2008 International Workshop on Nonlinear Circuits and Signal Processing
NCSP'08, Gold Coast, Australia, March 6-8, 2008

BANDWIDTH EXTENSION OF CELLULAR PHONE SPEECH BASED ON MAXIMUM LIKELIHOOD ESTIMATION WITH GMM

Wataru Fujitsuru†, Hidehiko Sekimoto†, Tomoki Toda†, Hiroshi Saruwatari† and Kiyohiro Shikano†

†Graduate School of Information Science, Nara Institute of Science and Technology, Nara, 631–0101 Japan
E-mail: † wataru-f, hidehi-s, tomoki, sawatari, shikano @is.naist.jp

Abstract

Bandwidth extension is a useful technique for reconstructing wideband speech from only narrowband speech. As a typical conventional method, a bandwidth extension algorithm based on minimum mean square error (MMSE) with a Gaussian mixture model (GMM) has been proposed [ ]. Although the MMSE-based method has reasonably high conversion-accuracy, there still remain some problems to be solved: 1) inappropriate spectral movements are caused by ignoring a correlation between frames, and 2) the converted spectra are excessively smoothed by the statistical modeling. In order to address these problems, we propose a bandwidth extension algorithm based on maximum likelihood estimation (MLE) considering dynamic features and the global variance (GV) with a GMM. A result of a subjective test demonstrates that the proposed algorithm outperforms the conventional MMSE-based algorithm.

1. Introduction

The use of cellular phones enables us to easily communicate with each other through speech. In general, the cellular phone speech is limited to a narrowband signal up to 3.4 kHz. Although narrowband speech is capable of intelligible communication, its sound quality is not high enough. Specifically considerable quality degradation is observable at several phonemes such as fricatives and plosives which have important energy distribution beyond 3.4 kHz. In order to realize higher-quality speech communication, a wideband speech codec has been developed [ ]. Essentially it needs more information than the narrowband speech codec. It is no doubtful that to realize wideband speech communication while not increasing information is more convenient.

Bandwidth extension is a useful technique for reconstructing wideband speech from only narrowband speech. Several statistical approaches to bandwidth extension based on a spectral mapping have been studied. A codebook mapping method [ ] is an approach based on hard clustering and discrete mapping. A reconstructed feature vector at each frame is determined by quantizing the narrowband speech feature vector to the nearest centroid vector of the narrowband codebook and substituting it with a corresponding wideband centroid vector of the mapping codebook. One of more sophisticated approaches allowing soft clustering and continuous mapping is a probabilistic bandwidth extension method based on a Gaussian mixture model (GMM) [ ]. The basic mapping algorithm has originally been proposed for voice conversion [ ]. Most of the conventional GMM-based methods perform the mapping based on MMSE criterion [,]. It has been reported that the mapping performance is further improved by modeling dynamic characteristics of a spectral sequence [,]. Moreover, an approach of combining mapping and coding processes has been studied [ ].

Recently the performance of voice conversion with a GMM was significantly improved by maximum likelihood estimation (MLE) considering dynamic features and the global variance (GV) [ ]. It is expected that it also causes the performance improvement of bandwidth extension. This paper proposes bandwidth extension based on MLE considering dynamic features and the GV. The effectiveness of considering dynamic features and the GV is demonstrated by a subjective evaluation.

This paper is organized as follows. Section 2 describes the conventional method of the bandwidth extension algorithm based on GMM. Section 3 describes the proposed bandwidth extension based on MLE. Section 4 describes the flow of the bandwidth extension. Section 5 describes an experimental evaluation. Finally, we summarize this paper in Section 6.

