Segmental effects on the prosody of voice quality

H. Pfitzinger

Christian-Albrechts-University, Leibnizstr. 10, 24118 Kiel, Germany
hpt@phonetik.uni-muenchen.de

Voice quality variability is due to supra-segmental influences but also to segmental factors like phoneme class, vowel quality, nasalization, airstream mechanism etc. These factors determine a rather unexplored micro-prosodic phenomenon: the phone-intrinsic voice quality which causes voice quality coarticulation and voice quality transitions in fluent speech. I subsume all these phenomena under the high-frequency components of prosody. Since high-frequency and low-frequency components (supra-segmentals) of voice quality prosody are superposed and thus encoded, the main goal of the present investigation is to separate them and make both accessible to speech research. In 2003 a holistic voice quality parameter extractor was introduced by Mokhtari, Pfitzinger & Ishi [10]: It applies a principal component analysis to a database of glottal-flow waveforms for the purpose of analysing and reconstructing all underlying glottal-flow waveforms from just a few principal components. By applying this basic principle to a large corpus of 92,167 manually segmented glottal-flow waveforms of 44 speakers I dramatically improved its applicability to any speech signals. Subsequent high-pass and low-pass filtering of the resulting voice quality parameters yielded phone-intrinsic voice quality parameter sets as well as slowly varying voice quality parameter contours. It turned out that 99.44% of the observed variance is explained by the first six principal components.

1 Introduction

The unique timbre of an individual human voice is the result of the static nature as well as the dynamic control of both the glottal behaviour and the vocal tract shape. Generally when referring to the term *voice quality* the properties of glottal behaviour are addressed, especially the precise vibration of the vocal folds which results in the glottal air flow and, after passing through the vocal tract and subsequent radiation at the lips, in the acoustic pressure waveform which is the speech signal.

Childers & Wong 1994 [3] emphasize the presence of interaction between the laryngeal activity and the vocal tract, meaning that the two components of the source-filter-model of speech production are neither fully separable nor independent of each other. The glottis and the vocal tract are interacting. Despite this fact Linear Predictive Coding (LPC) [9] is widely used to decompose the speech signal into a 'quasi' source signal and short-term 'quasi' transfer functions of the vocal tract represented by autoregressive filter coefficients. LPC-based inverse filtering yields a quasi-excitation signal and quasi-stationary filter coefficients and thus enables meaningful modifications to the vocal tract parameterization. Many recent approaches to speech morphing and voice conversion successfully use this approach [2, 14].

However, glottal air flow parameterization is a more heterogeneous topic. During several decades, glottal flow (derivative) models which combine basic mathematical functions to approximate the waveshape, approaches based on physical modelling, and data driven approaches arose and disappeared. This situation is euphemistically characterized as a modest breakthrough.

In 2006 [16] I stated that voice quality as one of the prosodic dimensions has its supra-segmental and segmental manifestation, i.e. has low- and high-frequency components which are almost independently. The consequence is to further decompose the parameterized glottal flow, which is the topic of the present study.

Koreman, Boves & Cranen 1992 [6] used an electroglottograph and a Rothenberg mask to measure and analyse the influence of linguistic variations on the dynamics of the voice source characteristics.

Swerts & Veldhuis 2001 [18] investigated voice quality changes as a function of the intonation contour. Their analyses are based on several productions of the vowel [a] provided with different intonation patterns. They found that F0 covaries with the amplitude difference of the first two harmonics and that this amplitude difference interacts with the open quotient as well as the skewness of the glottal pulse when analysed by means of the Liljencrants-Fant-model [4].

These and other findings motivate a more holistic approach to the parameterization of the glottal flow. In 2003 a new model was introduced by Mokhtari, Pfitzinger & Ishi [10]: It applies a principal component analysis (PCA) to a database of glottal-flow waveforms for the purpose of analysing and reconstructing all underlying glottal-flow waveforms from just a few principal components. This approach was successfully applied to the problem of laryngeal voice quality conversion [11] which in the past was addressed, rather, by means of e.g. code(-book) excited linear prediction (CELP) [2]. Therefore, it serves as the preferred analysis framework during the current investigation.

