Non-Audible Murmur Enhancement based on Statistical Conversion using Air- and Body-Conductive Microphones in Noisy Environments

Yusuke Tajiri, Kou Tanaka, Tomoki Toda, Graham Neubig, Sakriani Sakti, Satoshi Nakamura

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

{tajiri.yusuke.tk0, ko-t, tomoki, neubig, ssakti, s-nakamura}@is.naist.jp

Abstract

Non-Audible Murmur (NAM) is an extremely soft whispered voice detected by a special body-conductive microphone called a NAM microphone. Although NAM is a promising medium for silent speech communication, its quality is significantly degraded by its faint volume and spectral changes caused by body-conductive recording. To improve the quality of NAM, several enhancement methods based on statistical voice conversion (VC) techniques have been proposed, and their effectiveness has been confirmed in quiet environments. However, it can be expected that NAM will be used not only in quiet, but also in noisy environments, and it is thus necessary to develop enhancement methods that will also work in these cases. In this paper, we propose a framework for NAM enhancement using not only the NAM microphone but also an air-conductive microphone. Air- and body-conducted NAM signals are used as the input of VC to estimate a more naturally sounding speech signal. To clarify adverse effects of external noises on the performance of the proposed framework and investigate a possibility to alleviate them by revising VC models, we also implement noise-dependent VC models within the proposed framework. Experimental results demonstrate that the proposed framework yields significant improvements in the spectral conversion accuracy and listenability of enhanced speech under both quiet and noisy environments.

Index Terms: silent speech communication, Non-Audible Murmur, statistical voice conversion, air- and body-conducted speech signals, noisy conditions

1. Introduction

Speech communication plays a principal role in our daily life as the most efficient human communication method. In recent decades, we have come to be able to talk with each other beyond limitations of distance and location thanks to the advancement of information technologies such as mobile phones. However, there still exist some situations where we hesitate to talk with others. For instance, we have difficulty in talking about private information in a crowd; or speaking itself would sometimes annoy others in quiet environments.

Recently, *silent speech interfaces* have attracted attention as a technology to achieve a new style of speech communication [1]. They enable us to talk with each other while keeping silent. As one of the silent speech interfaces, Non-Audible Murmur (NAM) microphone has been developed [2]. **Figure 1** shows the setting position and structure of the NAM microphone. NAM is an extremely soft whispered voice, and it is hardly heard by people around the speaker because of its faint volume. The NAM microphone was originally designed to detect such an extremely soft speech through the soft tissue of the speaker's head. However, the quality of body-conducted NAM signals is significantly degraded in particular by spectral changes caused by an essential mechanism of body-conduction, such as influence of low-pass characteristics of the soft tissue and lack of radiation characteristics from lips. To address this issue, statistical voice conversion (VC) techniques [3, 4] for body-conducted speech enhancement have been proposed [5]. In these approaches, acoustic features of body-conducted speech are converted into those of normal speech or a whispered voice. It has been reported that NAM enhancement methods based on these approaches significantly improve the speech quality and intelligibility of NAM. On the other hand, their effectiveness has been confirmed by experimental evaluations using only body-conducted NAM signals recorded in a soundproof room. It can be expected that NAM will be used in noisy environments, such as in a crowd. Although the NAM microphone is more robust against external noises compared to an air-conductive microphone, body-conducted NAM signals will still suffer from them. Conversion accuracy of the statistical VC techniques would be significantly degraded if acoustic mismatches were observed between training and testing conditions. It is thus necessary to develop enhancement methods that will also work in these cases.

