Acoustic Compensation Method for Accepting Different Recording Devices in Body-Conducted Voice Conversion

Daisuke Deguchi*, Hironori Doi*, Tomoki Toda *, Hiroshi Saruwatari* and Kiyohiro Shikano*

* Graduate School of Information Science, Nara Institute of Science and Technology, Ikoma Nara 630-0192 Japan
E-mail:{daisuke-d, hironori-d, tomoki, sawatari, shikano}@is.naist.jp Tel:+81-743-72-5288

***Abstract*—This paper presents an acoustic compensation method in body-conducted speech conversion that automatically compensates for acoustic differences caused by changes in recording conditions. An enhancement process for body-conducted speech recorded with a Non-Audible Murmur (NAM) microphone has successfully applied a statistical voice conversion technique. Speech waveforms are generated from acoustic parameters of normal speech estimated from those of body-conducted speech with a conversion model previously trained using stereo data of those two types of speech. This framework suffers from mismatched conditions between training and conversion processes. To alleviate this issue, an unsupervised acoustic compensation method based on constrained maximum likelihood linear regression (CMLLR) has been proposed and its effectiveness has been reported in the compensation of acoustic differences caused by attachment location of the NAM microphone. This paper further applies the CMLLR-based acoustic compensation method to the compensation of acoustic differences caused by different recording devices and evaluates its effectiveness. The experimental results demonstrate that the proposed method effectively reduces quality degradation and converted speech caused by differences in recording devices as well as attachment location of the NAM microphone.**

I. Introduction

Recently cellular phones have enabled us to communicate with each other conveniently. However, noisy environments such as a crowd make it difficult to smoothly convey speech. To address this issue, the use of body-conducted speech for speech communication [1], [2] has been proposed. A Non-Audible Murmur (NAM) microphone[2], that is a body-conductive microphone, is placed on the neck below the ear and detects various types of speech such as non-audible murmurs, whispers, and normal speech through the soft tissue of the head. This approach is robust against external noise due to the noise-proof structure of body-conductive microphones. However, body-conducted speech causes severe quality degradation due to essential mechanisms of body conduction such as the lack of radiation characteristics from lips and the influence of the low-pass characteristics of soft tissue.

To improve the quality of body-conducted speech, body-conducted voice conversion has been proposed [3]. This technique is based on statistical voice conversion techniques [4], [5]. A conversion model is trained in advance using a parallel data set consisting of utterance-pairs of body-conducted speech and normal speech (i.e., air-conducted speech) uttered by the same speaker. A Gaussian mixture model (GMM) of joint probability density of speech parameters of those two types of speech is effectively used as the conversion model. The trained model is capable of converting body-conducted speech parameters into normal speech parameters without the use of linguistic information.

One weakness of body-conducted voice conversion is that severe quality degradation of the converted speech is caused by mismatched acoustic conditions between training and conversion. The recording of body-conducted speech with a NAM microphone is sensitive to various conditions such as the attachment location of the NAM microphone, settings of the amplifier, the type of NAM microphone, and so on. In the practical use of a NAM microphone, it is almost impossible to keep these conditions consistent. Moreover, the NAM microphone is still under development and various types of recording devices would may be developed. Therefore, it techniques for compensating for acoustic differences of body-conducted speech caused by changes of various recording conditions must be developed.

In our previous work [6], an acoustic compensation method based on constrained maximum likelihood linear regression (CMLLR) [7] has been proposed to compensate for acoustic differences caused by a change in the attachment location of the NAM microphone. This method estimates feature-space transforms for compensating for the acoustic differences in a completely unsupervised manner using only the acoustics of body-conducted speech. It is worthwhile to investigate the effectiveness of this method in various recording conditions beyond simply changing the location of the NAM microphone.

In this paper, we apply the CMLLR-based acoustic compensation method to the compensation of acoustic differences caused by different recording devices. Three different types of NAM microphone and their own amplifiers are used for recording body-conducted speech. The experimental results demonstrate that the proposed compensation method is effective for compensating the acoustic differences caused by the different recording devices as well as the different attachment location of NAM microphone.

II. Body-Conducted Speech

Body-conducted speech has a different spectral structure from that of air-conducted speech [6]. In particular, higher frequency components of body-conducted speech are severely attenuated. Consequently body-conducted speech usually sounds muffled.