2. MMSE-based bandwidth extension [ ]

2.1. Training

Let $\boldsymbol{x}_t$ and $\boldsymbol{y}_t$ be D_x-dimensional narrowband and D_y-dimensional wideband feature vectors at frame t, respectively. Joint probability density of the narrowband and wideband feature vectors is modeled by a GMM as follows:

$$P(\boldsymbol{z}_t|\theta) = \sum_{m=1}^{M} \omega_m \mathcal{N}(\boldsymbol{z}_t; \boldsymbol{\mu}_m^{(z)}, \boldsymbol{\Sigma}_m^{(z)}), \tag{1}$$

where $\boldsymbol{z}_t$ is a joint vector $[\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top]^\top$. The notation $^\top$ denotes transposition of the vector. The number of mixture components is M. The weight of the m-th mixture component is ω_m. The normal distribution with $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is denoted as $\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$. A parameter set of the GMM is θ, which consists of weights, mean vectors and the covariance matrices for individual mixture components. The mean vector $\boldsymbol{\mu}_m^{(z)}$ and the covariance matrix $\boldsymbol{\Sigma}_m^{(z)}$ of the m-th mixture component are written as

$$\boldsymbol{\mu}_m^{(z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(z)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}, \tag{2}$$

where $\boldsymbol{\mu}_m^{(x)}$ and $\boldsymbol{\mu}_m^{(y)}$ are the mean vector of the m-th mixture component for the narrowband and that for the wideband, respectively, and $\boldsymbol{\Sigma}_m^{(xx)}$ and $\boldsymbol{\Sigma}_m^{(yy)}$ are the covariance matrix of the m-th mixture component for the narrowband and that for the wideband, respectively. The matrix $\boldsymbol{\Sigma}_m^{(yx)}$ is the cross-covariance matrix of the m-th mixture component for the narrowband and wideband.

The GMM is trained with the EM algorithm using the joint vectors in a training set. This training method robustly estimates model parameters compared with the least squares estimation [ ].

2.2. MMSE-based conversion

The conventional method performs the conversion based on MMSE as follows:

$$\begin{aligned}\hat{\boldsymbol{y}}_t &= E[\boldsymbol{y}_t|\boldsymbol{x}_t] \\ &= \int P(\boldsymbol{y}_t|\boldsymbol{x}_t,\theta)\,\boldsymbol{y}_t d\boldsymbol{y}_t \\ &= \sum_{m=1}^{M} P(m|\boldsymbol{x}_t,\theta)\boldsymbol{E}_{m,t}, \end{aligned} \tag{3}$$

where

$$P(m|\boldsymbol{x}_t,\theta) = \frac{\omega_m \mathcal{N}(\boldsymbol{x}_t;\boldsymbol{\mu}_m^{(y)},\boldsymbol{\Sigma}_m^{(xx)})}{\sum_{j=1}^{M}\omega_j \mathcal{N}(\boldsymbol{x}_t;\boldsymbol{\mu}_j^{(x)},\boldsymbol{\Sigma}_j^{(xx)})}, \tag{4}$$

$$\boldsymbol{E}_{m,t} = \boldsymbol{\mu}_m^{(y)} + \boldsymbol{\Sigma}_m^{(yx)}(\boldsymbol{\Sigma}_m^{(xx)})^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(y)}). \tag{5}$$

2.3. Problems

Although MMSE has reasonably high conversion-accuracy, there still remain some problems to be solved: 1) inappropriate spectral movements are caused by ignoring a correlation between frames, and 2) the converted spectra are excessively smoothed by the statistical modeling.

3. MLE-based bandwidth extension

In order to solve two main problems of the conventional method, we proposed the bandwidth extension based on MLE considering dynamic features and the GV. Inter-frame correlation is considered for realizing the converted parameter trajectory with appropriate dynamic characteristics. The over-smoothing effects are alleviated by considering the GV capturing one of characteristics of the trajectory.

3.1. MLE considering dynamic features (ML) [ ]

We use $2D_x$-dimensional narrowband speech feature vector $\boldsymbol{X}_t = [\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top]^\top$ and $2D_y$-dimensional wideband speech feature vector $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$, consisting of static and dynamic features at frame t. Their time sequences are written as $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$ and $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \boldsymbol{Y}_2^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$, respectively. A time sequence of the converted static feature vectors $\hat{\boldsymbol{y}} = \left[\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top\right]^\top$ is determined as follows:

$$\begin{aligned}\hat{\boldsymbol{y}} &= \arg\max_{\boldsymbol{y}} P(\boldsymbol{Y}|\boldsymbol{X},\Theta) \\ &\quad subject to \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \end{aligned} \tag{6}$$

where $\boldsymbol{W}$ is a conversion matrix that extends a sequence of the static feature vectors $\boldsymbol{y}$ into that of the static and dynamic features vectors.

