Computationally Efficient Body-Conducted Voice Conversion with Original Excitation Signals

Daisuke Deguchi, Tomoki Toda, Hironori Doi, Hiroshi Saruwatari, and Kiyohiro Shikano
Graduate School of Information Science, Nara Institute of Science and Technology, Nara, Japan
E-mail: tomoki@is.naist.jp Tel/Fax: +81-743-72-5261

***Abstract*—In this paper, we propose a computationally efficient method of body-conducted voice conversion. A body-conducted voice is robust against to external noise but its voice quality is severely degraded by mechanisms of body-conduction. The conventional body-conducted voice conversion method effectively enhances the body-conducted voice by converting both spectral and excitation features. On the other hand, its computational cost is relatively high. To significantly reduce the computational cost while keeping the enhanced voice quality as high as possible, we propose a conversion method of using an original excitation signal of the body-conducted voice and computationally efficient feature extraction. The effectiveness of the proposed method is confirmed in the objective and subjective evaluations.**

I. Introduction

Towards noise-robust human-to-human speech communication, there have been several attempts to explore sensing devices as alternatives to the air-conductive microphone. It has been reported that speech signals detected by the bone-conductive microphone can be effectively used to enhance speech sounds under heavy noise conditions [1]. As one of the sensing devices to detect body-conducted voices, we focus on the nonaudible murmur (NAM) microphone [2]. It was originally developed to detect extremely soft murmur called NAM, which is so quiet that people around the speaker barely hear its emitted sound. Placed on the neck below the ear, the NAM microphone can effectively detect air vibrations in the vocal tract from the skin through only the soft tissues of the head. High-quality body-conductive recording of various types of speech, such as a very soft murmur as NAM and a normal voice, is possible from this position because the conduction through obstructions, such as bones whose acoustic impedance is different from that of soft tissues, is avoided. It is also robust against external noise owing to its noise-proof structure like in other body-conductive microphones. One serious drawback of the NAM microphone is that severe degradation of speech quality is caused by essential mechanisms of body conduction. Therefore, its speech quality improvements are essential if it is used in human-to-human speech communication.

To improve speech quality of the body-conducted voice detected with the NAM microphone, body-conducted voice conversion based on a statistical voice conversion technique [3], [4] has been proposed [5]. Joint probability density functions of the acoustic features of the body-conducted voice and those of the normal voice are modeled by Gaussian mixture models (GMMs). By using these GMMs, the acoustic features of the body-conducted voice are converted to those of the normal voice in a probabilistic manner. In the conventional conversion system, STRAIGHT [6] is used as a high-quality analysis-synthesis method to extract the acoustic features as accurately as possible and its mixed excitation model [7] is also used to convert an excitation signal as well. They are very effective for improving the converted voice quality but the computational cost of STRAIGHT analysis is relatively high. To develop a real-time conversion system, it is required to reduce the computational cost as much as possible. This requirement becomes more essential if only the limited resources are available to implement the system.

In this paper, we present a computationally efficient body-conducted voice conversion system. By assuming that the acoustic differences in the excitation signal between the body-conducted voice and normal voice less affect the converted voice quality, the original excitation signal of the body-conducted voice is used in synthesis without any modifications. Several methods are proposed to considerably reduce the computational cost while keeping the converted voice quality as high as possible. To demonstrate the effectiveness of the proposed system, both objective and subjective experimental evaluations are conducted.

II. Body-Conducted Voice Conversion

There are several acoustic differences between the body-conducted voice detected with the NAM microphone and the normal voice detected with the air-conductive microphone. High-frequency components of the body-conducted voice are severely attenuated by lack of radiation characteristics from lips and effect of low-pass characteristics of the soft tissues. Consequently, some phonemes with large power at high-frequency bands, such as unvoiced fricatives, often lose their own specific spectral structures. Waveform power differences are also noticeable. It is hard to compensate for these acoustic differences using simple modifications, such as global linear transformation. Moreover, aperiodic components, which capture noisy strength on each frequency band of the excitation signal [7], are also quite different between the body-conducted voice and the normal voice. These complicated acoustic differences are well dealt with in statistical voice conversion from the body-conducted voice into the normal voice uttered by the same speaker. The conversion process of the conventional body-conducted voice conversion system [5] is shown in **Figure 1**.

