Estimation of Voice Source and Vocal Tract Characteristics Based on Multi-frame Analysis

Yoshinori Shiga and Simon King

Centre for Speech Technology Research
University of Edinburgh, Edinburgh, U.K.
yoshi@cstr.ed.ac.uk

Abstract

This paper presents a new approach for estimating voice source and vocal tract filter characteristics of voiced speech. When it is required to know the transfer function of a system in signal processing, the input and output of the system are experimentally observed and used to calculate the function. However, in the case of source-filter separation we deal with in this paper, only the output (speech) is observed and the characteristics of the system (vocal tract) and the input (voice source) must simultaneously be estimated. Hence the estimate becomes extremely difficult, and it is usually solved approximately using oversimplified models. We demonstrate that these characteristics are separable under the assumption that they are independently controlled by different factors. The separation is realised using an iterative approximation along with the Multi-frame Analysis method, which we have proposed to find spectral envelopes of voiced speech with minimum interference of the harmonic structure.

1. Introduction

Various reports have been given on simultaneous estimation of the characteristics of voice source and vocal tract. The approaches in those reports can broadly be divided into two types. In one type of approach, approximating the source waveform using a rather simple model, the methods estimate a small number of model parameters that determine the shape of the source waveform and parameters that express the vocal tract transfer characteristic[1][2]. In the other type of approach, approximating the vocal tract characteristic using a rather simple model, the methods estimate the source waveform by filtering the speech signal through the inverse of the vocal tract characteristic[3][4]. In either case, one of the characteristics is simplistically modelled and, under the restriction of the model, the other characteristic is found. From the viewpoint of filter design in acoustics where both the input and output of a system are observed to find the transfer characteristic of the system, it seems almost impossible to estimate the characteristics of the input (the voice source) and the system (the vocal tract) simultaneously only from the output (speech). Hence conventional methods must rely on approximation using oversimplified models.

In this paper, we deal with the problem of simultaneous estimation of voice source and vocal tract characteristics. We demonstrate how these two characteristics are separated using the assumption that they are controlled separately by different factors. In order to achieve this separation, we employ an iterative method along with *Multi-frame Analysis* (MFA)[5], which we have proposed to estimate spectral envelopes of voiced sound with minimum interference of the harmonic structure.

The paper is organised as follows: In section 2, our MFA method is briefly explained, following which the simultaneous estimation of source and filter characteristics is described in section 3. After an experiment is described in section 4, we discuss its results in section 5.

2. Multi-frame analysis (MFA)

2.1. Outline

In order to resolve the conventional problem that spectral envelope estimation is seriously affected by the harmonic structure of voiced sound[6], we employ many portions of speech vocalised using the same vocal tract shape. Each of the portions having usually different F_0, we can obtain a sufficient number of harmonics at various frequencies to form a spectral envelope which corresponds to the vocal tract transfer function. The envelope is estimated by approximating all the harmonic spectra of all the portions using a cepstrum based on a least-squares criterion. Because of the use of multiple frames in spectral envelope analysis, we call this approach *Multi-frame Analysis* (MFA). For the above speech *portions*, we currently adopt *analysis frames*.

2.2. Least squares estimate of the spectral envelope

We employ the *cepstrum* as an expression of the spectral envelope for the purpose of smoothing and interpolating harmonic spectra of multiple frames. Let $X(e^{j\omega})$ denote the Fourier transform of speech waveform, and $c[n]$ its cepstrum. Then, the following relation holds:

$$\ln |X(e^{j\omega})| = \sum_{n=-\infty}^{\infty} c[n] \cos(\omega n) \tag{1}$$

Based on Eq.(1), we determine a cepstrum which best fits the amplitude of all the harmonics included in M speech frames using the least squares method. This can be considered the expansion of the cepstrum estimation in [7].