2 Method

The aim of the present study is to measure and analyse the effect of speaker and phoneme class on the prosody of voice quality as represented by the holistic PCA parameterization of laryngeal excitation [10].

2.1 Speech database

This investigation is based on a new speech database originally designed for the evaluation of automatic GCI detection (*glottal closure instant*, also called *glottal excitation impulse* or *glottal epoch*). Therefore it consists of 112 phonetically rich German sentences and 100 English sentences especially selected to provide an above-average number of voice onsets and offsets.

Each part of the database was produced by 11 female and 11 male native speakers, respectively. Thus, a total of 44 speakers participated in the recordings. An EGG (electroglottograph) was used to record laryngeal activity. The delay between the speech signal and the EGG signal was reduced to approximately ± 0.05 ms via cross-correlation based delay compensation [17]. The speech signals together with the EGG signals of 396 utterances (9 sentences times 44 speakers) were subjected to automatic GCI detection with careful manual re-adjustment of all marks. Additionally, the speech material was segmented on a phone level using German and English SAMPA symbols, respectively.

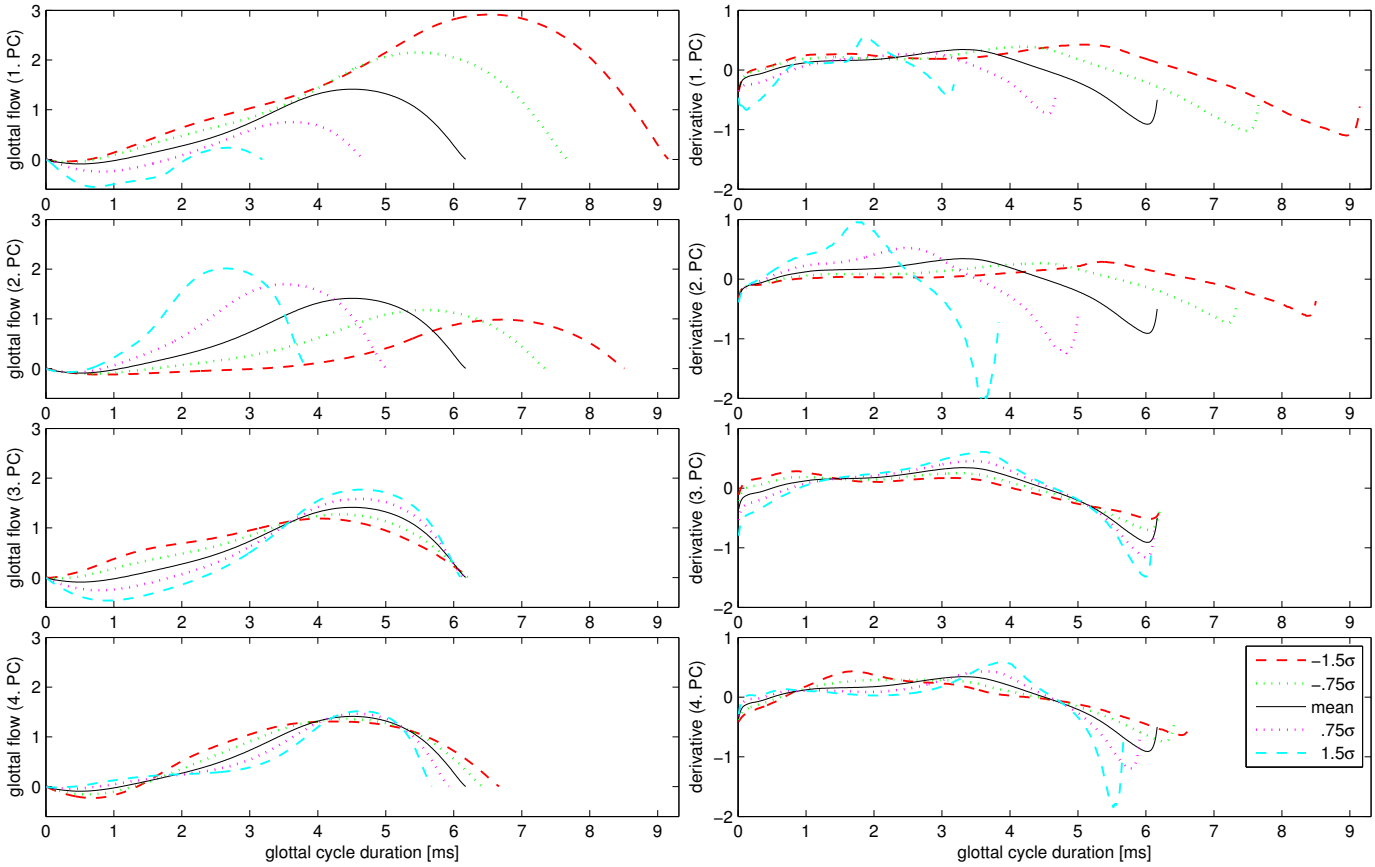

Figure 1: Effect of the 1st to 4th principal component on the mean glottal volume-velocity waveform *(left)* and its derivative *(right)*. The mean cycle *(solid)* is averaged over 92,167 manually segmented glottal cycles of 44 speakers.

2.2 Glottal excitation parameterisation

The quasi-excitation signal was extracted via inverse filtering: Standard autocorrelation-based LPC with a 98th order polynomial, a pre-emphasis coefficient of -0.995, a Kaiser window ($\beta = 5$) with 20 ms duration, and a step width of 5 ms was applied to the speech signal. The resulting filter coefficients were samplewise interpolated [13] and then used to perform