SUMMARY This paper describes a novel statistical bandwidth extension (BWE) technique based on a Gaussian mixture model (GMM) and 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 speech data with an arbitrary frequency bandwidth whereas the conventional methods perform the conversion from fixed narrow-band data. In the proposed method, we train a GMM with SBM parameters extracted from full-band spectra in advance. According to the bandwidth of input signal, the trained GMM is reconstructed to the GMM of the joint probability density between low-band SBM and high-band SBM components. Then high-band SBM components are estimated from low-band SBM components of the input signal based on the reconstructed GMM. Finally, BWE is achieved by adding the spectra decoded from estimated high-band SBM components to the ones of the input signal. To construct the full-band signal from the narrow-band one, we apply this method to log-amplitude spectra and aperiodic components. Objective and subjective evaluation results show that the proposed method extends the bandwidth of speech data robustly for the log-amplitude spectra. Experimental results also indicate that the aperiodic component extracted from the upsampled narrow-band signal realizes the same performance as the restored and the full-band aperiodic components in the proposed method.

key words: speech enhancement, voice conversion, bandwidth extension, sub-band basis spectrum model, Gaussian mixture model

1. Introduction

For a couple decades, bandwidth extension (BWE) has been widely studied [1] in order to improve speech quality and intelligibility. Classically, BWE is used in telephones in order to map telephone-band (300–3400 Hz) speech to wide-band (50–7000 Hz) one. Nowadays, BWE is widely applied to not only telephones but also audio codec systems to achieve high quality sound [2], hearing-aid of telephone speech not only telephones but also audio codec systems to achieve high quality sound [2], hearing-aid of telephone speech.

In this paper, we propose a GMM-based BWE using a sub-band basis spectrum model (SBM) [25] as the acoustic feature [26]. The 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 [27]. SBM parameters denoting weights for sub-band basis vectors represent a specific acoustic space in the frequency domain and it can be applied for frequency domain processing easily. In the proposed method, we train a single GMM with non-parallel data consisting of only full-band SBM parameters in advance, and then we construct the GMM of the joint probability density between low-band SBM and high-band SBM components from the GMM of the full-band SBM. Based on the constructed GMM, we estimate high-band SBM parameters from the input low-band SBM ones. Combining the low-band and high-band parameters, we restore full-band
signals. Thus, the BWE is performed for any input signal with an arbitrarily narrow-bandwidth. In order to construct the restored full-band signal, this paper applies the proposed BWE to both log-amplitude spectra and aperiodic components [28].

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

2. Related Work

2.1 Sub-Band Basis Spectrum Model

Figure 1 shows an overview of the SBM. $s_t$ is a $K$-dimensional speech spectrum vector of the $t$th frame. Here, encode and decode mean representing and restoring the spectrum with and from the SBM parameters, respectively. $c_t$ is an $N$-dimensional SBM parameter vector that represents weights of the sub-band basis vectors. $s_t$ can be described by $c_t = [c_t(1), c_t(2), \cdots, c_t(N)]^T$ ($T$: transposition) and sub-band basis vectors $\Phi = [\phi_1, \phi_2, \cdots, \phi_N]$ as follows:

$$s_t = \Phi c_t, \quad (1)$$

Here, the $n$th sub-band basis vector $\phi_n$ is defined as follows:

$$\phi_n(k) = \begin{cases} 
\frac{1}{2} - \frac{1}{2} \cos \left( \frac{\omega_n - \tilde{\Omega}(n-1)}{(\tilde{\Omega}(n) - \tilde{\Omega}(n-1)) \pi} \right), & \tilde{\Omega}(n-1) \leq \omega_n < \tilde{\Omega}(n) \\
\frac{1}{2} - \frac{1}{2} \cos \left( \frac{\omega_n - \tilde{\Omega}(n)}{(\tilde{\Omega}(n+1) - \tilde{\Omega}(n)) \pi} \right), & \tilde{\Omega}(n) \leq \omega_n < \tilde{\Omega}(n+1) \\
0, & \text{otherwise},
\end{cases} \quad (2)$$

where, $k$ and $n$ represent indexes of spectral dimension and SBM dimension, respectively. $\omega_n$ means $k\pi/K$ [rad]. $\tilde{\Omega}(n)$ is the frequency scale formulated by

