VOICE CONVERSION FOR VARIOUS TYPES OF BODY TRANSMITTED SPEECH

Tomoki Toda, Keigo Nakamura, Hidehiko Sekimoto†, Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology (NAIST), Japan

{tomoki, kei-naka, shikano}@is.naist.jp

ABSTRACT

In this paper, we review our proposed statistical voice conversion approaches to enhancing various types of body transmitted speech captured with non-audible murmur (NAM) microphone. Body transmitted speech conversion is a potential technique to bring a new paradigm to human-to-human speech communication. In addition to our previously proposed methods of enhancing body transmitted unvoiced speech for silent speech communication and of enhancing body transmitted artificial speech for speaking aid, we further propose conversion methods of enhancing body transmitted voiced speech for noise robust speech communication. An experimental result demonstrates that the proposed methods yield significant improvements in quality of body transmitted voiced speech.

Index Terms— body transmitted speech, voice conversion, noise robust speech communication, silent speech communication, speaking aid

1. INTRODUCTION

Explosive diffusion of a cellular phone has expanded the capability of speech communication. It enables people to talk with each other even if each of them is in a different place. On the other hand, this excellent device makes us aware of essential problems of speech communication. We have some situations where we face difficulties in uttering. For instance, we need to utter speech loudly under heavy noisy conditions, we have trouble privately talking in the crowd, or talking voices are often recognized as "noise" by other persons around the speaker in special situations (e.g., in the meeting or in the library). Moreover, some people cannot utter due to removal of their organs to generate speech sounds, e.g., laryngectomees. It is eagerly desired to develop technologies to overcome these inherent problems of speech communication and to make speech communication more convenient in any situations and for any persons.

To make speech communication more effective, there have been studied several attempts to explore sensing devices alternative to air microphone. Subramanya *et al*. [1] have proposed a speech enhancement method using an air-and-bone conductive microphone. Bone conductive speech signals are effectively used to enhance speech sounds under heavy noisy conditions. As an attempt to support private speech communication, several devices such as a throat microphone [2], electromyography (EMG) [3], and ultrasound imaging [4] have been studied as *silent speech interfaces*. These new recording devices suggest promising directions for bringing a new paradigm to speech communication.

As one of the microphones to detect body-transmitted speech such as the bone-conductive microphone and the throat microphone, Nakajima *et al*. [5] have developed non-audible murmur (NAM) microphone inspired by a stethoscopic. NAM is defined as very small murmur that is so quiet that the people around the speaker hardly hear it. NAM microphone can capture various types of speech such as not only NAM but also the other types of speech including whisper and normal speech. Its external noise barrier performance is as high as the other body-conductive microphones. Moreover, its usability is better compared with other devices such as EMG or ultrasound systems. Considering these advantages, we focus on NAM microphone as one of the promising devices.

Although NAM microphone enables us to talk in various types of body transmitted speech according to situations, e.g., NAM for *silent speech communication* or body-transmitted normal speech for noise robust speech communication, it has some serious problems for use in speech communication. One of the biggest problems is the quality degradation of body transmitted speech caused by essential mechanisms of body transmission such as lack of radiation characteristics from lips and influence of low-pass characteristics of the soft tissues. Therefore, quality improvements of body transmitted speech are inevitable to use it as a medium for human-to-human communication.

Statistical voice conversion [6], which has originally been proposed for speaker conversion, is one of useful techniques to enhance the body transmitted speech. This technique converts voice characteristics of input speech into those of another output speech while keeping linguistic information unchanged. A statistical model capturing correlations between acoustic features of input and output voices is previously trained using a small amount of parallel data consisting of utterance pairs of these voices. The resulting model allows the conversion from any sample of the input into that of the output using only acoustic information. It is well known that a Gaussian mixture model (GMM), which is a simple probabilistic model, works reasonably well in this framework [7]. This technique is indeed effective for body transmitted speech enhancement because of its ability of real time conversion processing.