In this paper, three types of different NAM microphone and their own amplifiers shown in Figure 1 are used for recording

Proceedings of the Second APSIPA Annual Summit and Conference, pages 502–505,
Biopolis, Singapore, 14-17 December 2010.

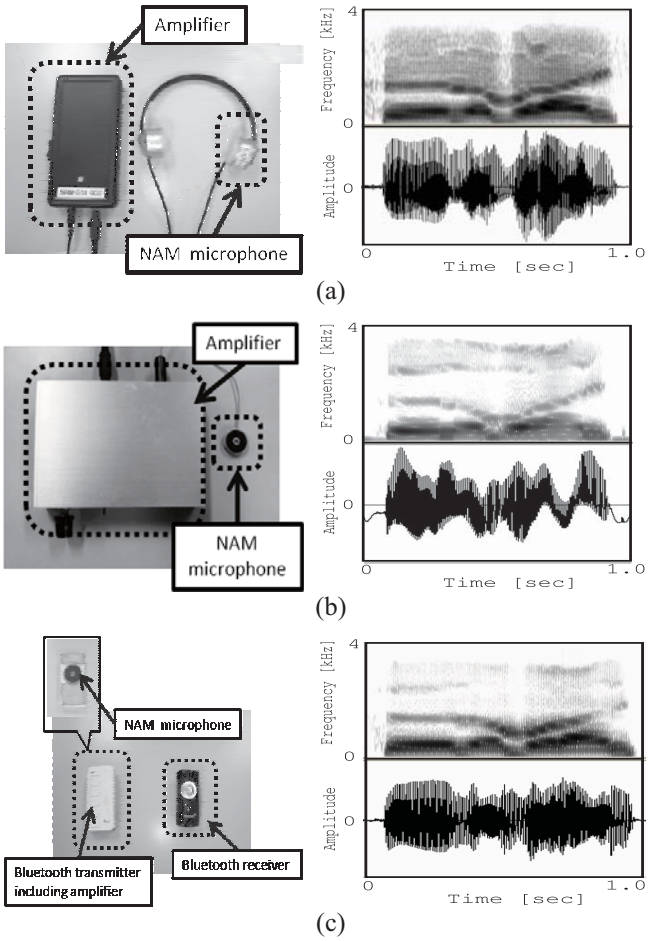

Fig. 1. Three types of NAM microphone, (a) wired-type with neckband, (b) wired-type without neckband, and (c) wireless-type, and an example of spectrogram of body-conducted speech recorded with each NAM microphone.

body-conducted speech. The use of different types of NAM microphone and their amplifiers causes noticeable changes in the acoustic features of body-conducted speech. These acoustic changes are caused by differences in various conditions such as the attachment location of the NAM microphone and the way of attaching the NAM microphone on the body as well as differences the recording device.

III. Body-conducted speech conversion

A. Feature Extraction of Body-Conducted Speech

As a source feature, we employ a spectral segment vector. Let $\boldsymbol{x}_t$ is a mel-cepstral vector at a frame t. We construct a concatenated vector $\boldsymbol{c}_t = [\boldsymbol{x}_{t-n}^\top \cdots \boldsymbol{x}_t^\top \cdots \boldsymbol{x}_{t+n}^\top]^\top$ over the current $\pm n$ frames, where the symbol $\top$ indicates transpose. And then, the spectral segment vector $\boldsymbol{X}_t$ at frame t is extracted by PCA as follows:

$$\boldsymbol{X}_t = \boldsymbol{D}\boldsymbol{c}_t - \boldsymbol{d}, \tag{1}$$

where $\boldsymbol{D}$ is the transformation matrix of PCA, and $\boldsymbol{d} = \boldsymbol{D}\bar{\boldsymbol{c}}$. The vector $\bar{\boldsymbol{c}}$ is the mean vector of $\boldsymbol{c}_t$ within all training data for PCA.

As a target feature, we employ the joint static and dynamic feature vector $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top \ \Delta\boldsymbol{y}_t^\top]^\top$, where $\boldsymbol{y}_t$ is the static mel-cepstral vector, and $\Delta\boldsymbol{y}_t$ is the delta mel-cepstral vector of the target data at frame t.

