9th ISCA Speech Synthesis Workshop
13-15 Sep 2016, Sunnyvale, USA

Nonaudible murmur enhancement based on statistical voice conversion and noise suppression with external noise monitoring

Yusuke Tajiri[1], *Tomoki Toda*[1]

[1]Graduate School of Information Science, Nagoya University, Japan
tajiri.yusuke@g.sp.m.is.nagoya-u.ac.jp, tomoki@is.nagoya-u.ac.jp

Abstract

This paper presents a method for making nonaudible murmur (NAM) enhancement based on statistical voice conversion (VC) robust against external noise. NAM, which is an extremely soft whispered voice, is a promising medium for silent speech communication thanks to its faint volume. Although such a soft voice can still be detected with a special body-conductive microphone, its quality significantly degrades compared to that of air-conductive voices. It has been shown that the statistical VC technique is capable of significantly improving quality of NAM by converting it into the air-conductive voices. However, this technique is not helpful under noisy conditions because a detected NAM signal easily suffers from external noise, and acoustic mismatches are caused between such a noisy NAM signal and a previously trained conversion model. To address this issue, in this paper we apply our proposed noise suppression method based on external noise monitoring to the statistical NAM enhancement. Moreover, a known noise superimposition method is further applied in order to alleviate the effects of residual noise components on the conversion accuracy. The experimental results demonstrate that the proposed method yields significant improvements in the conversion accuracy compared to the conventional method.

Index Terms: silent speech communication, nonaudible murmur, statistical voice conversion, noise suppression, external noise monitoring, normalization of noise conditions

1. Introduction

The recent advancement of information technologies, such as mobile phones, enables us to talk with others while not sharing the same environments. This newly developed speech communication style has openly reminded us that there exist some situations where we hesitate to talk with others; *e.g.*, we have difficulty in talking about private information in a crowd; or speaking itself would sometimes annoy others in quiet environments. To address this issue, *silent speech interfaces* [1] have recently attracted attention as a technology to make it possible for us to talk with each other without the necessity of emitting an audible acoustic signal. To detect silent speech, several sensing devices have been explored as alternatives to a usual air-conductive microphone, such as body-conductive microphones [2, 3], electromyography [4], ultrasound imaging [5], and others.

As one of the body-conductive microphones capable of detecting silent speech, we focus on nonaudible murmur (NAM) microphone [3]. This microphone was originally developed to detect an extremely soft whispered voice called NAM, which is so quiet that people around the speaker barely hear its emitted sound. Such a soft voice is detected through only the soft tissues of the head using the NAM microphone attached to the neck below the ear. The NAM microphone is also more robust against external noise than standard air-conductive microphones thanks to its noise-proof structure. However, severe quality degradation is always caused by acoustic changes resulting from body-conduction [6].

To improve the speech quality of NAM, statistical voice conversion (VC) techniques [7, 8] have been successfully applied to NAM enhancement [9]. In this approach, acoustic features of NAM are converted into those of air-conducted natural speech, such as a normal voice or a whispered voice, making it possible to significantly improve the voice quality and intelligibility of NAM. However, there still remain some issues to be addressed in order to make it possible to practically use NAM for silent speech communication. Although the NAM microphone is robust against external noises, it cannot completely block external noise signals. In particular, when detecting NAM, its body conducted speech signal significantly suffers from external noise owing to its faint volume. Such a noisy NAM signal causes significantly large acoustic mismatches between training and conversion conditions in the statistical VC, making the enhancement process completely fail [10]. Model adaptation techniques will be helpful to alleviate these acoustic mismatches, but it is not straightforward to accurately adapt the conversion model to arbitrary noisy conditions. Therefore, it is worthwhile to develop a front-end noise suppression technique robust against any external noisy condition for reducing the external noise components in the noisy NAM signals as much as possible.

