Improving Body Transmitted Unvoiced Speech with Statistical Voice Conversion

Mikihiro Nakagiri†, Tomoki Toda, Hideki Kashioka, Kiyohiro Shikano

Graduate School of Information Science, Nara Institute of Science and Technology, Japan

tomoki@is.naist.jp

Abstract

The conversion method from Non-Audible Murmur (NAM) to ordinary speech based on the statistical voice conversion (NAM-to-Speech) has been proposed towards realization of "silent speech telephone." Although NAM-to-Speech converts NAM to intelligible voices with similar quality to speech, there is still a large problem, i.e., difficulties of the F_0 estimation from unvoiced speech. In order to avoid this problem, we propose a conversion method from NAM to whisper that is a familiar and intelligible unvoiced speech (NAM-to-Whisper). Moreover, we enhance NAM-to-Whisper so that multiple types of body-transmitted unvoiced speech such as NAM and Body Transmitted Whisper (BTW) are accepted as input voices. We evaluate the performance of the proposed conversion method. Experimental results demonstrate that 1) intelligibility and naturalness of NAM are significantly improved by NAM-to-Whisper, 2) NAM-to-Whisper outperforms NAM-to-Speech, and 3) we can train a single conversion model successfully converting both NAM and BTW to the target voice.

Index Terms: silent speech telephone, body transmitted unvoiced speech, voice conversion, F_0 estimation, whisper

1. Introduction

Cellular phones have enabled us to communicate with each other by speech whenever and wherever. However, it has caused a problem. Speech is recognized as "noise" by the other persons around a speaker in some situations such as a meeting. In order to address this problem, we aim to realize "silent speech telephone" allowing speech communication annoying nobody in any situation based on Non-Audible Murmur (NAM) sensor and voice conversion.

Nakajima et al. [1] found that air vibrations in the vocal tract can be captured with a special acoustic sensor called NAM microphone from a position behind the ear through only the soft tissues of a head. Because of evading the transmission through obstructions such as bones whose acoustic impedance is quite different from that of the soft tissues, this position allows a high-quality recording of various types of body transmitted speech such as normal speech and considerably small whisper. Therefore, we focus on NAM microphone rather than other sensor devices such as bone microphone [2] and throat microphone [3].

We can actually talk in NAM while keeping silent. However, it is hard to directly use NAM as a medium for human communication because of its less intelligibility and unfamiliar sounds. Toda and Shikano [4] proposed the statistical voice conversion method from NAM to ordinary speech (NAM-to-Speech) for addressing this problem. In advance, we train Gaussian mixture models (GMMs) for representing correlations between acoustic features of NAM and those of speech using a small number of parallel data of NAM and speech, e.g., 50 utterance-pairs. Any sample of the features of NAM is converted to that of speech using the trained GMMs. NAM-to-Speech is a very useful technique to improve NAM in view of both intelligibility and voice quality. However, there is a remaining problem, i.e., the converted speech has unnatural prosody caused by difficulties of the estimation from an acoustic spectral sequence of NAM to an F_0 contour of speech. Unfortunately, we have an impression that an achievement of the F_0 estimation with acceptable quality is quite difficult.

We propose the voice conversion method from NAM to whisper in this paper. We can avoid the F_0 estimation by using not speech but whisper as the target speech. Whisper is a familiar unvoiced speech and has enough intelligibility and naturalness. In addition, we enhance NAM-to-Whisper so that the conversion system widely accepts various kinds of body-transmitted unvoiced speech such as NAM and Body Transmitted Whisper (BTW). Experimental results demonstrate that NAM-to-Whisper works very well and both NAM and BTW are converted to the target speech using the single conversion model.