In order to effectively reduce the computational cost, we approximate the likelihood as follows,

$$P(\boldsymbol{Y}|\boldsymbol{X},\Theta) \simeq P(\boldsymbol{m}|\boldsymbol{X},\Theta)P(\boldsymbol{Y}|\boldsymbol{X},\boldsymbol{m},\Theta), \tag{7}$$

where $\boldsymbol{m}$ is a mixture component sequence $[m_1, m_2, \cdots, m_t]$. After determining the sub-optimum mixture sequence $\hat{\boldsymbol{m}}$ written as

$$\hat{\boldsymbol{m}} = \arg\max_{\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{X},\Theta), \tag{8}$$

we determine $\hat{\boldsymbol{y}}$ that maximizes the approximated likelihood function as follows:

$$\begin{aligned}\hat{\boldsymbol{y}} &= \arg\max_{\boldsymbol{y}} P(\hat{\boldsymbol{m}}|\boldsymbol{X},\Theta)P(\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{m}},\Theta) \\ &= \left(\boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} \boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)}, \end{aligned} \tag{9}$$

where

$$\boldsymbol{E}_{\hat{\boldsymbol{m}}}^{(Y)} = \left[\boldsymbol{E}_{\hat{m}_1,1}^{(Y)}, \cdots, \boldsymbol{E}_{\hat{m}_t,t}^{(Y)}, \cdots, \boldsymbol{E}_{\hat{m}_T,T}^{(Y)}\right], \tag{10}$$

$$\boldsymbol{D}_{\hat{\boldsymbol{m}}}^{(Y)^{-1}} = \mathrm{diag}\left[\boldsymbol{D}_{\hat{m}_1}^{(Y)^{-1}}, \cdots, \boldsymbol{D}_{\hat{m}_t}^{(Y)^{-1}}, \cdots, \boldsymbol{D}_{\hat{m}_T}^{(Y)^{-1}}\right]. \tag{11}$$

A GMM parameter set Θ is estimated in advance with training data in the same manner as the converted method.

3.2. MLE considering GV (MLGV) [ ]

The GV of the static feature vectors in each utterance is written as

$$\boldsymbol{v}(\boldsymbol{y}) = [v(1), v(2), \cdots, v(D)]^\top \tag{12}$$

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \frac{1}{T}\sum_{\tau=1}^{T} y_\tau(d)\right)^2 \tag{13}$$

where $y_t^{(d)}$ is the d-th component of the target static feature at frame t.

The following likelihood function consisting of two probability densities for a sequence of the wideband feature vectors and for the GV of the wideband static feature vectors is maximized,

$$\begin{aligned} L &= \alpha \log P(\boldsymbol{Y}|\boldsymbol{X},\hat{\boldsymbol{m}},\Theta) + \log P(\boldsymbol{v}(\boldsymbol{y})|\Theta_v), \\ &\quad subject\ to\ \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \end{aligned} \tag{14}$$

where $P(\boldsymbol{v}(\boldsymbol{y})|\Theta_v)$ is modeled by the normal distribution. A set of model parameters Θ_v consists of the mean vector $\boldsymbol{u}^{(v)}$ and the covariance matrix $\boldsymbol{\Sigma}^{(vv)}$ for the GV vector $\boldsymbol{v}(\boldsymbol{y})$, which is also estimated in advance with training data. The constant α is a likelihood weight. In this paper, it is set to $\frac{1}{2T}$. Note that the GV likelihood works as a penalty term for the over-smoothing.