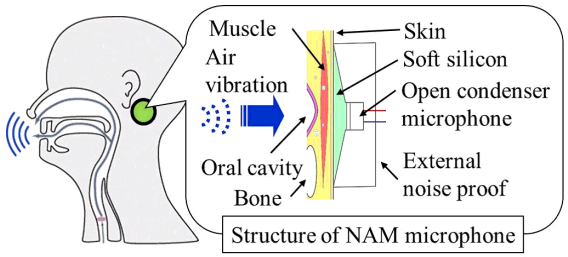


Figure 1: *Setting position and structure of NAM microphone.*

In this paper, we propose a new framework for NAM enhancement using not only the NAM microphone but also an air-conductive microphone. Air- and body-conducted NAM signals are simultaneously used as the input of the statistical VC-based NAM enhancement processing. To clarify adverse effects of external noises on the performance of the proposed framework and investigate a possibility to alleviate them by revising VC models, we also implement noise-dependent VC models within the proposed framework. Experimental results demonstrate that the proposed methods yield significant improvements in the spectral conversion accuracy and listenability of enhanced speech under both quiet and noisy environments.

2. NAM Enhancement Methods Based on Statistical Voice Conversion [5]

In the statistical NAM enhancement methods, acoustic features of NAM are converted into those of normal speech or a whispered voice. In this section, we briefly describe the procedure of these enhancement methods.

2.1. Training Process

Let us assume a source static feature vector $\boldsymbol{x}_t = [x_t(1), \cdots, x_t(D)]^\top$ and a target static feature vector $\boldsymbol{y}_t = [y_t(1), \cdots, y_t(D)]^\top$ at frame t, respectively. In the statistical VC techniques for the body-conducted speech enhancement, to compensate for lost acoustic characteristics at some phonemes due to body conduction, a segment feature $\boldsymbol{X}_t = \boldsymbol{A}[\boldsymbol{x}_{t-L}^\top, \cdots, \boldsymbol{x}_t^\top, \cdots, \boldsymbol{x}_{t+L}^\top]^\top + \boldsymbol{b}$ is calculated from current one $\pm L$ frames, where $\boldsymbol{A}$ and $\boldsymbol{b}$ are determined by principal component analysis (PCA). As a target speech parameter vector, a joint static and dynamic feature vector $\boldsymbol{Y}_t = [\boldsymbol{y}_t^\top, \Delta\boldsymbol{y}_t^\top]^\top$ is extracted. Using a parallel training data set consisting of time-aligned source and target feature 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 source and target feature vector is modeled with a Gaussian mixture model (GMM) as follows:

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

where $\mathcal{N}(\cdot; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$, and m is the mixture component index. 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.

To improve the converted speech quality, a global variance (GV) which is the variance of the static feature vectors over an utterance is considered. The probability density of the GV $\boldsymbol{v}(\boldsymbol{y})$ is modeled with a Gaussian distribution as follows:

$$P(\boldsymbol{v}(\boldsymbol{y})|\boldsymbol{\lambda}^{(v)}) = \mathcal{N}(\boldsymbol{v}(\boldsymbol{y}); \boldsymbol{\mu}^{(v)}, \boldsymbol{\Sigma}^{(v)}). \tag{2}$$

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

2.2. Conversion Process

In the conversion process, the source features are converted into the target features based on maximum likelihood estimation. First, the suboptimum mixture component sequence $\hat{\boldsymbol{m}}$ is determined by

$$\hat{\boldsymbol{m}} = \underset{\boldsymbol{m}}{\operatorname{argmax}}\, P(\boldsymbol{m}|\boldsymbol{X}, \boldsymbol{\lambda}). \tag{3}$$

Then, the converted static feature vector sequence is determined by

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}}\, P(\boldsymbol{Y}|\boldsymbol{X}, \hat{\boldsymbol{m}}, \boldsymbol{\lambda}) P(\boldsymbol{v}(\boldsymbol{y})|\lambda^{(v)})^\omega \quad \text{subject to } \boldsymbol{Y} = \boldsymbol{W}\boldsymbol{y} \tag{4}$$

where $\boldsymbol{W}$ is a window matrix to extend the static feature vectors to the joint static and dynamic feature vectors, and ω is the GV likelihood weight.