Some enhancement methods for body-conducted speech additionally using the air-conducted noisy speech signal detected with a usual air-conductive microphone have been proposed, *e.g.*, the direct filtering method [11] and the statistical enhancement method [12], although these methods actually deal with another problem, *i.e.*, speech enhancement under heavy noisy conditions. Inspired by these methods, we have proposed a noise suppression method based on external noise monitoring using the air-conductive microphone [13]. Unlike the bone-conducted speech enhancement methods, the proposed method uses the air-conductive microphone to detect only the external noise signal, leveraging a property of NAM (i.e., its faint volume). The detected external noise signal is effectively used as a reference signal to suppress the corresponding noise components in the noisy NAM signal. It has been reported that this method is capable of significantly improving the quality of NAM signals under various types of noisy conditions.

In this paper, we propose a statistical NAM enhancement method robust against external noise additionally using the noise suppression method based on external noise monitoring as the front-end noise suppression processing. Because it is

http://dx.doi.org/10.21437/SSW.2016-9

still difficult to perfectly suppress the external noise components in the noisy NAM signal, some noise components usually remain after the noise suppression. To alleviate their adverse effects on the statistical NAM enhancement, we further apply a known noise superimposition method, which is a simple and effective way to reduce relatively small acoustic mismatches by normalizing arbitrary noisy conditions [14]. Our experimental results demonstrate that the proposed method yields significant improvements in conversion accuracy.

2. Conventional NAM enhancement method based on statistical voice conversion

There have been proposed two main frameworks for converting NAM into air-conducted natural speech with the statistical VC technique [9], 1) conversion into a normal voice (NAM2SP) and 2) conversion into a whispered voice (NAM2WH), as shown in **Figure 1**. In NAM2SP, it is necessary to estimate not only spectral features but also excitation features, such as F_0 pattern and aperiodicity. On the other hand, in NAM2WH, it is necessary to estimate only spectral features because the whispered voice is totally unvoiced speech like NAM. It has been reported that 1) NAM2WH basically outperforms NAM2SP in terms of naturalness and intelligibility because the conversion process in NAM2WH is much easier than that in NAM2SP, effectively reducing possible quality degradation caused by conversion errors [9], but 2) voiced speech tends to be more intelligible than unvoiced speech under noisy conditions (*i.e.*, assuming that external noise exists in a listener's side), and therefore, NAM2SP outperforms NAM2WH in terms of intelligibility in such a condition [15]. Thus it is worthwhile to study both of these two frameworks.

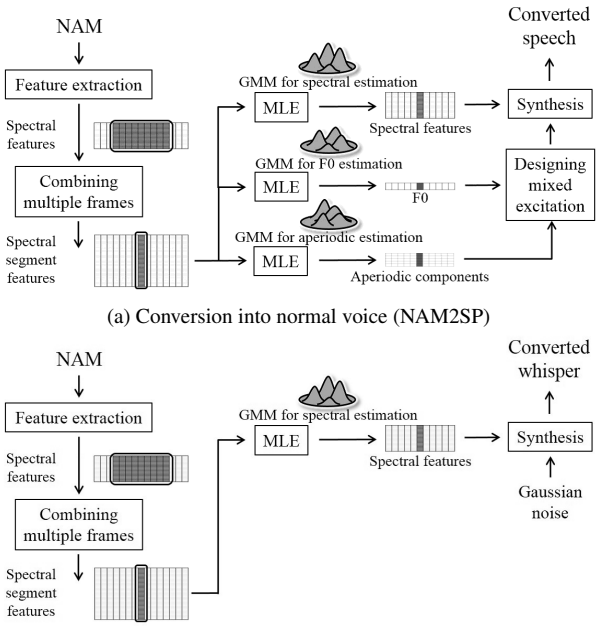
In these statistical NAM enhancement framework, a conversion model is trained in advance using a parallel dataset consisting of utterance pairs of NAM and the target air-conducted natural speech. The trained conversion model is used to convert arbitrary utterances in NAM. More details are described below.