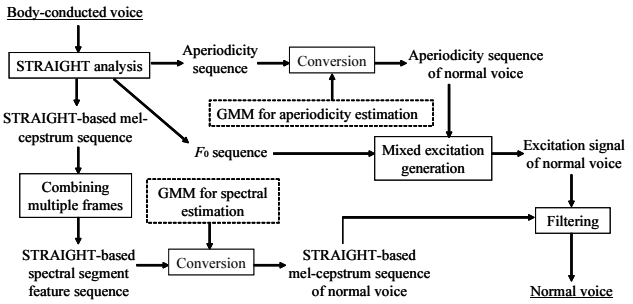

Fig. 1. Conversion process of conventional body-conducted voice conversion system (referred as **system A**). The excitation signal is modeled with STRAIGHT mixed excitation and its aperiodic components are converted.

STRAIGHT analysis is employed to extract time-varying speech parameters of the body-conducted voice. In this paper, mel-cepstrum is used as a spectral feature and F_0 and aperiodic components are used as excitation features. To compensate for the spectral structures often lost at several phonemes through the body-conductive recording, a spectral segment feature vector is extracted at each frame and is used as the source feature vector in the spectral conversion [5]. The spectral segment feature vector $\boldsymbol{X}_t$ at frame t is calculated as

$$\boldsymbol{X}_t = \boldsymbol{A}\left[\boldsymbol{c}_{t-L}^\top, \boldsymbol{c}_{t-L+1}^\top, \cdots, \boldsymbol{c}_t^\top, \cdots, \boldsymbol{c}_{t+L}^\top\right]^\top + \boldsymbol{b}, \quad (1)$$

where $\boldsymbol{c}_t^{(x)} = \left[c_t^{(x)}(1), \cdots, c_t^{(x)}(D)\right]^\top$ is a D-dimensional mel-cepstrum vector of the body-conducted voice at frame t. The transformation matrix $\boldsymbol{A}$ and the bias vector $\boldsymbol{b}$ are determined with principal component analysis (PCA) to extract the lower dimensional spectral segment feature vector by removing redundant components of $2L+1$ mel-cepstrum vectors.

Two GMMs are used in the body-conducted voice conversion: a GMM for spectral conversion and a GMM for aperiodicity conversion. Using parallel data of the body-conducted voice and normal voice as training data, which are recorded simultaneously with the NAM microphone and an air-conductive microphone, joint feature vectors of the source and target features are developed. For the spectral conversion, the spectral segment feature vector of the body-conducted voice $\boldsymbol{X}_t$ and a joint static and dynamic mel-cepstral feature vector of the target normal voice $\boldsymbol{Y}_t = \left[\boldsymbol{c}_t^{(y)\top}, \Delta\boldsymbol{c}_t^{(y)\top}\right]^\top$ are concatenated frame by frame to develop the joint feature vector $\left[\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top\right]^\top$. Using the joint feature vectors, a joint probability density function of the source and target features $P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t | \boldsymbol{\lambda}^{(X,Y)}\right)$ is modeled by a GMM, where $\boldsymbol{\lambda}^{(X,Y)}$ is a parameter set of the GMM. For the aperiodicity conversion, the joint feature vector is developed by concatenating static and dynamic features of aperiodic components of the body-conducted voice and those of the normal voice, and then its joint probability density function is modeled by the other GMM.

In conversion, given a time sequence of the source feature vectors $\boldsymbol{X} = \left[\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_T^\top\right]^\top$, a time sequence of the target static feature vectors $\boldsymbol{y} = \left[\boldsymbol{c}_1^{(y)}, \cdots, \boldsymbol{c}_T^{(y)}\right]^\top$ is determined by maximizing the conditional probability density function $P\left(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}^{(X,Y)}\right)$ of a time sequence of the target static and dynamic feature vectors $\boldsymbol{Y} = \left[\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top$. This estimation is performed under a constraint between static and dynamic feature vectors given by $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y}$, where $\boldsymbol{W}$ is a transformation matrix from the static feature vector sequence to the joint static and dynamic feature vector sequence. The spectral conversion and the aperiodicity conversion are performed independently. The global variance is also considered in the spectral conversion [4]. On the other hand, F_0 values with unvoiced/voiced (U/V) information extracted from the body-conducted voice are not converted.

