ACCURATE SPECTRAL ENVELOPE ESTIMATION FOR ARTICULATION-TO-SPEECH SYNTHESIS

Yoshinori Shiga and Simon King

Centre for Speech Technology Research, University of Edinburgh, U.K.

yoshi@cstr.ed.ac.uk

ABSTRACT

This paper introduces a novel articulatory-acoustic mapping in which detailed spectral envelopes are estimated based on the cepstrum, inclusive of the high-quefrency elements which are discarded in conventional speech synthesis to eliminate the pitch component of speech. For this estimation, the method deals with the harmonics of multiple voiced-speech spectra so that several sets of harmonics can be obtained at various pitch frequencies to form a spectral envelope. The experimental result shows that the method estimates spectral envelopes with the highest accuracy when the cepstral order is 48-64, which suggests that the higher-order coefficients are required to represent detailed envelopes reflecting the real vocal-tract responses.

1. INTRODUCTION

The objective of this study is to realise articulatory modification on the acoustic characteristics of speech whilst maintaining aspects of the signal relating to speaker identity, and with the high signal quality required for speech synthesis. For achieving this, this paper deals with the following two related points at issue together: 1) a mapping of articulation to the vocal-tract transfer function (VTTF) using the actual measurement of articulators; 2) accurate VTTF estimation based on the articulatory data for high-quality speech synthesis.

Kaburagi et al. [1] have 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, this approach is considered an alternative to acoustically simulated vocal-tract modelling which has been widely investigated [2]. In [1], the capability of their method is demonstrated in producing intelligible speech by employing LSP and multipulse excitation. However, because of the use of this common parameterisation, their synthesised speech has as many artefacts as the speech of conventional speech synthesis has.

With respect to parameterisation, speech representation derived from spectral peaks at harmonic frequencies of voiced speech has attracted attention widely in speech technology. Gu et al. [3] have proposed feature extraction for speech recognition based on the *Perceptual Harmonic Cepstral Coefficients* (PHCC), and confirmed by experiments that PHCC outperforms standard cepstral representation. A main idea of PHCC is that, in the process of extracting the coefficients, voiced speech is sampled at harmonic locations in the frequency domain. Such harmonic-based parameterisation has also been used in the field of speech coding since the early 90's for perceptually efficient encoding [4, 5].

It must be noted that, whereas the harmonic peaks have an important role in human auditory perception, only those peaks reflect the VTTF since voiced speech, due to its quasi-periodicity, only has energy at frequencies corresponding to integral multiples of the fundamental frequency (F_0). For this reason, similar techniques [6, 7, 8] which trace the harmonic peaks have been applied to text-to-speech synthesis in order to obtain spectral envelopes corresponding to the VTTFs. A recently developed high-quality vocoder, STRAIGHT [9], also exploits harmonic peaks, into which a bilinear surface is interpolated in the three-dimensional space composed of time, frequency and spectral power.

However, it has been pointed out that the harmonic structure of voiced speech interferes with identifying spectral envelopes that precisely reflect VTTFs. Since voiced speech consists of line spectra in the frequency domain, it is theoretically impossible to know the real characteristics at frequencies where no harmonic exists [10]. Therefore even the spectral envelopes from the above harmonic-based estimation are still inaccurate for representing actual VTTFs, because sections except harmonic peaks in the estimated envelope are interpolated and do not reflect the real VTTF.

This fact becomes a problem in speech synthesis where speech needs to be generated at various F_0s different from the original. In order to synthesise high-quality speech it is required to estimate spectral characteristics not only at harmonic peaks but also between the peaks. Moreover, any operation, such as averaging, on such a speech representation blurs spectral envelopes because, in the operation, *reliable* characteristics observed at harmonic locations and *unreliable* characteristics interpolated are both treated equivalently. We believe that, for these reasons, speech degrades during conventional parameter-based speech synthesis.

For resolving these problems, we have proposed a method for estimating spectral envelopes of voiced speech based on the diverse harmonic structures of multiple short-time speech signals produced under almost the same articulatory configuration [11]. The method is expected to obtain detailed spectral envelopes reflecting the responses of the intricate vocal tract, which conventional analysis is unable to estimate due to the interference of the harmonic structure. In the process of estimating the envelopes, the method also produces a mapping of articulation to spectral envelopes, and consequently we can realise high-quality articulatory-acoustic conversion applying the envelopes precisely estimated.

In this paper, we introduce two types of mapping functions based on piecewise constant approximation (which we have already proposed in [11]) and piecewise linear approximation. After examining these functions theoretically, we closely investigate the performance of both mappings through some experiments.

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

2. ARTICULATORY-ACOUSTIC MAPPING

2.1. Articulatory data

The data used in this study is a MOCHA (Multi-CHannel Articulatory) corpus [12]. The corpus is composed of 460 TIMIT sentences uttered by a female speaker (fsew0), and includes parallel acoustic-articulatory information which was recorded using a Carstens Electromagnetic Articulograph 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 data are 16 kHz and 0.5 kHz respectively (see [12] for details).