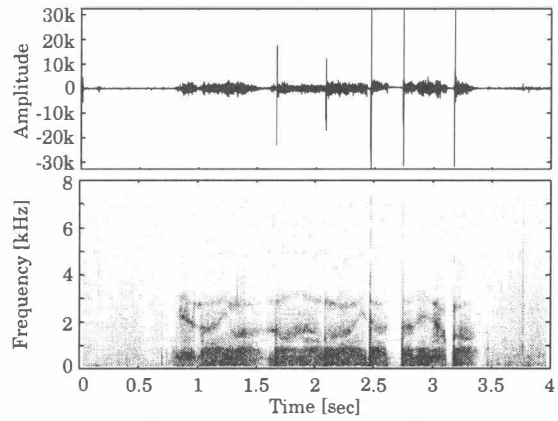
The paper is organized as follows. In **Section 2**, we describe body transmitted unvoiced speech. In **Section 3**, the conversion methods for improving NAM are described. In **Section 4**, the conversion method accepting both NAM and BTW as an input is described. In **Section 5**, experimental evaluations are described. Finally, we summarize this paper in **Section 6**.

2. Body Transmitted Unvoiced Speech: NAM and BTW

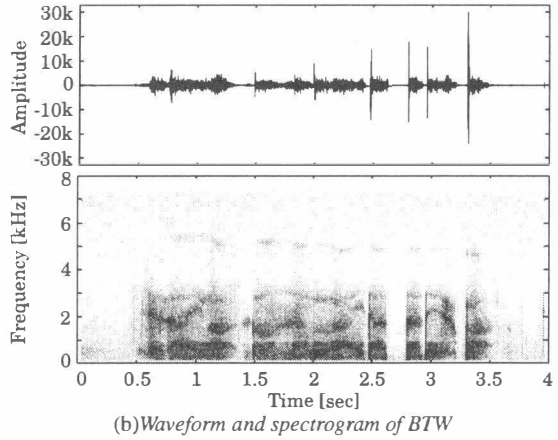
We focus on two kinds of body transmitted unvoiced speech, i.e., NAM and BTW, in this paper. NAM is defined as articulated respiratory sounds without vocal-fold vibration transmitted through the soft tissues of the head [1]. Anyone around a speaker hardly hears NAM because its power is extremely small. On the other hand, BTW is defined as whisper transmitted through the soft tissues of the head. We can communicate with some limited number of persons nearby using whisper because its power is enough large. In order to generate unvoiced source signals with enough power, we generally use the turbulent noise of expiratory air produced by the stricture of the glottis in uttering whisper. NAM often becomes whisper especially under noisy environments because we need to hear own voice for speaking.

Figure 1 shows an example of waveforms and spectrograms of NAM and BTW. Those signals are recorded the NAM microphone called Open Condenser Mediated with Soft Silicone (OCMSS) type [5]. We can see differences of not only a total power but also frequency components around 5 kHz between NAM and whisper. They might be caused by the stricture of the glottis. Higher frequency components of body transmitted speech usually disappear because it doesn't include radiation characteristics from lips and it is affected by low-pass characteristics of the body transmission. Consequently, some phonemes with large power on higher frequency bands such as unvoiced fricatives often lose their spe-

†Presently, with Matsushita Electric Industrial Co., Ltd., Japan.



(a)*Waveform and spectrogram of NAM*



(b)*Waveform and spectrogram of BTW*

Figure 1: *An example of waveforms and spectrograms of body transmitted unvoiced speech.*

cific characteristics.

3. Voice Conversion Methods for NAM

We employ the statistical voice conversion method with a GMM [6] for improving intelligibility and naturalness of NAM.

3.1. Conversion with Maximum Likelihood Estimation [7]

We train a GMM for representing the joint probability density of input and output features using around 50 utterance-pairs of input and output voices [8]. Once we train the GMM, we can convert any sample of the input feature to that of the output feature. In the conversion, we determine a time sequence of conditional probability density functions (pdfs) of the output features for the given input feature sequence based on the GMM. And then, we estimate the output feature sequence that maximizes likelihoods of the conditional pdfs. The conversion accuracy is improved by maximizing likelihoods on both static and dynamic features with respect to the output static features considering an explicit relationship between those features. Furthermore, the naturalness of converted speech is significantly improved by considering global variance (GV) of the output feature trajectory [7].