In order to maximize the likelihood L with respect to $\boldsymbol{y}$, we employ a steepest descent algorithm using the first derivative,

$$\begin{aligned}\frac{\partial L}{\partial \boldsymbol{y}} &= \alpha\left(-\boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}}\boldsymbol{W}\boldsymbol{y} + \boldsymbol{W}^\top \overline{\boldsymbol{D}^{(Y)^{-1}}\boldsymbol{E}^{(Y)}}\right) \\ &\quad + \left[\boldsymbol{v}_1'^\top, \boldsymbol{v}_2'^\top, \cdots, \boldsymbol{v}_t'^\top, \cdots, \boldsymbol{v}_T'^\top\right]^\top, \end{aligned} \tag{15}$$

$$\boldsymbol{v}_t' = [v_t'(1), v_t'(2), \cdots, v_t'(d), \cdots, v_T'(D)]^\top, \tag{16}$$

$$v_t'(d) = -\frac{2}{T}\boldsymbol{p}_v^{(d)^\top}(\boldsymbol{v}(\boldsymbol{y}) - \boldsymbol{\mu}_v)\left(y_t(d) - \frac{1}{T}\sum_{\tau=1}^{T} y_\tau(d)\right). \tag{17}$$

The vector $\boldsymbol{p}_v^{(d)}$ is the d-th column vector of $\boldsymbol{P}_v = \boldsymbol{\Sigma}^{(vv)^{-1}}$.

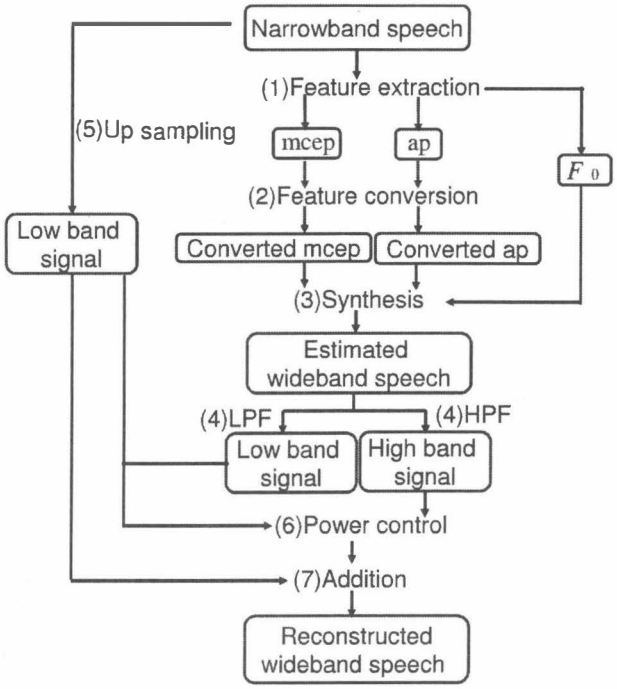

Figure 1: Bandwidth extension system. "mcep" denotes the mel-cepstrum and "ap" denotes the aperiodic component.

4. Details of bandwidth extension process

Figure 1 shows a flow of the bandwidth extension.

Step 1 Extracting F_0, mel-cepstrum and aperiodic component [ ] as speech features from the narrowband speech signal.

Step 2 Converting mel-cepstrum and aperiodic component of the narrowband speech into those of the wideband speech.

Step 3 Generating STRAIGHT mixed excitation [ ] using the extracted F_0 and the converted aperiodic component, and then synthesizing the estimated wideband speech with MLSA filter [ ] based on the converted mel-cepstrum.

Step 4 The estimated wideband speech is separated into a low-band signal and a high-band signal with a low-pass filter (LPF) and a high-pass filter (HPF).

Step 5 The input narrowband speech is converted into a low-band signal with up-sampling.

Step 6 Power of the estimated high-band signal is adjusted so that power of the estimated low-band signal is equal to that of the input low-band signal.

Step 7 Reconstructing the wideband speech by adding the resulting high-band signal and the low-band signal.

5. Experimental evaluations

In order to demonstrate the effectiveness of the proposed method, we conducted a subjective evaluation.

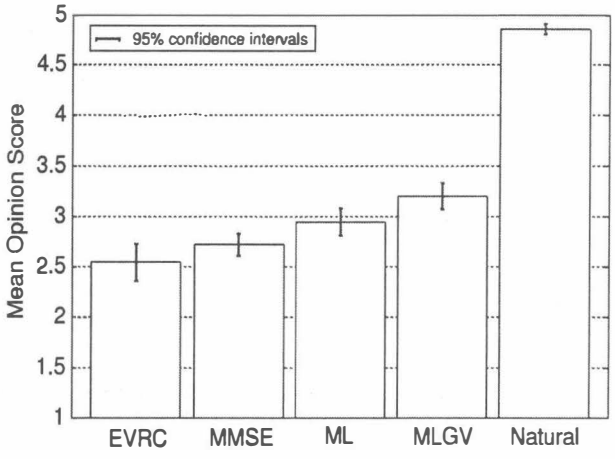

Figure 2: Result of subjective evaluation.

