Introducing a Japanese multi-talker database of laryngeal voice qualities[*]

○ Parham MOKHTARI, Daisuke MORIKAWA
（Toyama Prefectural University）

1 Introduction

Voice quality is an integral part of both verbal and nonverbal human interaction. As a baseline, a talker's habitual voice quality (i.e., *modal voice*) carries idiosyncratic information such as identity and age. Beyond this, long-term changes in voice quality can signal changes in health, mood or attitude; short-term changes, consciously or not, convey paralinguistic information, e.g. emotion.

Laver [1] presented a framework for describing and categorizing voice quality, based on auditory impressions and physiological considerations. In this broad view, voice quality is determined by laryngeal settings (e.g., *whisper*, *falsetto*, etc.), supralaryngeal settings (e.g., *spread lips*, *nasal voice*, etc.), and overall tension. The laryngeal component was reviewed and 7 typical examples shown of the glottal flow estimated by inverse filtering of the speech signal [2]. Examples of estimated glottal flow were also shown in the context of relating voice quality with vocal effort and emotions [3]. However, there does not appear to be any widely available multi-talker recording.

While emotion-related and speech synthesis technologies have advanced remarkably in the last two decades, the use of voice quality for more natural human-machine interfaces is still challenging. We believe that a carefully recorded database may help bridge the gap between theoretical descriptions and technical progress.

For example, even considering only laryngeal voice quality, it is not yet clear: (i) how vocal effort modulates different voice qualities; (ii) to what extent the typical voice quality descriptions are modified by individuality (talker differences); and in the first place, (iii) to what extent naive (not expert nor professional) talkers are able to produce the theoretical range of voice qualities.

To help answer these and related questions, here we introduce a new Japanese database of laryngeal voice quality recordings.

2 Speech Database

2.1 Participants & Recording Tools

For version 1 of the database, 16 participants (10 male, 6 female; ages 19 – 22) were recruited among our University's students. Each participant read an explanation of the experiment (which was approved by the University's ethics committee), signed a consent form, and was remunerated at the end of the session, which took about 1 hr.

Recordings took place in a soundproof room (3.24 × 3.58 × 2.30 m) with a background noise level of 16 dB. The talker stood, with their lips about 30 cm from a microphone (B&K 4190) secured on a single-pole tripod stand placed near the room's center. The microphone cable was fed to a conditioning amplifier (B&K Nexus) set to unity gain. The conditioned signal was fed to an analog-to-digital converter (RME ADI-2 Pro) set to 44.1 kHz sampling rate and 24 bits/sample, which in turn was connected via USB to a notebook PC (Dell XPS 13 2-in-1 9310) with a solid-state hard-disk. The digitized signal was recorded and saved in mono WAV format.

The first author was also standing in the room, at least 1 m diagonally behind the microphone, to instruct each participant as described next.

2.2 Recording Protocol & Speech Materials

Each talker was encouraged to relax and read as naturally as possible, text presented on a large card held up at eye level by the instructor. Starting with their normal, habitual voice (*modal voice*), the talker read the Japanese word /hai/, sustained vowels /a:/, /i:/, /ɯ:/, /e:/, /o:/, and vowel sequence /aiɯeo/ with normal vocal effort,

[*] 喉頭声質の日本語マルチトーカーデータベースの紹介
モクタリ パーハム, 森川 大輔 （富山県立大学）

Table 1 Voice qualities that participants were asked to produce, and their laryngeal settings. AT: adductive tension, MC: medial compression, LT: longitudinal tension, p: passive.

	AT	MC	LT
Modal Voice (地声)			
Whisper (ささやき声)	↓	↑	
Whispery Voice (ささやくような声)	↓	↑	
Falsetto (裏声)	↑	↑	↑p
Creak (きしる声)	↑	↑	↓
Creaky Voice (きしるような声)	↑	↑	↓
Breathy Voice (気息性の声)	↓	↓	↓
Tense Voice (緊張した声)	↑	↑	
Harsh Voice (ざらざらした声)	⇈	⇈	

as if reading to the instructor. Next, the same material was read with lower vocal effort (softly) as if reading to someone standing close. Next, the same material was read with higher vocal effort (loudly) as if reading to someone far away.

Next, at a normal vocal effort, the talker read a paragraph of Japanese text (taking about 1 min.): a version of the fable "The North Wind and the Sun" (北風と太陽) that had been modified [4] to include all the Japanese consonant phonemes and their major allophones.

The entire protocol described above was then repeated for each of the remaining 8 laryngeal voice qualities listed in Table 1, including an overall *tense voice*. These were chosen from Laver's [1] framework to cover a wide range of physiological possibilities while keeping to relatively simple types, i.e., avoiding many compound types. Prior to each one, the instructor briefly trained the talker with simple explanations and live demonstrations, encouraging the talker to practice imitating the instructor's voice quality as far as they were able, in their own way.

Regarding phonetic realization and prosody including vocal effort, no specific instructions were given other than those described above; each participant thus rendered the speech material with idiosyncratic timing, intonation, emphasis, and pronunciation. If a major mistake was made or a non-speech event such as coughing occurred during the reading, the paragraph would be restarted; otherwise, the fable recordings do occasionally contain some minor disfluencies.

2.3 Segmentation & Post-Processing

For each recording session, the 22 segments (7 isolated utterances × 3 vocal efforts + the fable) in each of the 9 voice qualities, were demarcated and saved as separate WAV files. Segmentation was done manually with the aid of the speech waveform, its spectrogram, and by audition.

Finally, the waveform of every segment was negated to correct for the negative polarity of the microphone, then high-pass filtered at cut-off 40 Hz with a linear-phase FIR filter (2482 taps = 56 ms) to suppress low-frequency ambient noise.

3 Voice Qualities Intended vs Produced

Despite having chosen a subset of 9 basic voice qualities, as expected of naive participants, not every talker was able to successfully produce every voice quality. Some talkers had particular difficulty with *creak* which was sometimes produced as modal voice with low fundamental frequency, and *harsh voice* which was sometimes produced as tense laryngo-pharyngalized whispery voice. Nevertheless, every talker's best attempts are included in the database. To clarify the difference between intention and production, each recorded segment was listened to carefully and labelled in terms of the voice quality actually produced. This auditory evaluation will be made available as part of the database.

4 Conclusions

Our motivations and methods for recording a new Japanese database of laryngeal voice quality were described. The database version 1 will soon be made available online. Future versions may include more talkers and new materials. We hope it may be useful for the research community.

Acknowledgements

This work was partially supported by a grant from the Casio Science Promotion Foundation.

References

[1] Laver, *The phonetic description of voice quality*, Cambridge University Press, 1980.

[2] Kasuya & Yang, JASJ **51**(11), 869-875, 1995.

[3] Klasmeyer & Sendlmeier, in *Voice Quality Meas.* (Kent & Ball, eds.), 339-357, 2000.

[4] Hiki *et al.*, Proc. ICPhS, 871-873, 2011.