# ACOUSTIC COMPENSATION METHODS FOR BODY TRANSMITTED SPEECH CONVERSION

*Daisuke Miyamoto, Keigo Nakamura, Tomoki Toda, Hiroshi Saruwatari, and Kiyohiro Shikano*

Graduate School of Information Science, Nara Institute of Science and Technology
{ daisuke-m, kei-naka, tomoki, sawatari, shikano }@is.naist.jp

## ABSTRACT

Statistical voice conversion is very effective for enhancing body transmitted speech recorded with Non-Audible Murmur (NAM) microphone. In this method, a probabilistic model to convert body transmitted speech into natural speech is trained previously. Because acoustic characteristics of body transmitted speech is sensitive to recording conditions such as a location of NAM microphone, significant degradation of the conversion performance is often caused in practical situations by acoustic mismatches between training and conversion processes. To alleviate this problem, we propose unsupervised acoustic compensation methods for body transmitted voice conversion. Experimental results demonstrate that the proposed methods significantly reduce the quality degradation of converted speech caused by the acoustic mismatches.

***Index Terms—*** Acoustic Compensation, CSMAPLR, CMLLR, CMS, Body Transmitted Voice Conversion

## 1. INTRODUCTION

Recently a cellular phone has enabled us to communicate with each other very conveniently. It makes people aware that using a cellular phone is problematic in some situations, e.g., under extremely heavy noisy conditions or under very quiet conditions (e.g., in a library). In order to alleviate these essential problems of speech communication, several body-conductive microphones, of which one good property is external noise robustness, have been developed [1, 2].

As one of the promising body-conductive microphones, we have focused on Non-Audible Murmur (NAM) microphone [2]. This microphone is attached on the skin behind the user's ear as shown in Fig. 1. One of advantages of NAM microphone is to record various types of body transmitted speech such as normal speech and considerably small whisper. However, body transmitted speech is usually distorted due to some factors, i.e., the lack of radiation characteristics from lips, low-pass characteristics of the body transmission, and so on.

In order to improve the quality of body transmitted speech, the body transmitted speech enhancement based on statistical voice conversion has been proposed [3, 4]. In this method, a conversion model is trained in advance using a parallel data set consisting of utterance pairs of the body- and air-transmitted voices. The trained model allows the conversion from body transmitted speech into the target air-transmitted speech without any linguistic information. This method dramatically improves the quality of the body transmitted speech under the same recording conditions between training and conversion processes. However, we empirically know that the acoustic characteristics of body transmitted speech are severely affected by recording conditions such as the attaching location of NAM microphone, the gain setting of an amplifier, and so on. Resulting acoustic variations of body transmitted speech, which is much larger than those of air-transmitted speech, would cause significant quality degradation of the converted voice due to the mismatched conversion model for input acoustic characteristics.

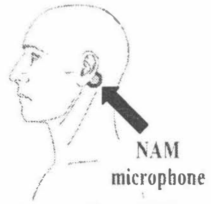


**Fig. 1.** Attaching location of NAM microphone.

In order to address this problem, we propose unsupervised acoustic compensation methods for body transmitted voice conversion. In this paper, we deal with the conversion from body transmitted ordinary speech (BTOS), which is defined as normal speech recorded with NAM microphone, into air transmitted natural speech. First we investigate an impact of the acoustic variations caused by different positions of NAM microphone on the conversion performance. And then, we apply Cepstrum Mean Subtraction (CMS) [5], Constrained Maximum Likelihood Linear Regression (CMLLR) [6], and Constrained Structural Maximum A Posteriori Linear Regression (CSMAPLR) [7] to the acoustic compensation for the body transmitted voice conversion. Experimental results demonstrate that 1) significant quality degradation of the converted speech is caused by the attaching position change of NAM microphone and 2) the proposed compensation methods are very effective for alleviating the quality degradation.

This paper is organized as follows. In Section 2, we describe the conversion method from BTOS into natural speech. Acoustic compensation methods for the body transmitted voice conversion are explained in Section 3, and these methods are experimentally evaluated in Section 4. Finally, this paper is summarized in Section 5.

