GMM-based bandwidth extension using sub-band basis spectrum model

Yamato Ohtani, Masatsune Tamura, Masahiro Morita and Masami Akamine

Knowledge Media Laboratory, Corporate Research & Development Center, Toshiba Corporation, Japan

{yamato.ohtani,masatsune.tamura,masahiro.morita,masa.akamine}@toshiba.co.jp

10.21437/Interspeech.2014-534

Abstract

This paper describes a novel GMM-based bandwidth extension (BWE) method based on a sub-band basis spectrum model (SBM), in which each dimensional component represents a specific acoustic space in the frequency domain. The proposed method can achieve the BWE from a speech data with an arbitrary frequency bandwidth while the conventional methods perform the conversion from a fixed narrowband data. In the proposed method, we train a GMM with SBM parameters extracted from wideband spectra in advance. An input signal with a limited frequency band is converted into a wideband signal by estimating high-band SBM components from low-band SBM components of the input signal based on the GMM. The results of some objective and subjective evaluations show that the proposed method extends bandwidth of speech data robustly.

Index Terms: speech synthesis, voice conversion, bandwidth extension, sub-band basis spectrum model, Gaussian mixture model.

1. Introduction

In recent years, speech synthesis techniques have been used for several applications such as speaking-aid systems [1, 2] and speech translation systems [3, 4]. These applications often require such synthesis systems that can generate users' voices. In general, recorded speech data from a user is necessary for constructing a speech synthesis system of the user's voice. The recorded speech data may not have frequency bands wide enough to realize high-quality speech synthesis because of limitations of users' recording devices such as microphones and AD converters. In these devices, high-band components are often lost. For the purpose of high-quality speech synthesis, it is therefore important to compensate for the lost frequency band of the speech data.

Bandwidth extension (BWE) has been studied widely [5], which is used in mobile phones to restore wideband audio signals to improve intelligibility. BWE methods using a Gaussian mixture model (GMM) have been proposed [6, 7]. They are based on voice conversion techniques [8, 9, 10]. In this framework, a GMM is trained with parallel data consisting of narrowband and wideband acoustic features (e.g. mel-cepstrum) in advance. In reconstruction, an input narrowband acoustic feature sequence is converted into the wideband one. In order to apply the conventional GMM-based method to an arbitrary bandwidth signal we need multiple parallel data of narrow band and wide band signals, and to train multiple GMMs in advance.

In this paper, we propose a GMM-based BWE using a sub-band basis spectrum model (SBM) [11] as the acoustic feature. SBM represents amplitude and phase characteristics of speech spectra by a linear combination of sub-band basis vectors, which are determined based on the result of a sparse coding method [12]. Each component of the SBM parameter which denotes weights for sub-band basis vectors represents a specific acoustic space in the frequency domain and it can be applied for frequency domain processing easily. By focusing on above characteristics, the proposed method trains a single GMM with non-parallel data consisting only wideband SBM parameters in advance, and then converts any input signal with an arbitrary bandwidth into a wideband signal by estimating high-band SBM parameters from the input low-band SBM parameters with the trained GMM.

The rest of this paper is organized as follows. Section 2 shows related techniques of the proposed method. The details of the proposed method are described in Section 3. Section 4 shows the results of evaluation experiments. Finally, we conclude this paper in Section 5.

2. Related work

2.1. Sub-band basis spectrum model

Figure 1 shows an overview of the SBM. $\boldsymbol{s}_t$ is a K-dimensional speech spectrum vector of the t^{th} frame. $\boldsymbol{c}_t$ is a N-dimensional SBM parameter vector which represents weights of the sub-band basis vectors. $\boldsymbol{s}_t$ can be described by $\boldsymbol{c}_t$ and sub-band basis vectors $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \cdots, \boldsymbol{\phi}_N]$ as follows:

$$\boldsymbol{s}_t = \boldsymbol{\Phi}\boldsymbol{c}_t. \tag{1}$$

Here, the n^{th} sub-band basis vector $\boldsymbol{\phi}_n$ is defined as follows:

$$\phi_n(k) = \begin{cases} 0.5 - 0.5\cos\left(\dfrac{k - \tilde{\Omega}(n-1)}{\tilde{\Omega}(n) - \tilde{\Omega}(n-1)}\pi\right), \\ \qquad \tilde{\Omega}(n-1) \le k < \tilde{\Omega}(n) \\ 0.5 - 0.5\cos\left(\dfrac{k - \tilde{\Omega}(n)}{\tilde{\Omega}(n+1) - \tilde{\Omega}(n)}\pi + \dfrac{\pi}{2}\right), \\ \qquad \tilde{\Omega}(n) \le k < \tilde{\Omega}(n+1) \\ 0 \qquad otherwise, \end{cases} \tag{2}$$

where, $\tilde{\Omega}(n)$ [rad] means the frequency scale formulated by

$$\tilde{\Omega}(n) = \begin{cases} \Omega_n + 2\tan^{-1}\dfrac{\alpha\sin\Omega_n}{1 - \alpha\cos\Omega_n}, & 0 \le n < N_w \\ \dfrac{n - N_w}{N - N_w}\pi + \dfrac{\pi}{2}, & N_w \le n < N. \end{cases} \tag{3}$$

Note that α, Ω_n [rad] and N_w represent a warping parameter, $n\pi/N_w$ [rad] and the dimension index that satisfies $\tilde{\Omega}(N_w) = \pi/2$ [rad], respectively. An encoded SBM parameter $\tilde{\boldsymbol{c}}_t$ is optimized by minimizing the error between the original spectrum and the decoded one as follows:

$$\tilde{\boldsymbol{c}}_t = \underset{\boldsymbol{c}_t}{\operatorname{argmin}} \|\boldsymbol{s}_t - \boldsymbol{\Phi}\boldsymbol{c}_t\|. \tag{4}$$

In the SBM framework, an SBM parameter for the log-amplitude spectrum is obtained by non-negative least squares [13] and that for the phase spectrum is calculated by general least squares. In this paper, we only apply SBM to the amplitude spectrum.

Copyright © 2014 ISCA

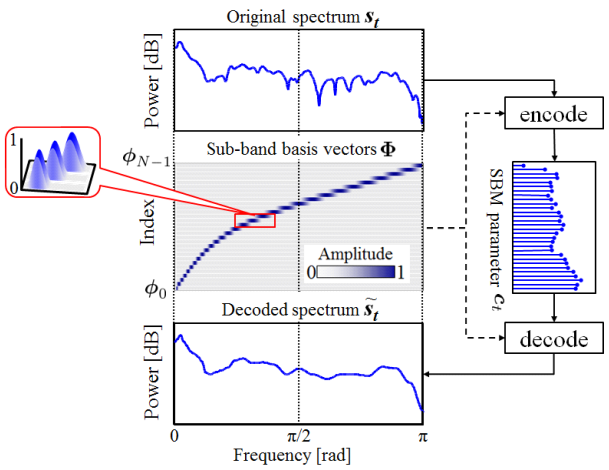

Figure 1: Overview of sub-band basis spectrum model

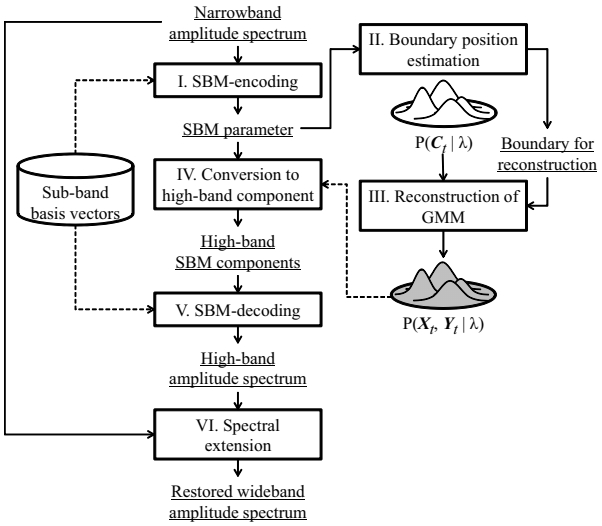

Figure 2: Overview of the proposed BWE process. Indeces I to VI indicate steps.