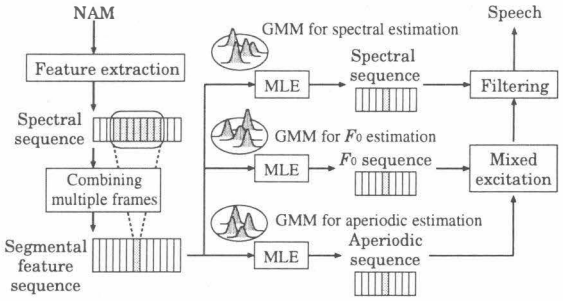


Figure 2: *Conversion Process of NAM-to-Speech.*

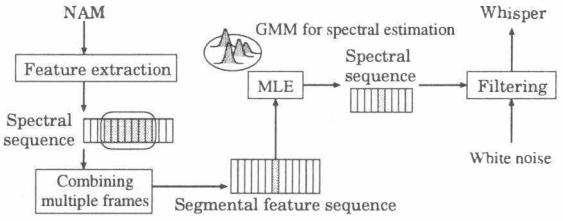


Figure 3: *Conversion Process of NAM-to-Whisper.*

3.2. NAM-to-Speech [4]

A conversion process of NAM-to-Speech is shown in **Figure 2**. We can use only spectral features as an input feature because NAM is unvoiced. In order to synthesize speech, we need to estimate not only spectral features but also source features such as F_0s from NAM spectra.

We construct a spectral segment feature at each frame by concatenating spectral vectors at current, preceding and succeeding frames for compensating for lost characteristics at some phonemes as mentioned in **Section 2**. We use three GMMs for converting the segment feature of NAM to three speech features, i.e., the spectrum, the F_0 including unvoiced/voiced (U/V) information, and an aperiodic component capturing noisy strength on each frequency band of the source signal. We design a mixed excitation based on the estimated F_0 and aperiodic components [11]. And then, we synthesize the converted speech by filtering the mixed excitation with the estimated spectra.

3.3. NAM-to-Whisper

A conversion process of NAM-to-Whisper is shown in **Figure 3**. In order to synthesize whisper, we need to estimate only spectral features from NAM spectra because we just use white noise as the source signal of unvoiced speech. The spectral estimation is performed in the same manner as in the NAM-to-Speech.

Main advantage of this conversion is that we don't have to estimate F_0s. Another advantage is that spectral characteristics of whisper are more similar to those of NAM than those of speech. **Figure 4** shows standard deviations of mel-cepstral coefficients of speech, whisper, and NAM. We can see that the deviations of whisper are smaller than those of speech especially at the 0^{th} and 1^{st} coefficients. The large deviations of speech might be caused by large variations of a total power and a spectral tilt of speech due to U/V changes. NAM-to-Speech seems more difficult process than NAM-to-Whisper even in the spectral estimation.

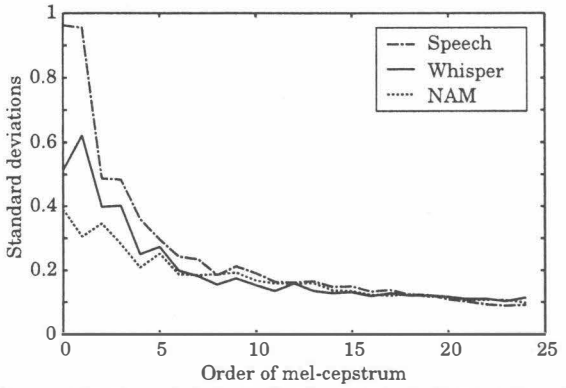


Figure 4: *Standard deviations of mel-cepstral coefficients. We calculate them at speech frames except for silence frames of 50 utterances.*

4. Simultaneous Modeling of NAM and BTW

It is practical to deal with not only NAM but also other speaking styles, e.g., BTW, as an input form of "silent speech telephone" because we select an appropriate speaking style according to situations such as noisy conditions.

One solution is that users switch a proper conversion system according to the current speaking style. In order to cope with NAM and BTW, we need to use two conversion systems for NAM-to-Whisper and for converting BTW to Whisper (BTW-to-Whisper). The model for BTW-to-Whisper is trained using a parallel data set consisting of utterance-pairs of BTW and whisper in the same manner as in the model for NAM-to-whisper.

Another solution is to construct a single conversion system accepting multiple input speaking styles. This is more convenient for users than the former solution. We train a GMM for simultaneously modeling the joint probability of acoustic features of NAM and whisper and that of BTW and whisper using two parallel data sets, i.e., utterance-pairs of NAM and whisper and those of BTW and whisper. The resulting GMM converts the acoustic features of both NAM and BTW into those of whisper. It is possible to apply the simultaneous modeling to the conversion from NAM and BTW to speech as well.