3. Proposed Framework for NAM Enhancement Robust against External Noise

3.1. Stereo NAM Enhancement based on Statistical VC using Air- and Body-Conductive Microphones

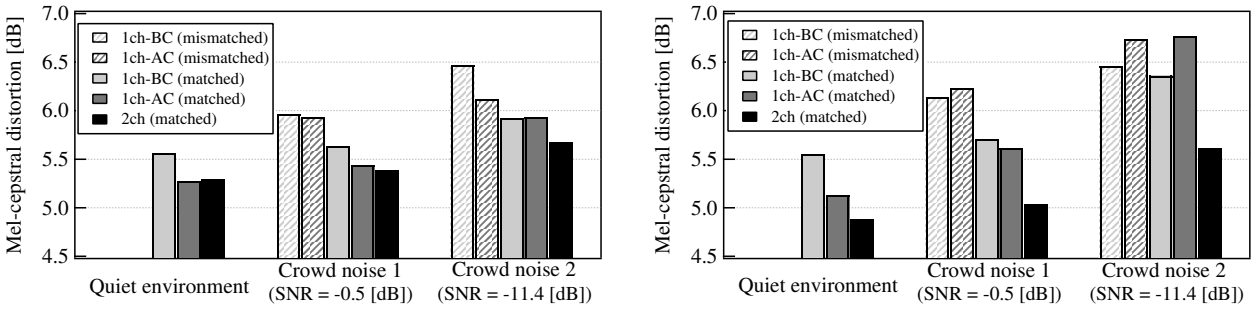
Although NAM is an extremely soft whispered voice, it can still be detected with an air-conductive microphone in a quiet environment by setting it close to speaker's mouth. Because the air-conducted NAM signal is not affected by body-conduction, its spectral feature is more similar to that of the target natural speech compared to that of the body-conducted NAM signal. However, the air-conducted NAM signal is easily deteriorated by external noises. On the other hand, although the spectral feature of the body-conducted NAM signal is quite different from that of the target natural speech, the body-conductive recording is much more robust against external noises than the air-conductive recording because of a noise-proof structure of the NAM microphone as shown in **Figure 1**.

Considering these complementary features of the body-conducted NAM signal and the air-conducted NAM signal, we propose a stereo NAM enhancement method based on the statistical VC techniques using both the NAM microphone and the air-conductive microphone. Let $\boldsymbol{x}_t^{(\mathrm{a})}$ and $\boldsymbol{x}_t^{(\mathrm{b})}$ be a static feature vector of the air- and body-conducted NAM at frame t, respectively. First, the segment feature is separately extracted from each of the air- and body-conducted NAM signals in the same manner as described in **Section 2.1**. Then, the segment feature vectors $\boldsymbol{X}_t^{(\mathrm{a})}$ and $\boldsymbol{X}_t^{(\mathrm{b})}$ are concatenated together into a single feature vector $\boldsymbol{X}_t^{(\mathrm{a,b})} = [\boldsymbol{X}_t^{(\mathrm{a})\top}, \boldsymbol{X}_t^{(\mathrm{b})\top}]^\top$ to be used as the source feature vector for the statistical VC. The remaining training and conversion processes are the same as in the conventional method. Note that the air- and body-conducted NAM signals are synchronized because they are recorded simultaneously.

3.2. Implementation of Noise-Dependent Conversion Models

In noisy environments, external noises come to be mixed in not only the air-conducted NAM signal but also the body-conducted NAM signal although the NAM microphone is robust against external noises. Therefore, spectral features of both the air- and body-conducted NAM signals are collapsed more or less. These acoustic changes easily cause severe deterioration of the converted speech quality in the statistical VC due to the acoustic mismatches between training and testing conditions. To address this issue, it will be helpful to adapt the conversion models to the noisy NAM signals.

