Source-Filter Separation for Articulation-to-Speech Synthesis

Yoshinori Shiga, Simon King

Centre for Speech Technology Research, University of Edinburgh
2 Buccleuch Place, Edinburgh EH8 9LW, U.K.

yoshi@cstr.ed.ac.uk

Abstract

In this paper we examine a method for separating out the vocal-tract filter response from the voice source characteristic using a large articulatory database. The method realises such separation for voiced speech using an iterative approximation procedure under the assumption that the speech production process is a linear system composed of a voice source and a vocal-tract filter, and that each of the components is controlled independently by different sets of factors. Experimental results show that the spectral variation is evidently influenced by the fundamental frequency or the power of speech, and that the tendency of the variation may be related closely to speaker identity. The method enables independent control over the voice source characteristic in our articulation-to-speech synthesis.

1. Introduction

Kaburagi et al. [1] first reported a technique to synthesise speech from articulator positions based on the search of a database composed of pairs of articulatory and acoustic data. For elucidating the speech production mechanism, such an approach is considered an alternative to acoustically simulated vocal-tract modelling which has been widely investigated (e.g. [2]). In [1], the capability of their method for producing intelligible speech is demonstrated by employing LSP and multipulse excitation; however, speech synthesised by their method has many artefacts and the speech quality is not sufficiently high.

As a main cause of degradation in speech quality, we point out that the method searches the articulatory-acoustic database based only on articulator positions. Clearly, such a search method causes temporal discontinuities in acoustic parameters if speech has a tendency to change spectrum depending on factors other than the articulator positions, such as the fundamental frequency (F_0) and speech power. There have actually been many reports, e.g. [3], that variation in the F_0 and power of speech mainly affects the glottal source signal and consequently influences speech signal. It is therefore essential for speech synthesis to modulate output speech based on these factors in addition to the articulatory settings.

We have studied a similar approach to Kaburagi's for converting articulation into speech based on an articulatory-acoustic mapping obtained from an articulatory database [4, 5]. In order to deal with the above problem, we have been examining a method for separating out the variation of the source [6]. In our method, the separation is achieved using an iterative approximation procedure under the assumption that the speech production process is a linear system where the voice source and vocal tract are cascaded, and that each of the components is independently controlled by different sets of factors.

This paper reports in detail the results of applying the separation to two different speech corpora, from one female speaker and one male speaker.

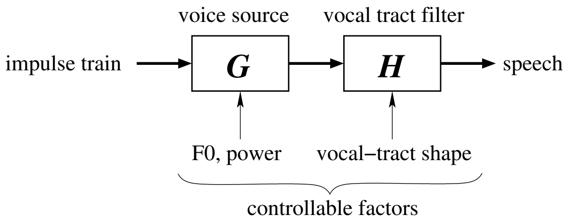


Figure 1: *Speech production model*

2. Source-filter separation

2.1. Outline

From the viewpoint of filter design in signal processing, where both the input and output of a system are observed to find the transfer function of the system, it is theoretically difficult to estimate the characteristics of the input (voice source) and system (vocal tract) simultaneously from the output (speech), which is the only observable signal.

However, variation in the transfer function of one component can be approximately separated if speech production can be modelled as a linear system composed of cascaded components, as in figure 1, and the transfer function of each component is controlled by a set of factors which is uncorrelated with those controlling the other component. The separation can be achieved by iterative approximation using a large corpus with the controlling factors of both components well represented. We accordingly apply the following assumptions to the source-filter separation:

1. The speech production process is modelled by a linear system composed of two cascaded components: voice source and vocal tract filter.
2. The voice source changes depending only on the F_0 and cepstral coefficient c_0.
3. The vocal-tract filter response changes depending only on the articulator positions.