In synthesis, the converted excitation signal $e^{(\hat{y})}(n)$ is generated with STRAIGHT mixed excitation using the converted aperiodic components and the extracted F_0 values with U/V information. Then, the converted voice signal $s^{(\hat{y})}(n)$ is generated by filtering the excitation signal with the converted spectral filter $h^{(\hat{y})}(n)$. This filtering process is written as

$$S^{(\hat{y})}(z) = H^{(\hat{y})}(z)E^{(\hat{y})}(z), \quad (2)$$

where $S^{(\hat{y})}(z)$, $E^{(\hat{y})}(z)$, and $H^{(\hat{y})}(z)$ are Z-transforms of $s^{(\hat{y})}(t)$, $e^{(\hat{y})}(t)$, and $h^{(\hat{y})}(t)$, respectively. Using Mel Log Spectral Approximation (MLSA) filter [8], $H^{(\hat{y})}(z)$ is given by

$$H^{(\hat{y})}(z) = \exp\sum_{d=0}^{D} c_t^{(\hat{y})}(d)\tilde{z}^{-d}, \quad (3)$$

where $c_t^{(\hat{y})}(d)$ is the d^{th} coefficient of the converted mel-cepstrum at frame t and $\tilde{z}^{-1}$ is the all-pass filter for frequency warping.

III. Body-Conducted Voice Conversion with Original Excitation Signals

STRAIGHT analysis is capable of extracting highly accurate spectral parameters by removing periodic components of the excitation signal from spectral envelope using F_0 information. Moreover, STRAIGHT mixed excitation enables to convert the excitation signal. On the other hand, STRAIGHT analysis is computationally expensive and STRAIGHT mixed excitation sometimes suffers from errors of the F_0 extraction and U/V estimation. To address these issues, a computationally efficient conversion method using the original excitation signals is proposed.

A. Use of Original Excitation Signal

The original excitation signal of the body-conducted voice is well approximated with a residual signal $r^{(x)}(n)$ if it is extracted by inversely filtering the body-conducted voice signal $s^{(x)}(n)$ with the sufficiently accurate spectral filter $h^{(x)}(n)$, such as given by mel-cepstrum extracted from the body-conducted voice with STRAIGHT. The inverse filtering process is given by

$$R^{(x)}(z) = \frac{1}{H^{(x)}(z)}S^{(x)}(z), \quad (4)$$

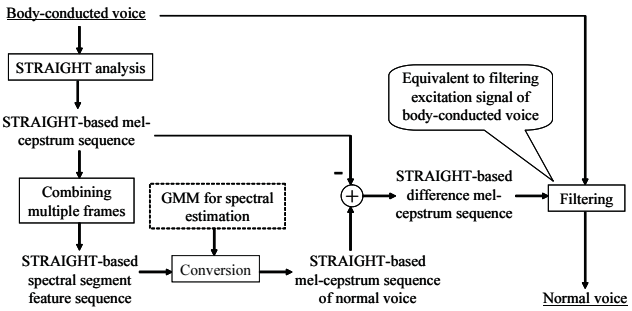

Fig. 2. Conversion process of body-conducted voice conversion system (referred as system **B**) using original excitation signals.

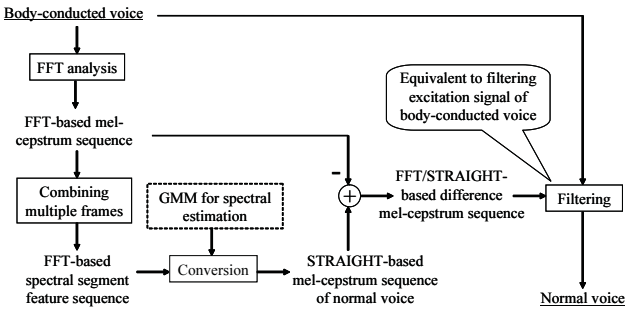

Fig. 3. Conversion process of computationally efficient body-conducted voice conversion system (referred as system **C**) using original excitation signals.

where $R^{(x)}(z)$, $S^{(x)}(z)$, and $H^{(x)}(z)$ are Z-transforms of $r^{(x)}(n)$, $s^{(x)}(n)$, and $h^{(x)}(n)$, respectively. In the proposed conversion method, the residual signal is used as the excitation signal in synthesizing the converted voice as follows:

$$S^{(\hat{y})}(z) = H^{(\hat{y})}(z)R^{(x)}(z) = \frac{H^{(\hat{y})}(z)}{H^{(x)}(z)}S^{(x)}(z). \tag{5}$$

This is equivalent to filtering the body-conducted voice signal with the difference spectral filter given by

$$\frac{H^{(\hat{y})}(z)}{H^{(x)}(z)} = \frac{\exp\sum_{d=0}^{D} c_t^{(\hat{y})}(d)\tilde{z}^{-d}}{\exp\sum_{d=0}^{D} c_t^{(x)}(d)\tilde{z}^{-d}} = \exp\sum_{d=0}^{D} c_t^{(\hat{y}-x)}(d)\tilde{z}^{-d}, \tag{6}$$

$$c_t^{(\hat{y}-x)}(d) = c_t^{(\hat{y})}(d) - c_t^{(x)}(d). \tag{7}$$

This conversion process is shown in **Figure 2**. Although the excitation signal is not converted, this process is free from errors of the F_0 extraction and U/V estimation.