FIR-based inverse filtering. The energy values of the LPC were ignored in order to pass on the amplitude variation of the speech signal to the white signal. Adaptive pre-emphasis [15] was not used since, at the actual sampling frequency of 96 kHz, 1st-order emphasis is not sufficient to compensate for all possible configurations of the spectral slope.

The resulting quasi-excitation signals were cut on a glottal cycle level according to the manually corrected GCI marks, yielding 92,167 glottal cycles. Following source-filter theory of speech production [9, p.6] these signals correspond to the glottal flow derivative while the glottal flow is achieved via subsequent mathematical integration. To make start and end samples equal for each glottal cycle and thus avoid any skewness, prior to integration the corresponding average amplitude value was subtracted from each glottal flow derivative.

Then, 300 two-dimensional t-parameters (known from computer graphics) were equidistantly distributed along the path of the waveform of each of the 92,167 glottal flow derivative cycles. The extraordinarily high number of parameters was chosen in anticipation of speech synthesis to retain roughly the first 150 harmonics and thus considering a spectral bandwidth of 15 kHz at a hypothetical fundamental frequency of 100 Hz.

These parameters were submitted to principal component analysis (PCA). But while in previous studies [10, 11] the approach was restricted to single-speaker data, to 30 parameter pairs, and up to 19,665 glottal cycles, in the present study I significantly extended this approach towards speaker- and language-independence.

3 Results

Fig. 2 shows that six principal components explain more than 99% of the local glottal flow variation. The effect of the first four principal components on the mean glottal flow cycle and its derivative is shown in Fig. 1 which substantially differs from Fig. 1 in [11] because, among other details, PC1 and PC2 are roughly exchanged.