2.2. Piecewise constant approximation

The source and filter characteristics are approximated locally by a constant value. For such a piecewise constant approximation, the following two types of clustering are applied to the same corpus:

- Based on the articulatory data, all the voiced frames are divided into K clusters (articulatory clusters) C_h^i (i = 1, 2, 3, ..., K), so that each of the clusters consists of frames with similar articulatory positions.
- Based on their F_0 and c_0 values, all the voiced frames are divided into L clusters (source clusters) C_g^j (j = 1, 2, 3, ..., L), so that each of the clusters consists of frames with similar F_0 and c_0 values.

LBG clustering [7] is adopted to identify frames with similar values for a particular controlling factor.

The spectral envelope estimation in this study puts emphasis on harmonic peaks in the spectrum of voiced speech in the same manner as some methods [8, 9, 10] successful in speech technology. In addition, the method inhibits an adverse effect of harmonic structure on the spectral envelope estimation by using the spectra of multiple speech frames vocalised with similar articulator settings, and consequently is able to estimate very detailed spectral envelopes (see [5] for details).

2.3. Iterative estimation procedure

The proposed method alternatively discovers the complex cepstra $\boldsymbol{c}_h^{(i)}$ and $\boldsymbol{c}_g^{(j)}$, which represent the frequency characteristics of vocal tract and voice source, according to the following iterative procedure. Here we define $\boldsymbol{h}_k$, a harmonic vector of frame k (k = 1, 2, 3, ..., M), as follows:

$$\boldsymbol{h}_k = \left[h_k^{(-N_k)}\ h_k^{(-N_k+1)}\ h_k^{(-N_k+2)}\ \cdots\ h_k^{(N_k-1)}\ h_k^{(N_k)}\right]^T$$

where $h_k^{(l)}$ represents the natural logarithm of the observed complex spectrum of the l-th harmonic at frequency $f_k^{(l)}$ in analysis frame k, and N_k indicates the number of harmonics in frame k.

Step 1: For each articulatory cluster C_h^i, the cepstrum $\boldsymbol{c}_h^{(i)}$ is calculated by applying the spectral envelope estimation to the harmonic vectors $\{\boldsymbol{h}_k|\text{frame}\,k \in C_h^i\}$. (the first approximation)

Step 2: The procedure is terminated if the following E, the sum of squared approximation errors, converges:

$$E = \sum_{k=1}^{M} \boldsymbol{\delta}_k^H \mathbf{W}_k \boldsymbol{\delta}_k \tag{1}$$

where ^{H} denotes Hermite transpose operation and $\mathbf{W}_k$ is a weighting matrix (see [5] for details). The vector $\boldsymbol{\delta}_k$ is defined as

$$\boldsymbol{\delta}_k = \boldsymbol{h}_k - \boldsymbol{b}_k - \mathbf{B}_k\left[\boldsymbol{c}_h^{(R_h(k))} + \boldsymbol{c}_g^{(R_g(k))}\right]$$

$$i = R_h(k) \Longleftrightarrow \text{frame } k \in C_h^i$$

$$j = R_g(k) \Longleftrightarrow \text{frame } k \in C_g^j$$

where

$$\boldsymbol{b}_k = \begin{bmatrix} d_k + j2\pi\tau_k f_k^{(-N_k)} \\ d_k + j2\pi\tau_k f_k^{(-N_k+1)} \\ \vdots \\ d_k + j2\pi\tau_k f_k^{(N_k)} \end{bmatrix} \tag{2}$$

$$\mathbf{B}_k = \begin{bmatrix} e^{-j(-p)\Omega_k^{(-N_k)}} & \cdots & e^{-jp\Omega_k^{(-N_k)}} \\ \vdots & \ddots & \vdots \\ e^{-j(-p)\Omega_k^{(N_k)}} & \cdots & e^{-jp\Omega_k^{(N_k)}} \end{bmatrix}$$

$$\Omega_k^{(l)} = 2\pi f_k^{(l)} T.$$

The power offset d_k and time delay τ_k in (2) are obtained during the spectral envelope estimation [5].