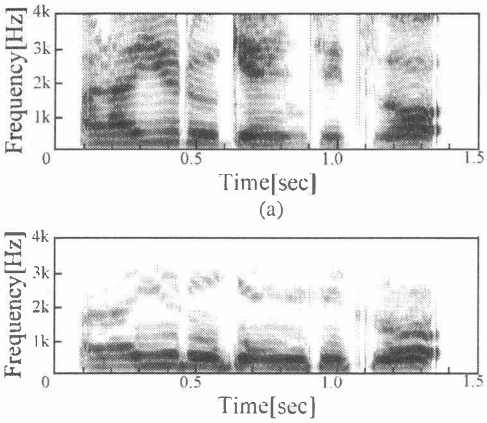
## 2. BODY TRANSMITTED VOICE CONVERSION

### 2.1. Body Transmitted Ordinary Speech (BTOS)

Fig. 2 shows an example of spectrograms of speech recorded with a headset microphones and those with NAM microphone. There are differences of spectral structures between air-conductive speech and BTOS. In particular, higher frequency components of the body transmitted voices are usually attenuated. Consequently BTOS sounds very muffled. Furthermore, some phonemes with large power on higher frequency bands such as unvoiced fricatives often lose their specific acoustic cue.

Thanks to MIC SCOPE for supporting this research in part.

978-1-4244-2354-5/09/$25.00 ©2009 IEEE



(a)

(b)

**Fig. 2**. An example of spectrograms of ordinary speech recorded (a) with air-conductive microphone and (b) with NAM microphone.

### 2.2. Acoustic Features

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 the target features, we employ the concatenated 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 feature vector, and $\Delta\boldsymbol{y}_t$ is the delta feature vector of target data at frame $t$.

### 2.3. Feature Conversion Based on Maximum Likelihood [4]

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

Let $\boldsymbol{X} = [\boldsymbol{X}_1^\top \cdots \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top \cdots \boldsymbol{Y}_T^\top]^\top$ are 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 $\boldsymbol{m} = [m_1 \cdots m_T]^\top$ 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

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

where $\boldsymbol{E}$ is a window matrix to expand the static feature sequence into the static and dynamic feature sequence. Furthermore, the quality of the converted voice is dramatically improved by considering global variance of the converted feature [4].

## 3. INTRODUCING ACOUSTIC COMPENSATION INTO BODY TRANSMITTED VOICE CONVERSION

When NAM microphone is used in a practical situation, it seems impossible to record under the completely same conditions, e.g., the exactly same location of NAM microphone. Therefore, it is inevitable to cope with the acoustic variations of BTOS caused by such a change of recording conditions. As a practical and convenient way, we propose unsupervised acoustic compensation methods using only the source features without any linguistic constraints and any target features.

### 3.1. CMS

CMS [5] effectively compensates static acoustic features characterized by multiplicative distortions.

We apply CMS to the mel-cepstral vector before extracting the spectral segment feature $\boldsymbol{X}_t$ described in Section 2.2. The vector $\boldsymbol{x}'_t$ processed by CMS is given by $\boldsymbol{x}'_t = \boldsymbol{x}_t - \bar{\boldsymbol{x}}_t$, where $\bar{\boldsymbol{x}}_t$ is the mean vector of $\boldsymbol{x}_t$ within an utterance. The final spectral segment feature vector $\boldsymbol{X}'_t$ is extracted as follows :

$$\boldsymbol{X}'_t = \boldsymbol{D}'\boldsymbol{c}'_t - \boldsymbol{d}', \tag{8}$$

where $\boldsymbol{c}'_t = [\boldsymbol{x}'^\top_{t-n} \cdots \boldsymbol{x}'^\top_t \cdots \boldsymbol{x}'^\top_{t+n}]^\top$ and $\boldsymbol{d}' = \boldsymbol{D}'\bar{\boldsymbol{c}}'$. The matrix $\boldsymbol{D}'$ and the vector $\bar{\boldsymbol{c}}'$ are the PCA transformation matrix and the mean vector of $\boldsymbol{c}'_t$ within all training utterances for PCA, respectively.

### 3.2. CMLLR

CMLLR [6] reduces the mismatch between the model and adaptation data. CMLLR estimates multiple linear transformations in individual regression classes, which are dynamically defined according to the amount of adaptation data using a regression tree, by maximizing the likelihood of the model for the adaptation data [8].