2.2. Clustering in the articulatory space

After normalising each dimension of the articulatory vectors which are composed of the articulator positions extracted from the corpus by frame, we apply LBG clustering [13] to all the normalised vectors, and group them into K clusters, C^i (i = 1, 2, 3, ..., K). Then we estimate articulatory-acoustic mapping functions for each cluster, as is described in the following sections.

2.3. Speech representation

We adopt the *cepstrum* as an expression of the spectral envelope for the purpose of approximating harmonic peaks of multiple speech spectra. The cepstrum is adequate to represent both zeros and poles with a small number of coefficients, while on the other hand the all-zero model, such as PSOLA [14] (the model of which is explained with impulse-excited FIR filter [15]), demands a number of coefficients (taps) to describe the detailed spectral envelopes so that more training data and computational complexity are required to obtain the optimal coefficients. The cepstrum is, in addition, a frequency-domain representation and thus has good interpolation properties. These merits allow the cepstrum to be widely applied in the field of speech technology (e.g. [16]).

3. PIECEWISE CONSTANT MAPPING

The clustering in the articulatory space makes each cluster include speech frames with comparatively similar articulatory settings. If we assume those settings identical in a cluster, the acoustical characteristics of the vocal tract can be assumed constant within the cluster. Under this assumption, the problem is reduced to estimating one unique spectral envelope for every cluster. We accordingly use the different harmonic structures of the multiple frames to form a spectral envelope [11].

3.1. Estimating the envelopes of amplitude spectra

Let us determine a cepstrum which best fits the log-amplitude of all the harmonics of speech frames belonging to cluster i, using the least squares method. This can be considered an extension of the cepstrum estimation in [6, 7] to the analysis of multiple frames.

Let $a_k^{(l)}$ denote an observed natural log amplitude of the l-th harmonic (l = 1, 2, 3, ..., N_k) at frequency $f_k^{(l)}$ within the speech frame k, and T the sampling period. Then, the sum of squared approximation errors for the amplitude of all the harmonics of all the frames is expressed as

$$E_a^{(i)} = \sum_{k \in C^i} \sum_{l=1}^{N_k} \frac{w(f_k^{(l)})}{N_k} \left(a_k^{(l)} - d_k - \sum_{n=-p}^{p} c_a^{(i)}[n] \cos \Omega_k^{(l)} n \right)^2 \quad (1)$$

where $c_a^{(i)}[n]$ indicates the n-th cepstral coefficient and $\Omega_k^{(l)} = 2\pi f_k^{(l)} T$. In (1) we have introduced two weighting factors, $w(f)$ for attaching importance to the lower frequency band, and $1/N_k$ for evaluating each frame equally regardless of the number of harmonics. The offset d_k adjusts the overall power of each frame so as to minimise the error $E_a^{(i)}$. Equation (1) is expressed in terms of vectors and matrices as

$$E_a^{(i)} = \sum_{k \in C^i} (\boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_a^{(i)})^T \mathbf{W}_k (\boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_a^{(i)}) \quad (2)$$

where $\boldsymbol{c}_a^{(i)} = [\ c_a^{(i)}[0], c_a^{(i)}[1], c_a^{(i)}[2], \cdots, c_a^{(i)}[p]\ ]^T$, and $\boldsymbol{y}_k = [\ a_k^{(1)} - d_k, a_k^{(2)} - d_k, a_k^{(3)} - d_k, \cdots, a_k^{(N_k)} - d_k\ ]^T$. The matrices $\mathbf{P}_k$ and $\mathbf{W}_k$ are as follows:

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

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

Equation (2) can be solved by reducing it to a problem of weighted least squares. The cepstrum coefficients $\boldsymbol{c}_a^{(i)}$ can be found by solving the following normal equation:

$$\left(\sum_{k \in C^i} \mathbf{P}_k^T \mathbf{W}_k \mathbf{P}_k \right) \boldsymbol{c}_a^{(i)} = \sum_{k \in C^i} \mathbf{P}_k^T \mathbf{W}_k \boldsymbol{y}_k. \quad (3)$$

The offset d_k is then calculated as

$$d_k = \frac{\sum_{l=1}^{N_k} w(f_k^{(l)}) \left(a_k^{(l)} - 2\sum_{n=1}^{p} c_a^{(i)}[n] \cos\Omega_k^{(l)} n \right)}{\sum_{l=1}^{N_k} w(f_k^{(l)})}. \quad (4)$$

Practically, we obtain the cepstrum according to the following procedure: 1) Substitute $\mathbf{0}$ for $\boldsymbol{c}_a^{(i)}$ (initial value); 2) Obtain d_k ($k \in C^i$) using (4); 3) Calculate $E_a^{(i)}$ using (2) and terminate the procedure if $E_a^{(i)}$ converges; 4) Find $\boldsymbol{c}_a^{(i)}$ by solving (3); 5) Substitute 0 for $c_a^{(i)}[0]$ (power normalization); 6) Return to Step 2.