Table 1: *Data sets used in the experiments*

corpus	number of frames		
	train	test	total
fsew0 (female speaker)	78876	8332	87208
msak0 (male speaker)	65859	6807	72666

Step 3: For all the harmonics, the difference between the observed harmonics $\boldsymbol{h}_k$ and the harmonics calculated from $\boldsymbol{c}_h^{(i)}$ is obtained as follows:

$$\boldsymbol{p}_k = \boldsymbol{h}_k - \mathbf{B}_k \boldsymbol{c}_h^{(R_h(k))}.$$

Thereby the residual $\boldsymbol{p}_k$ reflects the variation of the source characteristic.

Step 4: For each source cluster C_g^j, the cepstrum $\boldsymbol{c}_g^{(j)}$ is calculated by applying the spectral envelope estimation to $\{\boldsymbol{p}_k|\text{frame}\ k \in C_g^j\}$.

Step 5: For all the harmonics, the difference between the observed harmonics $\boldsymbol{h}_k$ and the harmonics calculated from $\boldsymbol{c}_g^{(j)}$ is obtained as follows:

$$\boldsymbol{q}_k = \boldsymbol{h}_k - \mathbf{B}_k \boldsymbol{c}_g^{(R_g(k))}.$$

Step 6: For each articulatory cluster C_h^i, the cepstrum $\boldsymbol{c}_h^{(i)}$ is calculated by applying the spectral envelope estimation to $\{\boldsymbol{q}_k|\text{frame}\ k \in C_h^i\}$.

Step 7: Return to step 2.

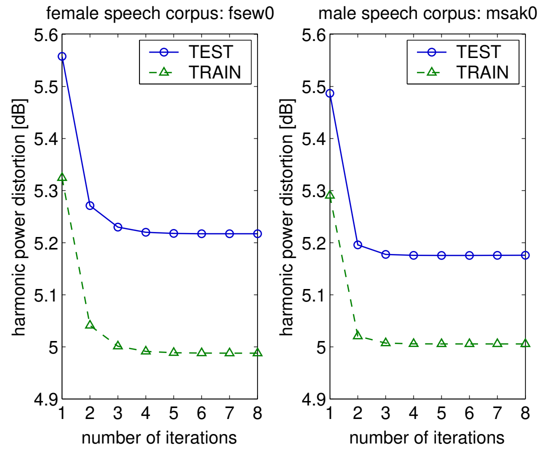
3. Experiments

3.1. Data and procedure

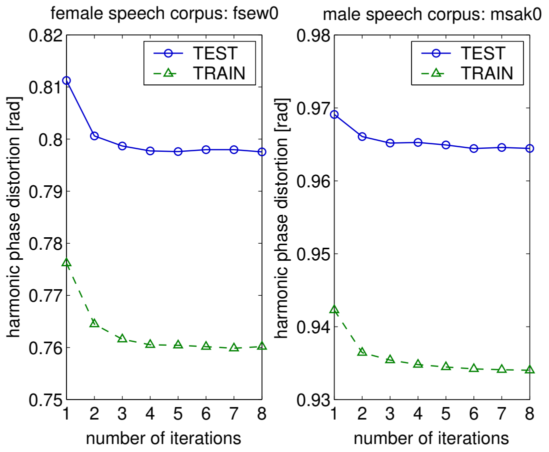
The data used in this study are two speakers from the MOCHA (Multi-CHannel Articulatory) set of corpora [11]: a female speaker (fsew0) and a male speaker (msak0). Each of the corpora is composed of 460 TIMIT sentences from a single speaker, and includes parallel acoustic-articulatory information which was recorded using a Carstens Electromagnetic Articulograph (EMA) system at Queen Margaret University College, Edinburgh. The articulatory information comprises the positions of the upper and lower lips, lower incisor, tongue tip, tongue blade, tongue dorsum and velum. The sampling rates of the acoustic waveform and articulatory trajectories are 16 kHz and 0.5 kHz respectively.

