51.

Laryngeal voice quality conversion by glottal waveshape PCA

Parham Mokhtari, Hartmut Pfitzinger, Carlos Toshinori Ishi, and Nick Campbell
JST/CREST-ESP Project, HIS Labs, ATR, Keihanna Science City, Japan
parham@atr.jp

Abstract

This paper describes a new approach to automatic conversion of laryngeal voice quality. A speaker's voice-quality space is first modelled by principal components analysis (PCA) of a large variety of carefully-measured glottal pulses. Voice quality conversion of a new utterance is then achieved by warping each glottal pulse-shape along principal dimensions.

1 Motivations

It is generally accepted that *voice quality* (Laver, 1980) plays an important role in conveying a speaker's emotions, moods, attitudes, and other personal characteristics. For example in everyday speech, common utterances are sometimes repeated with essentially the same linguistic content but with such differences in voice quality as to convey quite different meanings or nuances. Moreover, variation in voice quality is one of the key elements that distinguishes natural, human speech from even the most advanced systems of speech synthesis by computer.

Indeed, while the state-of-the-art in speech synthesis has achieved acceptable levels of segmental intelligibility and prosodic naturalness, there remains a lot to be desired by way of *expressive variety*. A critical role in that regard is played specifically by the *laryngeal* component of voice quality: in Laver's (1980) framework, voice qualities (such as breathy voice, creaky voice, harsh voice, and so on) can be regarded as arising from long-term *settings* of the laryngeal musculature. However, despite the recent surge of interest as displayed for example at last year's Geneva workshop on voice quality (cf. Mokhtari et al., 2003), there remains a relative paucity of both empirical data and methods to allow voice quality variation in speech synthesis.

Our motivations in this study are therefore to extend our earlier, empirically-guided quantification of laryngeal voice quality, and to show how such a model of a speaker's voice quality space may be used in an automatic conversion system that allows flexible control of voice quality, and thereby indirectly, of an important aspect of the expressivity of speech.

2 A Model of Voice Quality Space

As noted above, we have recently proposed a new method of quantifying a speaker's *voice quality space* (cf. also Mokhtari, 2003, for a more complete account). In particular, it is well-known that differences in voice quality arise at the production level by different modes or manners of laryngeal vibration, which at the acoustic level map to differences in the characteristics of the *glottal airflow*. In contrast with previous models of the glottal airflow that are controlled by salient but pre-defined parameters (e.g., the LF model of Fant et al., 1985), we proposed an empirically-defined model whose parameters are the first few (e.g., four) components of a principal component analysis (PCA) of a time-versus-amplitude representation of single glottal-airflow pulses, measured by inverse-filtering of speech recorded in a variety of different voice qualities by a given speaker.

In our previous work we avoided many of the recurrent difficulties of voice quality labelling and inverse-filtering, not only by restricting measurements to the more "reliable centres" in the speech stream (Mokhtari & Campbell, 2002), but also by using Laver's (1980) recordings which are prototypical examples of a distinct set of voice qualities. By contrast, in this study we apply the same methods to derive a model of the voice quality space of a speaker (adult, female, native speaker of Japanese) whose recordings include a far more natural range of qualities used in everyday interactions with family members, friends, company employees, and so on (Campbell & Mokhtari, 2003).

Naturally, the sheer amount of data (comprising more than 7 hours of speech) precludes manual intervention and calls for completely automatic analyses. On the other hand, as such spontaneous data present a great challenge for robust analysis, it is rarely attempted in voice quality research. Nevertheless, while acknowledging that a proportion of our automatically-measured glottal pulses will probably contain analysis artefacts (such as formant ripples), it is hoped that the large number of pulses thereby obtained will enable more reliable measurement of pulse-shape statistics than would otherwise be possible using a smaller database. Moreover, such a large database allows us to check the consistency of our analyses, by comparing results obtained separately for each of four nonoverlapping subsets of the data.

Results obtained for the first subset, comprising about 100 minutes of speech and yielding 19,665 reliable centres, are shown in Fig. 1. Although each pulse was characterised by 60 parameters (30 equally-spaced time-amplitude pairs), the top *four* principal components (PCs) together account for 74.5% of the original variance. The first two PCs, accounting for 40.9% and 19.8% of the variance respectively, show roughly