2.2. Parameter conversion based on Gaussian mixture model

2.2.1. Gaussian mixture model

The GMM-based parameter conversion framework uses a GMM which models joint probability density of the input vector $\boldsymbol{x}_t$ and the output vector $\boldsymbol{y}_t$ [9]. The probability density function of the GMM is written by

$$\mathrm{P}(\boldsymbol{x}_t, \boldsymbol{y}_t|\lambda) = \sum_{m=1}^{M} w_m \mathcal{N}\left(\left[\boldsymbol{x}_t^\top, \boldsymbol{y}_t^\top\right]^\top; \boldsymbol{\mu}_m^{(x,y)}, \boldsymbol{\Sigma}_m^{(x,y)}\right), \quad (5)$$

where, $\top$ is transposition of the vector and λ represents the model parameter set of the GMM. $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes a Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. w_m represents the weight for the m^{th} component. The m^{th} mean vector $\boldsymbol{\mu}_m^{(x,y)}$ and covariance matrix $\boldsymbol{\Sigma}_m^{(x,y)}$ are respectively described as

$$\boldsymbol{\mu}_m^{(x,y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(x)} \\ \boldsymbol{\mu}_m^{(y)} \end{bmatrix}, \boldsymbol{\Sigma}_m^{(x,y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(xx)} & \boldsymbol{\Sigma}_m^{(xy)} \\ \boldsymbol{\Sigma}_m^{(yx)} & \boldsymbol{\Sigma}_m^{(yy)} \end{bmatrix}. \quad (6)$$

In the case of the conventional GMM-based BWE, a GMM is trained with a parallel data consisting of narrowband acoustic feature as the input vector and the wideband feature as the output vector by using the EM algorithm.

2.2.2. Parameter conversion

As the parameter conversion, there are two types of methods to estimate the output vector $\boldsymbol{y}$, minimum mean square error (MMSE) estimation [8] and maximum likelihood estimation (MLE) [10].

In the MMSE method, the t^{th} vector $\hat{\boldsymbol{y}}_t$ is given as the conditional expectation $\mathrm{E}\left[\hat{\boldsymbol{y}}_t|\boldsymbol{x}_t\right]$:

$$\hat{\boldsymbol{y}}_t = \mathrm{E}\left[\hat{\boldsymbol{y}}_t|\boldsymbol{x}_t\right] = \sum_{m=1}^{M} \mathrm{P}\left(m|\boldsymbol{x}_t, \lambda\right) \boldsymbol{E}_{m,t}, \quad (7)$$

where,

$$\mathrm{P}\left(m|\boldsymbol{x}_t, \lambda\right) = \frac{w_m \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_m^{(x)}, \boldsymbol{\Sigma}_m^{(xx)}\right)}{\sum_{l=1}^{M} w_l \mathcal{N}\left(\boldsymbol{x}_t; \boldsymbol{\mu}_l^{(x)}, \boldsymbol{\Sigma}_l^{(xx)}\right)}, \quad (8)$$

$$\boldsymbol{E}_{m,t} = \boldsymbol{\Sigma}_m^{(yx)} \boldsymbol{\Sigma}_m^{(xx)^{-1}} \left(\boldsymbol{x}_t - \boldsymbol{\mu}_m^{(x)}\right) + \boldsymbol{\mu}_m^{(y)}. \quad (9)$$

On the other hand, the MLE method estimates parameter sequence considering dynamic features. In this technique, we use the input vector $\boldsymbol{X}_t = \left[\boldsymbol{x}_t^\top, \Delta\boldsymbol{x}_t^\top\right]^\top$ and the output vector $\boldsymbol{Y}_t = \left[\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top\right]^\top$, which consist of static and dynamic features at the t^{th} frame. Note that the dynamic feature is defined as $\Delta\boldsymbol{x}_t = 0.5(\boldsymbol{x}_{t+1} - \boldsymbol{x}_{t-1})$. The GMM $\mathrm{P}\left(\boldsymbol{X}_t, \boldsymbol{Y}_t|\lambda\right)$ is trained using a parallel data set of the above input and output vectors. The estimated vector sequence $\hat{\boldsymbol{y}} = \left[\hat{\boldsymbol{y}}_1^\top, \hat{\boldsymbol{y}}_2^\top, \cdots, \hat{\boldsymbol{y}}_T^\top\right]^\top$ is determined by the following equation:

$$\begin{aligned} \hat{\boldsymbol{y}} &= \arg\max_{\boldsymbol{y}} \mathrm{P}\left(\boldsymbol{Y}|\boldsymbol{X}, \lambda\right) \\ &\text{Subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}, \end{aligned} \quad (10)$$

where, $\boldsymbol{W}$ denotes the conversion matrix from the static vector sequence to the static and dynamic vector sequence.