Let a_k^l and f_k^l denote, respectively, an observed log-amplitude and frequency of the l-th harmonic ($l = 1, 2, ..., N_k$) within the speech frame $k(= 1, 2, ..., M)$, and T the sampling period. Then, the total error ε, the sum of the squares of the approximation error for the amplitude of all the harmonics of all the frames, is expressed as:

$$\varepsilon = \sum_{k=1}^{M} \sum_{l=1}^{N_k} \frac{w(f_k^l)}{N_k} \left[a_k^l - d_k - c[0] - 2\sum_{n=1}^{p} c[n] \cos(\Omega_k^l n) \right]^2 \tag{2}$$

where

$$\Omega_k^l = 2\pi f_k^l T$$

In Eq.(2) we introduce a weighting function, $w(f)$ for attaching importance to the lower frequency band, and $(1/N_k)$ for evaluating each frame equally regardless of the number of harmonics. The factor d_k is an offset that adjusts the total speech amplitude of each frame so as to minimise the error ε. Eq.(2) can be solved by reducing it to a problem of weighted least squares. The normal equation is expressed in terms of vectors and matrices as:

$$\left(\sum_{k=1}^{M} \boldsymbol{B}_k^T \boldsymbol{W}_k \boldsymbol{B}_k\right) \boldsymbol{c} = \sum_{k=1}^{M} \boldsymbol{B}_k^T \boldsymbol{W}_k (\boldsymbol{a}_k - d_k \boldsymbol{u}_k) \qquad (3)$$

where the vector $\boldsymbol{c}$ indicates 0- to p-order cepstral coefficients,

$$\boldsymbol{c} = [\ c[0]\ c[1]\ c[2]\ \cdots\ c[p]\ ]^T$$

both $\boldsymbol{a}_k$ and $\boldsymbol{u}_k$ are N_k dimension vectors,

$$\boldsymbol{a}_k = [\ a_k^1\ \ a_k^2\ \ a_k^3\ \cdots\ a_k^{N_k}\ ]^T,\ \boldsymbol{u}_k = [\ 1\ \ 1\ \ 1\ \ \cdots\ \ 1\ ]^T$$

and

$$\boldsymbol{B}_k = \begin{bmatrix} 1 & 2\cos(\Omega_k^1) & 2\cos(2\Omega_k^1) & \cdots & 2\cos(p\Omega_k^1) \\ 1 & 2\cos(\Omega_k^2) & 2\cos(2\Omega_k^2) & \cdots & 2\cos(p\Omega_k^2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & 2\cos(\Omega_k^{N_k}) & 2\cos(2\Omega_k^{N_k}) & \cdots & 2\cos(p\Omega_k^{N_k}) \end{bmatrix}$$

$$\boldsymbol{W}_k = \frac{1}{N_k} \begin{bmatrix} w(f_k^1) & & & \\ & w(f_k^2) & & \mathbf{0} \\ & \mathbf{0} & \ddots & \\ & & & w(f_k^{N_k}) \end{bmatrix} \qquad (4)$$

By solving Eq.(3), the cepstrum coefficients $\boldsymbol{c}$ can be found. Using $\boldsymbol{c}$ obtained, the factor d_k is calculated as:

$$d_k = \frac{\boldsymbol{u}_k^T \boldsymbol{W}_k (\boldsymbol{a}_k - \boldsymbol{B}_k \boldsymbol{c})}{\boldsymbol{u}_k^T \boldsymbol{W}_k \boldsymbol{u}_k} \qquad (5)$$

From the above, we can obtain a cepstrum which best approximates all the harmonic amplitude spectra of all the M frames according to the following procedure: 1) Substitute $\mathbf{0}$ for $\boldsymbol{c}$ (initial value); 2) Obtain d_k using Eq.(5); 3) Calculate ε of Eq.(2), and terminate the procedure if ε converges; 4) Find $\boldsymbol{c}$ by solving Eq.(3); 5) Substitute 0 for $c[0]$ (amplitude normalization); 6) Return to 2.

3. Simultaneous estimation of source and filter responses

3.1. Outline

Fig.1 shows the linear model of speech production assumed in this study, which consists of two cascade components: a voice source $\boldsymbol{G}$ and a vocal tract filter $\boldsymbol{H}$. If the transfer function of each component is controlled by a factor/factors independent of that/those of the other component, the function of one component can be separated from that of the other by iterative approximation using a large corpus well-balanced with those factors of both the components.

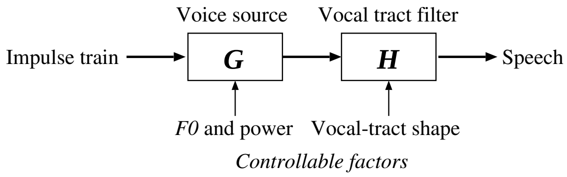

Figure 1: *Speech production model*

3.2. Controllable factors

The controllable factors must be observable. For the factor of the filter component, we can use data of articulators' positions measured using an electromagnetic articulograph (EMA) system[8]. For the factor of the source, we select F_0 and power of speech according to an early observation[9] that the waveform of the voice source varies depending on F_0 and power of speech. Thus we adopt the following two assumptions:

A. The filter frequency response $\boldsymbol{H}$ changes depending only on the vocal tract shape (or articulators' setting).