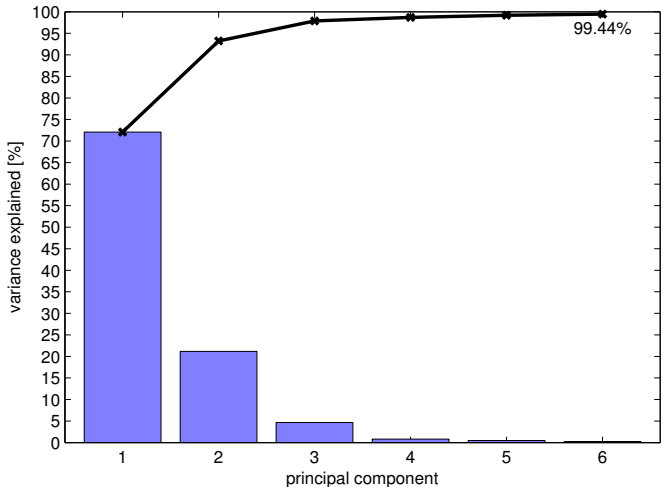

Figure 2: Amount of variance of all 92,167 glottal cycles explained by the first 6 principal components.

To assess the relevance of the principal components of the glottal flow with regard to speaker variability and phoneme class the first and second principal components are projected into a two-dimensional space.

3.1 Inter-speaker effect

Fig. 3 shows that in an PC1/PC2-space speakers are perfectly clustered according to their gender. But this is not remarkable since mean F0 of female versus male speakers does not overlap albeit not as clearly clustered as in the principal components based scatter plot. As also can be seen in Fig. 3, the standard deviation as well as the circumference of the gender-specific distribution of principal components of female speakers is smaller than that of male speakers.

It is possible that this difference is due to the roughly equal F0 range of female and male speakers in the logarithmic frequency domain, since the covariation of some principal components with glottal period duration then leads to a larger variation when the period is longer. But there is evidence that this is not a sufficient explanation since the standard deviations of all speakers have roughly the same magnitude, only the female speakers are less widely distributed in the PC1-PC2-space. One could speculate that the voice qualities of female voices are more similar than male voices.

3.2 Segmental effect

Fig. 4 shows the influence of phoneme class on the position of phones in the PC1/PC2-space for speaker *25*.

Across speakers the arrangements of the positions of phones in the PC1/PC2-space are non-uniform. A possible reason is that acoustic variation of actual phonetic realisations within a phoneme class is to a large extent due to speaker-specific production strategies [7, 12].

But it is worth noting that within a speaker similar phones are clustered: on the left side of Fig. 4 are labial (v, m) and alveolar (d, n, l) phones, at the top are high front vowels (I, i:, i), at the bottom open vowels (A:, aI, aU) and in the center are central vowels (@, @U, e@, eI). Phones represented by less than 50 glottal cycles are omitted from this diagram.

Generally, phone-dependent variation of the white signal is regarded as a weakness of the LPC since it is basically a linear pole-modelling method. Consequently, zeros are assigned to the excitation signal. However, in [15] I investigated the influence of various, also nonlinear, inverse filtering techniques on the residual signal. It turned out, albeit informally, that none of the investigated methods significantly reduced phone-dependency, which supports the present assumption that the excitation signal is to a certain extent phone-specific.

3.3 Macro- and micro-prosody

The scores of the first three principal components were transformed into time-domain signals, i.e. the samples of each glottal cycle of the original excitation signal were replaced by its corresponding PCA score leading to, for each principle component, an individual time function synchronous to the speech signal.

As an example how our PCA-based glottal flow model represents the local variation of voice quality, Fig. 5 shows the utterance "*His blood grew hot with rage at the thought*" with F0- and amplitude contours as well as the first three principal component contours.

A remarkable property of the resulting contours is that successive values are not randomly fluctuating but highly correlated. This gives rise to expect reliable trends and assume meaningfulness.