3. Bandwidth extension using sub-band basis spectrum model and GMM

We propose a BWE method based on a single GMM that can be applied to speech data with an arbitrary frequency bandwidth. The GMM is trained using only a wideband speech data parametrized with SBM.

Figure 2 shows an overview of the proposed framework. In the proposed framework, narrowband spectra are extracted from the narrowband signal which is upsampled to the sampling rate of the wideband signal. We train a GMM $\mathrm{P}\left(\boldsymbol{C}_t|\lambda\right)$ with the $2N$-dimensional static and dynamic wideband SBM parameter vector $\boldsymbol{C}_t = [\boldsymbol{c}_t^\top, \Delta\boldsymbol{c}_T^\top]^\top$ in advance.

In the first step, using Eq. (4), we extract the narrowband SBM parameter from narrowband input spectrum.

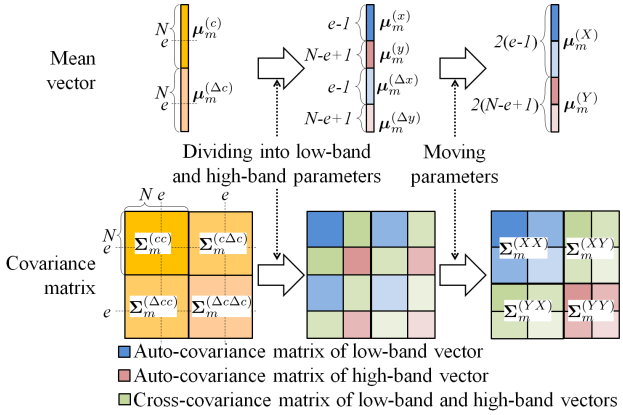

Figure 3: Reconstruction of the GMM.

Next, for reconstructing the trained GMM, we determine the boundary demarcating the high-frequency and the low-frequency bands based on the narrowband SBM parameter. Here, the SBM parameter has a similar shape to the original spectrum and thus the amplitudes of its higher frequency components are much smaller than the ones of lower components. By focusing on this, the boundary e is obtained finding the frequency where the difference between adjacent SBM parameters is maximized.

After determination of the boundary, the trained GMM $\mathrm{P}\left(\boldsymbol{C}_t|\lambda\right)$ is changed to the GMM with joint probability density of low-band and high-band parameters $\mathrm{P}\left(\boldsymbol{X}_t, \boldsymbol{Y}_t|\lambda\right)$ by moving mean vectors and covariance matrices of SBM parameters as shown in Figure 3. By this, we can obtain the GMM for the BWE which is composed of $2(e-1)$-dimensional $\boldsymbol{X}_t$ and $2(N-e+1)$-dimensional $\boldsymbol{Y}_t$.

In parameter conversion, we construct input data $\boldsymbol{X}$ by extracting the low-band SBM components from the narrowband SBM parameters based on the boundary e and adding the dynamic feature of the low-band SBM components. The high-band SBM components $\tilde{\boldsymbol{y}}$ are obtained by converting the input data $\boldsymbol{X}$ based on Eq. (10). Then the high-band spectra are generated from the high-band SBM components $\tilde{\boldsymbol{y}}$ using sub-band basis vectors based on Eq. (1).

Finally, we restore a wideband spectrum from the input narrowband spectrum and decoded high-band spectrum. Figure 4 shows how the wideband spectrum is restored in the proposed method. First we apply a one-sided hanning window to the edge of each spectrum. The window size is determined based on the bandwidth of the sub-band based vector located on the boundary e. Then we add the decoded high-band spectrum to the narrowband spectrum.

