Estimating Detailed Spectral Envelopes Using Articulatory Clustering

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

This paper presents an articulatory-acoustic mapping where detailed spectral envelopes are estimated. During the estimation, the harmonics of a range of F_0 values are derived from the spectra of multiple voiced speech signals vocalized with similar articulator settings. The envelope formed by these harmonics is represented by a cepstrum, which is computed by fitting the peaks of all the harmonics based on the weighted least square method in the frequency domain. The experimental result shows that the spectral envelopes are estimated with the highest accuracy when the cepstral order is 48-64 for a female speaker, which suggests that representing the real response of the vocal tract requires high-quefrency elements that conventional speech synthesis methods are forced to discard in order to eliminate the pitch component of speech.

1. Introduction

Speech representation derived from spectral peaks at harmonic frequencies of voiced speech has attracted attention widely in speech technology. Gu et al. [1] have recently proposed novel 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. In the field of speech coding, such harmonic-based spectral-envelope estimation has been used since the early 90's for perceptually efficient encoding [2, 3].

It must be noted that, whereas the harmonic peaks have an important role in human auditory perception, only those peaks reflect the vocal tract transfer function (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 [4, 5] 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* [6], also exploits harmonic peaks, into which a bilinear surface is interpolated in the three-dimensional space composed of time, frequency and spectral power.

It has been pointed out, however, that the harmonic structure interferes with identifying spectral envelopes that precisely reflect VTTFs [7]. Hence 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.

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, we deal with the following two related points in this paper: 1) a mapping of articulation to the VTTF using the actual measurement of articulators; 2) accurate VTTF estimation based on the articulatory data for high-quality speech synthesis.

In order to resolve the above problem of spectral envelope estimation, we have proposed a method based on the diverse harmonic structures of multiple short-time speech signals produced under almost the same articulatory condition [8]. 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: piecewise constant mapping and piecewise linear mapping. After examining these functions theoretically, we closely investigate the performance of both mappings through experiments.

2. Articulatory-acoustic mapping

2.1. Articulatory data

The data used in this study is a *MOCHA (Multi-CHannel Articulatory)* corpus [9]. 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 (articulatory vector) 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.

2.2. 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, and in addition, is a frequency-domain representation and thus has good interpolation (smoothing) properties. Because of these merits the cepstrum is widely applied in the field of speech technology (e.g. [10]).

2.3. Clustering in the articulatory space

We partition the articulatory space and obtain a mapping function for each cluster so that articulatory-acoustic conversion becomes possible. Specifically, after normalising each dimension of the articulatory vectors, we apply *LBG clustering* [11] to all the normalised vectors and divide them into K clusters (articulatory clusters), C^i (i = 1, 2, 3, ..., K), and then estimate articulatory-acoustic mapping functions for each cluster.

3. Piecewise Constant Mapping

3.1. Outline

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 therefore be assumed constant within the cluster. Under this assumption, the problem is reduced to estimating one unique spectral envelope for every cluster, and accordingly we can collect the different harmonic structures of the multiple frames to form a spectral envelope.

3.2. Estimating the envelopes of amplitude spectra

Let us determine a cepstrum which best fits the 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 [4, 5] to the analysis of multiple frames.

Let $a_k^{(l)}$ denote an observed 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

$$E_a^{(i)} = \sum_{k \in C^i} (\boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_a^{(i)})^T \mathbf{W}_k (\boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_a^{(i)}) \qquad (1)$$

where the vector $\boldsymbol{c}_a^{(i)}$ indicates a cepstrum vector for cluster i, $\boldsymbol{c}_a^{(i)} = [\ c_a^{(i)}[0], c_a^{(i)}[1], c_a^{(i)}[2], \cdots, c_a^{(i)}[p]\ ]^T$, and $\boldsymbol{y}_k$ is an N_k dimension vector, $\boldsymbol{y}_k = [\ a_k^{(1)} - d_k, a_k^{(2)} - d_k, a_k^{(3)} - d_k, \cdots, a_k^{(N_k)} - d_k\ ]^T$. The offset d_k adjusts the overall amplitude of each frame so as to minimise the error $E_a^{(i)}$. The matrices $\mathbf{P}_k$ and $\mathbf{W}_k$ are as follows:

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

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

where $\Omega_k^{(l)} = 2\pi f_k^{(l)} T$. 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. Equation (1) can be solved for the cepstrum $\boldsymbol{c}_a$ by reducing it to a problem of weighted least squares. The offset d_k is then calculated as

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

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 (2); 3) Calculate $E_a^{(i)}$ using (1) and terminate the procedure if $E_a^{(i)}$ converges; 4) Find $\boldsymbol{c}_a^{(i)}$ by solving the normal equation; 5) Substitute 0 for $c_a^{(i)}[0]$ (power normalization); 6) Return to step 2.

3.3. Estimating the envelopes of phase spectra

The spectral envelopes of phase can be 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 in the speech frame k. Then, the sum of squared approximation errors for the phases of all the harmonics of frames belonging to cluster i is expressed as

$$E_p^{(i)} = \sum_{k \in C^i} (\boldsymbol{\vartheta}_k - \mathbf{Q}_k \boldsymbol{c}_p^{(i)})^T \mathbf{W}_k (\boldsymbol{\vartheta}_k - \mathbf{Q}_k \boldsymbol{c}_p^{(i)}) \qquad (3)$$

where the vector $\boldsymbol{c}_p^{(i)}$ indicates a cepstral vector for cluster i, $\boldsymbol{c}_p^{(i)} = [\ c_p^{(i)}[1], c_p^{(i)}[2], c_p^{(i)}[3], \cdots, c_p^{(i)}[p]\ ]^T$. The matrix $\mathbf{Q}_k$ is as follows:

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

The vector $\boldsymbol{\vartheta}_k$ is an N_k-dimensional vector, $\boldsymbol{\vartheta}_k = [\ \vartheta_k^{(1)}, \vartheta_k^{(2)}, \vartheta_k^{(3)}, \cdots, \vartheta_k^{(N_k)}\ ]^T$, where $\vartheta_k^{(l)}$ 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. 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

$$\varphi_i(f_k^{(l)}) = \frac{\phi_i(f_k^{(l)})}{\left|\phi_i(f_k^{(l)})\right|} \qquad (4)$$

$$\phi_i(f) = \frac{\sum_{k \in C^i} \sum_{l=1}^{N_k} G(f_k^{(l)} - f) e^{j(\theta_k^{(l)} - 2\pi f_k^{(l)} \tau_k)} / N_k}{\sum_{k \in C^i}^{M} \sum_{l=1}^{N_k} G(f_k^{(l)} - f) / N_k}. \qquad (5)$$

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

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

Equation (3) can be solved for the cepstrum $\boldsymbol{c}_p^{(i)}$ by reducing it to a problem of weighted least squares. 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 (3) and terminate the procedure if $E_p^{(i)}$ converges; 4) Find $\boldsymbol{c}_p^{(i)}$ by solving the normal equation; 5) Return to step 2.

4. Piecewise Linear Mapping

4.1. Outline

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 a noticeable error. 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.

4.2. Piecewise linear approximation

The cepstra $\boldsymbol{c}_a^{(i)}$ and $\boldsymbol{c}_p^{(i)}$ in (1) and (3) are represented by the linear transformation of L-dimensional articulatory vector $\boldsymbol{x}_k$ as follows:

$$\boldsymbol{c}_a^{(i)} = \boldsymbol{q}^{(i)} + \mathbf{U}^{(i)}\boldsymbol{x}_k, \;\; \boldsymbol{c}_p^{(i)} = \boldsymbol{r}^{(i)} + \mathbf{V}^{(i)}\boldsymbol{x}_k \tag{7}$$