B. Reducing Computational Cost

In the statistical voice conversion, it is important to use a highly accurate spectral parameter for the target feature since it directly affects the converted voice quality. On the other hand, the spectral parameter for the source feature does not directly affect the converted voice quality as far as it is still useful as an explanatory variable to estimate the target features. Therefore, simple FFT-based spectral analysis is used to quickly extract the source mel-cepstrum. This conversion process is shown in **Figure 3**. Note that a GMM for the spectral conversion needs to be trained with the joint feature vectors consisting of the spectral segment feature vectors developed from the FFT-based mel-cepstra of the body-conducted voice and the STRAIGHT-based mel-cepstrum feature vectors of the normal voice. This conversion process is very computationally efficient since STRAIGHT analysis is avoided.

One drawback of this conversion process is that the extraction accuracy of the source mel-cepstrum still affects the filtering process even if it does not affect the conversion accuracy of the target mel-cepstrum. Since the FFT-based spectral analysis is sensitive to the periodicity of excitation signal, the source mel-cepstrum often captures harmonic components of the periodic excitation signal. This causes the degradation of converted voice quality because the difference spectral filter given by Eq. (6) is contaminated by the periodicity of excitation signal.

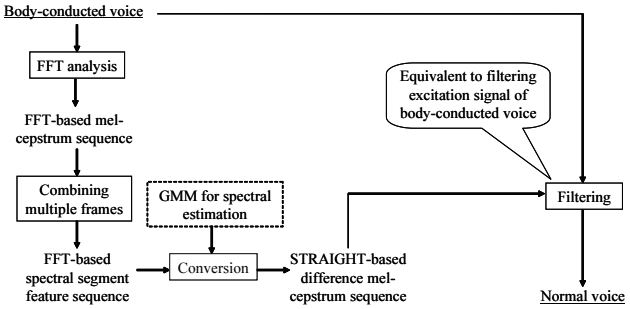

Fig. 4. Conversion process of computationally efficient body-conducted voice conversion system (referred as system **D**) using original excitation signals and direct estimation of difference mel-cepstrum sequence.

TABLE I
COMPARISON OF BODY-CONDUCTED VOICE CONVERSION SYSTEMS SHOWN IN **Figures 1**, **2**, **3**, AND **4**

System index	Excitation	Analysis	Conversion target
A (conventional)	w/ conversion of aperiodicity	STRAIGHT	Target mel-cepstrum and aperiodicity
B	w/o conversion	STRAIGHT	Target mel-cepstrum
C	w/o conversion	FFT	Target mel-cepstrum
D	w/o conversion	FFT	Difference mel-cepstrum

To address this issue, the conversion to a difference mel-cepstrum is proposed. In training, the difference mel-cepstrum $\boldsymbol{c}_t^{(y-x)} = \boldsymbol{c}_t^{(y)} - \boldsymbol{c}_t^{(x)}$ is calculated by subtracting the STRAIGHT-based mel-cepstrum of the body-conducted voice from the STRAIGHT-based mel-cepstrum of the normal voice. Thus, the difference mel-cepstrum does not capture the periodic components of the excitation signal thanks to STRAIGHT analysis. A GMM is trained with the joint feature vectors consisting of the spectral segment feature vectors based on the FFT-based mel-cepstra of the body-conducted voice and the difference mel-cepstral feature vectors. This conversion process is shown in **Figure 4**.

Table I shows a comparison of the body-conducted voice conversion systems depicted in **Figures 1**, **2**, **3**, and **4**. The systems **B**, **C**, and **D** use the original excitation signals. The computational costs of the systems **C** and **D** are much lower than those of the systems **A** and **B**. The system **D** estimates the difference mel-cepstrum to keep the spectral filter from capturing the periodicity of excitation signal.

IV. Experimental Evaluation

A. Experimental Conditions

The body-conducted voice and normal voice were simultaneously recorded by detecting a natural voice uttered by a Japanese male speaker with the NAM microphone and an air-conductive microphone. The sampling frequency was set to 8 kHz. A sentence set consisting of 100 phonetically balanced sentences was recorded with each of three different types of the NAM microphone. In each set, 50 sentences were used for training and the remaining 50 sentences were used for evaluation.