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

Note that $\alpha$, $\Omega_n$ [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. The encoded SBM parameter $\tilde{c}_t$ for the log-amplitude spectrum is obtained based on the non-negative least squares (NNLS) [29] criterion, i.e., minimizing the error between the original spectrum and the decoded one frame-by-frame.

$$\tilde{c}_t = \arg\min_{c_t} \| s_t - \Phi c_t \|_2,$$

Subject to $c_t \geq 0$, \quad (4)

where, $\| \cdot \|_2$ means the Euclidean norm. Note that the SBM parameter for the phase spectrum is calculated using the standard least squares. Whereas the original SBM method focuses on the log-amplitude spectrum and phase spectrum, we apply SBM to the log-amplitude spectrum and aperiodic component. Here, the aperiodic component is defined as the ratio between harmonic and non-harmonic components in each frequency band of a speech signal [30]. The value 0 means complete periodic and 1 is purely non-periodic.

The SBM parameter for the aperiodic component is obtained based on the NNLS-based criterion.

In this paper, we use SBM parameters as acoustic features, and train GMMs of the log-spectrum and aperiodic component, respectively. The proposed BWE described in Sect. 3 is achieved by using these GMMs.

2.2 Parameter Conversion Based on Gaussian Mixture Model

2.2.1 Gaussian Mixture Model

The traditional GMM-based parameter conversion method uses a GMM that models joint probability density of the input vector $x_t$ and the output vector $y_t$ [23]. The probability density function of the GMM is written by

$$P(x_t, y_t|\lambda) = \sum_{m=1}^{M} w_m N \left( \begin{bmatrix} x_t^T, y_t^T \end{bmatrix}^T, \mu_m^{(x,y)}, \Sigma_m^{(x,y)} \right), \quad (5)$$

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

$$\mu_m^{(x,y)} = \begin{bmatrix} \mu_m^{(x)} \\ \mu_m^{(y)} \end{bmatrix}, \quad \Sigma_m^{(x,y)} = \begin{bmatrix} \Sigma_m^{(x,x)} & \Sigma_m^{(x,y)} \\ \Sigma_m^{(y,x)} & \Sigma_m^{(y,y)} \end{bmatrix}. \quad (6)$$

In the case of the conventional GMM-based BWE, a GMM
is trained with a parallel data consisting of narrow-band acoustic features as the input vector and full-band features as the output vector by using the EM algorithm. To achieve BWE, the conventional method first converts input narrow-band acoustic features into the full-band ones using the trained GMM as described in Sect. 2.2.2. The higher-frequency components of the converted full-band spectra are added to input narrow-band spectra and then we obtain restored acoustic features. This paper employs this conventional GMM-based BWE as the reference technique.

2.2.2 Parameter Conversion

There are two types of methods to estimate the output vector \( \mathbf{y} \): minimum mean square error (MMSE) estimation [22] and maximum likelihood estimation (MLE) [24].

In the MMSE method, the \( n \)th vector \( \hat{y}_n \) is given as the conditional expectation \( E[\hat{y}_n|x_n] \):

\[
\hat{y}_n = E[\hat{y}_n|x_n] = \sum_{m=1}^{M} P(m|x_n, \lambda) E_{m,n},
\]

(7)

where,

\[
P(m|x_n, \lambda) = \frac{u_m N(\mathbf{x}_n; \mu_m^{(X)}, \Sigma_m^{(XX)})}{\sum_{l=1}^{M} u_l N(\mathbf{x}_n; \mu_l^{(X)}, \Sigma_l^{(XX)})},
\]