We have proposed several conversion methods based on a state-of-the-art GMM-based voice conversion technique [8] to enhance body transmitted speech recorded with NAM microphone. For *silent speech communication*, a conversion method of enhancing body transmitted unvoiced speech [9, 10] has been proposed. To develop a new speaking aid system for total laryngectomees, a conversion method of enhancing body transmitted artificial speech has also been proposed [11]. Although there still remain some problems to be solved, these methods promise to cause new speech communication styles.

In this paper, we review our previous research on voice conversion for body transmitted unvoiced and artificial speech. Moreover, we extend this framework to a body transmitted voiced speech enhancement for noise robust speech communication. Experimental results demonstrate that the proposed conversion methods cause significant improvements in quality of body transmitted voiced speech.

† Presently, with OMRON Corporation, Japan.

The authors are grateful to Prof. Hideki Kawahara of Wakayama University, Japan, for permission to use the STRAIGHT analysis-synthesis method. This research was supported in part by MIC SCOPE.

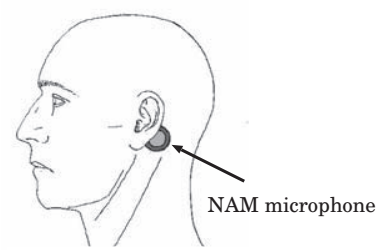

Fig. 1. Setting position of NAM microphone.

2. VARIOUS TYPES OF BODY TRANSMITTED SPEECH DETECTED WITH NAM MICROPHONE

Air vibrations in the vocal tract can be captured with NAM microphone from a position behind the ear shown in **Figure 1** through only the soft tissues of a head. This position allows a high-quality recording of various types of body transmitted speech such as normal speech, whisper, and NAM, by evading the transmission through obstructions such as bones whose acoustic impedance is quite different from that of the soft tissues. Although quality of body transmitted speech recorded with NAM microphone is relatively higher than that with other body transmitted microphones, it is still much lower than natural speech quality. One of main factors causing the quality degradation is lack of high frequency components due to body transmission. Some phonemes with large power on higher frequency bands such as unvoiced fricatives often lose their specific acoustic characteristics.

In this section, we describe three types of body transmitted speech, which are successfully recorded with an improved version of NAM microphone [12].

2.1. Body Transmitted Voiced Speech

Body transmitted voiced speech (BTOS) is defined as natural speech transmitted through the soft tissues of the head. BTOS is very effective under heavy noisy conditions due to a sound proof property of NAM microphone.

Small body transmitted voiced speech (SBTOS) is defined as small speech recorded with NAM microphone. We often talk private things in small voice if there are other persons nearby. Such a small speech is easily masked with external noise sounds. SBTOS is a useful medium in such a situation.

We can extract F_0 values and unvoiced/voiced information from these two types of body transmitted voiced speech. Their spectral and aperiodic excitation features are quite different from those of air transmitted voiced speech.

2.2. Body Transmitted Unvoiced Speech

NAM is defined as articulated respiratory sounds without vocal-fold vibration transmitted through the soft tissues of the head [5]. Because its power is extremely small, NAM is hardly heard by anyone around a speaker.

Body transmitted whisper (BTW) is defined as whisper recorded with NAM microphone [10]. We generally produce the turbulent noise of expiratory air by the stricture of the glottis in uttering whisper to generate large unvoiced sounds because a power of whisper is large enough for us to communicate with persons nearby. Note that NAM often becomes whisper under non-quiet environments because we need auditory feedback to articulate properly.

Body transmitted unvoiced speech is a possible medium for *silent speech communication*. However, it is hard to directly use it for human-to-human speech communication because of its less intelligible and unfamiliar sounds.

2.3. Body Transmitted Artificial Speech

An electrolarynx is an external sound source generator for laryngectomees to speak without vocal-fold vibration. A conventional electrolarynx needs to generate loud enough sound source signals to make articulated sounds (i.e., artificial speech) audible. These sound source signals sound quite mechanical and would be perceived as noise by other persons. NAM microphone provides a solution for this problem, i.e., generating very small signals and detecting the considerably small artificial speech. The artificial speech recorded with NAM microphone is called **body transmitted artificial speech**.