2.1. Training process

Let us assume a source static feature vector $\boldsymbol{x}_\tau$ (*e.g.*, a spectral parameter of NAM) and a target static feature vector $\boldsymbol{y}_\tau$ (*e.g.*, a spectral parameter, an aperiodic parameter, or F_0 parameter of the target air-conducted natural speech) at frame τ. As the source feature vector, a segment feature $\boldsymbol{X}_\tau = \boldsymbol{A}[\boldsymbol{x}_{\tau-L}^\top, \cdots, \boldsymbol{x}_\tau^\top, \cdots, \boldsymbol{x}_{\tau+L}^\top]^\top + \boldsymbol{b}$ is calculated from current one $\pm L$ frames, where $\boldsymbol{A}$ and $\boldsymbol{b}$ are determined from the training data (*e.g.*, using principal component analysis (PCA)). As the target feature vector, a joint static and dynamic feature vector $\boldsymbol{Y}_\tau = [\boldsymbol{y}_\tau^\top, \Delta\boldsymbol{y}_\tau^\top]^\top$ is extracted. Using the time-aligned source and target feature vectors $\left\{[\boldsymbol{X}_1^\top, \boldsymbol{Y}_1^\top]^\top, \cdots, [\boldsymbol{X}_N^\top, \boldsymbol{Y}_N^\top]^\top\right\}$ developed with the training data, the joint probability density of the source and target feature vectors is modeled with a Gaussian mixture model (GMM) as follows:

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

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 shown as $\boldsymbol{\lambda}$, which consists of the mixture-dependent weights w_m, mean vectors $\boldsymbol{\mu}_m^{(X,Y)}$, and full covariance matrices $\boldsymbol{\Sigma}_m^{(X,Y)}$ for individual mixture components.



(a) Conversion into normal voice (NAM2SP)

(b) Conversion into whispered voice (NAM2WH)

Figure 1: *Conversion process in statistical body-conducted soft speech enhancement.*

Let us also assume the global variance (GV) vector $\boldsymbol{v}(\boldsymbol{y})$, which is calculated as the variance values at individual dimensions over the target static feature vector sequence $\boldsymbol{y} = [\boldsymbol{y}_1^\top, \cdots, \boldsymbol{y}_T^\top]^\top$ [8]. Its probability density 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 conversion process, the trajectory-wise conversion method based on maximum likelihood estimation considering the GV [8] is used to determine the target static feature vector sequence $\boldsymbol{y}$ from the given source feature vector sequence $\{\boldsymbol{X}_1, \cdots, \boldsymbol{X}_T\}$. First, the suboptimum mixture component sequence $\hat{\boldsymbol{m}}$ is determined as follows:

$$\hat{\boldsymbol{m}} = \{\hat{m}_1, \cdots, \hat{m}_T\} = \underset{\boldsymbol{m}}{\operatorname{argmax}} \prod_{\tau=1}^{T} P(m_\tau|\boldsymbol{X}_\tau, \boldsymbol{\lambda}). \tag{3}$$

Then, the converted static feature vector sequence is determined as follows:

$$\hat{\boldsymbol{y}} = \underset{\boldsymbol{y}}{\operatorname{argmax}} \prod_{\tau=1}^{T} P(\boldsymbol{Y}_\tau|\boldsymbol{X}_\tau, \hat{\boldsymbol{m}}_\tau, \boldsymbol{\lambda}) P(\boldsymbol{v}(\boldsymbol{y})|\lambda^{(v)})^\omega \tag{4}$$

$$\text{subject to } \left[\boldsymbol{Y}_1^\top, \cdots, \boldsymbol{Y}_T^\top\right]^\top = \boldsymbol{W}\boldsymbol{y} \tag{5}$$

where $\boldsymbol{W}$ is a linear transform to extend the static feature vector sequence to the joint static and dynamic feature vector sequence [16], and ω is the GV likelihood weight.