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

$$\hat{\boldsymbol{X}}_t = \boldsymbol{A}_r\boldsymbol{X}_t + \boldsymbol{b}_r = \boldsymbol{W}_r\boldsymbol{\xi}(t), \tag{9}$$

where $\boldsymbol{W}_r$ is the extended transform in the regression class $r$, $[\boldsymbol{b}_r\ \boldsymbol{A}_r]$, 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 a likelihood of the marginal distribution for the adaptation source data $\boldsymbol{X}$ is maximized as follows:

$$\hat{\boldsymbol{W}}_r = \underset{\boldsymbol{W}_r}{\arg\max} \int P(\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{W}_r, \boldsymbol{\lambda}) d\boldsymbol{Y}. \tag{10}$$

Because the probability density is modeled by a GMM, EM algorithmis is employed. The updated transformation matrix $\hat{\boldsymbol{W}}_r$ is given by

$$\boldsymbol{w}_{ri} = (\alpha \boldsymbol{c}_i + \boldsymbol{k}^{(ri)}) \boldsymbol{G}^{(rii)-1}, \tag{11}$$

where $\boldsymbol{w}_{ri}$ and $\boldsymbol{c}_i$ are the $i$-th row vector of $\hat{\boldsymbol{W}}_r$ and the extended cofactor row vector of $\boldsymbol{A}_r$, and $\alpha$ is found by solving a quadratic equation [6]. Then $\boldsymbol{k}^{(ri)}$ and $\boldsymbol{G}^{(rii)}$ are given by

$$\boldsymbol{G}^{(rij)} = \sum_{m=1}^{M_r} p_m(i,j) \sum_{t=1}^{T} \gamma_m(t) \boldsymbol{\xi}(t) \boldsymbol{\xi}(t)^\top, \tag{12}$$

$$\boldsymbol{k}^{(ri)} = \sum_{m=1}^{M_r} \boldsymbol{p}_m(i) \boldsymbol{\mu}_m \sum_{t=1}^{T} \gamma_m(t) \boldsymbol{\xi}(t)^\top - \sum_{j=1, j\neq i}^{d} \boldsymbol{w}_j \boldsymbol{G}^{(rij)}, \tag{13}$$

where $\boldsymbol{p}_m(i)$ and $p_m(i,j)$ are the $i$-th row vector and the $(i,j)$-th element of the covariance matrix $\boldsymbol{\Sigma}_m^{(XX)}$, respectively. $M_r$ and $\gamma_m(t)$ are the number of mixture components in class $r$ and 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

$$\hat{\boldsymbol{\mu}}_m^{(Z)} = \begin{bmatrix} \boldsymbol{A}'_r \boldsymbol{\mu}_m^{(X)} - \boldsymbol{b}'_r \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \tag{14}$$

$$\hat{\boldsymbol{\Sigma}}_m^{(Z)} = \begin{bmatrix} \boldsymbol{A}'_r \boldsymbol{\Sigma}_m^{(XX)} \boldsymbol{A}'^\top_r & \boldsymbol{A}'_r \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} \boldsymbol{A}'^\top_r & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix}, \tag{15}$$

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 $\boldsymbol{A}'_r = \boldsymbol{A}_r^{-1}$ and $\boldsymbol{b}'_r = \boldsymbol{A}'_r \boldsymbol{b}_r$. Parameters of all mixture components are adapted using corresponding transforms.

If using Maximum Likelihood Linear Regression (MLLR) [6] instead of CMLLR, it is difficult to robustly estimate multiple transforms because $\boldsymbol{G}^{(rii)}$ shown by Eq. (12) is easy to be a rank deficient matrix because $\boldsymbol{G}^{(rii)}$ is calculated by a weighted sum of mean vectors in MLLR. Because the number of mixture components is very limited in the GMM-based voice conversion, this problem often happens when increasing the number of regression classes.

### 3.3. CSMAPLR

It is essentially difficult to robustly estimate multiple transforms by the CMLLR estimation when using a small amount of adaptation data. An over-fitting problem easily happens especially in the unsupervised adaptation of the GMM. In order to alleviate this problem, we employ the CSMAPLR [7].

The CSMAPLR transform is estimated, so that a likelihood of the marginal distribution for the adaptation source data $\boldsymbol{X}$ is maximized as follows:

$$\hat{\boldsymbol{W}}_r = \underset{\boldsymbol{W}_r}{\operatorname{argmax}} \int P(\boldsymbol{X}, \boldsymbol{Y} | \boldsymbol{W}_r, \boldsymbol{\lambda}) P(\boldsymbol{W}_r) d\boldsymbol{Y}, \tag{16}$$

where the prior distribution function $P(\boldsymbol{W}_r)$, which is the matrix variate normal distribution, is defined by