These signals contain segmental and supra-segmental variation. Therefore, a lowpass-filter with a cutoff frequency of 2 Hz was used to extract the slowly varying prosodic information on voice quality, while the residual contains the micro-prosodic variation caused by phone-

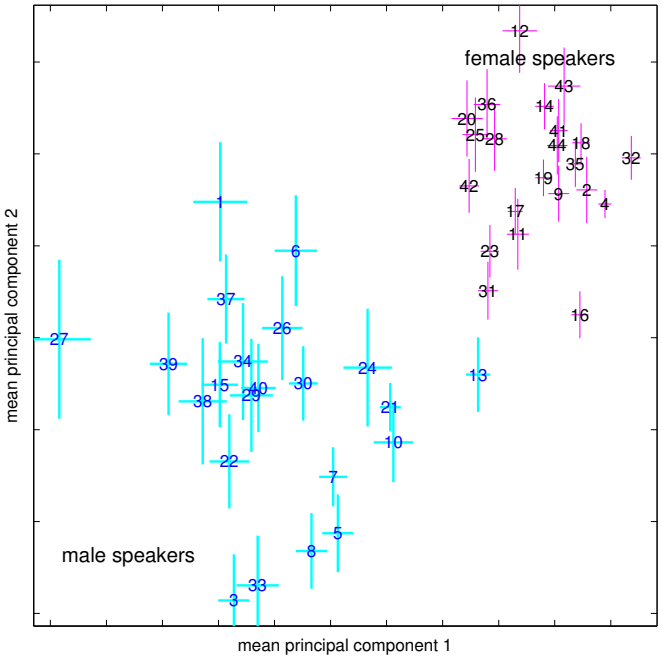

Figure 3: Scatter plot of mean PC1 versus mean PC2. PCs are averaged for each speaker separately. Underlying horizontal and vertical lines represent standard deviations (male: *light bold lines*).

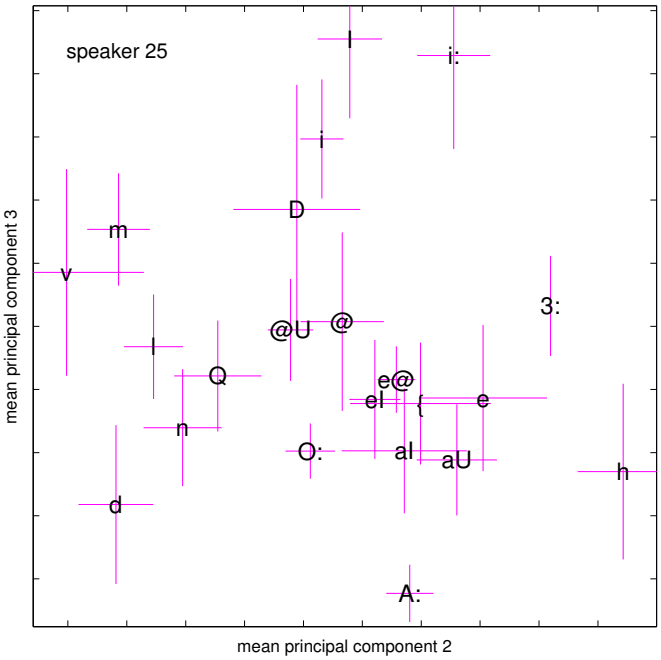

Figure 4: Scatter plot of mean PC2 versus mean PC3 for female speaker *25*. PCs are averaged for each SAMPA phone separately. Underlying horizontal and vertical lines represent standard deviations.