Voiced sections were first extracted from the corpus and used to build a set of pairs of harmonic spectra and articulator positions. We estimated the harmonic spectra from the speech waveform using the weighted least squares method in [12]. The width and spacing of the time window (Hanning) were 20 ms and 8 ms respectively. We downsampled the articulatory information to the same spacing of 8 ms. Out of the data obtained, we set 10% of the sentences (46 sentences) aside for testing, and used the remaining 90% (414 sentences) for training. Details of the data sets are given in table 1.

All the voiced frames were divided into 512 articulatory clusters (K = 512) and 128 source clusters (L = 128) using LBG clustering. The order of cepstrum was set to 48 for the vocal tract characteristic, and 32 for the voice source characteristic. These numbers were established from the results of preliminary experiments. Finally, according to the procedure in section 2.3, iterative approximation was performed to find the complex cepstra, $\boldsymbol{c}_h^{(i)}$ and $\boldsymbol{c}_g^{(j)}$, for each articulatory and source cluster.



(a) harmonic power distortion



(b) harmonic phase distortion

Figure 2: *Number of iterations vs. harmonic distortion*

In order to evaluate estimation accuracy, we introduced two types of distortions, *harmonic power distortion* D_a and *harmonic phase distortion* D_p, defined as

$$D_a = \frac{20}{\ln 10}\sqrt{\frac{1}{M}\sum_{k=1}^{M}\boldsymbol{\delta}_{Rk}^T \mathbf{W}_k \boldsymbol{\delta}_{Rk}}$$

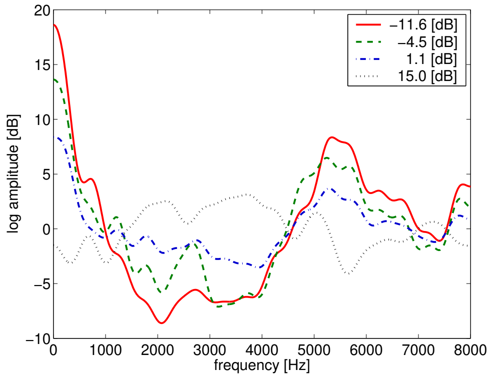
$$D_p = \sqrt{\frac{1}{M}\sum_{k=1}^{M}\boldsymbol{\delta}_{Ik}^T \mathbf{W}_k \boldsymbol{\delta}_{Ik}}$$

where $\boldsymbol{\delta}_{Rk}$ is a vector each of whose elements is the real part of the corresponding element of $\boldsymbol{\delta}_k$ in equation (1), and $\boldsymbol{\delta}_{Ik}$ is a vector each of whose elements is the imaginary part of the corresponding element of $\boldsymbol{\delta}_k$. Both of the distortions were computed in step 2 of the procedure in section 2.3.

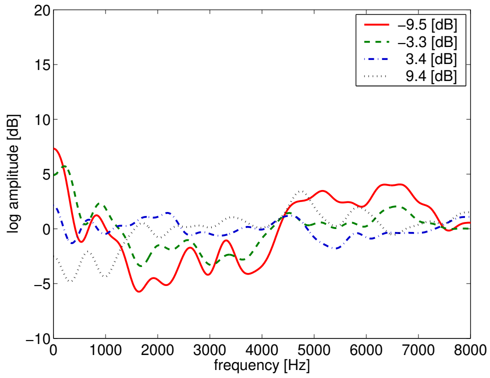
3.2. Results and discussion

Figure 2(a) shows the relationship between the number of iterations and harmonic power distortion. Figure 2(b) shows the relationship between the number of iterations and harmonic phase distortion. As is evident from these graphs, these distortions decrease as the process is iterated, for both power and phase.