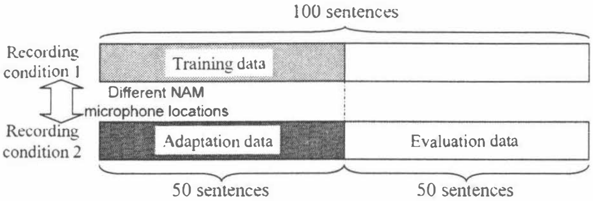


**Fig. 3**. Data sets for experiments.

$$P(\boldsymbol{W}_r) \propto |\boldsymbol{\Psi}|^{-(d+1)/2} |\boldsymbol{\Phi}|^{-d/2} \cdot \exp\left[ -\frac{1}{2} \operatorname{tr}\{ (\boldsymbol{W}_r - \boldsymbol{H})^\top \boldsymbol{\Psi}^{-1} (\boldsymbol{W}_r - \boldsymbol{H}) \boldsymbol{\Phi}^{-1} \} \right], \tag{17}$$

where $d$ is the number of the dimension of the source feature, and $\boldsymbol{\Psi}$, $\boldsymbol{\Phi}$ and $\boldsymbol{H}$ are the hyperparameters for this distribution. The transformation matrix of the parent's node in the regression tree is used as $\boldsymbol{H}$. We fix $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}$ to $\boldsymbol{\Psi}^{-1} = C \cdot \boldsymbol{I}$ and $\boldsymbol{\Phi}^{-1} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the unit matrix, respectively. The scaling of $P(\boldsymbol{W}_r)$ is controlled by only a scalar coefficient $C$. The transformation matrix $\boldsymbol{W}_r$ is updated as follows:

$$\boldsymbol{w}_{ri} = (\alpha \boldsymbol{c}_i + \boldsymbol{n}^{(ri)}) \boldsymbol{V}^{(rii)-1}. \tag{18}$$

Note that $\boldsymbol{n}^{(ri)}$ and $\boldsymbol{V}^{(rii)}$ are given by

$$\boldsymbol{V}^{(rii)} = \boldsymbol{G}^{(rii)} + C \cdot \boldsymbol{I}, \tag{19}$$

$$\boldsymbol{n}^{(ri)} = \boldsymbol{k}^{(ri)} + C \cdot \boldsymbol{h}(i), \tag{20}$$

where $\boldsymbol{h}(i)$ is the $i$-th row vector of $\boldsymbol{H}$.

## 4. EXPERIMENTAL EVALUATION

### 4.1. Experimental Conditions

We simultaneously recorded BTOS and air-conductive speech. Two Japanese male speakers uttered in two different conditions, in which only NAM microphone location was different (left or right side of the neck) and the others were kept as constant as possible. First, each speaker uttered 100 phonetically balanced sentences while fixing the attaching location of NAM microphone. Next, he switched the attaching side of NAM microphone, and then, he again uttered the same sentences One speaker uttered one sentence set (including 100 × 2 sentences) and the other speaker uttered two different sentence sets. Therefore, totally three sentence sets were recorded. Sampling frequency was 8 kHz.

In order to evaluate the conversion performance in the mismatched conditions, each sentence set was used as shown in Fig. 3. In the training of the conversion model, 50 sentences in the first recording condition were used. Different 50 sentences in the second recording condition were used as the test data. In the CMLLR or CSMAPLR compensation, the transformation matrices were estimated using a part of 50 sentences in the second recording condition, which were not included in the test data, as the adaptation data. Note that these adaptation data were used to train the conversion model in the matched condition. In the CMS compensation, a cepstral mean vector calculated from each sentence was used for the same sentence in the training. On the other hand, a cepstral mean vector calculated from the previous sentence was used in the evaluation assuming a practical situation. These evaluation processes were again conducted by swapping data in the first recording condition for those in the second one.

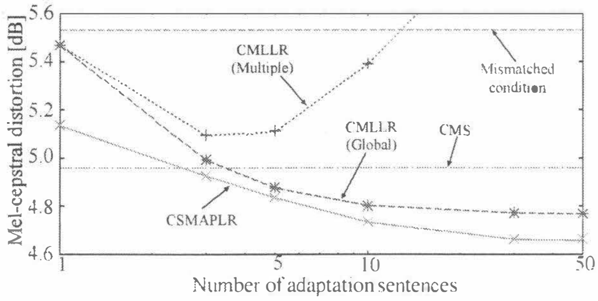


**Fig. 4**. Mel-cepstral distortion with power coefficient as a function of the number of adaptation sentences. CMLLR (Global) and (Multiple) show the result of CMLLR compensation with the global transform and that with the multiple transforms. The mel-capstral distortion in the matched condition is 3.93 dB.