B. Feature Conversion Based on Maximum Likelihood Estimation of Parameter Trajectory [5]

The joint probability density of the source and target feature vectors is modeled by a GMM as follows:

$$P(\boldsymbol{Z}_t|\boldsymbol{\lambda}) = \sum_{m=1}^{M} w_m \mathcal{N}(\boldsymbol{Z}_t; \boldsymbol{\mu}_m^{(Z)}, \boldsymbol{\Sigma}_m^{(ZZ)}), \tag{2}$$

where $\boldsymbol{Z}_t$ is the joint feature vector $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top \ \boldsymbol{Y}_t^\top]^\top$. The symbol $\mathcal{N}()$ indicates the normal distribution. The number of mixture components is M. $\boldsymbol{\lambda}$ is the model parameter set including w_m, $\boldsymbol{\mu}_m^{(Z)}$, and $\boldsymbol{\Sigma}_m^{(ZZ)}$, which are the weight, mean vector, and covariance matrix of the m-th mixture component, respectively. $\boldsymbol{\mu}_m^{(Z)}$ and $\boldsymbol{\Sigma}_m^{(ZZ)}$ are represented by

$$\boldsymbol{\mu}_m^{(Z)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \tag{3}$$

$$\boldsymbol{\Sigma}_m^{(ZZ)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}, \tag{4}$$

where the matrices $\boldsymbol{\Sigma}_m^{(XX)}$ and $\boldsymbol{\Sigma}_m^{(YY)}$ are the covariance matrix of the m-th mixture component of the source and that of the target, respectively. The matrices $\boldsymbol{\Sigma}_m^{(XY)}$ and $\boldsymbol{\Sigma}_m^{(YX)}$ are the cross covariance matrices of the m-th mixture component between the source and target. These covariance matrices are completely full.

Let $\boldsymbol{X} = [\boldsymbol{X}_1^\top; \ \cdots \ ; \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top; \ \cdots \ ; \boldsymbol{Y}_T^\top]^\top$ be time sequences of the source and the target features, respectively. The converted static feature vector sequence is determined so that the following approximated conditional probability density is maximized.

$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \simeq P(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}) P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}), \tag{5}$$

where $\boldsymbol{m} = \{m_1, \ \cdots, \ m_T\}$ is a mixture component sequence. First, suboptimum mixture component sequence $\hat{\boldsymbol{m}}$ is determined by

$$\hat{\boldsymbol{m}} = \arg\max_{\boldsymbol{m}} P(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}). \tag{6}$$

And then, the converted static feature vector $\hat{\boldsymbol{y}}$ is obtained by

$$\hat{\boldsymbol{y}} = \arg\max_{\boldsymbol{y}} P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}), \quad \text{subject to } \boldsymbol{Y} = \boldsymbol{E}\boldsymbol{y}, \tag{7}$$

where $\boldsymbol{E}$ is a window matrix to extend the static feature vector sequence into the joint static and dynamic feature vector sequence. Furthermore, the quality of the converted voice is dramatically improved by considering the global variance of the converted feature [5].

IV. Unsupervised Acoustic Compensation Method Based on CMLLR

To compensate for the acoustic differences caused by changes in recording conditions, the CMLLR-based unsupervised adaptation method [6] is adopted. In this method, to effectively reduce the mismatch between the model and the adaptation data, the CMLLR transform is estimated using only body-conducted speech and the previously trained GMM of joint probability density.

We apply the CMLLR transformation to the source features to compensate for their acoustic variations. The transformed source feature vector is given by

$$\hat{\boldsymbol{X}}_t = \boldsymbol{A}\boldsymbol{X}_t + \boldsymbol{b} = \boldsymbol{W}\boldsymbol{\xi}(t), \tag{8}$$

where $\boldsymbol{W}$ is the extended transform, $[\boldsymbol{b}\ \boldsymbol{A}]$, and $\boldsymbol{\xi}(t)$ is the extended source feature vector, $[1\ \boldsymbol{X}_t^\top]^\top$.