5.1. Experimental conditions

We used 16 kHz sampled natural speech of 4 Japanese speakers (2 males, 2 females) as the wideband speech. The 3.4 kHz narrowband speech was prepared by down-sampling the wideband speech and then passing it through EVRC (Enhanced Variable Rate Codec) [ ]. The training data was 50 sentences from subset A of ATR's phonetically balanced sentence database. The evaluation data was 50 sentences from subset B of ATR's phonetically balanced sentence database. For narrowband, we used the 16-dimensional mel-cepstral coefficients from the mel-cepstral analysis [ ]. For wideband, we used the 24-dimensional mel-cepstral coefficients from the STRAIGHT analysis [ ]. We used the averaged aperiodic components [ ] on three frequency bands (0 to 1, 1 to 2, 2 to 4 kHz) for narrowband and those on five frequency bands (0 to 1, 1 to 2, 2 to 4, 4 to 6, 6 to 8 kHz) for wideband. The frame shift was 5 ms. The number of mixture components of the GMM for mel-cepstral conversion was set to 64. The number of mixture components of the GMM for aperiodic components conversion was set to 4. Speaker dependent models were evaluated.

We conducted an opinion test on speech quality. An opinion score was set to a 5-point scale (5: excellent, 4: good, 3: fair, 2: poor, 1: bad). The evaluated speech samples consisted of EVRC, MMSE, ML, MLGV, and wideband natural speech (Natural). The listeners were eight Japanese adult man and woman.

5.2. Experimental results

Figure 2 shows a result of the opinion test. There is no significant difference between EVRC and MMSE. On the other hand, the proposed method ML is significantly better than both EVRC and MMSE. The reconstructed wideband speech with the highest speech quality was obtained by considering the GV as well in the proposed method. An example of spectral sequences of the narrowband speech, the reconstructed speech and the natural wideband speech is shown in Figure 3. The proposed method estimates spectral envelopes considering inter-frame correlation while alleviating the over-smoothing effects.