(8)

\[
E_{m,n} = m \left[ \Sigma_m^{(xx)} \right]^{-1} \left( \mathbf{x}_n - \mu_m^{(X)} \right) + \mu_m^{(Y)}.
\]

(9)

On the other hand, estimating a parameter sequence, the MLE method considers dynamic features. Let us use the input vector \( \mathbf{X}_t = [\mathbf{x}_t^T, \Delta \mathbf{x}_t^T]^T \) and the output vector \( \mathbf{Y}_t = [\mathbf{y}_t^T, \Delta \mathbf{y}_t^T]^T \), which consist of static and dynamic features at the \( t \)th frame. The dynamic feature is defined as \( \Delta \mathbf{x}_t = 0.5(\mathbf{x}_{t+1} - \mathbf{x}_{t-1}) \). The GMM \( P(\mathbf{X}_t, \mathbf{Y}_t | \lambda) \) is trained using a parallel data set of the above input and output vectors. The vector sequence \( \hat{y} = [\hat{y}_1^T, \hat{y}_2^T, \cdots, \hat{y}_T^T] \) is determined as follows:

\[
\hat{y} = \arg \max_y P(\mathbf{Y}|\mathbf{X}, \lambda)
\]

\[
= \arg \max_y \sum_{m=1}^{M} P(m|\mathbf{X}, \lambda) P(\mathbf{Y}|m, \mathbf{X}, \lambda),
\]

(10)

Subject to \( \mathbf{Y} = \mathbf{W} \mathbf{y} \),

where, \( \mathbf{X} = [\mathbf{x}_1^T, \mathbf{x}_2^T, \cdots, \mathbf{x}_T^T]^T \), \( \mathbf{Y} = [\mathbf{y}_1^T, \mathbf{y}_2^T, \cdots, \mathbf{y}_T^T]^T \), \( m = [m_1, m_2, \cdots, m_T] \) and \( \mathbf{W} \) respectively denote sequences of input vector, output vector and mixture components, and a matrix to convert the static features to a concatenation of static and dynamic features. The probabilities \( P(m|\mathbf{X}_t, \lambda) \) and \( P(\mathbf{Y}|m, \mathbf{X}_t, \lambda) \) are defined as follows:

\[
P(m_t|\mathbf{X}_t, \lambda) = \frac{u_m N(\mathbf{x}_t; \mu_m^{(X)}, \Sigma_m^{(XX)})}{\sum_{l=1}^{M} u_l N(\mathbf{x}_t; \mu_l^{(X)}, \Sigma_l^{(XX)})},
\]

(11)

\[
P(\mathbf{Y}|m_t, \mathbf{X}_t, \lambda) = N(\mathbf{Y}_t; E_{m_t}^{(Y)}, D_{m_t}^{(Y)}),
\]

(12)

where,

\[
E_{m_t}^{(Y)} = \Sigma_m^{(XX)} \Sigma_m^{(YY)}^{-1} (\mathbf{x}_t - \mu_m^{(X)}) + \mu_m^{(Y)},
\]

(13)

\[
D_{m_t}^{(Y)} = \Sigma_m^{(YY)} - \Sigma_m^{(XX)} \Sigma_m^{(YY)}^{-1} \Sigma_m^{(XY)}.
\]

(14)

Note that we employ an approximation method [24] instead of Eq. (10) to find a suboptimum mixture sequence. In this case, the vector sequence \( \hat{y} \) is determined by

\[
\hat{y} = \arg \max_y P(\mathbf{Y}|\hat{m}, \mathbf{X}, \lambda),
\]

(15)

Subject to \( \mathbf{Y} = \mathbf{W} \mathbf{y} \),

where, \( \hat{m} = [\hat{m}_1, \hat{m}_2, \cdots, \hat{m}_T] \) is the suboptimum mixture sequence determined as

\[
\hat{m}_t = \arg \max_m P(m|\mathbf{X}_t, \lambda).
\]