3.2. Estimating the envelopes of phase spectra

The spectral envelopes of phase are obtained in a similar manner, but we need to take care about the unwrapping problem of phase.

Let $\theta_k^{(l)}$ denote an observed wrapped phase of the l-th harmonic within the speech frame k. Then, the sum of squared approximation errors for the phases of all the harmonics of all the frames belonging to cluster i is expressed as

$$E_p^{(i)} = \sum_{k \in C^i} \sum_{l=1}^{N_k} \frac{w(f_k^{(l)})}{N_k} \left(\vartheta_k^{(l)} + \sum_{n=-p}^{p} c_p^{(i)}[n] \sin \Omega_k^{(l)} n \right)^2 \quad (5)$$

where $c_p^{(i)}[n]$ indicates the n-th cepstral coefficient and $\vartheta_k^{(l)}$ is defined by

$$\vartheta_k^{(l)} = \arg \varphi_i(f_k^{(l)}) + \mathrm{ARG}\left[e^{j(\theta_k^{(l)} - 2\pi f_k^{(l)} \tau_k)} \varphi_i(f_k^{(l)})^* \right].$$

The operator $\mathrm{ARG}[X]$ represents wrapping of phase X, and the symbol $*$ the complex conjugate operation. The time delay τ_k adjusts the global tilt of the phase spectrum so as to minimise the error $E_p^{(i)}$. The function $\varphi_i(f)$ represents the moving average of the phase $\{\theta_k^{(l)} - 2\pi f_k^{(l)}\tau_k\}$ (for all the harmonics of all the frames in cluster i) along the frequency axis in the complex spectral domain under a weighting factor $1/N_k$, and is expressed as

$$\varphi_i(f) = \frac{\phi_i(f)}{|\phi_i(f)|}$$

$$\phi_i(f) = \frac{\sum_{k\in C^i}\sum_{l=1}^{N_k} G(f_k^{(l)} - f)e^{j(\theta_k^{(l)} - 2\pi f_k^{(l)}\tau_k)}/N_k}{\sum_{k\in C^i}\sum_{l=1}^{N_k} G(f_k^{(l)} - f)/N_k}.$$

The function $G(f)$ indicates a moving average window. For the initial value of $\varphi_i(f_k^{(l)})$, we adopt the following minimum phase spectrum calculated from the cepstrum $\boldsymbol{c}_a^{(i)}$ which has already been obtained for the amplitude envelope:

$$\varphi_i(f_k^{(l)}) = -2\sum_{n=1}^{p} c_a^{(i)}[n]\sin\Omega_k^{(l)}n. \tag{6}$$

In terms of vectors and matrices, (5) is expressed as

$$E_p^{(i)} = \sum_{k\in C^i}(\boldsymbol{\vartheta}_k - \mathbf{Q}_k\boldsymbol{c}_p^{(i)})^T\mathbf{W}_k(\boldsymbol{\vartheta}_k - \mathbf{Q}_k\boldsymbol{c}_p^{(i)}) \tag{7}$$

where $\boldsymbol{c}_p^{(i)} = [\ c_p^{(i)}[1], c_p^{(i)}[2], c_p^{(i)}[3], \cdots, c_p^{(i)}[p]\ ]^T$ and $\boldsymbol{\vartheta}_k = [\ \vartheta_k^{(1)}, \vartheta_k^{(2)}, \vartheta_k^{(3)}, \cdots, \vartheta_k^{(N_k)}\ ]^T$. The matrix $\mathbf{Q}_k$ is as follows:

$$\mathbf{Q}_k = (-2)\cdot\begin{bmatrix} \sin\Omega_k^{(1)} & \sin 2\Omega_k^{(1)} & \cdots & \sin p\Omega_k^{(1)} \\ \vdots & \vdots & \vdots & \vdots \\ \sin\Omega_k^{(N_k)} & \sin 2\Omega_k^{(N_k)} & \cdots & \sin p\Omega_k^{(N_k)} \end{bmatrix}.$$

Thus, the cepstrum $\boldsymbol{c}_p^{(i)}$ can be found by solving the following normal equation:

$$\left(\sum_{k\in C^i}\mathbf{Q}_k^T\mathbf{W}_k\mathbf{Q}_k\right)\boldsymbol{c}_p^{(i)} = \sum_{k\in C^i}\mathbf{Q}_k^T\mathbf{W}_k\boldsymbol{\vartheta}_k. \tag{8}$$

The delay τ_k can be calculated on the basis of the cross-correlation which is computed by the inverse Fourier transform of the cross-spectrum $\{\exp[j\theta_k^{(l)}]\cdot\varphi_i(f_k^{(l)})^*\}$ (l = 1, 2, 3, ..., N_k).