It is indeed difficult to generate a natural F_0 contour from the external devices. Consequently, body transmitted artificial speech doesn't usually have useful F_0 and unvoiced/voiced information.

3. BASIC VOICE CONVERSION ALGORITHM BASED ON MAXIMUM LIKELIHOOD ESTIMATION

Here we describe a conversion method based on maximum likelihood estimation of speech parameter trajectories considering a global variance (GV) [8] as one of the state-of-the-art voice conversion methods.

3.1. Training Process

Let us assume an input static feature vector $\boldsymbol{x}_t = [x_t(1), \cdots, x_t(D_x)]^\top$ and an output static feature vector $\boldsymbol{y}_t = [y_t(1), \cdots, y_t(D_y)]^\top$ at frame t, respectively. In order to compensate for lost characteristics at some phonemes due to body transmission [10], we use a segment feature $\boldsymbol{X}_t = \boldsymbol{W}_x[\boldsymbol{x}_{t-L}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+L}^\top]^\top + \boldsymbol{b}_x$ extracted over several frames ($t \pm L$) as an input speech parameter vector, where $\boldsymbol{W}_x$ and $\boldsymbol{b}_x$ are determined by PCA in this paper. As an output speech parameter vector, we use $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ consisting of both static and dynamic feature vectors.

Using parallel training data set consisting of time-aligned input and output parameter vectors $[\boldsymbol{X}_1^\top, \boldsymbol{Y}_1^\top]^\top$, $[\boldsymbol{X}_2^\top, \boldsymbol{Y}_2^\top]^\top$, $\cdots$, $[\boldsymbol{X}_T^\top, \boldsymbol{Y}_T^\top]^\top$, the joint probability density of the input and output parameter vectors is modeled by a GMM [13] as follows:

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

$$\boldsymbol{\mu}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y)} \end{bmatrix}, \qquad \boldsymbol{\Sigma}_m^{(X,Y)} = \begin{bmatrix} \boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)} \end{bmatrix} \quad (2)$$

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The mixture component index is m. The total number of mixture components is M. A parameter set of the GMM is $\boldsymbol{\lambda}$, which consists of weights w_m, mean vectors $\boldsymbol{\mu}_m^{(X,Y)}$ and full covariance matrices $\boldsymbol{\Sigma}_m^{(X,Y)}$ for individual mixture components.

The probability density of the GV of the output static feature vectors over an utterance is also modeled by a Gaussian distribution,

$$P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}) = \mathcal{N}(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(v)}) \quad (3)$$

where the GV $\boldsymbol{v}(\boldsymbol{y}) = [v(1), \cdots, v(D_y)]^\top$ is calculated by

$$v(d) = \frac{1}{T}\sum_{t=1}^{T}\left(y_t(d) - \frac{1}{T}\sum_{\tau=1}^{T} y_\tau(d)\right)^2. \quad (4)$$

A parameter set $\boldsymbol{\lambda}^{(v)}$ consists of a mean vector $\boldsymbol{\mu}^{(v)}$ and a diagonal covariance matrix $\boldsymbol{\Sigma}^{(v)}$.

3.2. Conversion Process

Let $\boldsymbol{X} = [\boldsymbol{X}_1^\top, \cdots, \boldsymbol{X}_t^\top, \cdots, \boldsymbol{X}_T^\top]^\top$ and $\boldsymbol{Y} = [\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_t^\top, \cdots, \boldsymbol{Y}_T^\top]^\top$ be a time sequence of the input parameter vectors and that of the output parameter vectors, respectively. The converted static feature sequence $\hat{\boldsymbol{y}} = [\hat{\boldsymbol{y}}_1^\top, \cdots, \hat{\boldsymbol{y}}_t^\top, \cdots, \hat{\boldsymbol{y}}_T^\top]^\top$ is determined by maximizing a product of the conditional probability density of $\boldsymbol{Y}$ given $\boldsymbol{X}$ and the GV probability density as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})^\omega P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}) \quad \text{subject to } \boldsymbol{Y} = \boldsymbol{W}_y \boldsymbol{y} \quad (5)$$

where $\boldsymbol{W}_y$ is a window matrix to extend the static feature vector sequence to the parameter vector sequence consisting of static and dynamic features. A balance between $P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda})$ and $P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)})$ is controlled by the weight ω ($= 1/2T$ in this paper).