4. Evaluations

4.1. Speaker-independent and speaker-dependent bandwidth extension

We evaluated speaker-independent and speaker-dependent bandwidth conversions of the proposed method. We used a wideband speech data from 8 male and 8 female speakers to train a speaker-independent GMM (SI-GMM) with a full covariance matrix. Each speaker uttered the same 87 sentences. The first 50 utterances were used for training and the remaining 37 were used for testing. As target speakers, we used 10 speakers consisting of 5 males and 5 females not included in the SI-GMM training. We created speaker-dependent GMMs (SD-GMMs) for the target speakers. The first 50 utterances of

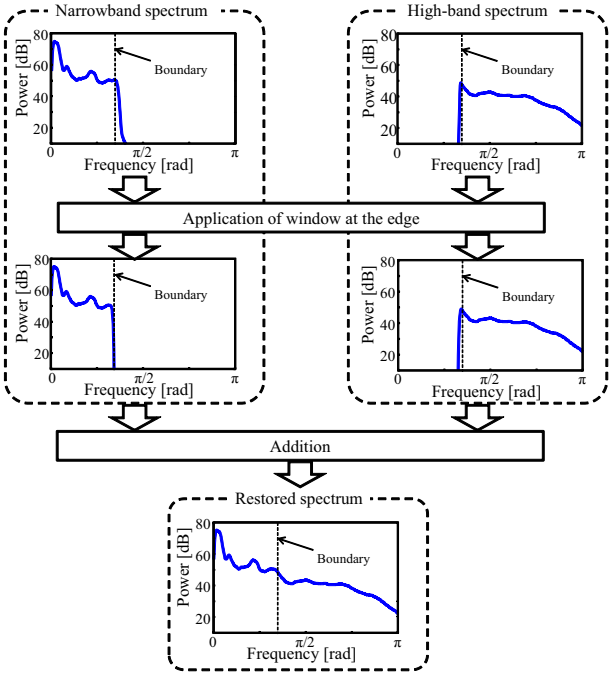

Figure 4: Reconstruction of amplitude spectrum in the proposed method.

Table 1: Spectral types for evaluations. "NO_BWE" means spectrum not applied bandwidth extension.

Label	GMM type	Estimation criterion
SI_MMSE	SI-GMM	MMSE
SI_MLE	SI-GMM	MLE
SD_MMSE	SD-GMM	MMSE
SD_MLE	SD-GMM	MLE
NO_BWE	no use	no use

each speaker were used for training and the remaining 37 were used for testing. The number of mixtures ranges from 2 to 256 for the SI-GMMs and from 2 to 64 for the SD-GMMs.

Recorded natural speech data sampled at 22.05 [kHz] was used as the original wideband speech. We used 1024-point FFT spectra analyzed in a pitch synchronous manner. Each GMM was trained with 80-dimensional static and 80-dimensional dynamic SBM parameters in the proposed method. In synthesis, we used a mixed excitation signal generated from the fundamental frequencies and aperiodic components which were extracted from the wideband speech.

The objective evaluation employed a log-spectral distance (LSD) defined as follows:

$$\mathrm{LSD} = \sum_{k=1}^{K} \frac{\sqrt{(l_t(k) - \hat{l}_t(k))^2}}{K} \ [\mathrm{dB}], \tag{11}$$

where, $l_t(k)$ and $\hat{l}_t(k)$ denote k^{th} log-spectral component of the original wideband and the reconstructed signals, respectively. In subjective evaluation, we conducted 5-level mean opinion score (MOS) tests (1: bad and 5: excellent) for speech quality using a crowdsourcing-based evaluation system. We selected 3 utterances from the test data set at random and each listener evaluated 30 samples in each condition. The number of listeners was 20.

We evaluated four types of bandwidth extension shown in Table 1 for three types of narrowband spectra with 4, 6 and 8

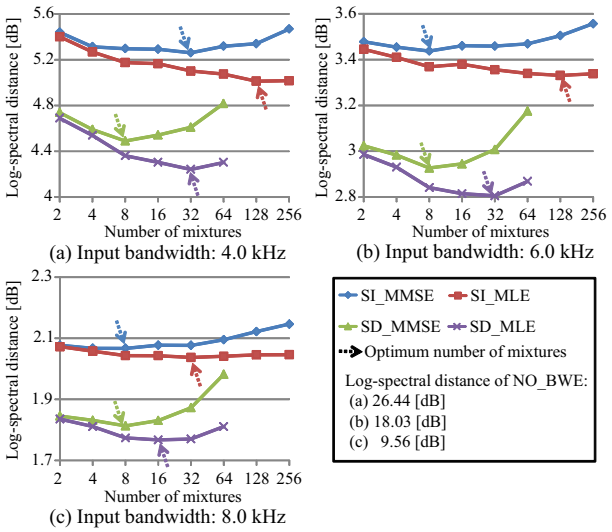

Figure 5: Results of objective test of BWE methods. Each LSD value is averaged over 10 target speakers.