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

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

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

The problem is now reduced to finding these matrices and vectors. Substituting (7) into (1) and (3) and rewriting the formulae, we obtain the following equations:

$$E_a^{(i)} = \sum_{k \in C^i} (\boldsymbol{y}_k - \boldsymbol{\Gamma}_k \boldsymbol{u}_k^{(i)})^T \mathbf{W}_k (\boldsymbol{y}_k - \boldsymbol{\Gamma}_k \boldsymbol{u}_k^{(i)}) \tag{8}$$

$$E_p^{(i)} = \sum_{k \in C^i} (\boldsymbol{\vartheta}_k - \boldsymbol{\Delta}_k \boldsymbol{v}_k^{(i)})^T \mathbf{W}_k (\boldsymbol{\vartheta}_k - \boldsymbol{\Delta}_k \boldsymbol{v}_k^{(i)}) \tag{9}$$

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

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

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

Having the same form as (1) and (3), (8) and (9) 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 procedure

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 [12], 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.

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

$$D_a = \frac{20}{\ln 10}\sqrt{\frac{1}{M}\sum_k (\boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_a^{(R(k))})^T \mathbf{W}_k (\boldsymbol{y}_k - \mathbf{P}_k \boldsymbol{c}_a^{(R(k))})}$$

$$D_p = \sqrt{\frac{1}{M}\sum_k (\boldsymbol{\vartheta}_k - \mathbf{Q}_k \boldsymbol{c}_p^{(R(k))})^T \mathbf{W}_k (\boldsymbol{\vartheta}_k - \mathbf{Q}_k \boldsymbol{c}_p^{(R(k))})}$$

$$i = R(k) \Longleftrightarrow \text{frame } k \in C^i$$

where M denotes the number of frames evaluated. These distortions for the training set were computed in the training process using equation (1) and (3) for the piecewise constant mapping, and using equation (8) and (9) for the piecewise linear mapping. We calculated the distortions for the test data as follows: first, the nearest neighbour method chooses one of the articulatory clusters based on the Euclidean distance between each of the cluster centroid vectors and each of the articulatory vectors to be tested, and then the distortions are calculated using the cepstral coefficients of the chosen cluster for the piecewise constant mapping, and 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 with 0 Hz mean and 4 kHz standard deviation, and a Gaussian window with 100 Hz standard deviation, respectively.

5.2. Results and discussion

Shown in figure 1 is the estimation error of the piecewise constant mapping. As in this figure, the error for the test data set has the minimum values in the case of cepstral order 48 (3.0 ms in quefrency) and 512 clusters for amplitude, and in the case of order 64 (4.0 ms) and 256 clusters for phase, where the error values are 5.56 dB and 0.807 rad. Figure 2 shows the result of the piecewise linear mapping. The error has 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 error values are 5.18 dB and 0.778 rad.

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 estimation errors were minimised. These results indicate that representing spectral envelopes reflecting real VTTFs requires cepstral coefficients of high quefrency range, which are usually discarded in conventional speech synthesis to eliminate the pitch component of speech.

Also, as is evident from figure 1 and 2, the piecewise linear mapping is more accurate and requires a smaller number of clusters than the piecewise constant mapping. The piecewise linear mapping is therefore more suitable than the piecewise constant mapping to represent relationship between the articulator positions and the cepstrum, and we may consider that the relationship is locally almost linear.

Moreover, for both proposed mapping functions the error of the phase spectrum indicates the variance of phase in each frequency band. This is useful information especially for multiband-type speech synthesis to control the phase of each of the frequency bands for the purpose of reducing the buzziness.

6. 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.