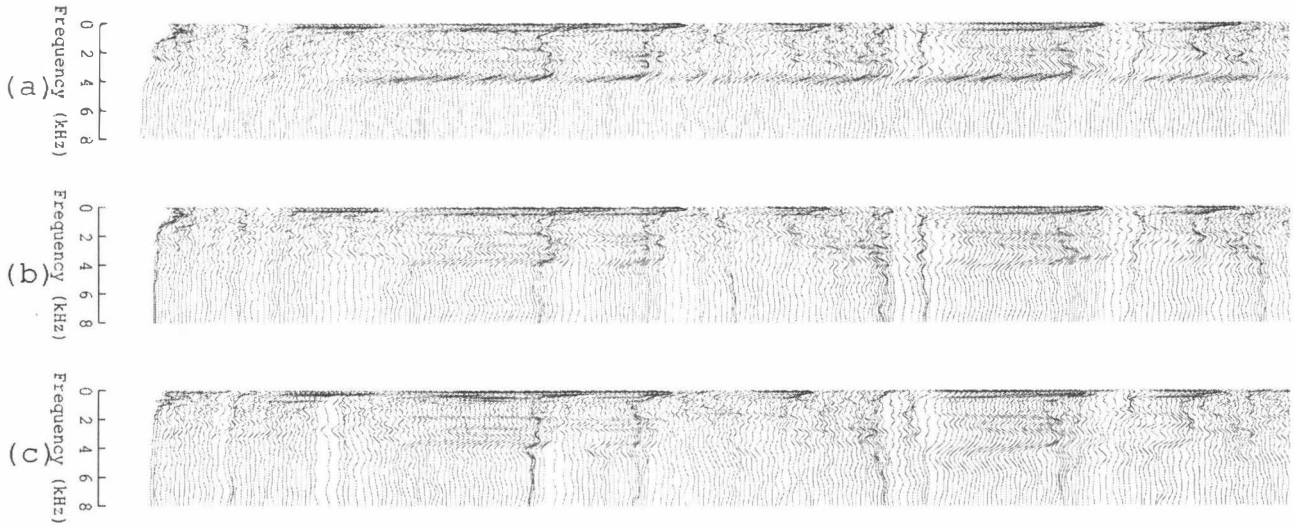

Figure 3: An example of spectral sequences of narrowband speech, (a) converted speech by the ML using the GV , (b) spectra of natural speech, (c) for a sentence fragment, "/ jy u: j i ts U sh I t e i k u /".

6. Conclusions

We proposed bandwidth extension based on maximum likelihood estimation (MLE) with a Gaussian mixture model (GMM) considering dynamic features and the global variance (GV). A result of the subjective evaluation demonstrated that the speech quality of the narrowband speech is significantly improved by the proposed method. We plan to deal with a speaker independent model, online processing and noise robustness.

7. Acknowledgments

This research was supported in part by e-Society project and KDDI collaborative research.

References

[1] B. Bessette, R. Salami, R. Lefebvre, M. Jelinek, J. Potola-Pukkila, J. Vainio, H. Mikkola and K.Jarvinen, "The adaptive multirate wideband speech codec (AMR-WB)," *IEEE Trans.*, Vol. 10, No. 8, pp. 620–636, 2002.

[2] Y. Yoshida, and M. Abe, "An Algorithm to Reconstruct Wideband Speech From Narrowband Speech Based on Codebook Mapping," *Proc. ICSLP94*, pp. 1591–1593, 1994.

[3] K.Y. Park and H.S. Kim. "Narrowband to wideband conversion of speech using GMM based transformation," *Proc. ICSLP*, pp. 1847–1850, Istanbul, June, 2000.

[4] Y. Stylianou, O. Cappe, E. Moulines, "Coutinuous probabilistic transform for voice conversion," *IEEE Trans.*, Speech and Audio Processing, Vol. 6, No. 2, pp. 131–142, 1998.

[5] M.L. Seltzer, A. Acero, and J. Droppo, "Robust Bandwidth Extension of Noise-corrupted Narrowband Speech," *Proc. ICSLP*, pp. 1509–1512, 2005.

[6] S. Yao and C.F. Chan, "Block-based Bandwidth Extension of Narrowband Speech Signal by using CDHMM," *Proc. ICASSP*, pp. I793–I796, 2005.

[7] S.Y Yao and C.F. Chan, "Speech bandwidth enhancement using state space speech dynamics," *Proc. ICASSP2006*, pp. I489–I492, 2006.

[8] Y. Agiomyrgiannakis and Y. Stylianou, "Combined Estimation/coding of Highband Spectral Envelopes for Speech Spectrum Expansion," *Proc. ICASSP2004*, pp. 469–472, 2004.

[9] T. Toda, A.W. Black, and K. Tokuda, "Voice Conversion Based on Maximum-Likelihood Estimation of Spectral Parameter Trajectory," *IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2234, 2007.

[10] A. Kain and M.W. Macon. "Spectral voice conversion for text-to-speech synthesis," *Proc. ICASSP*, Seattle, U.S.A,. pp. 285–288, May 2004.

[11] H. Kawahara, Jo Estill and O. Fujimura, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," *Proc. MAVEBA*, Sep. 13–15, Firentze Italy, 2001.

[12] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, Vol. 27, No. 3-4, pp. 187–207. 1999.

[13] T. Fukada, K. Tokuda, T. Kobayashi, and S. Imai, "An adaptive algorithm for mel-cepstral analysis of speech," *Proc. ICASSP*, Vol 1, pp. 137–140, San Francisco, USA, Mar. 1992.

[14] T.V. Ramabadran, J.P. Ashley and M.J. McLaughlin, "Background Noise Suppression for Speech Enhancement and Coding,", *IEEE Workshop on Speech Coding and Tel.*, Pocono Manor, PA, pp. 43–44, 1997.