Shown in figure 3 is the estimated variation in the power spectrum of the voice source depending on the c_0 value. In this figure,

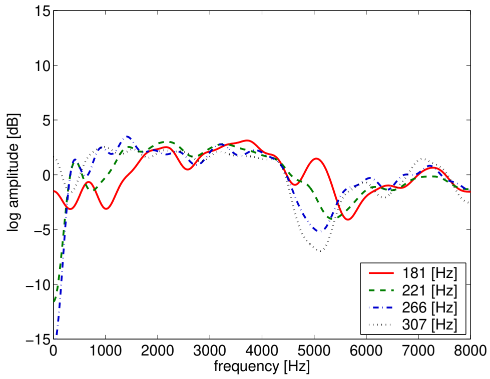


(a) female speech corpus: fsew0

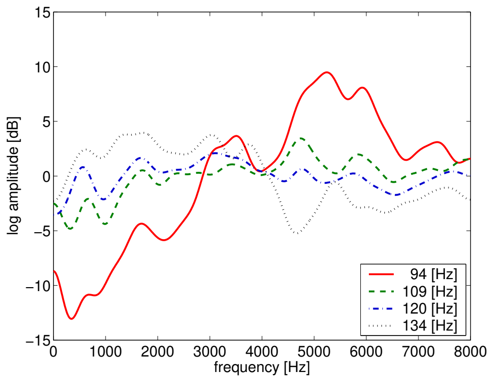


(b) male speech corpus: msak0

Figure 3: *Variation in the source characteristics depending on* c_0



(a) female speech corpus: fsew0



(b) male speech corpus: msak0

Figure 4: *Variation in the source characteristics depending on* F_0

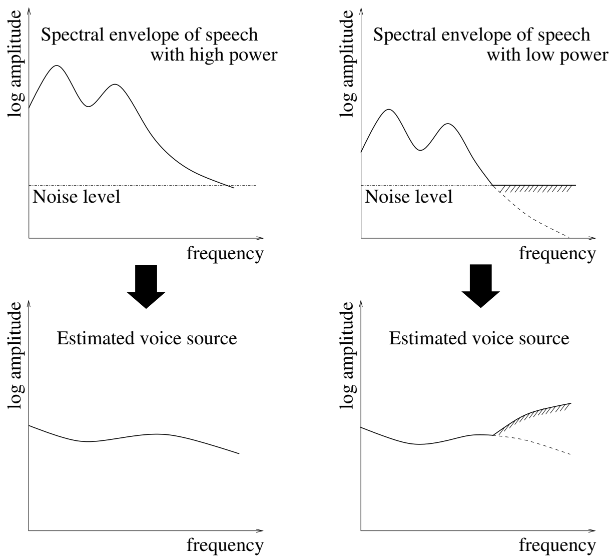


Figure 5: *Detected noise-level in the high frequency band*

c_0 is expressed using relative power in dB. The female voice has larger spectral variation than the male voice, and the lowering of c_0 (i.e. speech power) increases relative power in the low frequency range around F_0 and in the high frequency range above 4 kHz. The former increasing tendency in the low frequency range is very much in agreement with reports that the glottal waveform becomes more sinusoidal in the case of low voice power [3].

We think that the increase in power above 4 kHz indicates a relative rise of the noise level. Since the speech spectrum is generally inclined at 6 dB/oct, as in figure 5, the spectrum in the high frequency range becomes buried under the noise level, as the speech power decreases. The noise in the high frequency band is accordingly detected as spectral change caused by the lowering of coefficient c_0. As is obvious from these results, we must therefore be aware that what is obtained by our method is not the actual voice source characteristic, but the spectral variation due to c_0 and/or F_0.

Figure 4 shows the estimated variations in the power spectrum of the voice source depending on the F_0 value. Contrary to the result in figure 3, the male voice has larger spectral variation than the female voice in this case. Thus the spectral variation caused by F_0 or c_0 differs across speakers, and we consider that the tendency of the variation is closely related to speaker identity. To clarify this we need to accumulate more analysis results for other speakers and to investigate how much those spectral changes influence human auditory perception.