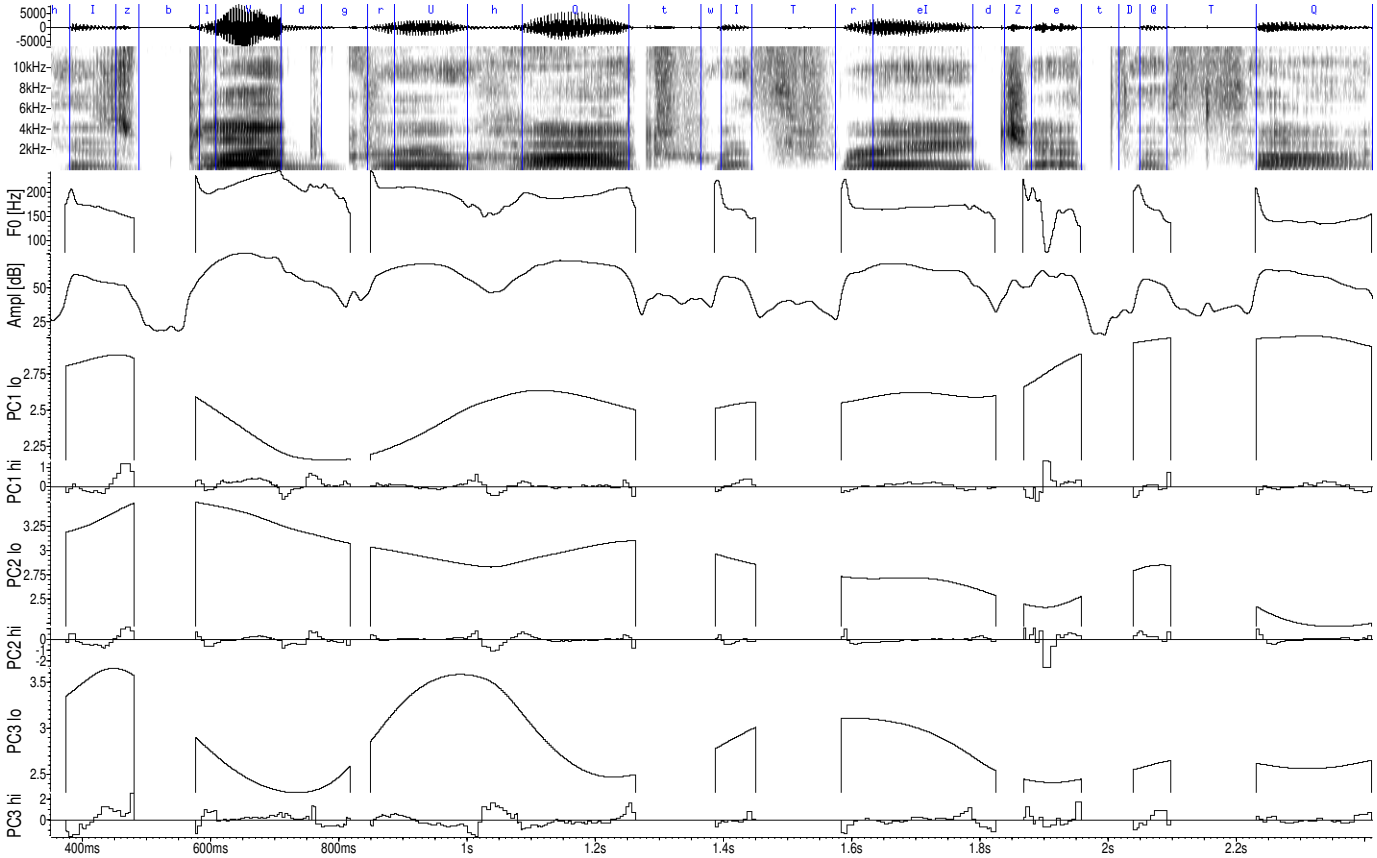

Figure 5: "*His blood grew hot with rage at the thought*" (female speaker *23*) with manual phone segmentation. From top to bottom: speech signal, F0, amplitude, low- and high-frequency components of the 1st, 2nd, and 3rd PCs.

specific voice quality settings, by coarticulatory effects during phone-class changes, and also by voice-quality jitter/shimmer.

Although the meaning of the supra-segmental PC contours is far less obvious than e.g. F0 or amplitude contours, three observations should be mentioned: 1) the 2nd PC reflects the global declination, 2) the sentence focus *blood* is accompanied by a minimum of the contours of the 1st and 3rd PC, and 3) non-prominent stretches of the utterance such as the initial word and the last three words reach the highest values of PC1.