To perform unsupervised compensation, the CMLLR transform is estimated so that the likelihood of the marginal distribution for the adaptation source data $\boldsymbol{X} = [\boldsymbol{X}_1, \cdots, \boldsymbol{X}_T]$ is maximized as follows:

$$\hat{\boldsymbol{W}} = \arg\max_{\boldsymbol{W}} \prod_{t=1}^{T} \int P(\boldsymbol{X}_t, \boldsymbol{Y}_t | \boldsymbol{W}, \boldsymbol{\lambda}) d\boldsymbol{Y}_t. \tag{9}$$

Because the probability density is modeled by a GMM, EM algorithm is employed. In each M-step, an iterative row-by-row update is performed for determining the updated transformation matrix [7], [8]. The i-th row vector of $\hat{\boldsymbol{W}}$ is given by

$$\boldsymbol{w}_i = (\alpha \boldsymbol{c}_i + \boldsymbol{k}^{(i)}) \boldsymbol{G}^{(ii)-1}, \tag{10}$$

where $\boldsymbol{c}_i$ is the extended cofactor row vector of $\boldsymbol{A}$ and α is found by solving a quadratic equation [7]. $\boldsymbol{k}^{(i)}$ and $\boldsymbol{G}^{(i)}$ are given by

$$\boldsymbol{G}^{(ij)} = \sum_{m=1}^{M} p_m(i,j) \sum_{t=1}^{T} \gamma_m(t) \boldsymbol{\xi}_m(t) \boldsymbol{\xi}_m(t)^\top, \tag{11}$$

$$\boldsymbol{k}^{(i)} = \sum_{m=1}^{M} \boldsymbol{p}_m(i) \boldsymbol{\mu}_m \sum_{t=1}^{T} \gamma_m(t) \boldsymbol{\xi}_m(t)^\top - \sum_{j=1, j\neq i}^{d} \boldsymbol{w}_j \boldsymbol{G}^{(ij)}, \tag{12}$$

where $\boldsymbol{p}_m(i)$ and $p_m(i,j)$ are the i-th row vector and the (i,j)-th element of the inverse covariance matrix $\boldsymbol{\Sigma}_m^{(XX)-1}$, respectively. $\gamma_m(t)$ is the posterior probability of the m-th mixture component given $\boldsymbol{X}_t$.

When applying the CMLLR transformation in the model-space, the adapted model parameters are given by

$$\hat{\boldsymbol{\mu}}_m^{(Z)} = \begin{bmatrix} \boldsymbol{A}' \boldsymbol{\mu}_m^{(X)} - \boldsymbol{b}' \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \tag{13}$$

$$\hat{\boldsymbol{\Sigma}}_m^{(Z)} = \begin{bmatrix} \boldsymbol{A}' \boldsymbol{\Sigma}_m^{(XX)} \boldsymbol{A}'^\top & \boldsymbol{A}' \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{A}'^\top & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}, \tag{14}$$

where $\hat{\boldsymbol{\mu}}_m^{(Z)}$ and $\hat{\boldsymbol{\Sigma}}_m^{(Z)}$ are the adapted mean vector and covariance matrix of the m-th mixture component, respectively. Note that $\boldsymbol{A}' = \boldsymbol{A}^{-1}$ and $\boldsymbol{b}' = \boldsymbol{A}'\boldsymbol{b}$. A global transform is used in this paper.

V. Experimental Evaluations

A. Experimental Conditions

Body-conducted speech and air-conducted speech uttered by one Japanese male speaker were simultaneously recorded with a NAM microphone and a headset microphone. Using each type of NAM microphone and its amplifier shown in Figure 1, 100 phoneme-balanced sentences were recorded. Consequently, three sets of parallel data of body-conducted speech (a, b, and c) and air-conducted speech (A, B, and C) were developed. Because each data set was recorded in a different recording session, the voice quality of the recorded speech samples were slightly different from each other. Fifty sentences were used for training or adaptation and the remaining fifty sentences were used for conversion in each data set.