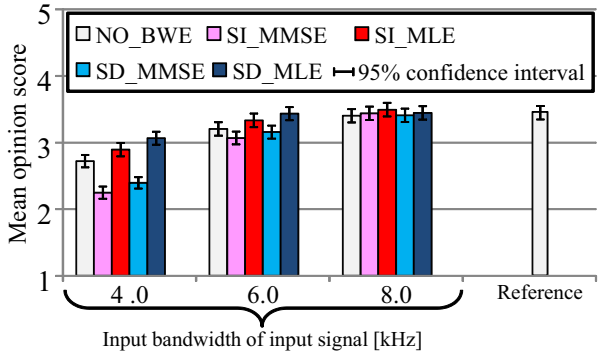

Figure 6: Result of subjective test of BWE methods in MOS test.

[kHz] cut-off frequencies. Figure 5 shows the results of LSD. We can see from this figure that the proposed methods perform the bandwidth extension well to reduce the LSDs. Obviously, speaker-dependent extension is better than speaker-independent one. It is also seen that the MLE criterion is better than MMSE.

Figure 6 shows the result of subjective evaluation. The reference is a synthetic speech generated in an analysis-synthesis manner where an excitation signal is fed to a synthesis filter extracted from the original wideband FFT spectrum. In this evaluation, the GMMs had 16 mixture components. In the cases of 4.0 and 6.0 [kHz] input bandwidths, the speech quality is improved by the bandwidth extension using "SI_MLE" and "SD_MLE." However, MOS scores for "SI_MMSE" and "SD_MMSE" are lower than "NO_BWE." The quality degradation is considered as a result of discontinuities between frames of the high-frequency components. Figure 6 indicates that the proposed methods produce high-quality speech almost equivalent to the wideband reference signal for the 8.0 kHz input bandwidth.

4.2. Comparison to the conventional GMM-based methods

We compared the proposed method "SI_MLE" to conventional GMM-based methods. For the sake of comparison, we created three SI-GMMs for the input bandwidths of 4, 6, and 8 [kHz]. We used three sets of parallel data of narrowband and wideband mel-cepstra represented with 25-dimentional static including power and their dynamic features for training. These SI-GMMs of the conventional method were called "MCEP4," "MCEP6," and "MCEP8," respectively. The boundary estimation described in Section 3 was used in the proposed method. We calculated LSDs for the input spectra of 4.0, 5.0, 6.0, 7.0 and 8.0 [kHz]. The number of mixtures was 16 and the MLE-based parameter conversion was used in this evaluation both for the proposed and conventional methods.

Table 2: Results of the objective test for comparison between conventional and proposed methods. Underlines represent matched conditions where the input bandwidths are the same for training and testing, and the bold numbers mean best scores for each input.

	Bandwidth of input spectra [kHz]				
Model type	4.0	5.0	6.0	7.0	8.0
MCEP4	**4.79**	6.40	7.06	5.95	3.79
MCEP6	19.57	9.24	**3.27**	3.14	2.41
MCEP8	23.08	16.74	11.93	6.64	**1.95**
Proposed (SI_MLE)	4.97	**4.19**	3.51	**2.72**	2.03

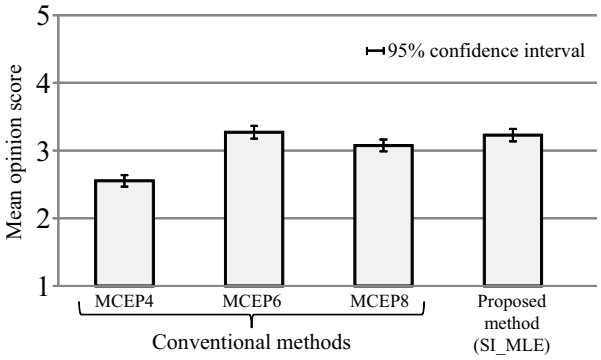

Figure 7: Result of subjective test for the comparison of the conventional and proposed methods.

Table 2 shows the results of the objective tests. We can see that the conventional methods work well for the matched conditions, but does not for mismatched conditions. On the other hand, the proposed methods work for any input spectrum.