B. The source frequency response $\boldsymbol{G}$ changes depending only on the speech F_0 and power.

3.3. Data clustering

LBG clustering[10] is employed to identify frames with similar value for a particular controllable factor. According to assumption A above, the controllable factor of the vocal tract filter is the shape of the vocal tract. All the voiced frames included in the corpus are divided, based on the articulatory information, 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 settings. According to assumption B, the controllable factors of the voice source are the F_0 and power of speech. All the voiced frames included in the corpus are divided, based on the F_0 and speech power information, 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 power values.

3.4. Simultaneous estimation

Let $\hat{S}_k$, $\hat{H}_k$ and $\hat{G}_k$ denote the log-amplitude frequency responses of speech, vocal tract and voice source, respectively, at an analysis frame k. Since we assume that the speech production system is linear, there is the following relation between these three:

$$\hat{S}_k = \hat{H}_k + \hat{G}_k$$

If we consider frames that belong to an articulatory cluster, the vocal tract response $\hat{H}_k$ becomes cluster specific because we assumed the vocal tract shape identical in each articulatory cluster. In addition, $\hat{G}_k$ can be divided into its mean and variation, which are indicated by $\hat{G}_{mean}$ and $\Delta\hat{G}_k$ respectively. If we express a cluster specific vocal tract response by $\hat{H}^{(i)}$, the equation can be written as:

$$\hat{S}_k = \hat{H}^{(i)} + \hat{G}_{mean} + \Delta\hat{G}_k \quad (k \in C_h^{(i)}) \qquad (6)$$

By taking averages of both sides, we obtain the following relation:

$$\frac{1}{N^{(i)}} \sum_{k \in C_h^{(i)}} \hat{S}_k = \hat{H}^{(i)} + \hat{G}_{mean} + \frac{1}{N^{(i)}} \sum_{k \in C_h^{(i)}} \Delta\hat{G}_k$$

where $N^{(i)}$ represents the number of frames included in the i-th cluster $C_h^{(i)}$. In this equation the third term on the right hand side can be eliminated if it varies with no correlation with the vocal tract shape. Since the operation of MFA is equivalent to averaging responses in a cluster, the first and the second terms approximately correspond to a spectral envelope obtained as the result of applying the MFA to the cluster. Let $\hat{P}^{(i)}$ be the spectral envelope obtained by applying the MFA to the cluster $C_h^{(i)}$, then,

$$\hat{P}^{(i)} \approx \hat{H}^{(i)} + \hat{G}_{mean} \tag{7}$$

If we substitute Eq.(7) into Eq.(6),

$$\hat{S}_k \approx \hat{P}^{(R_h(k))} + \Delta\hat{G}_k \tag{8}$$

where the following function was introduced:

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

If we consider frames that belong to a source cluster, the variation of the voice source response, $\Delta\hat{G}_k$, becomes cluster specific according to assumption B above. If we express the cluster specific variation of the source response by $\Delta\hat{G}^{(j)}$, Eq.(8) can be written as:

$$\hat{S}_k \approx \hat{P}^{(R_h(k))} + \Delta\hat{G}^{(j)} \quad (k \in C_g^{(j)}) \tag{9}$$

For each source cluster $C_g^{(j)}$, $\Delta\hat{G}^{(j)}$ is thus given as:

$$\Delta\hat{G}^{(j)} \approx \hat{S}_k - \hat{P}^{(R_h(k))} \quad (k \in C_g^{(j)}) \tag{10}$$

The variation $\Delta\hat{G}^{(j)}$ is predictable by applying MFA to each source cluster $C_g^{(j)}$. Let $\hat{Q}^{(j)}$ be the estimate, then $\hat{S}_k$ is consequently expressed as:

$$\hat{S}_k \approx \hat{P}^{(R_h(k))} + \hat{Q}^{(R_g(k))} \tag{11}$$

where we introduced:

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

If our corpus was large enough, we would find an identical $\hat{G}_{mean}$ for every articulatory cluster. However, $\hat{G}_{mean}$ is actually different for each articulatory cluster due to the bias of F_0 or power in the corpus. In order to minimise this difference, we reiterate the following estimates alternatively.