In this paper, we clarify adverse effects of the external noises on the performance of the proposed NAM enhancement framework. Moreover, we evaluate the performance of noise-dependent conversion models as an upper bound of the model adaptation techniques to investigate a possibility to alleviate the adverse effects by revising the conversion models. Assuming that the noisy environment in testing is known, only external noise signals are recorded using both the air-conductive microphone and NAM microphone. Then, air- and body-conducted noisy NAM signals are virtually generated by superimposing the recorded noise signals on clean NAM signals in individual channels. The generated noisy NAM signals are used as the input of the NAM enhancement based on statistical VC in both training and testing. Note that the time alignment between the noisy NAM signals and the target natural speech signals is



(a) CVWH (b) CVSP

Figure 2: *Mel-cepstral distortion of enhanced speech. The labels "mismatched" and "matched" indicate whether or not noise-dependent models are used in the statistical VC-based NAM enhancement processing.*

Table 1: *Conversion accuracy of excitation features.*

	U/V error rate [%]			F_0 correlation			Aperiodic distortion [dB]		
	1ch-BC	1ch-AC	2ch	1ch-BC	1ch-AC	2ch	1ch-BC	1ch-AC	2ch
Quiet environment	26.3	19.1	23.1	0.41	0.37	0.39	4.90	4.77	4.79
Crowd noise 1 (SNR = -0.5 [dB])	27.5	26.8	25.7	0.31	0.28	0.26	4.94	4.94	4.87
Crowd noise 2 (SNR = -11.4 [dB])	38.1	52.7	32.9	0.21	0.08	0.28	5.17	5.41	5.10

performed using a time-warping function determined with the clean NAM signals. As a result, noise-dependent GMMs capable of converting the acoustic features of the noisy NAM into those of the target natural speech.

4. Experimental Evaluations

4.1. Experimental Conditions

We simultaneously recorded air- and body-conducted NAM signals with a close-talk microphone and the NAM microphone in a sound-proof room. We also recorded two types of a crowd noise using the same microphone settings by presenting them from a loud speaker in the sound-proof room. One was recorded in a street and the other was recorded in an exhibition hall. Sound pressure levels of individual noises at the close-talk microphone were set to -0.5 dB(A) and -11.4 dB(A), respectively. The noise signals recorded with the close-talk microphone and the NAM microphone were randomly segmented to be superimposed on the air- and body-conducted NAM signals. We also recorded normal speech and a whispered voice by using a usual air-conductive microphone. We recorded 50 sentences of a phoneme balanced sentence set [6] uttered by one Japanese male speaker. We used 40 sentences for training and 10 sentences for evaluation. The sampling frequency was 16 kHz.

The 0^{th} through 24^{th} mel-cepstral coefficients were used as a spectral feature at each frame. FFT analysis, STRAIGHT analysis [7], and mel-cepstral analysis [8] were used for NAM, normal speech and a whispered voice, respectively. We used the 50-dimensional segment feature at each input frame extracted using PCA from current one $\pm$ 4 frames. As excitation features, we used a log-scaled F_0 extracted with STRAIGHT F_0 extractor [9] and aperiodic components [10] on five frequency bands, i.e., 0-1, 1-2, 2-4, 4-6, and 6-8 kHz [11]. The shift length was 5 ms. The number of mixture components was set to 32 for the spectral conversion, 16 for the F_0 conversion and 16 for the aperiodic conversion.

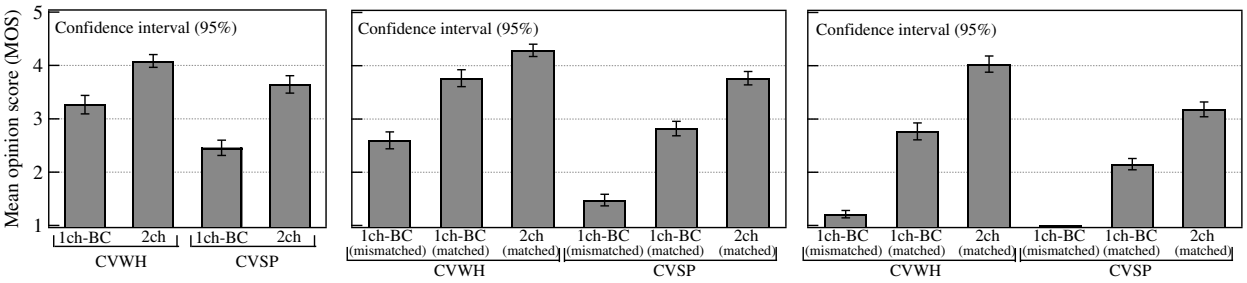
To investigate the effectiveness of the proposed methods, the following 8 kinds of speech samples were evaluated.

