speech Transmission Laboratory, Quarterly Progress and Status Report* **1**: 9-29.
- [257] Wakita, H. and Gray, A.H. (1975), "Numerical Determination of the Lip Impedance and Vocal Tract Area Functions", *IEEE Transactions on Acoustics, Speech, and Signal Processing* **23**: 574-580.
- [258] Watrous, R.L. (1991), "Current status of Peterson-Barney vowel formant data", *Journal of the Acoustical Society of America* **89**: 2459-2460.
- [259] Welch, P.D. and Wimpess, R.S. (1961), "Two Multivariate Statistical Computer Programs and Their Application to the Vowel Recognition Problem", *Journal of the Acoustical Society of America* **33**: 426-434.
- [260] Wood, S. (1979), "A radiographic analysis of constriction locations for vowels", *Journal of Phonetics* **7**: 25-43.
- [261] Wood, S. (1986), "The acoustical significance of tongue, lip, and larynx maneuvers in rounded palatal vowels", *Journal of the Acoustical Society of America* **80**: 391-401.
- [262] Yang, C.S. and Kasuya, H. (1994), "Accurate measurement of vocal tract shapes from magnetic resonance images of child, female and male subjects", *Proceedings of the 3rd International Conference on Spoken Language Processing*: 623-626.
- [263] Yang, C.-S. and Kasuya, H. (1995), "Vowel normalization revisited: integration of articulatory, acoustic, and perceptual measurements", *Proceedings of the 13th International Congress of Phonetic Sciences* **3**: 234-237.
- [264] Yang, C.-S. and Kasuya, H. (1996), "Speaker individualities of vocal tract shapes of Japanese vowels measured by magnetic resonance images", *Proceedings of the 4th International Conference on Spoken Language Processing*: 949-952.
- [265] Yegnanarayana, B. (1978), "Formant extraction from linear-prediction phase spectra", *Journal of the Acoustical Society of America* **63**: 1638-1640.
- [266] Yegnanarayana, B. and Reddy, D.R. (1979), "A distance measure based on the derivative of linear prediction phase spectrum", *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*: 744-747.
- [267] Yehia, H. and Itakura, F. (1994), "Determination of human vocal-tract dynamic geometry from formant trajectories using spatial and temporal Fourier analysis", *Proceedings of the International Conference on Acoustics, Speech and Signal Processing* **I**: 477-480.

- [268] Yehia, H. and Itakura, F. (1996), “A method to combine acoustic and morphological constraints in the speech production inverse problem”, *Speech Communication* **18**: 151-174.
- [269] Young, S. (1996), “A Review of Large-vocabulary Continuous-speech Recognition”, *IEEE Signal Processing Magazine* **13(5)**: 45-57.
- [270] Zahorian, S.A. and Jagharghi, A.J. (1993), “Spectral-shape features versus formants as acoustic correlates for vowels”, *Journal of the Acoustical Society of America* **94**: 1966-1982.
- [271] Zue, V.W. (1969), “A noniterative computation of vocal-tract area function”, unpublished M.S. Thesis, University of Florida, Gainesville, USA.
- [272] Zwicker, E. (1961), “Subdivision of the Audible Frequency Range into Critical Bands (Frequenzgruppen)”, *Journal of the Acoustical Society of America* **33**: 248.
- [273] Zwicker, E. and Terhardt, E. (1980), “Analytical expressions for critical-band rate and critical bandwidth as a function of frequency”, *Journal of the Acoustical Society of America* **68**: 1523-1525.