TABLE I
Several conditions evaluated in experiments. Body-conducted speech (a, b, and c) and air-conducted speech (A, B, and C) are simultaneously recorded in three different conditions using different recording devices ((a), (b), and (c) in Figure 1)

Condition	Parallel data set in training	Body-conducted speech in conversion
Fully-matched	a-A	a
	b-B	b
	c-C	c
Mismatched	a-A	b or c
	b-B	a or c
	c-C	a or b
Adapted	a-A	b or c (adapted to a)
	b-B	a or c (adapted to b)
	c-C	a or b (adapted to c)
Matched	a-B	a
	a-C	a
	b-A	b
	b-C	b
	c-A	c
	c-B	c

By changing data sets used in training and in conversion, various conditions were evaluated as shown in Table I. Three conversion models were independently built using three parallel data sets (a-A, b-B, and c-C). In the "Fully-matched" condition, the same data sets used in training were also used in conversion. Therefore, not only the recording conditions but also the recording sessions were consistent between training and conversion. In this condition, three sets of converted speech were generated. In the "Mismatched" condition, different data sets were used in conversion. Consequently, six sets of converted speech were generated. Acoustic differences caused by changes in various recording conditions such as different recording devices and different attachment locations of the NAM microphone were observed in this condition. In the "Adapted" condition, the proposed method was used to compensate for these differences. Again, six sets of converted speech were generated. Moreover, to show ideal results in the proposed compensation framework, the other conversion models were also trained using parallel data sets consisting of body-conducted speech and air-conducted speech recorded in different sessions. The other models were then used to convert the body-conducted speech recorded in the same session as used in training. This was called the "Matched" condition and again six sets of converted speech were generated. In this condition, there were some differences of voice quality in target air-conducted speech between training and conversion due to the different recording sessions.

The 0-th through 16-th mel-cepstral coefficients were adopted as a spectral parameter. As the input feature, a 34-dimensional spectral segment feature vector was extracted by PCA from a concatenated vector consisting of current and $\pm$ 4 frames of mel-cepstrum vectors. The sampling frequency was set to 8 kHz. The number of mixture components of GMM was set to 32. The frame shift was set to 5 ms.

In the objective evaluation, we evaluated spectral conversion accuracy with mel-cepstral distortion calculated from the first through 16-th mel-cepstral coefficients between the converted and target mel-cepstra. Converted speech samples in four con-

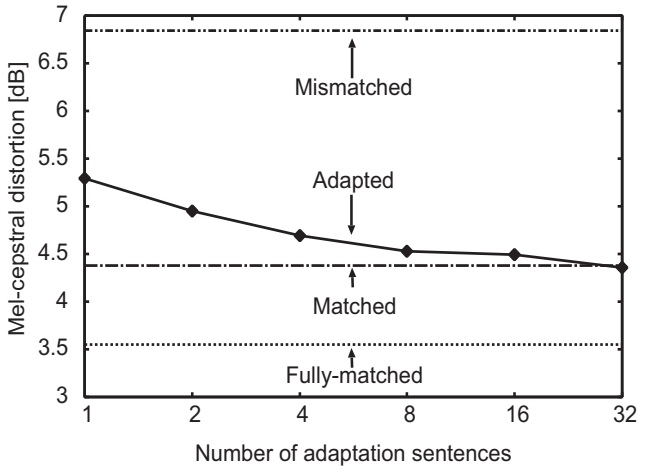

Fig. 2. Mel-cepstral distortion as a function of the number of adaptation sentences.

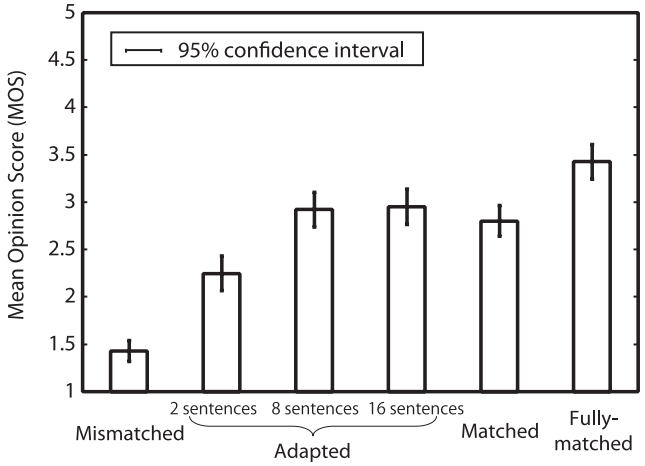

Fig. 3. Result of opinion test on speech quality.

ditions as shown in Table I were evaluated. In the "Adapted" condition, the number of adaptation sentences was set to 1, 2, 4, 8, 16, or 32, which were used for estimating the CMLLR transform.