4 Discussion

A possible explanation for the observed phone-specific voice quality variation might be twofold: *i)* the acoustic and mechanical coupling of the glottis and the vocal tract yields unavoidable interaction, and *ii)* the actively (but not necessarily intentionally) controlled changes in voice quality are needed to produce the appropriate phone quality. A typical example of the latter is post-aspiration in the Indian language Gujarati. Going that far, one could also regard the glottal plosive as a segmental voice quality change with strong voice quality coarticulation between the adjacent phones. This view is in accordance with the tenor of Firth's theory [5] of prosodies being more than only intonation contours, manifesting also in the segmental structure of speech, being one of the reasons for a specific segmental realization or, more generally, for segmental variation.

Obviously, F0, amplitude, and voice quality parameters like e.g. the normalized amplitude quotient (NAQ) [1] or the four Liljencrants-Fant-parameters [4] are not directly correlated with any of the glottal principal components. It is hard, if not impossible, to assign any impressionistic interpretation to one of the PCs. Even the most important PC, which explains 72% of the total variance, encodes at least F0, amplitude, and "waveform smoothness". But this should not be interpreted as a shortcoming of the method. On the contrary, the effect of the principal components on the mean glottal flow resulting in a specific waveshape, reveals the covariation of several important acoustic features.

Usually, they are measured and interpreted separately. This practice is very common when studying intonation. F0 is generally the only acoustic feature being analysed, and to make matters worse, the only parameter which is strongly manipulated in stimuli for studies on prosody perception. The well-known interactions between the glottis and the vocal tract are ignored just as much as the little-known F0-voice quality interactions.

In this regard, the holistic approach opens new possibilities to manipulate speech features in a more natural way. This approach maintains that acoustic parameters are single independent variables only in the acoustic domain, but not in the perceptual domain. For example, a stimulus with an algorithmically increased F0 gives the impression of an unnaturally large vocal tract the higher F0 becomes. Thus, an implicitly covarying vocal tract length would lead to the perception of only F0 changes. The same is expected to be true for F0 and voice quality. This approach allows for single independent variables in the perceptual domain and hence prepares for high-level and higher-knowledge voice quality modelling.

5 Conclusions

This paper presented methodological aspects as well as data and empirical analyses of the multi-dimensional nature of voice quality. 92,167 glottal cycles of 44 speakers were subjected to principal component analysis to condense multi-dimensionality down to less than 10 parameters. It turned out that 99.44% of the observed variance is explained by the first six principal components.

Voice quality is speaker-specific and phone-specific but also contains a prosodic component with its micro- and at the same time macro-prosodic properties. While until recently, voice qualities such as "creaky voice" have been regarded as a static setting [8], the present study enables detailed insights into the glottal flow variation along successive glottal cycles and suggests the separation of micro-prosodic effects, e.g. caused by adduction and abduction of the vocal folds at the onset and offset of voiced stretches of speech, from macro-prosodic or supra-segmental effects, caused by covariation with the F0 contour or the amplitude contour but also by phenomena such as the pre-final lengthening phase of an utterance which sometimes is accompanied by slowly increasing breathiness. The presented methodology could be regarded as both a new paradigm of covarying prosodies and a practical advice how to significantly reducing the only apparently infinite dimensionality of prosody as already outlined in 2006 [16].

A very small fraction of the unexplained variance is due to transient signals such as the bursts of voiced plosives which are currently not separable from the glottal excitation signal. Thus, at the present stage of the new analysis method for the prosody of voice quality, these transients occasionally cause outlier values in the micro-prosody of the voice quality but due to their singular nature the supra-segmental level is almost unaffected.

Acknowledgements

I am very grateful to Siemens AG, Munich for partly supporting this work. I am also indebted to Nils Ülzmann, Benjamin Hamer, Anja Hübner, Thomas Jacobsen, and Nina Redlingshöfer from IPDS, Kiel for their skilful manual epoch and phone segmentation.