5. Experimental Evaluations

We evaluated the performance of NAM-to-Whisper compared with that of NAM-to-Speech. Moreover, we performed objective evaluations for investigating the effect of the simultaneous modeling of NAM and BTW.

5.1. Experimental Conditions

We used NAM, BTW, whisper and speech uttered by a Japanese female speaker. There were 149 utterances in each. We used 64 utterances forthe training and remaining 85 utterances for the evaluations.

The 0^{th} through 24^{th} mel-cepstral coefficients were used as a spectral feature at each frame. The mel-cepstral analysis [9] was employed for unvoiced speech, i.e., NAM, BTW, and whisper. On the other hand, STRAIGHT analysis [10] was employed for speech. We used the 50 dimensional spectral segment feature compressed with PCA at each input frame [4]. As source features of speech, we used a log-scaled F_0 and aperiodic components on five frequency bands, i.e., 0-1, 1-2, 2-4, 4-6, and 6-8 kHz, which were used for designing the mixed excitation [11].

Table 1: *Result of intelligibility test.*

	Word correct [%]	Word accuracy [%]	Number of replays
Speech	94.65	94.13	1.91
Whisper	91.46	91.08	2.09
NAM	45.90	45.25	4.33
NAM-to-Speech	71.77	69.79	3.23
NAM-to-Whisper	75.85	75.71	3.03

We trained the following six conversion models:

NAM2Sp for converting NAM to speech,

NAM2Wh for converting NAM to whisper,

BTW2Sp for converting BTW to speech,

BTW2Wh for converting BTW to whisper,

(NAM+BTW)2Sp for converting NAM and BTW to speech,

(NAM+BTW)2Wh for converting NAM and BTW to whisper.

In the simultaneous modeling, i.e., **(NAM+BTW)2Sp** and **(NAM+BTW)2Wh**, we used double-sized training data compared with the others. We optimized the number of mixtures of each GMM and the number of concatenated frames for the spectral segment feature so that the feature conversion accuracy was minimized in the evaluation set.

5.2. Evaluations of NAM Conversion Methods

We performed perceptual evaluations on intelligibility and naturalness of the following five voices: 1) analysis-synthesized speech, 2) analysis-synthesized whisper, 3) analysis-synthesized NAM, 4) NAM-to-Speech voice with **NAM2Sp**, and 5) NAM-to-Whisper voice with **NAM2Wh**.

5.2.1. Intelligibility

We performed a perceptual test on the intelligibility by dictation. We allowed listeners to replay the same stimulus time after time. Ten Japanese listeners who have never listened to NAM participated in the test. We used 50 sentences in the evaluation set.

Table 1 shows word correct, word accuracy, and the average number of replays by listeners. The voice conversion considerably improves intelligibility of NAM. We can see an interesting result that NAM-to-Whisper causes more intelligible voices than NAM-to-Speech. This might be caused by the better spectral conversion accuracy in NAM-to-Whisper compared with in NAM-to-Speech as described in the following section. Moreover, NAM-to-Whisper is not affected by the U/V estimation error causing word insertion errors as shown in the result of NAM-to-Speech.

5.2.2. Naturalness in terms of human voices

We performed an opinion test on the naturalness in terms of not ordinary speech but human voices using a 5-point scale. Five Japanese listeners who have never listened to NAM participated in the test. Each listener evaluated 25 sentences randomly selected from the evaluation set for each voice.

Figure 5 shows a result of the test. NAM-to-Whisper causes an admirable improvement of the naturalness. The converted voice