In this paper, we employ the approximation with suboptimum mixture component sequence [8] to efficiently perform the conversion process. The conditional probability density, which is modeled by a GMM, is approximated with a single mixture component sequence $\boldsymbol{m} = \{m_1, \cdots, m_t, \cdots, m_T\}$ as follows:

$$P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\lambda}) \simeq P(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}) P(\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{m}, \boldsymbol{\lambda}) \quad (6)$$

First, the suboptimum mixture component sequence $\hat{\boldsymbol{m}}$ is determined by

$$\hat{\boldsymbol{m}} = \underset{\boldsymbol{m}}{\operatorname{argmax}} P(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}). \quad (7)$$

Then the converted static feature vector sequence is determined by

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} P(\boldsymbol{W}_y \boldsymbol{y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda})^\omega P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}) \quad (8)$$

where note $\boldsymbol{Y} = \boldsymbol{W}_y \boldsymbol{y}$. We iteratively update the converted sequence $\hat{\boldsymbol{y}}$ using the gradient method [8].

4. VOICE CONVERSION METHODS FOR BODY TRANSMITTED SPEECH

In this section, we describe a voice conversion method for enhancing each type of body transmitted speech.

4.1. Conversion for body transmitted voiced speech

In order to improve quality of BTOS, we propose a conversion method from BTOS into speech (**BTOS-to-Speech**). This method employs two GMMs, one for converting a spectral segment sequence of BTOS into a spectral sequence of speech, and the other for converting aperiodic components to design mixed excitation [14] of BTOS into those of speech, respectively. The extracted F_0 values are directly used for synthesizing the converted speech.

As an enhancement method for SBTOS, we propose a conversion method from SBTOS into speech (**SBTOS-to-Speech**). We have found that there are several mismatches between speech and small speech in the source features: 1) F_0 of small speech tends to be relatively lower than that of natural speech; and 2) some voiced phonemes are often devoiced in small speech. In order to cope with the first problem, F_0 is also converted with a global linear transform while the other speech features (i.e., a spectrum and aperiodic components) are converted in a similar manner as in BTOS-to-Speech. However, the second problem still remains in SBTOS-to-Speech. In order to avoid this problem, we also propose a conversion method from SBTOS into small speech (**BTOS-to-Small Speech**) as an alternative way. There are not any severely mismatched acoustics between input and output voices in this framework. Therefore, we need to convert only spectral and aperiodic features and can directly use the extracted F_0 values including voiced/unvoiced information as in the case of BTOS-to-Speech.

4.2. Conversion of body transmitted unvoiced speech [9, 10]

We have proposed the voice conversion method from NAM to speech (**NAM-to-Speech**) [9]. This conversion method employs three GMMs for converting a spectral segment sequence of NAM into a speech spectral sequence, an F_0 sequence including unvoiced/voiced information, and an aperiodic component sequence, respectively. Although the converted speech by this method sounds more intelligible and more similar voice quality to natural speech than the original NAM, unnatural prosody is always caused by difficulties of the F_0 contour estimation from acoustics of unvoiced speech.

To avoid this problem, we have proposed the voice conversion method from NAM to whisper (**NAM-to-Whisper**) [10]. Whisper is familiar unvoiced speech and has enough intelligibility and naturalness. This conversion uses a single GMM for converting a spectral segment sequence of NAM into a spectral sequence of whisper. White noise can be employed as an excitation signal. It has been reported that NAM-to-Whisper outperforms NAM-to-Speech in views of both intelligibility and naturalness of the converted speech.

The conversion method from BTW to whisper (**BTW-to-Whisper**) also works reasonably well [10]. Furthermore, a single GMM converting both NAM and BTW into whisper is effectively trained using two parallel data sets of NAM-and-whisper and BTW-and-whisper simultaneously.