Figure 7 shows the result of 5-level MOS tests for the input with 6.0 [kHz] bandwidth. This result shows that the proposed method "SI_MLE" has no significant difference in the speech quality with the matched condition "MCEP6" of conventional methods.

5. Conclusions

This paper has proposed a bandwidth extension (BWE) method using a Gaussian mixture model (GMM) and a sub-band basis spectrum model. We train a GMM using only wideband data and structure the GMM so that low-band and high-band components are separated. Unlike conventional GMM-based methods, the proposed method needs no parallel data for training. Experimental results show that the proposed method with a single GMM performs the bandwidth extension well for any input with arbitrary bandwidth.

Future work includes speaker adaptation [14, 15, 16] for improving speaker-independent extension.

6. References

[1] J. Yamagishi, C. Veaux, S. King, and S. Renals, "Speech synthesis technologies for individuals with vocal disabilities: voice banking and reconstruction," *Acoustical Science and Technology*, vol. 33, no. 1, pp. 1–5, Jan. 2012.

[2] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Speaking-aid systems using gmm-based voice conversion for electrolaryngeal speech," *Speech communication*, vol. 54, no. 1, pp. 134–146, Jan. 20102.

[3] A. Waibel, A. N. Jain, A. E. McNair, H. Saito, A. G. Hauptmann, and J. Tebelskis, "JANUS: a speech-to-speech translation system using connectionist and symbolic processing strategies," *Proc. ICASSP*, pp. 793–796, Apr. 1992.

[4] S. Matsuda, X. Hu, Y. Shiga, H. Kashioka, C. Hori, K. Yasuda, H. Okuma, M. Uchiyama, E. Sumita, H. Kawai, and S. Nakamura, "Multilingual speech-to-speech translation system: VoiceTra," *Mobile Data Management (MDM), 2013 IEEE 14th International Conference*, vol. 2, pp. 229–233, Jun. 2013.

[5] I. Bernd, S. Gerhard, and M. Wolfgang, "Bandwidth extension of speech signals," *Springer*, 2008.

[6] K. Park and H. Kim, "Narrowband to wideband conversion of speech using GMM based transformation," *Proc ICASSP*, vol. 3, pp. 1843–1846, 2000.

[7] W. Fujitsuru, H. Sekimoto, T. Toda, H. Saruwatari, and K. Shikano, "Bandwidth extension of cellular phone speech based on maximum likelihood estimation with GMM," *2008 RISP International Workshop on Nonlinear Circuits and Signal Processing*, Mar. 2008.

[8] Y. Stylianou, O. Cappé, and E. Moulines, "Continuous probabilistic transform for voice conversion," *IEEE Trans. Speech and Audio Processing*, vol. 6, no. 2, pp. 131–142, 1998.

[9] A. Kain and M. W. Macon, "Spectral voice conversion for text-to-speech synthesis," *Proc. ICASSP*, pp. 285–288, May 1998.

[10] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Trans. Audio, Speech and Language Processing*, vol. 15, no. 8, pp. 2222–2235, Nov. 2007.

[11] M. Tamura, T. Kagoshima, and M. Akamine, "Sub-band basis spectrum model for pitch-synchronous log-spectrum and phase based on approximation of sparse coding," *Proc. Interspeech 2010*, pp. 2046–2049, Sep. 2010.

[12] B. Olshausen and D. Field, "Emergence of simple-cell respective field properties by learning a sparse code for natural images," *Nature*, vol. 381, 1996.

[13] C. Lawson and R. Hanson, "Solving least squares problems," *SIAM classics in applied mathematics*, 1995 (first published by 1974).

[14] C.-H. Lee and C.-H. Wu, "MAP-based adaptation for speech conversion using adaptation data selection and non-parallel training," *Proc. Interspeech 2006 - ICSLP*, pp. 2254–2257, Sep. 2006.

[15] T. Toda, Y. Ohtani, and K. Shikano, "One-to-many and many-to-one voice conversion based on eigenvoices," *Proc. ICASSP*, vol. 4, pp. 1249–1252, Apr. 2007.

[16] A. Mouchtaris, J. der Spiegel, and P. Mueller, "Non-parallel training for voice conversion based on a parameter adaptation approach," *IEEE Trans. Audio, Speech and Language Processing*, vol. 14, no. 3, pp. 952–963, May 2006.