In subjective evaluation, we conducted an opinion test of speech quality. An opinion score was set to a 5-point scale (1: bad - 5: excellent). Seven listeners evaluated the sound quality of six types of converted speech: under the "Mismatched" condition; under the "Adapted" condition when using 2, 8, or 16 adaptation sentences; under the "Matched" condition, and under the "Fully-matched" condition. For each listener, 18 sentences were randomly selected from the test set and 108 samples of the converted speech were evaluated in total.

B. Experimental Results

Figure 2 shows the results of the objective evaluation. A large mel-cepstral distortion is observed in the converted speech under the "Mismatched" condition. This result shows that body-conducted voice conversion is very sensitive to the recording conditions. This degradation of conversion accuracy is effectively alleviated by the proposed acoustic compensation method, i.e., the "Adapted" condition. Even if only one sentence is used for adaptation, mel-cepstral distortion significantly decreases compared with the "Mismatched" condition. Mel-cepstral distortion gradually decreases according to an increase in the amount of adaptation data up to around 8 sentences which is close to an ideal result shown in the "Matched" condition. Note that the difference between "Matched" and "Fully-matched" is caused by the voice quality difference between different recording sessions.

Figure 3 shows the results of the subjective evaluation. We can see that changes of the recording devices and the attachment location of NAM microphone cause significant quality degradation. Our proposed compensation method effectively improves the converted speech quality. The use of only 8 adaptation sentences makes the converted speech quality equivalent to that in the "Matched" condition. We can also see a significant quality difference between the "Matched" and "Fully-matched" conditions. These results are consistent with those observed in the objective evaluation.

VI. Conclusion

This paper described an unsupervised acoustic compensation method based on CMLLR in body-conducted voice conversion to compensate for acoustic differences caused by different recording conditions. The proposed method has been applied to body-conducted voice conversion in mismatched conditions caused by the use of different recording devices as well as a change in the attachment location of a NAM microphone. The experimental results from both objective and subjective tests have demonstrated that the proposed compensation method is capable of effectively alleviating performance degradation in body-conducted voice conversion under mismatched conditions using only a few adaptation sentences.

VII. Acknowledgment

This work was supported in part by MIC SCOPE.

References

[1] Y. Zheng, Z. Liu, M.Sinclair, J. Droppo, L. Deng, A. Acero, and X.Huang, "Air- and Bone-Conductive Integrated Microphones for Robust Speech Detection and Enhancement", *Proc. ASRU*, pp. 249 – 254, 2003

[2] Y. Nakajima, H. Kashioka, N. Campbell and K. Shikano, "Non-Audible Murmur (NAM) Recognition", *IEICE Trans. Information and Systems*, Vol. E89-D, No. 1, pp. 1 – 8, 2006.

[3] T. Toda, K. Nakamura, H. Sekimoto and K. Shikano, "Voice Conversion for Various Types of Body Transmitted Speech", *Proc. ICASSP*, pp. 3061 – 3064, 2009.

[4] Y. Stylianou, O. Capp´e and E. Moulines, "Continuous Probabilistic Transform for Voice Conversion",*in Proc. IEEE* Trans. SAP, Vol. 6, No. 2, pp. 131 – 142, 1998.

[5] T. Toda, A. W. Black, and K. Tokuda, "Voice Conversion Based on Maximum Likelihood Estimation of Spectral Parameter Trajectory", *Trans. ASLP*, Vol. 15, No. 8, pp. 2222 – 2235, 2007.

[6] D. Miyamoto, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano, "Acoustic Compensation Mmethods for Body Transmitted Speech Conversion", *Proc. ICASSP*, pp. 3901 – 3904, 2009.

[7] M. J. F. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition", *Computer Speech and Language*, vol. 12, no. 2, pp. 75 – 98, 1998.

[8] K. C. Sim and M. J. F. Gales, "Adaptation of Precision Matrix Models on Large Vocablary Continuous Speech Recognition", *Proc. ICASSP*, 2005.