The 0^{th} through 16^{th} mel-cepstral coefficients were used as a spectral feature. Aperiodic components on five frequency bands (*i.e.*, 0-1, 1-2, 2-4, 4-6, and 6-8 kHz) were used as an excitation feature. The shift length was set to 5 ms. In the extraction of the spectral segment feature, a 34-dimensional vector was extracted at each frame from concatenated mel-cepstrum vectors at current and $\pm$ 4 preceding/succeeding frames. The number of mixture components of each GMM was set to 32.

Objective and subjective evaluations were conducted to compare the conversion performance of the systems **A**, **B**, **C**, and **D**. In the objective evaluation, mel-cepstral distortion between the converted voice signal and the target voice signal was used as an evaluation metric. In the subjective evaluation, the opinion test on voice quality was conducted using a 5-point opinion scale, such as 1: very bad, 2: bad, 3: fair, 4: good, and 5: excellent. Eight listeners participated in the test. Each listener evaluated 120 samples consisting of 30 samples for each system, which were randomly selected for each listener.

B. Experimental Results

Figure 5 shows the result of objective evaluation. The system **B** yields lower mel-cepstral distortion than the system **A**. This is because the conversion accuracy of the system **A** suffers from the errors of F_0 extraction and U/V estimation. The system **C** causes very large mel-cepstral distortion since the converted voice signal is directly affected by poor accuracy in mel-cepstrum extraction of the body-conducted voice. This issue is effectively addressed in the system **D**, the conversion accuracy of which is equivalent to that of the system **B**.

The result of the opinion test is also shown in **Figure 5**. The systems **B**, **C**, and **D** using the original excitation signals yield the better converted voice quality than the system **A**. This result implies that the converted voice quality is sensitive to the F_0 extraction error or the U/V estimation error rather than a mismatch of aperiodic components. The system **C** causes the significant degradation in the converted voice quality but this degradation is effectively alleviated by the system **D**. Consequently, the system **D** yields the best converted voice quality equivalent to that of the system **B** while keeping its computational cost much lower than that of the system **B**.

These results show the effectiveness of the system **D**. It could depend on individual speakers whether the system **D** outperforms the system **A**. We plan to evaluate the performance of these systems for various speakers.

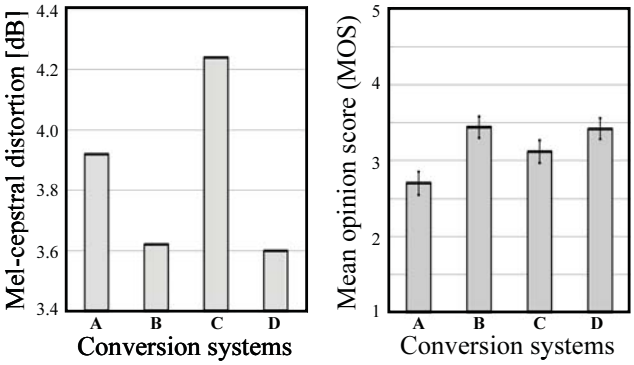

Fig. 5. Mel-cepstral distortion (left) and mean opinion score (right) for each body-conducted voice conversion system shown in **Figures 1**, **2**, **3**, and **4**.

V. Conclusions

In this paper, we proposed computationally efficient body-conducted voice conversion based on FFT-based spectral analysis, conversion from the source spectral feature to a difference spectral feature between the body-conducted voice and the normal voice, and the use of a residual signal of the body-conducted voice as an excitation signal in synthesis. The experimental results showed that the proposed conversion yields better converted voice quality than the conventional conversion since the proposed conversion is free from the errors of F_0 extraction and unvoiced/voiced estimation. Further investigation for various speakers' voices will be performed.

Acknowledgment

This work was supported in part by MEXT Grant-in-Aid for Young Scientists (A). The authors are grateful to Prof. Hideki Kawahara of Wakayama University, in Japan for permission to use the STRAIGHT analysis-synthesis method.

References

[1] A. Subramanya, Z. Zhang, Z. Liu, and A. Acero. Multisensory processing for speech enhancement and magnitude-normalized spectra for speech modeling. *Speech Communication*, Vol. 50, No. 3, pp. 228–243, 2008.

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

[3] 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.

[4] 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, 2007.

[5] T. Toda, K. Nakamura, H. Sekimoto, and K. Shikano. Voice conversion for various types of body transmitted speech. *Proc. ICASSP*, pp. 3601–3604, Taipei, Taiwan, Apr. 2009.

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

[7] H. Kawahara, J. Estill, and O. Fujimura. Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT. *Proc. MAVEBA*, Firenze, Italy, Sep. 2001.

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