3. Noise suppression method using external noise monitoring

3.1. External noise monitoring with air-conductive microphone

NAM is practically difficult to be detected with a usual air-conductive microphone because it is easily masked by external noise due to its faint volume. Therefore, by setting an air-conductive microphone away from the speaker's mouth, only the external noise signals can be detected. **Figure 2** shows an example of the air-conductive microphone and its setting position. Although the NAM signal is actually leaked into the air-conductive microphone from mouth, the signals detected with the air-conductive microphone placed as shown in **Figure 2** can be well approximated with only the external noise signals if the sound pressure level of the external noise is higher than 60 dBA as reported in [13]. It is also expected that this setting position of the air-conductive microphone close to the NAM microphone is helpful to detect the external noise signals corresponding to noise signals detected with the NAM microphone. Consequently, the mixing process of the observed body- and air-conducted signals in noisy environments is assumed as follows:

$$x_1(t) = s_1(t) + \sum_{u=0}^{U} a_t(u) s_2(t-u) \quad (6)$$

$$x_2(t) \approx s_2(t) \quad (7)$$

where $s_1(t)$ is a clean body-conducted NAM signal, $s_2(t)$ is an air-conducted external noise signal, and $\{a_t(0), \cdots, a_t(U)\}$ is an acoustic transfer function to transfer the air-conducted external noise signal into the body-conducted external noise signal.

3.2. Noise suppression based on semi-blind source separation

In the above mixing process, an estimation problem of a clean body-conducted NAM signal $s_1(t)$ is equivalent to the classical acoustic echo cancellation (AEC) problem [17]; *i.e.*, the observed air-conducted signal $x_2(t)$ and the acoustic transfer function $\{a_t(0), \cdots, a_t(U)\}$ correspond to a reference signal and an echo path, respectively. Semi-blind source separation (semi-BSS) can be effectively applied to this problem. Because the semi-BSS is an unsupervised estimation technology, it is not necessary to detect NAM activity sections. Therefore, it can avoid double-talk, which is a well-known problem in AEC.

Let us assume frequency components of the source signals $\boldsymbol{s}(\omega,\tau) = [s_1(\omega,\tau), s_2(\omega,\tau)]^\top$ and those of the observed signals $\boldsymbol{x}(\omega,\tau) = [x_1(\omega,\tau), x_2(\omega,\tau)]^\top$, where ω and τ show a frequency bin index and a time frame index, respectively. By further assuming that the acoustic transfer function is time-invariant, the mixing process given by Eqs. (6) and (7) is modeled as instantaneous mixture in the frequency domain as follows:

$$\boldsymbol{x}(\omega,\tau) = \boldsymbol{A}(\omega)\boldsymbol{s}(\omega,\tau) \quad (8)$$

where $\boldsymbol{A}(\omega)$ is a (2×2) time-invariant mixing matrix. In a standard BSS problem, a (2×2) un-mixing matrix $\boldsymbol{W}(\omega)$ is estimated with independent component analysis. On the other hand, in the noise monitoring problem, one of the two source signals (*i.e.*, $s_2(\omega,\tau)$) is known, and some elements of the un-mixing matrix $\boldsymbol{W}$ can be fixed as follows:

$$\boldsymbol{W}(\omega) = \begin{bmatrix} 1 & w_{12}(\omega) \\ 0 & 1 \end{bmatrix}. \quad (9)$$

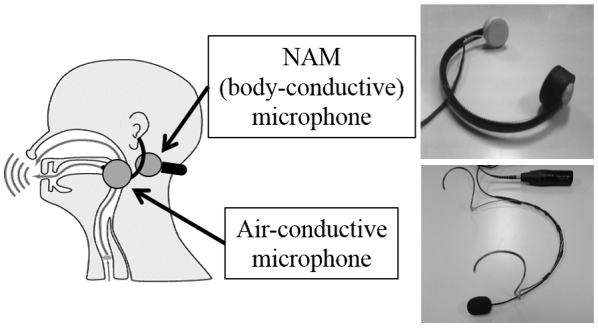


Figure 2: *Air- and body-conductive microphones and their setting positions.*

Therefore, only the component $w_{12}(\omega)$ needs to be estimated by maximizing independence between a separated NAM signal and the observed air-conducted signal. It is iteratively updated using natural gradient [18] as follows:

$$\Delta w_{12}(\omega) = \eta\{w_{12}(\omega) - \langle\varphi(y_1(\omega,\tau))\boldsymbol{y}^{\mathrm{H}}(\omega,\tau)\rangle_\tau [w_{12}(\omega), 1]^\top\} \quad (10)$$

$$w_{12}(\omega) \leftarrow w_{12}(\omega) + \Delta w_{12} \quad (11)$$

where $\boldsymbol{y}(\omega,\tau) = [y_1(\omega,\tau), s_2(\omega,\tau)]^\top$ is the separated signals, η is a step-size parameter, $\langle\cdot\rangle_\tau$ is a time average operator, and $\varphi(y_1(\omega,\tau))$ is a nonlinear function like polar function given by

$$\varphi(y_1(\omega,\tau)) = \tanh(|y_1(\omega,\tau)|)\exp(j\arg(y_1(\omega,\tau))). \quad (12)$$

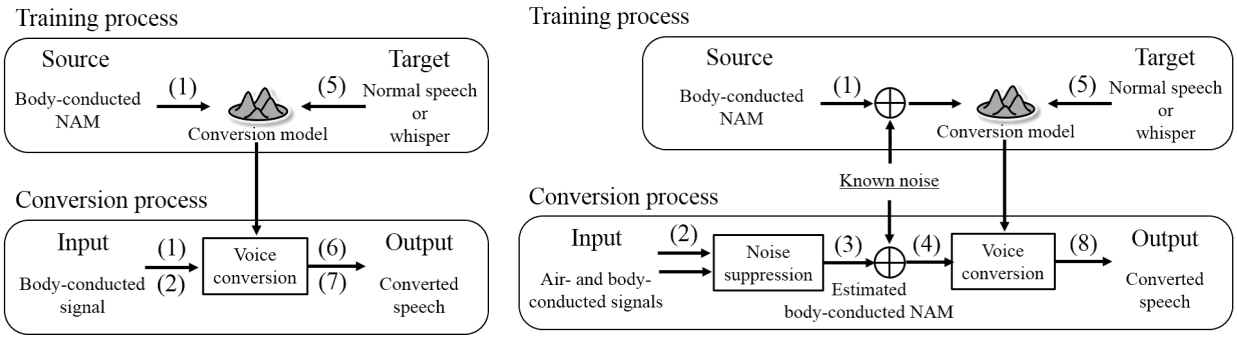
4. Proposed NAM enhancement method robust against external noise

To develop the NAM enhancement method robust against any noisy condition, we apply the noise suppression method based on external noise monitoring and the known noise superimposition method to the statistical NAM enhancement method as front-end processing. The framework of the proposed method is shown in **Figure 3**, which also shows that of the conventional method for comparison. Moreover, an example of spectrograms of individual signals observed or generated during the enhancement process is shown in **Figures 4** and **5** to demonstrate the effectiveness of each process.

4.1. Front-end process to normalize noisy conditions

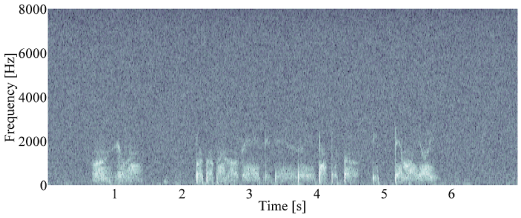
Acoustic characteristics of the noisy NAM signal are very different from those of the clean NAM signal because the NAM signal easily suffers from external noise under noisy conditions as shown in **Figure 4**, where the clean and noisy NAM signals are shown in **(a)** and **(b)**, respectively. To reduce these noise components from the noisy NAM signal, the semi-BSS-based noise suppression method using external noise monitoring is first applied to the noisy NAM signal. As shown in **Figure 4 (c)**, this method is capable of significantly reducing arbitrary time-variant noise components. However, remaining noise components are still observed in the processed noisy NAM signal.