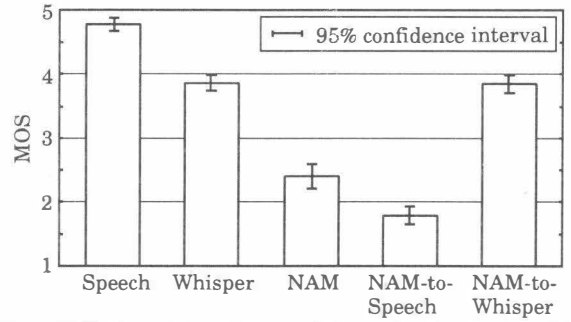


Figure 5: *Result of opinion test on the naturalness in terms of human voices.*

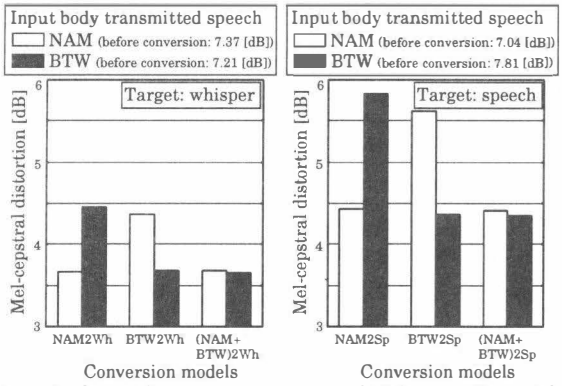


Figure 6: *Spectral conversion accuracy of each conversion model for body transmitted unvoiced speech.*

has the same naturalness as whisper. On the other hand, NAM-to-Speech causes the naturalness degradation. Although the NAM-to-Speech voice has much more similar voice quality to ordinary speech compared with NAM [4], it has an unnatural F_0 contour making listeners feel that it doesn't sound human voices very well.

From those results, it is demonstrated that NAM-to-Whisper is very useful method for improving NAM.

5.3. Evaluations of Simultaneous Modeling of NAM and BTW

We objectively evaluated the performance of each conversion model for investigating the effect of the simultaneous modeling.

Figure 6 shows mel-cepstral distortion when using each conversion model. Inconsistency of input speaking styles in the training and the conversion processes causes large degradation of the conversion accuracy. The simultaneous modeling is very effective for addressing that problem. As for the conversion into speech, we also shows the source feature conversion accuracies in **Fig. 7**. We can see the same results as in the spectral conversion.

It is shown that the mel-cepstral distortion in the conversion into whisper is much smaller than that in the conversion into speech. Namely, the spectral conversion between unvoiced voices is much easier than that between unvoiced and voiced ones.

6. Conclusions

We proposed a statistical conversion method from Non-Audible Murmur (NAM) to whisper (NAM-to-Whisper) for avoiding a problem of the conversion from NAM to Speech (NAM-to-Speech), i.e., difficulties of the F_0 estimation from unvoiced speech. We also proposed a training method of a single conversion model allowing both NAM and Body Transmitted Whisper (BTW) as an input. We evaluated the performance of the proposed methods. Experimental results demonstrated that 1) intelligibility and naturalness of NAM are considerably improved by NAM-to-Whisper, 2) NAM-to-Whisper outperforms NAM-to-Speech, and 3) both NAM and BTW are successfully converted to the target voice using the single conversion model.

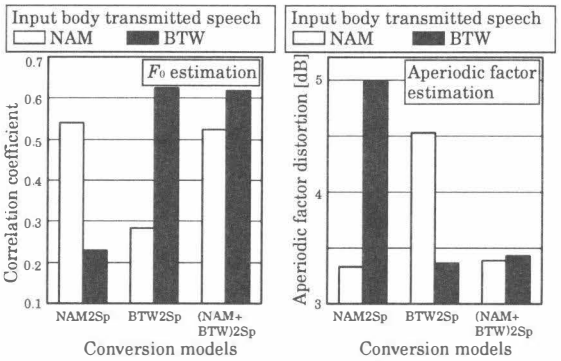


Figure 7: *Source feature conversion accuracies of each conversion model from body transmitted unvoiced speech to ordinary speech.*

Acknowledgment: This research was supported in part by MIC SCOPE-S and MEXT e-Society leading project.

7. References

[1] 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.

[2] Y. Zheng, Z. Liu, Z. Zhang, 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, St. Thomas, USA, Dec. 2003.

[3] 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.

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

[5] 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.

[6] 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.

[7] T. Toda, A.W. Black, and K. Tokuda. Spectral conversion based on maximum likelihood estimation considering global variance of converted parameter. *Proc. ICASSP*, Vol. 1, pp. 9–12, Philadelphia, USA, Mar. 2005.

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

[9] 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.

[10] 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.

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