(16)

Therefore, the estimated vector sequence \( \hat{y} \) is derived by

\[
\hat{y} = \left( W^T D_{\hat{m}}^{(Y) -1} W \right)^{-1} W^T D_{\hat{m}}^{(Y)} -1 E_{\hat{m}}^{(Y)},
\]

(17)

where,

\[
E_{\hat{m}}^{(Y)} = \left[ E_{\hat{m}_1}^{(Y)}, E_{\hat{m}_2}^{(Y)}, \cdots, E_{\hat{m}_T}^{(Y)} \right]^T,
\]

(18)

\[
D_{\hat{m}}^{(Y)} = \text{diag} \left[ D_{\hat{m}_1}^{(Y)}, D_{\hat{m}_2}^{(Y)}, \cdots, D_{\hat{m}_T}^{(Y)} \right].
\]

(19)

3. Bandwidth Extension Using Sub-Band Basis Spectrum Model with 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 full-band speech data parameterized with SBM.

Figure 2 shows an overview of the proposed method. In this method, narrow-band spectra and aperiodic components are extracted from the narrow-band signal that is upsampled to the sampling rate of the full-band signal. In advance, we train GMMs with the \( 2N \)-dimensional static and dynamic full-band SBM parameter vectors of spectrum and aperiodic components, which we call SBM-SP and SBM-AP respectively.

3.1 SBM-Encoding

In the first step, using Eq. (4), we extract the narrow-band SBM-SP and SBM-AP of narrow-band input spectrum and aperiodic component respectively.

3.2 Boundary Estimation

Next, we determine the boundary demarcating the high-frequency and the low-frequency bands based on the narrow-band SBM-SP. The SBM parameter has a similar shape to the original spectrum and thus the amplitudes of its
Fig. 2  Overview of the proposed BWE process. Indexes 1 to 6 indicate steps.

higher-frequency components are much smaller than those of lower-frequency components. Considering this we search for the boundary frequency \( e \) where the difference between adjacent SBM-SP parameters is maximized. This process is computationally expensive for finding \( e \) for all SBM-SP frames. In order to reduce the computational cost, we average a SBM-SP sequence except for silent frames and determine the boundary \( e \) from its averaged SBM-SP.

3.3 GMM Reconstruction

Once the boundary is determined, the trained GMM \( P(C_t|\lambda) \) (\( C_t \): static and dynamic SBM-SP/AP vector at the \( t \)th frame) is changed to the GMM with joint probability density of low-band and high-band components \( P(X_t,Y_t|\lambda) \) by moving mean vectors and covariance matrices of SBM parameters as shown in Fig. 3. By this, we can obtain the GMM for the BWE that is composed of \( 2(e - 1) \)-dimensional \( X_t \) and \( 2(N - e + 1) \)-dimensional \( Y_t \).

3.4 Conversion of Low-Band to Higher-Band Components

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

3.5 SBM-Decoding and Parameter Extension

Finally, we restore a full-band spectrum from the input narrow-band spectrum and decoded high-band spectrum. Figure 4 shows how the full-band 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. Then we add the decoded high-band spectrum to the narrow-band spectrum.

In the same manner, we restore the full-band aperiodic component.

4. Evaluations

4.1 Speaker-Independent and Speaker-Dependent Bandwidth Extension for Spectral Feature

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

Recorded natural speech data sampled at 22.05 kHz were used as the original full-band speech. All of the original full-band speech was high-pass filtered with 50 Hz cut-off frequency in order to reduce electric hum noise. Thus, the full-band in this paper means 50–11025 Hz. Speech spectra were derived by 1024-point pitch synchronous Fourier transform using our in-house tools [30]. When the SBM parameters were extracted, the warping parameter $\alpha$ was set to 0.35 according to [25]. Each GMM was trained with 80-dimensional static and 80-dimensional dynamic SBM-SP parameters. In synthesis, we used a mixed excitation signal generated from the fundamental frequencies and aperiodic components extracted from the full-band speech by a pitch-scaled harmonic filter (PSHF) [31].