According to the following procedure, we obtain the cepstrum representing the envelope of the phase spectrum: 1) Initialise $\varphi_i(f)$ using (6); 2) Find τ_k ($k \in C^i$) based on the cross-spectrum; 3) Calculate $E_p^{(i)}$ using (7) and terminate the procedure if $E_p^{(i)}$ converges; 4) Find $\boldsymbol{c}_p^{(i)}$ by solving (8); 5) Return to Step 2.

4. PIECEWISE LINEAR MAPPING

The piecewise constant assumption is clearly only a rough approximation. Because, in practice, articulation is not identical within a cluster and accordingly neither is the vocal tract response, such an approximation is likely to cause noticeable distortion. For more accurate estimation, a mapping function can be introduced per cluster which transforms articulatory vectors into acoustic features. We must, however, be aware that models with high complexity may estimate harmonic structure itself instead of the spectral envelope necessary. Here we choose a linear mapping, the complexity of which is considered low enough.

4.1. Piecewise linear approximation

The cepstra $\boldsymbol{c}_a^{(i)}$ and $\boldsymbol{c}_p^{(i)}$ in (2) and (7) are represented by the linear transformation of L-dimensional articulatory vector $\boldsymbol{x}_k$ as follows:

$$\boldsymbol{c}_a^{(i)} = \boldsymbol{q}^{(i)} + \mathbf{U}^{(i)}\boldsymbol{x}_k,\quad \boldsymbol{c}_p^{(i)} = \boldsymbol{r}^{(i)} + \mathbf{V}^{(i)}\boldsymbol{x}_k \tag{9}$$

where $\boldsymbol{q}^{(i)}$, $\boldsymbol{r}^{(i)}$, $\mathbf{U}^{(i)}$ and $\mathbf{V}^{(i)}$ consist of the coefficients of the linear transformation, and are defined as

$$\boldsymbol{q}^{(i)} = [q_0^{(i)}\ q_1^{(i)}\ q_2^{(i)}\ \cdots\ q_p^{(i)}]^T,\quad \boldsymbol{r}^{(i)} = [r_1^{(i)}\ r_2^{(i)}\ r_3^{(i)}\ \cdots\ r_p^{(i)}]^T$$

$$\mathbf{U}^{(i)} = \begin{bmatrix} u_{01}^{(i)} & \cdots & u_{0L}^{(i)} \\ \vdots & \ddots & \vdots \\ u_{p1}^{(i)} & \cdots & u_{pL}^{(i)} \end{bmatrix},\quad \mathbf{V}^{(i)} = \begin{bmatrix} v_{11}^{(i)} & \cdots & v_{1L}^{(i)} \\ \vdots & \ddots & \vdots \\ v_{p1}^{(i)} & \cdots & v_{pL}^{(i)} \end{bmatrix}.$$

The problem is now reduced to finding these matrices and vectors. Substituting (9) into (2) and (7) and rewriting the formulae, we obtain the following equations:

$$E_a^{(i)} = \sum_{k\in C^i}(\boldsymbol{y}_k - \mathbf{\Gamma}_k\boldsymbol{u}_k^{(i)})^T\mathbf{W}_k(\boldsymbol{y}_k - \mathbf{\Gamma}_k\boldsymbol{u}_k^{(i)}) \tag{10}$$

$$E_p^{(i)} = \sum_{k\in C^i}(\boldsymbol{\vartheta}_k - \mathbf{\Delta}_k\boldsymbol{v}_k^{(i)})^T\mathbf{W}_k(\boldsymbol{\vartheta}_k - \mathbf{\Delta}_k\boldsymbol{v}_k^{(i)}) \tag{11}$$

where $\boldsymbol{u}_k^{(i)} = [u_{01}^{(i)}\ u_{11}^{(i)}\ u_{21}^{(i)}\cdots u_{02}^{(i)}\ u_{12}^{(i)}\ u_{22}^{(i)}\cdots u_{pL}^{(i)} q_0^{(i)}\cdots q_p^{(i)}]^T$, $\boldsymbol{v}_k^{(i)} = [v_{11}^{(i)}\ v_{21}^{(i)}\ v_{31}^{(i)}\cdots v_{12}^{(i)}\ v_{22}^{(i)}\ v_{32}^{(i)}\cdots v_{pL}^{(i)}\ r_1^{(i)}\cdots r_p^{(i)}]^T$ and

$$\mathbf{\Gamma}_k = \left[x_1\mathbf{P}_k \vdots x_2\mathbf{P}_k \vdots x_3\mathbf{P}_k \vdots\vdots\cdots\vdots x_{L-1}\mathbf{P}_k \vdots x_L\mathbf{P}_k \vdots \mathbf{P}_k\right]$$

$$\mathbf{\Delta}_k = \left[x_1\mathbf{Q}_k \vdots x_2\mathbf{Q}_k \vdots x_3\mathbf{Q}_k \vdots\vdots\cdots\vdots x_{L-1}\mathbf{Q}_k \vdots x_L\mathbf{Q}_k \vdots \mathbf{Q}_k\right].$$

Having the same form as (2) and (7), (10) and (11) can be solved for $\boldsymbol{u}_k^{(i)}$ and $\boldsymbol{v}_k^{(i)}$ likewise during the same procedures as in section 3.