We evaluated four conversion models in the mismatched condition, i.e., 1) no compensation, 2) CMS, 3) CMLLR, or 4) CSMAPLR. We also evaluated the conversion model in 5) the matched condition. We conducted both the objective and the subjective evaluations. In the objective evaluation, the mel-cepstral distortion between the converted and target features was measured. In the subjective evaluation, a pair of the different two types of the converted speech was randomly presented to the listeners, and then they were asked which voice sounded more natural. Each listener evaluated every pair-combination of all types of the converted speech [1]. The number of listeners was 6 including 3 males and 3 females.

The 0-th through 16-th mel-cepstral coefficients were adopted as the spectral parameter. The 34-dimensional source segment feature was calculated from a current ±4 frames. The number of mixture components of a GMM was set to 64. In the CMLLR or CSMAPLR compensation, the hyperparameter $C$ and the threshold of occupancy in each regression class were set to $10^5$ and 1000, respectively.

### 4.2. Experimental Results

Fig. 4 shows a result of the objective evaluation. The acoustic variations due to the change of the location of the NAM microphone make mel-cepstral distortion significantly larger. The proposed CMS compensation effectively alleviates this degradation of the conversion performance. The proposed CMLLR compensation using the global transform causes further improvements of the conversion performance when using more than a few adaptation sentences. However, using multiple transforms causes the performance degradation because of the over-fitting problem. The proposed CSMAPLR compensation effectively alleviates this problem. Consequently, it always outperforms the CMLLR compensation.

Fig. 5 shows the preference score on speech quality. The change of the NAM microphone location causes the significant quality degradation. The proposed compensation methods effectively recover the converted speech quality. The CSMAPLR compensation using 10 adaptation sentences causes significantly better converted speech than the CMS compensation. These results are very similar to as observed in the objective evaluation.

[1] Several samples are available from
http://spalab.naist.jp/~tomoki/ICASSP/AdaptBTVC/index.html

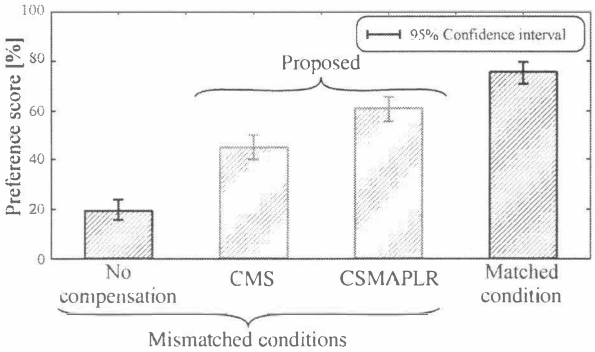


**Fig. 5**. Results of preference test on speech quality. 10 sentences were used for adaptation data in the CSMAPLR compensation.

## 5. CONCLUSION

We proposed unsupervised acoustic compensation methods for the body transmitted voice conversion based on CMS, CMLLR, or CSMAPLR. Experimental results of the objective and subjective tests demonstrated that the proposed compensation methods can effectively alleviate the degradation of the conversion performance in the mismatched condition between training and conversion processes. Moreover, they showed that CSMAPLR is the most effective among the proposed methods when using more than only a few adaptation data.

## 6. 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 Recognition", *IEICE Trans. Information and Systems*, Vol. E89-D, No. 1, pp. 1–8, 2006

[3] T. Toda and K. Shikano, "NAM-to-Speech Conversion with Gaussian Mixture Models", *Proc. INTERSPEECH*, pp. 1957–1960, 2005

[4] 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

[5] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification", J. Acoust. Soc. America, Vol. 55, No. 6, pp. 1304 – 1312, 1974

[6] 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

[7] Yuji Nakano, Makoto Tachibana, Junichi Yamagishi, and Takao Kobayashi, "Constrained Structural Maximum A Posteriori Linear Regression for Average-Voice-Based Speech Synthesis", *Proc. INTERSPEECH*, pp. 2286–2289, 2006

[8] M. J. F. Gales, "The Generation and Use of Regression Class Trees for MLLR Adaptation", Technical Report CUED/F-INFENG/TR 263, Cambridge University, 1996

[9] K. C. Sim and M. J. F. Gales, "Adaptation of Precision Matrix Models on Large Vocabulary Continuous Speech Recognition", *Proc. ICASSP*, pp. I-97–100, 2005