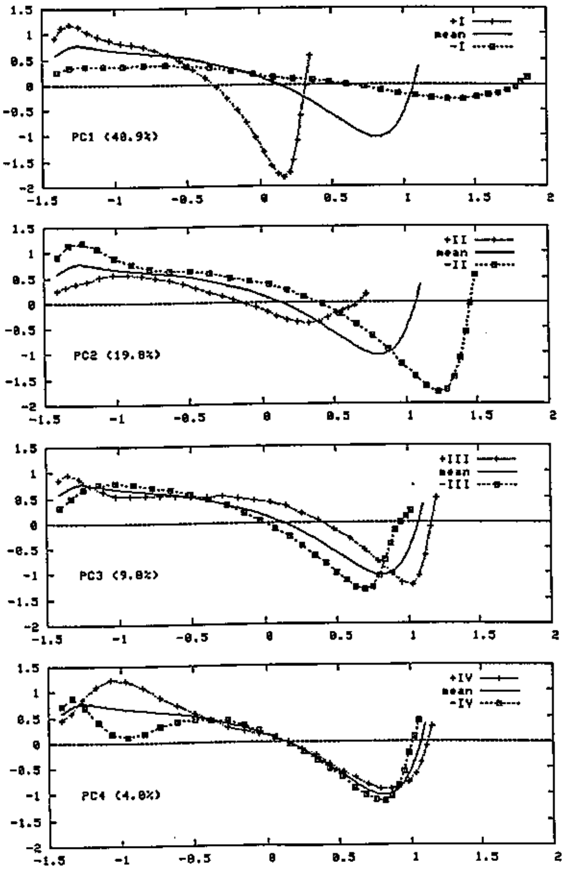

Figure 1. Mean and ±1σ in each of the first 4 principal components of the *glottal-flow derivative*, by PCA of 19,665 glottal cycles measured automatically in one female speaker's spontaneous speech data. Abscissae: standardised units of time; ordinates: standardised units of flow-amplitude per unit time.

声帯音源波形形状の主成分分析による声質変換
○ パーハム・モクタリ、ハートムット・フィッツィンガー、石井カルロス寿憲、ニック・キャンベル
JST/CREST-ESP Project, ATR-HIS Labs, Kyoto.

orthogonal effects of co-variations in the fundamental frequency F0 (inversely proportional to the pulse duration) and the strength of the main excitation (the negative peak of the flow-derivative). The third PC accounts for 9.8% of the variance, and appears to model a left-to-right skewing of the glottal pulse. The fourth PC explains 4.0% of the variance, and accounts for changes in the slope of the opening-phase. These PCs were found to be roughly in agreement with those obtained for two of the remaining three subsets of the same speaker's data, thus reasonably supporting the consistency of the results yielded by our automatic methods of analysis.

However, as hypothesised in previous work (Mokhtari, 2003), these underlying basis-functions are quantitatively different to those reported earlier. Indeed, differences between the two sets of results are certainly not surprising, given the widely different nature of the data: male vs. female, controlled vs. spontaneous, English vs. Japanese, and 77 carefully-selected vs. 19,665 automatically-measured glottal pulses. In fact, any similarities at all can be considered significant: these are to be found qualitatively in the modelling of variations in F0, strength of excitation, skewness, and opening-phase, all within the top four principal components.

3 Voice Quality Conversion

The principal components described above, can together be regarded as a four-parameter model allowing control of the *underlying dimensions of natural variability* in the shape and duration of the glottal-airflow derivative. In other words, these four parameters can be thought to span the space of laryngeal voice qualities measured in our speaker's data. We therefore propose an automatic system that is able to convert the voice quality of any of the speaker's utterances, according to any desired transformation in the four-parameter voice quality space.

The conversion system first locates boundaries of syllable nuclei in a given utterance. Within each nucleus, the first 4 formant frequencies and bandwidths are estimated with the help of a linear transformation of the cepstrum (Broad & Clermont, 1989), and the speech signal is inverse-filtered to remove the formant structure and thereby retain an estimate of the glottal-flow derivative (which includes the effects of both laryngeal source and lip-radiation). Boundaries of consecutive glottal pulses are then located by advancing forward to the first upward zero-crossing, from local minima approximately one period-length apart; detection errors near syllable boundaries are reduced by processing each syllable nucleus from the centre outward (Lea & Clermont, 1984).

Each glottal pulse is then warped (simultaneously in time and amplitude) by a fixed *perturbation function* that represents the desired transformation in the four-parameter voice quality space. This perturbation function is computed simply by subtracting the *prototype* pulse-shape corresponding to the main voice quality of the utterance, from a *target* pulse-shape specified interactively on a computer display. As both the prototype and target pulses are represented by 30 equi-distant time-amplitude pairs, the resulting perturbation function must be resampled at a new number of points before being applied to each of the glottal pulses measured in the utterance; furthermore, after warping, each pulse must be resampled at regular time-intervals corresponding to the original sampling frequency. Thus, to avoid aliasing, we take care to apply appropriate low-pass filters prior to resampling. Finally, each syllable nucleus is synthesised by re-imposing the original formants on the warped glottal waveform, and the entire utterance is made by concatenating the modified syllable nuclei with their adjacent (unmodified) segments.