- **BC-NAM**:
  Unprocessed body-conducted NAM
- **AC-NAM**:
  Unprocessed air-conducted NAM
- **1ch-BC-CVWH**:
  Whispered voice converted from body-conducted NAM
- **1ch-AC-CVWH**:
  Whispered voice converted from air-conducted NAM
- **1ch-BC-CVSP**:
  Normal speech converted from body-conducted NAM
- **1ch-AC-CVSP**:
  Normal speech converted from air-conducted NAM
- **2ch-CVWH**:
  Whispered voice converted from air- and body-conducted NAM
- **2ch-CVSP**:
  Normal speech converted from air- and body-conducted NAM

4.2. Objective Evaluations

The spectral conversion accuracy was evaluated using mel-cepstral distortion calculated using the 1^{st} through 24^{th} coefficients between the converted speech and the target natural speech. **Figure 2** shows the result. As the noise level increases, the mel-cepstral distortion also increases due to the effect of external noises on the air- and body-conducted NAM signals. This degradation tends to be reduced by using the noise-dependent models but we can still observe significantly large degradation in the second type of a crowd noise (Crowd noise 2). The proposed 2ch framework also using the noise-dependent models well alleviates this degradation. It also achieves almost the best spectral conversion accuracy in any condition.

In a comparison between the conversion from the air-conducted NAM and that from the body-conducted NAM using the noise-dependent models, we can observe that the conversion accuracy from the air-conducted NAM to normal speech



(a) Quiet environment (b) Crowd noise 1 (SNR = -0.5 [dB]) (c) Crowd noise 2 (SNR = -11.4 [dB])

Figure 3: *Listenability of enhanced speech.*

(**1ch-AC-CVSP**) gets worse than that from the body-conducted NAM to normal speech (**1ch-BC-CVSP**) when the SNR is -11.4 dB, but we cannot observe such a tendency in the conversion into a whispered voice. It is expected that the conversion into a whispered voice is more robust against the external noise than that into normal speech because similarity between the source and target spectral features is high as reported in [5].

The conversion accuracy of the excitation features was also evaluated. For the F_0 features, the U/V error rate and correlation coefficient between the converted and target F_0 contours were calculated. For the aperiodic components, the aperiodic distortion between the converted and target aperiodic components was calculated. **Table 1** shows the result. We can observe a similar tendency in the excitation conversion accuracy as observed in the spectral conversion accuracy, i.e., the conversion accuracy tends to be degraded as the external noise level increases and the proposed 2ch framework using the noise-dependent conversion models usually reduces such degradation.

4.3. Subjective Evaluations

The opinion test on listenability was conducted using a 5-point opinion scale, such as 1: very bad, 2: bad, 3: fair, 4: good, 5: excellent. The number of listeners was six. Each listener evaluated 22 samples for each enhancement method, i.e., 132 samples for each of the noisy environment and 88 samples for the quite environment in total.

Figure 3 shows the result. The listenability of enhanced speech is improved by using the noise-dependent model and is further improved by additionally using both the air- and body-conducted NAM signals. It is observed that the proposed method using the noise-dependent model and the 2ch NAM enhancement is effective for improving the listenability of enhanced speech in all environments. In a comparison between **2ch-CVWH** and **2ch-CVSP**, **2ch-CVWH** yields higher listenability than **2ch-CVSP** in all environments. It has been reported in [5] that the NAM-to-whisper conversion yields better enhancement speech than the NAM-to-speech conversion in the quiet environment. This tendency is also observed even in the noisy environments.

To verify whether or not the listenability of NAM is really improved by the proposed methods in each environment, another opinion test on listenability was conducted. In this opinion test, we evaluated the unprocessed air- and body-conducted NAM signals and **2ch-CVWH** as the best enhancement result as shown in the first opinion test.