4.3. Conversion for body transmitted artificial speech [11]

We have proposed a new speaking aid system for total laryngectomees based on three main techniques: 1) extremely small sound source signals; 2) NAM microphone to capture extremely small artificial speech; and 3) the voice conversion for enhancing body transmitted artificial speech [11]. This system would allow laryngectomees to utter more intelligible and natural speech while keeping the external sound source signals from annoying other persons. Whisper is employed as the output speech in the conversion as in the case of body transmitted unvoiced speech conversion.

Experimental results using artificial speech imitated by a non-laryngectomee have demonstrated that this system causes significant improvements in both naturalness and intelligibility compared with the conventional speaking aid system using an existing electrolarynx [15]. On the other hand, the different results have been observed in the experiments using artificial speech by an actual laryngectomee as described in [16]: naturalness is very significantly improved but intelligibility tends to degrade through the conversion.

5. EXPERIMENTAL EVALUATION IN BODY TRANSMITTED VOICED SPEECH CONVERSION

In order to demonstrate the effectiveness of the proposed body transmitted voiced speech conversion methods described in **Section 4.1**, we conducted an experimental evaluation.

5.1. Experimental Conditions

We recorded BTOS and SBTOS uttered by four Japanese speakers (two males and two females). Air-transmitted speech was simultaneously recorded during the recording of each speaking style. Each speaker uttered about 50 phoneme balanced sentences as training data and about 105 newspaper article sentences for evaluation. The sampling frequency was 8 kHz.

The 0^{th} through 16^{th} mel-cepstral coefficients were used as a spectral feature at each frame, which was extracted with STRAIGHT

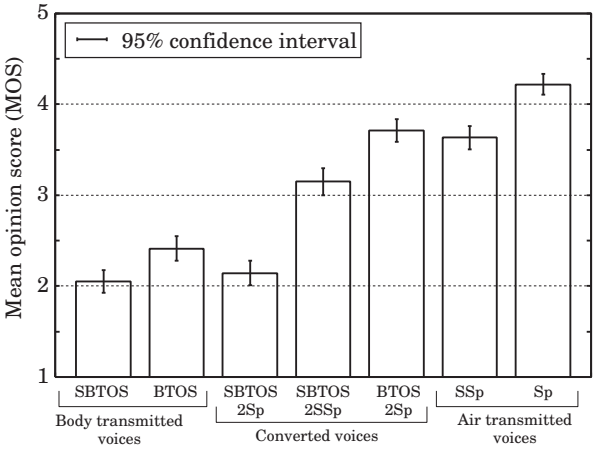

Fig. 2. Mean opinion score for each speech sample.

analysis [17]. We used the 34-dimensional segment feature at each input frame extracted using PCA from multiple frames around a current one. As source features, we used a log-scaled F_0 extracted with STRAIGHT F_0 extractor [18] and aperiodic components on five frequency bands, i.e., 0-1, 1-2, 2-4, 4-6, and 6-8 kHz [14]. The shift length was 5 ms.

We trained the following three conversion models: 1) **BTOS2Sp** for converting BTOS to speech; 2) **SBTOS2Sp** for converting SBTOS to speech; and 3) **SBTOS2SSp** for converting SBTOS to small speech. We optimized the number of mixture components of each GMM and the number of concatenated frames for extracting the segment feature so that the best feature conversion accuracy was achieved in the evaluation set. The optimum settings were 64 mixture components for the spectral conversion, 4 mixture components for the aperiodic conversion, and current$\pm$4 frames for extracting the segment feature.

We conducted an opinion test on speech quality using the converted speech samples from the above three conversion models as well as four analysis-synthesized speech samples of speech (**Sp**), small speech (**SSp**), **BTOS** and **SBTOS**. The number of listeners was ten. Each listener evaluated 35 samples consisting of five sentences for each of four speakers, i.e., 140 samples in total. These sentences were randomly selected for each speaker and each listener from the evaluation data.