To alleviate the adverse effect of these remaining noise components on the conversion accuracy in the statistical NAM enhancement, the known noise superimposition method is further applied to the processed NAM signal after the noise suppression processing. In this method, a pre-determined specific

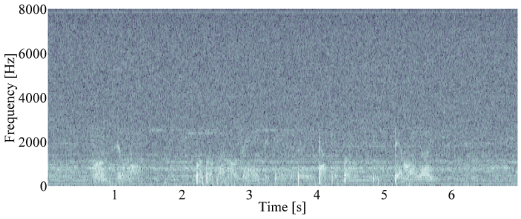


(a) Conventional method *(b) Proposed method*

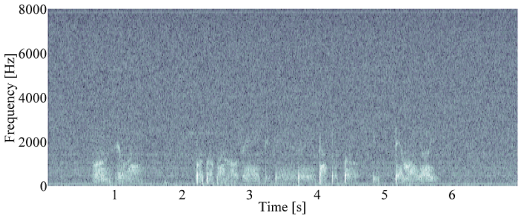
Figure 3: *Conventional and proposed NAM enhancement frameworks based on statistical VC. The proposed framework additionally uses noise suppression based on external noise monitoring and known noise superimposition as front-end processing.*



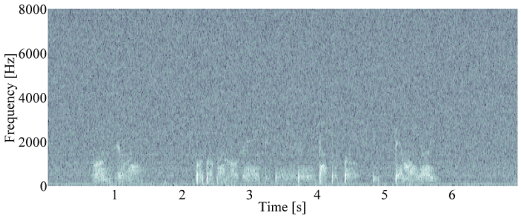
(a) *Clean NAM signal corresponding to (1) in* **Figure 3**



(b) *Noisy NAM signal (detected under 70 dBA booth noise condition) corresponding to (2) in* **Figure 3**

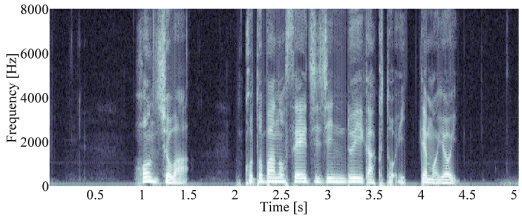


(c) *Processed noisy NAM signal after semi-BSS-based noise suppression (before known noise superimposition) corresponding to (3) in* **Figure 3**

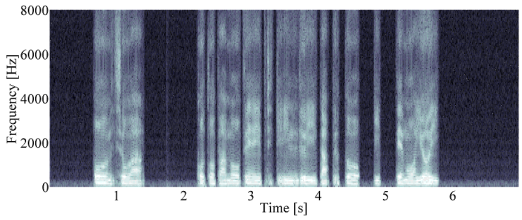


(d) *Processed noisy NAM signal after known noise superimposition (using 10 dB SNR of white noise) corresponding to (4) in* **Figure 3**

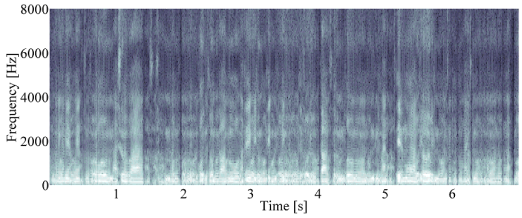
Figure 4: *Example of spectrograms of NAM signals.*



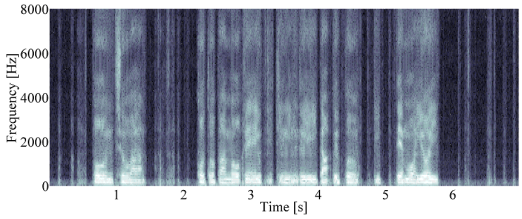
(a) *Target normal voice corresponding to (5) in* **Figure 3**



(b) *Normal voice converted from clean NAM signal corresponding to (6) in* **Figure 3**



(c) *Normal voice converted from noisy NAM signal (detected under 70 dBA booth noise condition) corresponding to (7) in* **Figure 3**



(d) *Normal voice converted from processed noisy NAM signal after front-end (semi-BSS and known noise superimposition) corresponding to (8) in* **Figure 3**

Figure 5: *Example of spectrograms of target and converted signals. Note that duration of the target signal is different from that of the converted ones.*

noise signal (*e.g.*, white noise in this paper) is superimposed on the processed NAM signal. The remaining noise components are masked by the superimposed noise components if power of the remaining noise components is smaller than that of the superimposed ones. Consequently, the noisy NAM signal detected under arbitrary noise conditions is well normalized to the noisy NAM signal detected under known noise conditions through the front-end processing. An example of the noisy NAM signal after the front-end processing is shown in **Figure 4 (d)**.