Figure 4 shows the result. In the quiet environment, **AC-NAM** actually has the highest listenability. However, as the noise level increases, the listenability of **AC-NAM** quickly degrades. On the other hand, although the listenability of **BC-NAM** is lower than that of **AC-NAM** in the quiet environment, its degradation caused by the increase of the external noise level is much smaller than that of **AC-NAM**. Consequently, the listenability of **AC-NAM** gets lower than that of **BC-NAM** when the SNR is -11.4 dB. The proposed method **2ch-CVWH** is capable of keeping the listenability consistently high. In the noisy environments, it can significantly improve the original listenability of air- and body-conducted NAM signals. From these results, the effectiveness of the proposed NAM enhancement method is confirmed.

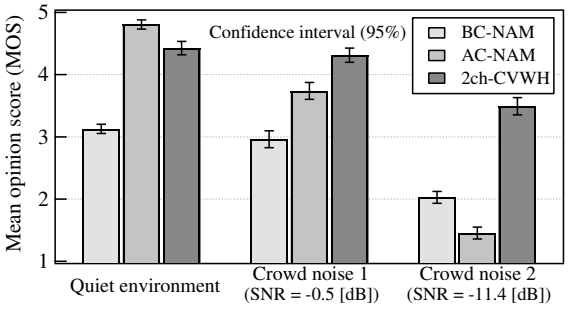


Figure 4: *Listenability of unprocessed NAM and enhanced speech.*

5. Conclusions

This paper has described a new framework for NAM enhancement using not only the NAM microphone but also an air-conductive microphone. In addition to the 2ch NAM enhancement, to investigate a possibility to alleviate the effect of external noises by revising VC models, we have implemented noise-dependent conversion models. Experimental results demonstrate that the proposed framework yields significant improvements in the spectral conversion accuracy and listenability of enhanced speech under both quiet and noisy environments. In this paper, we don't consider Lombard effect on NAM [12]. We plan to investigate the effectiveness of the proposed NAM enhancement method in a real noisy environment.

6. Acknowledgements

Part of this work was supported by JSPS KAKENHI Grant Numbers: 26280060.

7. References

[1] B. Denby, T. Schultz, K. Honda, T. Hueber, J. M. Gilbert, and J. S. Brumberg, "Silent speech interfaces," *Speech Communication*, vol. 52, no. 4, pp. 270–287, 2010.

[2] Y. Nakajima, H. Kashioka, N. Cambell, and K. Shikano, "Non-Audible Murmur (NAM) recognition," *IEICE Transactions on 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 Transactions on 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 Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[5] T. Toda, M. Nakagiri, and K. Shikano, "Statistical voice conversion techniques for body-conducted unvoiced speech enhancement," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 9, pp. 2505–2517, 2012.

[6] Y. Sagisaka, K. Takeda, M. Ave, S. Katagiri, T. Umeda, and H. Kuwabara, "A large-scale Japanese speech database," *First International Conference on Spoken Language Processing*, pp. 1089–1092, 1990.

[7] H. Kawahara, I. Masuda-Katsuse, and A. de Cheveigné, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, no. 3, pp. 187–207, 1999.

[8] K. Tokuda, T. Kobayashi, and S. Imai, "Adaptive cepstral analysis of speech," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 6, pp. 481–489, 1995.

[9] H. Kawahara, H. Katayose, A. de Cheveigné, and R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of F0 and periodicity," *Proc. EUROSPEECH*, vol. 99, no. 6, pp. 2781–2784, 1999.

[10] H. Kawahara, J. Estill, and O. Fujimaru, "Aperiodicity extraction and control using mixed mode excitation and group delay manipulation for a high quality speech analysis, modification and synthesis system STRAIGHT," *Proc. MAVEABA*, pp. 59–64, 2001.

[11] 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, 2006.

[12] T. Toda, K. Nakamura, T. Nagai, T. Kaino, Y. Nakajima, and K. Shikano, "Technologies for processing body-conducted speech detected with Non-Audible Murmur microphone," *Proc. INTERSPEECH*, pp. 632–635, 2009.