5. EXPERIMENTS

5.1. Data and method

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 speech waveform using the weighted least squares method [17], in which the width and spacing of the time window (Hanning) were 20 ms and 8 ms respectively. Accordingly we downsampled the articulatory information to the same spacing of 8 ms. Thereby 87208 voiced frames with parallel acoustic-articulatory information were obtained in total. We set 10% of the sentences (46 sentences including 8332 frames) aside for testing, and used the remaining 90% (414 sentences including 78876 frames) for training.

Estimation accuracy is evaluated only at harmonic frequencies where reliable characteristics can be observed. For this purpose we introduced two types of distortions: *harmonic power distortion* D_a and *harmonic phase distortion* D_p. They are defined by

$$D_a = \frac{20}{\ln 10}\sqrt{\frac{1}{M}\sum_{i=1}^{K}E_a^{(i)}},\quad D_p = \sqrt{\frac{1}{M}\sum_{i=1}^{K}E_p^{(i)}} \tag{12}$$

where M denotes the total number of frames included in all the clusters.

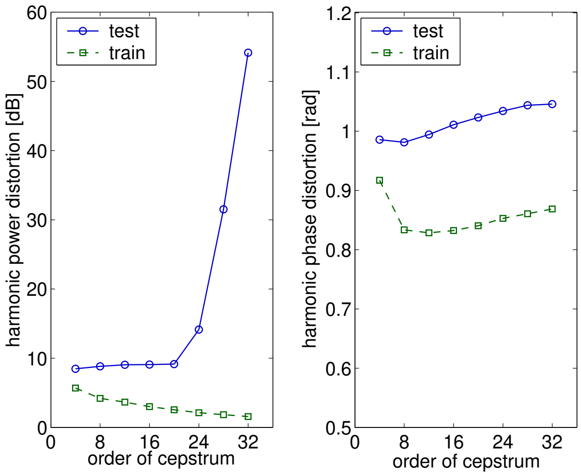

Fig. 1. *Harmonic distortion vs. order of cepstrum, when each cluster comprises only one frame*

5.2. Experiments with conventional criteria

Prior to the main experiments, we examined the tendency of estimation accuracy when mapping functions were obtained using criteria *in the cepstral domain* (instead of the harmonic-based criteria described in section 3 and 4). Such cepstral-domain criteria are generally used in conventional speech technology. With respect to the amplitude spectral envelope, the criteria are defined by

$$E_{PC}^{(i)} = \sum_{k \in C^i} (\boldsymbol{c}_k - \boldsymbol{c}_a^{(i)})^T (\boldsymbol{c}_k - \boldsymbol{c}_a^{(i)}) \tag{13}$$

$$E_{PL}^{(i)} = \sum_{k \in C^i} \left[\boldsymbol{c}_k - (\boldsymbol{q}^{(i)} + \mathbf{U}^{(i)}\boldsymbol{x}_k)\right]^T \left[\boldsymbol{c}_k - (\boldsymbol{q}^{(i)} + \mathbf{U}^{(i)}\boldsymbol{x}_k)\right] \tag{14}$$

for the piecewise constant mapping and for the piecewise linear mapping, respectively. In both equations, $\boldsymbol{c}_k$ represents the cepstrum (exclusive of a coefficient at the quefrency of 0 second) of the amplitude spectral envelope for speech frame k, which is computed using the cepstral estimation in [6]. In the case of the piecewise constant mapping, cepstrum $\boldsymbol{c}_a^{(i)}$ was computed based on criteria (13), and the distortions were obtained using equation (2) and (12). In the case of the piecewise linear mapping, $\boldsymbol{q}^{(i)}$ and $\mathbf{U}^{(i)}$ were computed based on criteria (14), and the distortions were obtained using equation (10) and (12). With respect to the phase spectral envelope, due to the unreliable phase-unwrapping [15], we used the minimum phase spectrum, which is derived from the cepstrum of the amplitude spectral envelope.

First, we examined the estimation distortions when every cluster has only one frame. In this case, distortions for the training data set correspond to errors caused by parameterisation, and distortions for the test set correspond to errors when the nearest-neighbour articulation is chosen in the articulatory-acoustic transformation. In other words, the former distortions represent the speech-quality deterioration in speech analysis-synthesis, and the latter represent the degradation caused mainly by the alteration of F_0 and the ensuing change of harmonic structure. As is obvious from the result shown in Fig. 1, the power distortion for the test data is almost constant up to order 20 (1.25 ms in quefrency), but around 24 (1.5 ms) the distortion rapidly increases. The main reason of this tendency is considered that harmonic structure comes to appear in the envelopes, smooth interpolation between harmonic peaks being failed, and consequently the distortion increased for the test data having different harmonic structure. Therefore order 20 (1.25 ms) is a limit in the usual cepstral analysis for the female voice used in the experiments, and synthetic speech deteriorates when higher order of cepstrum is used.