4.2. Conversion process under normalized noisy conditions

The noisy NAM signal after the front-end processing is converted to the target voice using the statistical NAM enhancement method. Note that the conversion model needs to be trained using not the clean NAM signals but the noisy NAM signals generated by adding the known noise signals to the clean NAM signals. The resulting conversion model is effectively used without any model adaptation processes under any noisy conditions in the proposed framework.

4.3. Effectiveness

Figure 3 shows an example of spectrograms of **(a)** the target normal voice, **(b)** the converted voice from the clean NAM signal, **(c)** the converted voice from the noisy NAM signal in the conventional method, and **(d)** the converted voice from the noisy NAM signal in the proposed method. Under the clean condition, the converted voice **(b)** is similar to the target voice **(a)**. However, under the noisy condition, the converted voice in the conventional method **(c)** has very different acoustic characteristics from those of the target voice **(a)** because of the acoustic mismatches between the clean NAM signal (shown in **Figure 4 (a)**) and the noisy NAM signal (shown in **Figure 4 (b)**). On the other hand, the proposed method is capable of significantly reducing the adverse effects of external noise and making the converted voice more close to the target voice than the conventional method although some acoustic differences are still observed in particular at silence frames.

5. Experimental evaluation

5.1. Experimental conditions

We simultaneously recorded clean NAM signals uttered by one Japanese male speaker simultaneously using the NAM microphone and the air-conductive microphone in a sound-proof room. We also recorded the following seven kinds of noise signals using the same microphone settings by presenting them from a loud speaker in the sound-proof room:

- **Babble50dB**: 50 dBA babble noise
- **Babble60dB**: 60 dBA babble noise
- **Babble70dB**: 70 dBA babble noise
- **Office50dB**: 50 dBA office noise
- **Crowd60dB**: 60 dBA crowd noise
- **Booth70dB**: 70 dBA booth noise
- **Station80dB**: 80 dBA station noise

The sound pressure levels of the individual noises were measured by a sound level meter placed at around the speaker's head. Babble noise indicating human speech-like noise were generated by superimposing 20 different speakers' speech signals. The recorded air- and body-conducted noise signals were superimposed on the clean air- and body-conducted NAM signals to simulate noisy NAM signals. Fifty sentences in the phoneme balanced sentence set were uttered in NAM. They were also uttered in a normal voice and a whispered voice by the same speaker. Forty utterances were used for the training of the conversion models in the statistical NAM enhancement. The remaining ten utterances were used for the test. The sampling frequency was set to 16 kHz.

In the statistical NAM enhancement, the 0^{th} through 24^{th} mel-cepstral coefficients were used as the spectral feature at each frame. FFT analysis, STRAIGHT analysis [19], and mel-cepstral analysis [20] were used for NAM, normal voices, and whispered voices, respectively. We used the 50-dimensional segment feature at each input frame extracted using PCA from the current $\pm$ 4 frames. As the excitation features, we used log-scaled F_0 value extracted with the STRAIGHT F_0 extractor [21] and aperiodic components [22] averaged on five frequency bands, 0-1, 1-2, 2-4, 4-6, and 6-8 kHz [23]. 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.

In the semi-BSS for the noise suppression, the window length of STFT was set to 64 ms and the shift length was set to 32 ms. The step-size parameter η was set to 0.01. The number of iterations was set to 200.