As our female Japanese speaker's data has not yet been labelled for voice quality, we turn to Laver's (1980) data which we used in our previous studies, and for which we already have prototype glottal pulses for 13 voice qualities (Mokhtari, 2003). The top panel in Fig. 2 shows a portion of the inverse-filtered signal corresponding to the vowel /a/ in the final syllable of the example in *modal* voice. The next three panels show the same portion of glottal-flow derivative signal after automatic warping of each pulse according to perturbation functions aimed at converting the speaker's voice quality from modal to *breathy voice*, *creaky voice*, and *falsetto*, respectively. From these visual examples it is clear that in all three cases the fundamental frequency has been modified in accord with expectations. While it is more complicated to visually confirm modifications in the details of the waveshapes, informal auditory judgments on these and other preliminary results using the female speaker's data are encouraging.

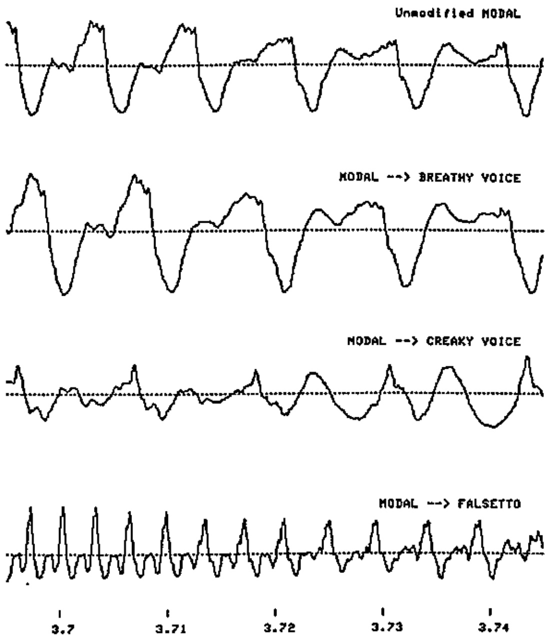

Figure 2. Three examples of laryngeal voice conversion, using Laver's (1980) data. Top: glottal-flow derivative in a portion of the original *modal* voice. Below: the same portion warped towards the target for *breathy voice*, *creaky voice*, and *falsetto*, resepectively. Abscissae: time in secs; ordinates: arbitrary units of flow-amplitude per unit time, shown on the same scale.

4 Conclusions

We have first extended our earlier results by computing a new four-parameter model of the laryngeal voice quality space of a female Japanese speaker's spontaneous speech, then outlined a method of voice quality conversion by analysis, transformation of glottal pulse-shapes, and resynthesis. We are currently evaluating the conversion results and increasing the robustness of the analysis algorithms. We also intend to use the four PCs as a compact acoustic labelling of voice quality, useful in selection of *affectively* appropriate units in concatenative synthesis of expressive speech.

Acknowledgements

We thank the Japan Science and Technology (JST) Corporation, and all members of the JST-CREST ESP project for their support and encouragement.

References

Broad, D.J. & Clermont, F. (1989). "Formant estimation by linear transformation of the LPC cepstrum", *J. Acoust. Soc. Am* 86 (5), 2013-2017.

Campbell, N. & Mokhtari, P. (2003). "Voice quality: the 4th prosodic dimension", in *Proc. 15th Int. Congress of Phonetic Sciences*, Barcelona, Spain, 2417-2420.

Fant, G., Liljencrants, J. & Lin, Q. (1985). "A four-parameter model of glottal flow", *STL-QPSR* (KTH) No. 4, 1-13.

Laver, J. (1980). *The phonetic description of voice quality*, Cambridge Univ. Press.

Lea, W.A. & Clermont, F. (1984). "Algorithms for acoustic prosodic analysis", in *Proc. IEEE-ICASSP*, 42.7.1-4.

Mokhtari, P. (2003). "Parameterisation and Control of Laryngeal Voice Quality by Principal Components of Glottal Waveforms", *J. Phonetic Soc. Japan* (音声研究), Vol. 7, No. 3.

Mokhtari, P. & Campbell, N. (2002). "Automatic detection of acoustic centres of reliability for tagging paralinguistic information in expressive speech", in *Proc. 3rd Int. Conf. Lang. Resources and Evaluation*, Las Palmas, Spain, 2015-2018.

Mokhtari, P., Pfitzinger, H.R. & Ishi, C.T. (2003). "Principal components of glottal waveforms: towards parameterisation and manipulation of laryngeal voice-quality", in *Proc. ISCA Tut. Res. Workshop on "Voice Quality: Functions, Analysis, and Synthesis"* (VOQUAL'03), Geneva, Switzerland, 133-138.