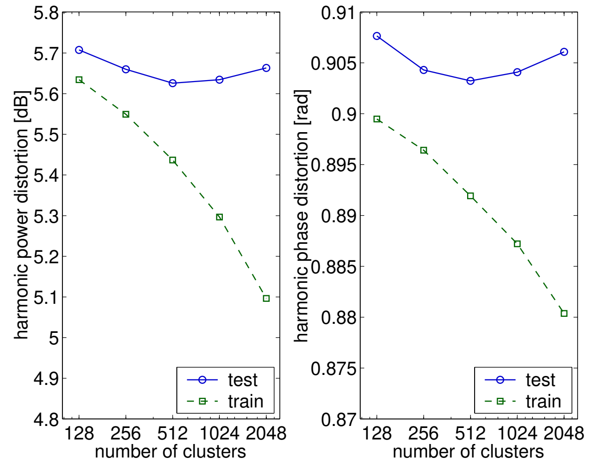

Fig. 2. *Harmonic distortion vs. number of clusters, based on cepstral domain criteria for the piecewise constant mapping*

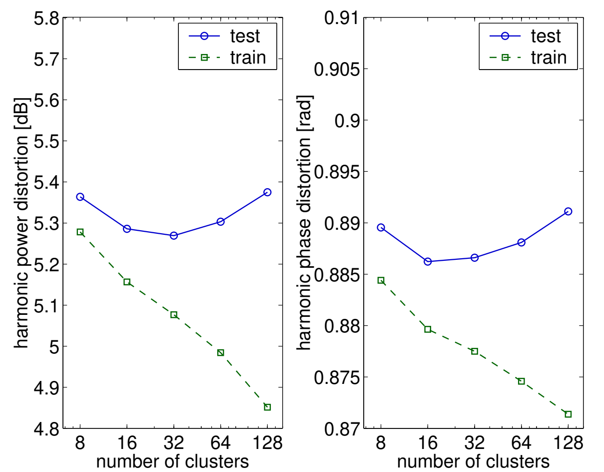

Fig. 3. *Harmonic distortion vs. number of clusters, based on cepstral domain criteria for the piecewise linear mapping*

Next, we examined relation between the number of clusters and the distortions, where the cepstral order was set to 20 according to the above result. The results are shown in Fig. 2 and 3. Power distortion for the test set has the minimum value in the case of 512 articulatory clusters for the piecewise constant approximation, where the distortions are 5.63 dB and 0.903 rad; and in the case of 32 clusters for the piecewise linear approximation, where the values are 5.27 dB and 0.887 rad.

5.3. Experiments with the proposed method

We examined the performance of the two mapping functions we discussed in section 3 and 4.

The distortions for the training set were calculated in the training process using equation (2), (7) and (12) for the piecewise constant mapping; and using equation (10), (11) and (12) for the piecewise linear mapping. The distortions for the test data were calcu-

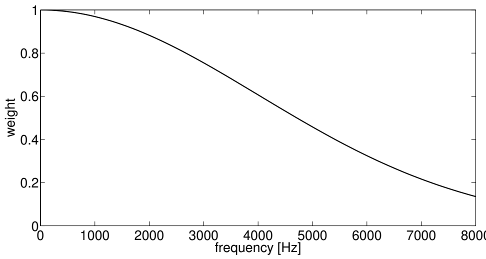

Fig. 4. *Weighting function $w(f)$*

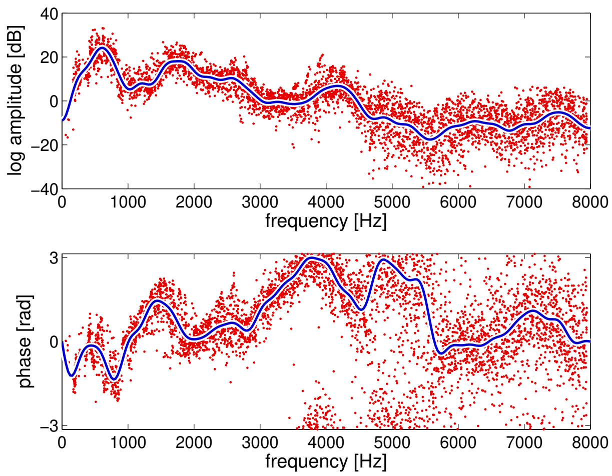

Fig. 5. *Spectral envelopes of one of the articulatory clusters computed using the piecewise constant approximation*

lated as follows: first, the nearest neighbour method chooses one of the articulatory clusters based on the Euclidean distance between the centroids of the clusters and a frame to be tested in the articulatory space, and then the distortions are calculated based on equation (2), (7) and (12) using the cepstral coefficients of the chosen cluster for the piecewise constant mapping; and based on equation (10), (11) and (12) using the linear mapping coefficients of the chosen cluster for the piecewise linear mapping. For the weighting function $w(f)$ and moving-average window $G(f)$ in section 3, we introduced a Gaussian distribution (Fig. 4) with 0 Hz mean and 4 kHz standard deviation, and a Gaussian window with 100 Hz standard deviation, respectively.