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

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

where, $l_t(k)$ and $\hat{l}_t(k)$ denote the $k^{th}$ log-spectral component of the original full-band and the reconstructed signals, respectively. In the subjective evaluation, we conducted 5-level mean opinion score (MOS) tests (1: bad, 2: poor, 3: fair, 4: good and 5: excellent) for the 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.

4.1.1 Results of Objective and Subjective Evaluations

We evaluated the four types of bandwidth extension shown in Table 1 for three types of narrow-band spectra with 4, 6 and 8 kHz cut-off frequencies. Figure 5 shows the results of LSD. In this evaluation, the number of mixtures ranges from 2 to 256 for the SI-GMMs and from 2 to 64 for the SD-GMMs. We can see from this figure that the proposed methods perform the bandwidth extension well to reduce the LSDs. Obviously, the speaker-dependent extension is better than the 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 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 for “NO_BWE.” The MMSE method performs the frame-by-frame conversion without time variations such as dynamic features. The quality degradation can be attributed to discontinuity between frames of the high-frequency components. Figure 6 indicates that the proposed methods produce high-quality speech almost equivalent to the full-band reference signal for the 8.0 kHz input bandwidth.

In addition, Figs. 5 and 6 show that the performance of SD-GMM is better than that of SI-GMM. This is because SI-GMMs include parameters affected by variations among training speakers. To improve the performance of SI-GMMs, we can apply speaker adaptation and speaker adaptive training [32]–[36] to the speaker-independent models.

4.1.2 Comparison with the Conventional GMM-Based Methods

We compared the proposed method “SI_MLE” with conventional GMM-based methods using objective and subjective measures. 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 narrow-band and full-band mel-cepstra represented with 25-dimentional static features including power and their dynamic features for training. These SI-GMMs of the conventional method are called “MCEP4,” “MCEP6,” and “MCEP8,” respectively. The boundary estimation described in Sect. 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 in the objective test.

In the subjective evaluation, we prepared test sets of full-band speech restored from input speech of 4.0, 5.0, 6.0, 7.0 and 8.0 kHz bandwidths and conducted listening tests on the crowdsourcing system. Each test set included the same utterances as those mentioned in Sect. 4.1. The number of listeners was 20 in each test set.

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 shows the results of the objective tests. We can see that the conventional methods work well for the matched conditions, but do not for the mismatched conditions.

Figure 7 shows the result of 5-level MOS tests for the input with 4.0 to 8.0 kHz bandwidths. These results show that in the case of the proposed method “SI_MLE” there is no significant difference in the speech quality from that in the case of the matched conditions “MCEP4,” “MCEP6,” and “MCEP8” with the conventional methods. In the cases of the mismatched conditions (b) and (d), the proposed method can get better quality than the conventional methods. However, when input bandwidth becomes wider than 7.0 kHz, there are no significant differences among the methods.

Figure 8 shows restored spectrograms of MCEP6 and the proposed method for 4.0, 6.0 and 8.0 kHz bandwidths. In the case of the matched condition (the middle of Fig. 8), we can see that MCEP6 achieves BWE correctly. However, MCEP6 does not recover high-band spectra for 4.0 kHz bandwidth. On the other hand, we can see that the proposed methods work well for any input spectrum.

4.2 Bandwidth Extension for Aperiodic Component

Finally, in order to evaluate the effectiveness of the proposed
BWE for aperiodic component, we compared the aperiodic component of the upsampled narrow-band signal and the full-band signal with the restored aperiodic component. The number of SBM-AP’s dimensions was 60. We trained a SI-GMM with the static and dynamic SBM-AP features of the same training speakers as the SI-GMM for SBM-SP, and employed the MLE-based parameter conversion for BWE. The number of the mixtures was 4, which was determined by the preliminary test. In synthesis, we employed restored spectra in each bandwidth. We conducted the preference test and test speakers and utterances were the same as those mentioned in Sect. 4.1. The number of listeners was 15.