References

[1] Alku, P.; Bäckström, T.; Vilkman, E. (2002). Normalized amplitude quotient for parametrization of the glottal flow. *J. of the Acoustical Society of America*, 112(2): 701–710.

[2] Childers, D. G. (1995). Glottal source modeling for voice conversion. *Speech Communication*, 16(2): 127–138.

[3] Childers, D. G.; Wong, C.-F. (1994). Measuring and modeling vocal source-tract interaction. *IEEE Trans. on Biometrical Engineering*, 41(7): 663–671.

[4] Fant, G.; Liljencrants, J.; Lin, Q.-g. (1985). A four-parameter model of glottal flow. Speech Transmission Lab., Quarterly Progress and Status Report 4, pp. 1–13, KTH, Stockholm.

[5] Firth, J. R. (1948). Sounds and prosodies. *Transactions of the Philological Society*, pp. 127–152.

[6] Koreman, J.; Boves, L.; Cranen, B. (1992). The influence of linguistic variations on the voice source characteristics. In *Proc. of ICSLP '92*, vol. 1, pp. 125–128, Banff; Kanada.

[7] Kuehn, D. P.; Moll, K. L. (1976). A cineradiographic study of VC and CV articulatory velocities. *J. of Phonetics*, 4: 303–320.

[8] Laver, J. (1980). *The phonetic description of voice quality.* Cambridge University Press, Cambridge.

[9] Markel, J. D.; Gray Jr., A. H. (1976). *Linear prediction of speech.* Communication and Cybernetics, 12. Springer-Verlag, Berlin, Heidelberg, New York.

[10] Mokhtari, P.; Pfitzinger, H. R.; Ishi, C. T. (2003). Principal components of glottal waveforms: Towards parameterisation and manipulation of laryngeal voice-quality. In *Proc. of the ISCA Tutorial and Research Workshop on Voice Quality: Functions, Analysis and Synthesis (Voqual'03)*, pp. 133–138, Geneva.

[11] Mokhtari, P.; Pfitzinger, H. R.; Ishi, C. T.; Campbell, N. (2004). Laryngeal voice quality conversion by glottal waveshape PCA. In *Proc. of the Spring 2004 Meeting of the Acoustical Society of Japan*, pp. 341–342, Atsugi; Japan.

[12] Pfitzinger, H. R. (2002). Intrinsic phone durations are speaker-specific. In *Proc. of ICSLP '02*, vol. 2, pp. 1113–1116, Denver.

[13] Pfitzinger, H. R. (2004). DFW-based spectral smoothing for concatenative speech synthesis. In *Proc. of ICSLP '04*, vol. 2, pp. 1397–1400, Korea.

[14] Pfitzinger, H. R. (2004). Unsupervised speech morphing between utterances of any speakers. In *Proc. of the 10th Australian Int. Conf. on Speech Science and Technology (SST 2004)*, pp. 545–550, Sydney.

[15] Pfitzinger, H. R. (2005). Influence of differences between inverse filtering techniques on the residual signal of speech. In *Fortschritte der Akustik (DAGA '05)*, vol. 1, pp. 223–224, Munich. Deutsche Physikalische Gesellschaft (DPG).

[16] Pfitzinger, H. R. (2006). Five dimensions of prosody: Intensity, intonation, timing, voice quality, and degree of reduction. In Hoffmann, R.; Mixdorff, H., eds., *Speech Prosody Abstract Book. Studientexte zur Sprachkommunikation*, vol. 40, pp. 6–9. TUDpress, Dresden.

[17] Pfitzinger, H. R.; Reichel, U. D. (2006). Delay compensation between the speech signal and the corresponding electroglottograph signal. In *Proc. of the AST Workshop*, pp. 63–74, Maribor; Slovenia.

[18] Swerts, M.; Veldhuis, R. (2001). The effect of speech melody on voice quality. *Speech Communication*, 33(4): 297–303.