- Find $\hat{P}^{(i)}$ by applying MFA to $\{\hat{S}_k - \hat{Q}^{(R_g(k))} | k \in C_h^{(i)}\}$
- Find $\hat{Q}^{(j)}$ by applying MFA to $\{\hat{S}_k - \hat{P}^{(R_h(k))} | k \in C_g^{(j)}\}$

We hereinafter consider $\hat{P}^{(i)}$ and $\hat{Q}^{(j)}$, obtained in the manner described above, to be the vocal tract response and voice source response respectively.

3.5. Iterative estimation procedure

The MFA alternatively discovers the cepstrum $\boldsymbol{c}_h^{(i)}$ and $\boldsymbol{c}_g^{(j)}$, which correspond to the frequency responses of vocal tract and voice source, according to the following iterative procedure:

Step 1: For each articulatory cluster $C_h^{(i)} (i = 1, 2, 3, ..., K)$, the cepstrum $\boldsymbol{c}_h^{(i)}$ is calculated by applying the MFA to the harmonic amplitudes $\{\boldsymbol{a}_k | k \in C_h^{(i)}\}$. (the first approximation)

Step 2: For all the harmonics, approximate errors between the observed log-amplitude $\boldsymbol{a}_k$ and the log-amplitude calculated from $\boldsymbol{c}_h^{(i)}$ are obtained as follows:

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

This equation corresponds to Eq.(10) and $\boldsymbol{p}_k$ accordingly includes the fluctuation of the source response.

Step 3: For each source cluster $C_g^{(j)} (j = 1, 2, 3, ..., L)$, the cepstrum $\boldsymbol{c}_g^{(j)}$ is calculated by applying the MFA to $\{\boldsymbol{p}_k | k \in C_g^{(j)}\}$.

Step 4: For all the harmonics, approximate errors between the observed log-amplitude $\boldsymbol{a}_k$ and the log-amplitude calculated from $\boldsymbol{c}_g^{(j)}$ are obtained as follows:

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

Step 5: For each articulatory cluster $C_h^{(i)} (i = 1, 2, 3, ..., K)$, $\boldsymbol{c}_h^{(i)}$ is calculated by applying the MFA to $\{\boldsymbol{q}_k | k \in C_h^{(i)}\}$.

Step 6: The estimation error ε in Eq.(2) is evaluated and the procedure is terminated if ε converges. If not, Step 2-5 is applied repeatedly.

4. Experiment

4.1. Data and method

Although the similarity of the vocal tract shape can be roughly estimated from its phonetic contexts, we employ data that tell us the actual shape with greater reliability.

The data used in the experiment is a corpus composed of 460 sentences uttered by a female speaker (fsew0). The corpus includes parallel acoustic-articulatory information, which was recorded using a Carstens EMA system at Queen Margaret University College, Edinburgh (see [8] for details). The sampling frequencies of the acoustic waveform and articulatory data 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 articulators' positions. The harmonic spectra were estimated from the waveform using the weighted least squares method in [11]. The width and spacing of the time window (Hanning) for this analysis were 20 ms and 8 ms respectively. The articulatory information was down-sampled to the same spacing of 8 ms. Thereby 87208 voiced frames with parallel acoustic-articulatory information were obtained in total.

All the voiced frames were divided into 2048 articulatory clusters and 64 source clusters using LBG clustering. These numbers have been established empirically by preliminary experiments.

Finally, according to the procedure in 3.5, iterative approximation was performed to find the pair of cepstra (order of 40) that correspond to $\boldsymbol{c}_h^{(i)}$ and $\boldsymbol{c}_g^{(j)}$. In this calculation, we employed a Gaussian distribution with 0 Hz mean and 2 kHz standard deviation for the weighting function $w(f)$ in Eq.(4).