5.2. Experimental Result

Figure 2 shows a result of the test. We can see that the two types of body transmitted voiced speech **BTOS** and **SBTOS** have poor quality. Quality of BTOS is significantly improved by **BTOS2Sp**. For SBTOS, **SBTOS2Sp** doesn't cause any quality improvements. One of main factors causing this result is inconsistency of voiced/unvoiced information between speech and small speech. **SBTOS2SSp** yields significant improvements in quality of SBTOS by effectively avoiding this problem. These results suggest that **BTOS2Sp** and **SBTOS2SSp** are very effective enhancement methods for individual types of body transmitted voiced speech.

6. CONCLUSIONS

This paper has described statistical voice conversion methods for enhancing various types of body transmitted speech recorded with non-audible murmur microphone. In addition to previously proposed conversion methods for body transmitted unvoiced and artificial speech, we have further proposed the conversion methods for body transmitted voiced speech. An experimental result has demonstrated that the proposed methods yield very significant improvements in quality of body transmitted voiced speech. Our next step is to develop low-delay conversion systems based on [19].

7. REFERENCES

[1] A. Subramanya, Z. Zhang, Z. Liu, 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] S-C. Jou, T. Schultz, and A. Waibel. Adaptation for soft whisper recognition using a throat microphone. *Proc. INTERSPEECH*, pp. 1493–1496, Jeju Island, Korea, 2004.

[3] L. Maier-Hein, F. Metze, T. Schultz, and A. Waibel. Session independent non-audible speech recognition using surface electromyography. *Proc. ASRU*, pp. 331–336, San Juan, Puerto Rico, Nov. 2005.

[4] T. Hueber, G. Chollet, B. Denby, G. Dreyfus, M. Stone. Continuous-speech phone recognition from ultrasound and optical images of the tongue and lips. *Proc. Interspeech*, pp. 658–661, Antwerp, Belgium, Aug. 2007.

[5] 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.

[6] H. Kuwabara and Y. Sagisaka. Acoustic characteristics of speaker individuality: control and conversion. *Speech Communication*, Vol. 16, No. 2, pp. 165–173, 1995.

[7] 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.

[8] T. Toda, A.W. Black, and K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235, Nov. 2007.

[9] T. Toda and K. Shikano. NAM-to-speech conversion with Gaussian mixture models. *Proc. INTERSPEECH*, pp. 1957–1960, Lisbon, Portugal, Sep. 2005.

[10] M. Nakagiri, T. Toda, H. Saruwatari, and K. Shikano. Improving body transmitted unvoiced speech with statistical voice conversion. *Proc. INTERSPEECH*, pp. 2270–2273, Pittsburgh, USA, Sep. 2006.

[11] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano. Speaking aid system for total laryngectomees using voice conversion of body transmitted artificial speech. *Proc. INTERSPEECH*, pp. 1395–1398, Pittsburgh, USA, Sep. 2006.

[12] Y. Nakajima, H. Kashioka, K. Shikano, and N. Campbell. Remodeling of the sensor for non-audible murmur (NAM). *Proc. INTERSPEECH*, pp. 389–392, Lisbon, Portugal, Sep. 2005.

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

[14] Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Maximum likelihood voice conversion based on GMM with STRAIGHT mixed excitation. *Proc. INTERSPEECH*, pp. 2266–2269, Pittsburgh, USA, Sep. 2006.

[15] K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano. Impact of various small sound source signals on voice conversion accuracy in speech communication aid for laryngectomees. *Proc. INTERSPEECH*, pp. 2517–2520, Antwerp, Belgium, Aug. 2007.

[16] K. Nakamura, T. Toda, Y. Nakajima, H. Saruwatari, and K. Shikano. Evaluation of speaking-aid system with voice conversion for laryngectomees toward its use in practical environments. *Proc. INTERSPEECH*, pp. 2209–2212, Brisbane, Australia, Sep. 2008.

[17] 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.

[18] H. Kawahara, H. Katayose, A.de Cheveigné, and R.D. Patterson. Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimatT. Todaion of F_0 and periodicity. *Proc. EUROSPEECH*, pp. 2781–2784, Budapest, Hungary, Sep. 1999.

[19] T. Muramatsu, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Low-delay voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *Proc. INTERSPEECH*, pp. 1076–1079, Brisbane, Australia, Sep. 2008.