Figure 5 shows a pair of spectral envelopes of a cluster computed from the cepstrum obtained by the piecewise constant approximation. The cepstral order was set to 48 for the result. In the figure, the solid lines indicate the envelopes of the amplitude spectrum (upper) and the phase spectrum (lower), while the dots represent $\{a_k^{(l)} - d_k\}$ of equation (1) in the upper graph, and $\vartheta_k^{(l)}$ of equation (5) in the lower. Shown in Fig. 6 are the harmonic distortions of the piecewise constant mapping. As in this figure, the distortions for the test data set have the minimum values in the case of cepstral order 48 (3.0 ms in quefrency) and 512 articulatory clusters for amplitude, and in the case of order 64 (4.0 ms) and 256 clusters for phase, where the distortions are 5.56 dB and 0.807 rad. Figure 7 shows the result of the piecewise linear mapping. The distortions have the minimum values in the case of order 64 (4.0 ms) and 32 clusters for amplitude and in the case of order 64 (4.0 ms) and 16 clusters for phase, where the values are 5.18 dB and 0.778 rad.

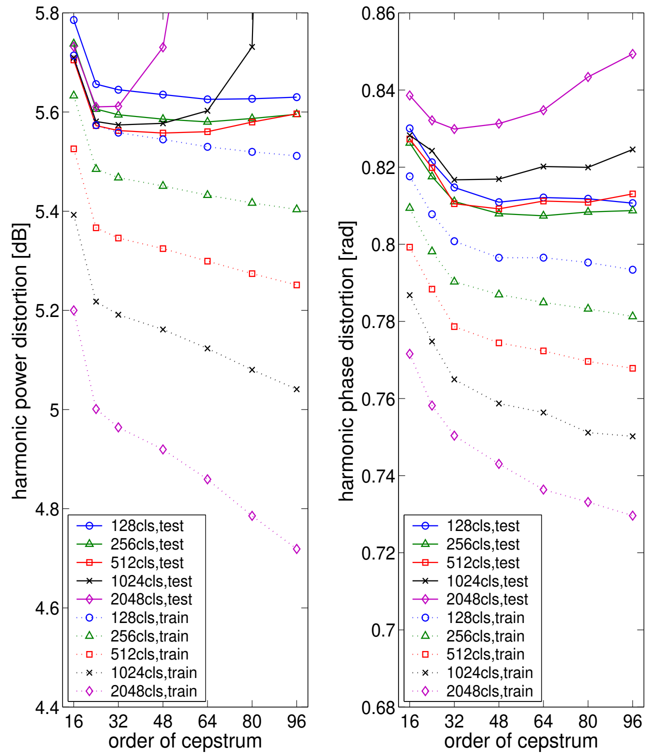

Fig. 6. *Harmonic distortion vs. order of cepstrum, in the case of the piecewise constant mapping*

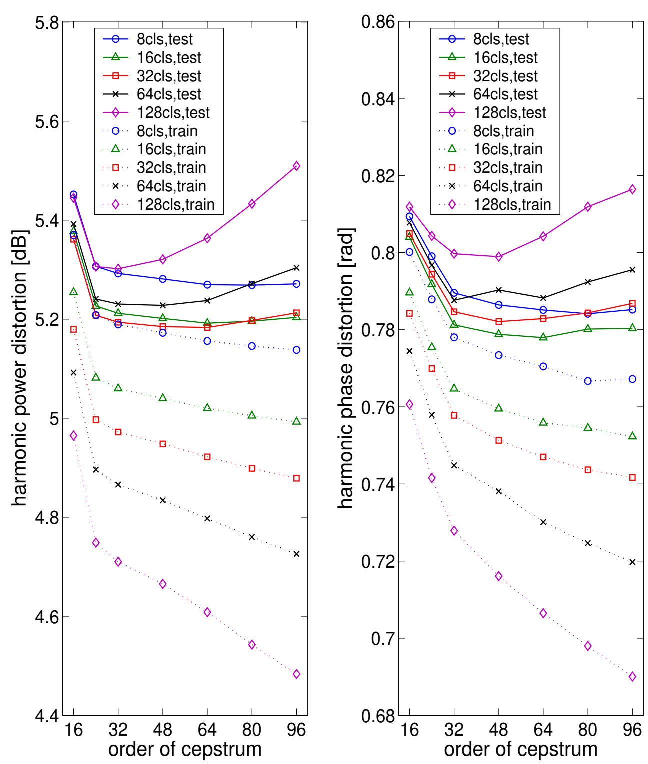

Fig. 7. *Harmonic distortion vs. order of cepstrum, in the case of the piecewise linear mapping*