We have confirmed that applying a source-filter separation [13], 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 a Gaussian mixture model, which we

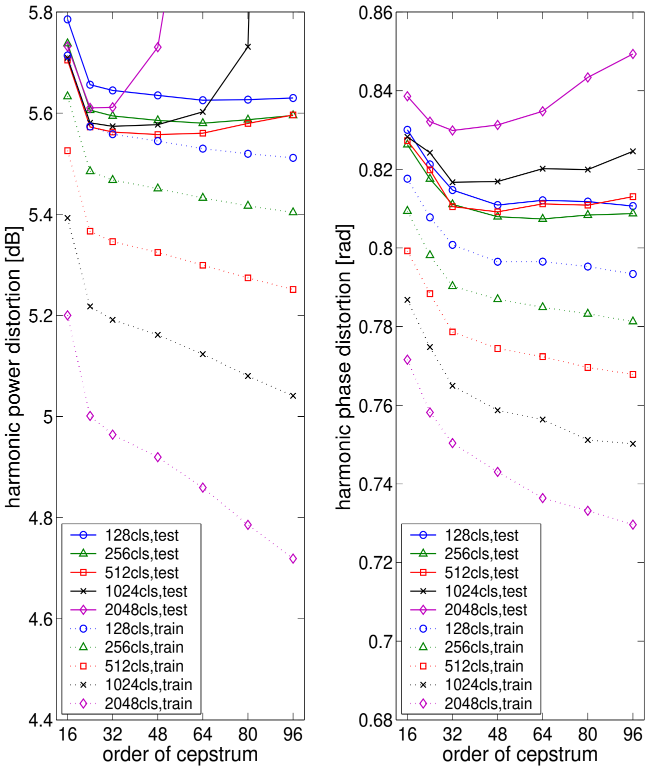

Figure 1: *Harmonic distortion vs. order of cepstrum, in the case of the piecewise constant mapping*

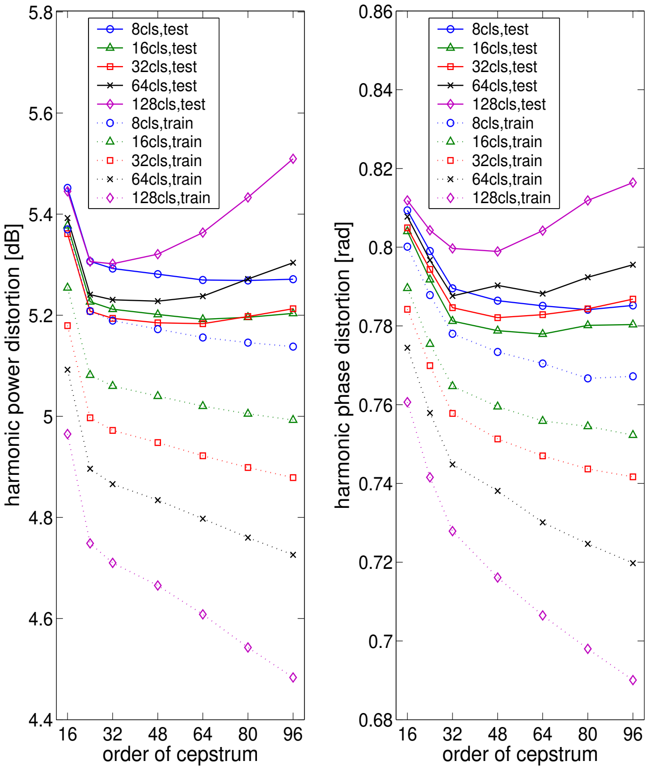

Figure 2: *Harmonic distortion vs. order of cepstrum, in the case of the piecewise linear mapping*

have already employed for the purpose of reducing acoustical discontinuity of output speech at the boundaries of clusters. Furthermore, we expect that applying this harmonic-peak smoothing to multiple *blocks* of speech in the three-dimensional space including time axis, in the same manner as of STRAIGHT [6], will enable our articulation-to-speech synthesis to produce temporally smoother, more natural-sounding speech.

Acknowledgements

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

References

[1] 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.

[2] 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.

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

[4] 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).

[5] 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.

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

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

[8] 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.

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

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

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

[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] Y. Shiga and S. King, "Source-filter separation for articulation-to-speech synthesis," in *Proc. ICSLP2004*, Jeju, Korea, Oct. 2004.