Table 3 shows the results of the preference test. From the results, we can see that “BWE” performance is almost equivalent to the “Full-band” one in each case. Therefore, BWE for aperiodic components performs correctly. However, “Narrow-band” scores are similar to “BWE” and “Full-band” scores, and there are no statistical differences at the 5% significant level among the preference tests.

To find the cause of such results, we compared full-band aperiodic components with upsampled narrow-band ones. Figure 9 shows the aperiodic components of full-band signal and upsampled narrow-band signal at a voiced interval. High-band aperiodic components of the upsampled narrow-band signal are similar to those of the full-band signal. Both mixed excitation signals in Fig. 10 show similar forms. From these observations, it can be seen that the synthetic speech using narrow-band aperiodic components has almost the same speech quality as that using full-band and restored ones.

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 full-band 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. The proposed method is applied to log-amplitude spectra and aperiodic components in order to build the full-band signal from the narrow-band one. Experimental results show that the proposed method with a single GMM performs the bandwidth extension well for any input with arbitrary bandwidth in the case of the spectra. The results also indicate that synthetic speech using the aperi-
odic component extracted from the upsampled narrow-band signal has almost the same speech quality as that using the full-band or restored aperiodic component.

References


Yamato Ohtani received the B.E. degree in engineering from Osaka University, Osaka, Japan, in 2005 and the M.E. and Ph.D. degrees in engineering from the Graduate School of Information Science, Nara Institute of Science and Technology (NAIST) in 2007 and 2010, respectively. He was an intern researcher at ATR Spoken Language Communication Research Laboratories, Kyoto, Japan from 2006 to 2009. He joined Corporate Research and Development center of Toshiba Corporation in 2010. His research interests include statistical approaches to speech processing such as speech synthesis and voice transformation. He received the Awaya Prize Young Researcher Award from ASJ in 2015. He is a member of ASJ, IPSJ, ISCA and IEEE.

Masatsune Tamura received his B.E., M.E. and Dr.Eng. degrees from Tokyo Institute of Technology in 1997, 1999 and 2002, respectively. Since 2002, he has been with the Corporate Research and Development Center of Toshiba Corporation. His research interests include speech synthesis. He is a member of the Acoustical Society of Japan and IEEE.

Masahiro Morita received the B.E. and M.E. degrees in biophysical engineering from Osaka University, Osaka, Japan, in 1994 and 1996, respectively. He joined the Corporate Research and Development Center of Toshiba Corporation in 1996. His research interest focuses on Text-to-Speech synthesis. He is a member of ASJ.

Masami Akamine received the B.E. degree from the University of the Ryukyus, Japan in 1979, the M.A. and Ph.D. degrees in electrical engineering from Tohoku University in 1982 and 1985, respectively. Since 1985 he has been with the Toshiba Research and Development Center. He is currently a Senior Fellow at Toshiba Research and Consulting, responsible for coordinating research programs among research groups in Kawasaki Japan, Cambridge UK and Beijing China. His research interests include speech coding, speech synthesis, automatic speech recognition and their applications. He was awarded as an outstanding researcher by the Minister of Education, Science and Technology Japan in 2001. He has received Technology Development Award from the Acoustic Society of Japan in 2002, Society Best Paper Award from IEICE Japan in 2003, Prime Minister’s Prize from the Japan Institute of Invention and Innovation in 2008, and Achievement Award from IEICE Japan in 2012. He was also honoured to receive Purple Ribbon Medal from the Emperor Japan in 2013. He is a senior member of IEEE and had served as a member of the Speech and Language Technical Committee for two years since 2012.