6. DISCUSSION

Discovered through the experiments are the following points:

1) For both of the introduced mappings, spectral envelopes are obtained with the highest accuracy when the cepstral order is 48-64 (3.0-4.0 ms in quefrency), where the distortions were minimised. The results suggest that, in order to represent spectral envelopes reflecting real VTTFs, cepstral coefficients of high quefrency range are necessary, which are usually discarded in conventional speech synthesis to eliminate the pitch component of speech;

2) The piecewise linear mapping is more accurate and requires a smaller number of clusters than the piecewise constant mapping. Therefore relation between the articulator positions and the cepstrum is considered locally linear rather than constant;

3) Evidently from the comparison of Fig. 2 and 6, and of Fig. 3 and 7, the estimation based on the cepstral-domain criteria leads to producing larger distortion than our proposed harmonic-based estimation. This may indicate the necessity of reconsidering the parameterisation used in the current speech technology;

4) The phase distortions of both of the proposed mappings showed much smaller values than those derived from the minimum phase of cepstrum. This may suggest a problem of phase prediction based on the minimum phase. Further experiments are necessary to examine how much such distortion in phase influences perception; and

5) For both proposed mapping functions, the variance of the errors of the phase spectrum indicates, as in Fig. 5, the degree of randomness of phase in each frequency band, which can be useful information for controlling phase of synthetic speech so as to reduce its buzziness.

7. CONCLUSIONS

We introduced an articulatory-acoustic mapping which enables the estimation of detailed spectral envelopes by dealing only with harmonic peaks of multiple voiced-speech spectra. The experimental results showed that the piecewise linear mapping is more suitable than the piecewise constant mapping to represent relationship between articulatory configuration and acoustic characteristics of speech represented by the cepstrum. Also, the results suggest that cepstral coefficients of higher quefrency range are required for estimating detailed envelope reflecting vocal tract filter characteristics, compared with the order used commonly in conventional speech synthesis.

We have confirmed that applying a source-filter separation [18], where the characteristics of the voice source are taken into account using F_0 and speech power, further improves the estimation accuracy and reduces the distortions of the piecewise linear mapping to 4.93 dB for power and 0.775 rad for phase. Moreover, the proposed harmonic-based estimation can also be applied to an articulatory-acoustic mapping based on the Gaussian mixture model, which we have already employed for the purpose of reducing acoustical discontinuity of output speech at the boundaries of clusters. As for the detail of these applications, we would like to report on the next opportunity.

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. ICSLP-98*, 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*, Seattle, USA., 1998, vol. II, p. 1283.

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

[4] 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*, Apr. 1993, vol. 2, pp. 439–442.

[5] A. El-Jaroudi and J. Makhoul, "Discrete all-pole modeling," *IEEE Trans. on signal processing*, vol. 39, no. 2, pp. 411–423, Feb. 1991.

[6] T. Nakajima and T. Suzuki, "Speech power spectrum envelope (PSE) analysis based on the F0 interval sampling," *IEICE Technical Report*, vol. SP86, no. 94, pp. 55–62, Jan. 1987, (in Japanese).

[7] 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.

[8] O. Cappé, J. Laroche, and E. Moulines, "Regularized estimation of cepstrum envelope from discrete frequency points," in *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*, Oct. 1995, pp. 213–216.

[9] H. Kawahara, "Speech representation and transformation using adaptive interpolation of weighted spectrum: vocoder revisited," in *Proc. ICASSP97*, Apr. 1997, vol. 2, pp. 1303–1306.

[10] R. D. Kent and C. Read, *The Acoustic Analysis of Speech*, Singular Publishing Group, 1992.

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

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

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

[14] E. Moulines and F. Charpentier, "Pitch synchronous waveform processing techniques for text-to-speech synthesis using diphones," *Speech Communication*, vol. 9, no. 5, pp. 453–467, 1990.

[15] X. Huang, A. Acero, and H.-W. Hon, *Spoken Language Processing — A Guide to Theory, Algorithm, and System Development*, Prentice Hall, 2001.

[16] Y. Shiga, Y. Hara, and T. Nitta, "A novel segment-concatenation algorithm for a cepstrum-based synthesizer," in *Proc. ICSLP94*, 1994, vol. 4, pp. 1783–1786.

[17] 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.

[18] Y. Shiga and S. King, "Estimation of voice source and vocal tract characteristics based on multi-frame analysis," in *Proc. Eurospeech2003*, Sept. 2003, vol. 3, pp. 1749–1752.