4.2. Results and discussion

Fig.2 shows voice source responses obtained from three source clusters with different F_0s and with powers that are close to their average value in the corpus. In addition, Fig.3 shows voice source responses obtained from three source clusters with different powers and with F_0s that are close to their average

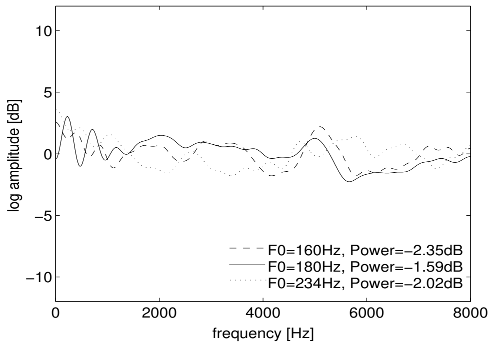

Figure 2: *Voice source frequency responses for different F_0s*

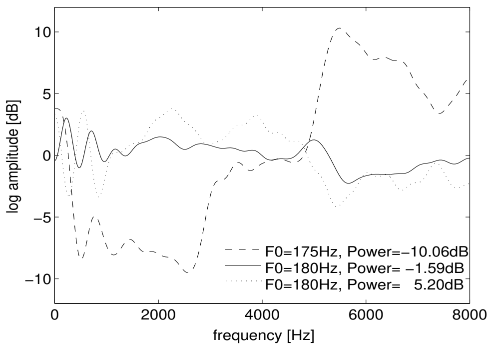

Figure 3: *Voice source frequency responses for different powers*

value in the corpus. In these figures, "F0" and "Power" indicate the fundamental frequency and speech power of the centroid of each source cluster. The voice-source response does not change significantly with F_0, though does with power when the power is low.

In Fig.3, the response with low speech power (-10.06 dB) shows a large amount of energy at the first harmonic (fundamental) and suppressed higher harmonics in the low frequency band ($<$ 3 kHz). This tendency is very much in agreement with reports (e.g. [9]) that the glottal waveform changes sinusoidally when the power of voice is low. Moreover, it can be seen in the same figure that the response with low speech power in the high frequency band ($>$ 5 kHz) increases more than 10 dB compared to the amplitude in the low frequency band. We think that it shows the relative increase of noise level since SN ratio reduces when voice power is low.

5. Conclusions

We introduced a new approach to estimating voice source and vocal tract characteristics simultaneously from voiced speech. After a theoretical examination, an experiment is made applying the approach to actual speech data.

Strictly speaking, it is clear from the explanation in section 3.4 that the approach does not completely separate the characteristics of voice source and vocal tract filter. However, it becomes possible to control voice source and vocal tract responses independently by using its result, which is useful in some areas of speech technology, such as speech synthesis. The responses can be determined automatically from the corpus using the method we presented here.

We informally carried out a speech re-synthesis test based on sinusoidal synthesis[12] using the cepstra obtained in the above experiment and the original F_0 contour, and confirmed that intelligible, high quality speech was generated for two speakers from the corpus.

6. Future work

In this paper we employed speech power as a factor to perform clustering in order to obtain the source clusters. However, we should probably use the power of *voice source*, obtained by filtering speech through the inverse of the vocal tract filter. This could be achieved by incorporating the clustering algorithm into the iteration procedure.

Moreover, we need to consider improving the clustering of the articulatory space. Although we employed a data clustering technique based on articulatory data, articulators' movements are sometimes perceptually important and sometimes less important depending on their positions. We are therefore currently examining a new approach to perform clustering in the articulatory space using a criterion in the acoustic (or perceptual) space.

Acknowledgements

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

References

[1] Hedelin, P., "A glottal LPC-vocoder," in *Proc. ICASSP84*, Mar. 1984, vol. 1, pp. 1.6.1–1.6.4.

[2] Fujisaki, H. and Ljungqvist, M., "Estimation of voice source and vocal tract parameters based on ARMA analysis and a model for the glottal source waveform," in *Proc. ICASSP87*, 1987, vol. 15.4, pp. 637–640.

[3] Wong, D. Y., Markel, J. D., and A. H. Gray, J., "Least squares glottal inverse filtering from the acoustic speech waveform," *IEEE Trans. Acoust. Speech Signal Process.*, vol. 27, pp. 350–355, 1979.

[4] Alku, P., "Glottal wave analysis with pitch synchronous iterative adaptive inverse filtering," *Speech Communication*, vol. 11, pp. 109–118, 1992.

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

[6] Makhoul, J., "Linear prediction: A tutorial review," *Proc. IEEE*, vol. 63, pp. 561–580, 1975.

[7] Galas, T. and Rodet, X., "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.

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

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

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

[11] Stylianou, Y., "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.

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