4. Conclusions

We investigated a method for separating out the variations in speech spectra due to the source characteristic from those due to the filter response, based on an iterative approximation procedure. The experimental result showed that the spectral variation was influenced by F_0 or c_0, and suggests that the tendency of the variation is closely related to speaker identity.

The proposed method statistically discovers variation in the voice source characteristic from a large articulatory corpus, and enables the independent control of the voice source characteristic in our articulation-to-speech conversion [5]. We have informally confirmed that intelligible, high-quality speech can be generated by sinusoidal synthesis [13] using harmonics reproduced from the mapping obtained.

As we have already reported, introducing a piecewise linear function for the articulatory-acoustic mapping in each cluster enables better approximation with a smaller number of clusters [5]. Further improvement in the estimation accuracy is expected in combination with the source-filter separation we have discussed in this paper. Moreover, we intend to apply this separation technique to text-to-speech synthesis by replacing the articulatory clustering with one based on phonetic context (or on the types of synthesis units) instead of the articulator positions measured by the EMA system.

Acknowledgements

In carrying out this research, the first author, Y. Shiga, is supported financially in part by the ORS Awards Scheme.

References

[1] T. Kaburagi and M. Honda, "Determination of the vocal tract spectrum from the articulatory movements based on the search of an articulatory-acoustic database," in *Proc. ICSLP98*, 1998, pp. 433–436.

[2] T. Yokoyama, N. Miki, and Y. Ogawa, "An interactive construction system of 3-D vocal tract shapes from tomograms," in *Proc. the 16th International Conference on Acoustics and 135th Meeting of the Acoustical Society of America*, vol. II, Seattle, USA., 1998, p. 1283.

[3] R. L. Miller, "Nature of the vocal cord wave," *J. Acoust. Soc. Am.*, vol. 31, no. 6, p. 667, 1959.

[4] Y. Shiga and S. King, "Estimating the spectral envelope of voiced speech using multi-frame analysis," in *Proc. Eurospeech2003*, vol. 3, Geneva, Switzerland, Sept. 2003, pp. 1737–1740.

[5] ——, "Accurate spectral envelope estimation for articulation-to-speech synthesis," in *Proc. 5th ISCA Speech Synthesis Workshop*, CMU, Pittsburgh, USA, June 2004, pp. 19–24.

[6] ——, "Estimation of voice source and vocal tract characteristics based on multi-frame analysis," in *Proc. Eurospeech2003*, vol. 3, Geneva, Switzerland, Sept. 2003, pp. 1749–1752.

[7] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.*, vol. COM-28, pp. 84–95, 1980.

[8] T. Galas and X. Rodet, "An improved cepstral method for deconvolution of source-filter systems with discrete spectra: Application to musical sounds," in *Proc. Int. Computer Music Conf.*, 1990, pp. 82–84.

[9] R. J. McAulay and T. F. Quatieri, "The application of subband coding to improve quality and robustness of the sinusoidal transform coder," in *Proc. ICASSP93*, vol. 2, Apr. 1993, pp. 439–442.

[10] L. Gu and K. Rose, "Perceptual harmonic cepstral coefficients as the front-end for speech recognition," in *Proc. ICSLP2000*, vol. 1, Oct. 2000, pp. 309–312.

[11] A. A. Wrench, "A new resource for production modelling in speech technology," in *Proc. Workshop on Innovations in Speech Processing*, Stratford-upon-Avon, 2001.

[12] Y. Stylianou, "Applying the harmonic plus noise model in concatenative speech synthesis," *IEEE Trans. Speech and Audio Processing*, vol. 9, no. 1, pp. 21–29, Jan. 2001.

[13] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. ASSP*, vol. 34, no. 4, pp. 744–754, Aug. 1986.