We examined the effectiveness of the known noise superimposition in the proposed method by controlling power of the superimposed white noise signals so as to set the resulting signal-to-noise ratio (SNR) of the NAM signal to 15 dB, 10 dB, and 5 dB. The SNR was set to the same value in training and conversion. We evaluated the final conversion accuracy in each setting and also that when not performing the known noise superimposition. Moreover, we also evaluated the performance of the following methods:

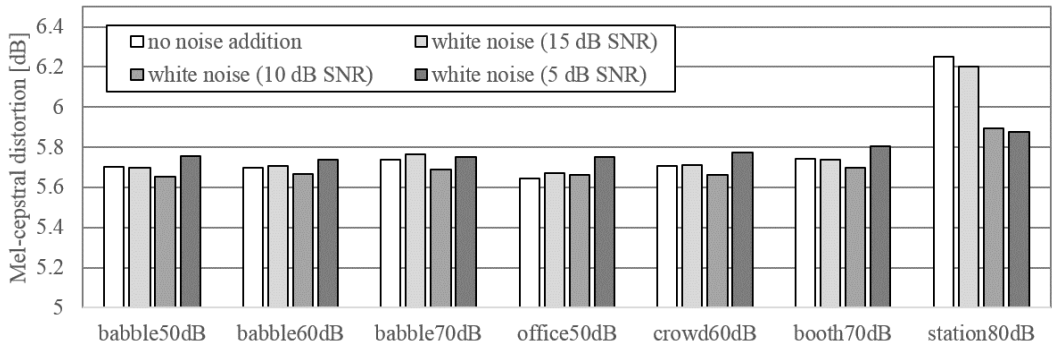
- **unprocessed**: the conventional method without any processing to deal with external noise
- **matched model**: the conventional method using the matched conversion model trained with the noisy NAM detected under the same noisy conditions as in the test
- **BSS w/ noise addition**: the proposed method

As an evaluation metric, the mel-cepstral distortion between the converted voice and the target voice was used. Both NAM2SP and NAM2WH were evaluated.

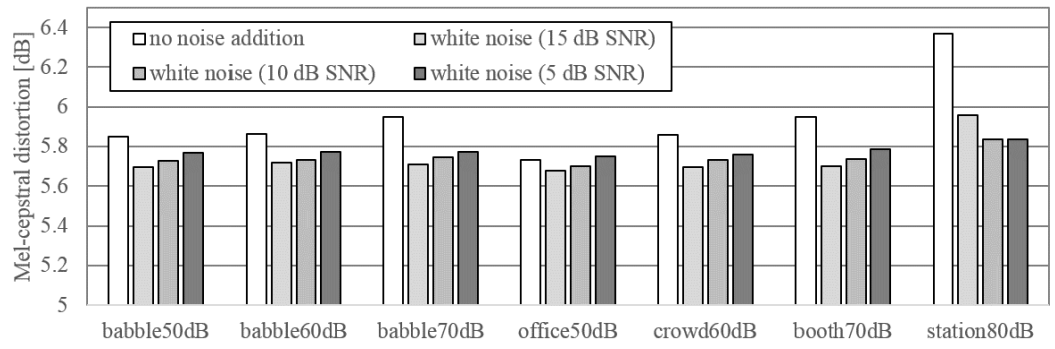
5.2. Experimental results

Figure 6 shows a result of the examination on the effectiveness of the known noise superimposition. In the station 80 dBA noisy condition, it is observed that the known noise superimposition yields significant performance improvements. It is interesting that its effectiveness is clearly observed in NAM2WH rather than in NAM2SP. In total, by setting the SNR to 10 dB, the known noise superimposition yields significantly better conversion accuracy or just keeps the conversion accuracy almost the same as that in no superimposition

Figure 7 shows a result of the comparison among the different methods. We can observe that **unprocessed** causes significant degradation in the conversion accuracy due to the adverse effect of the remaining noise components. The use of the matched model alleviates this adverse effect. However, its effectiveness tends to be smaller as the external noise level is higher. On the other hand, the proposed method yields good conversion accuracy over any noise conditions. Note that the proposed method can handle any noise conditions unlike the matched model.

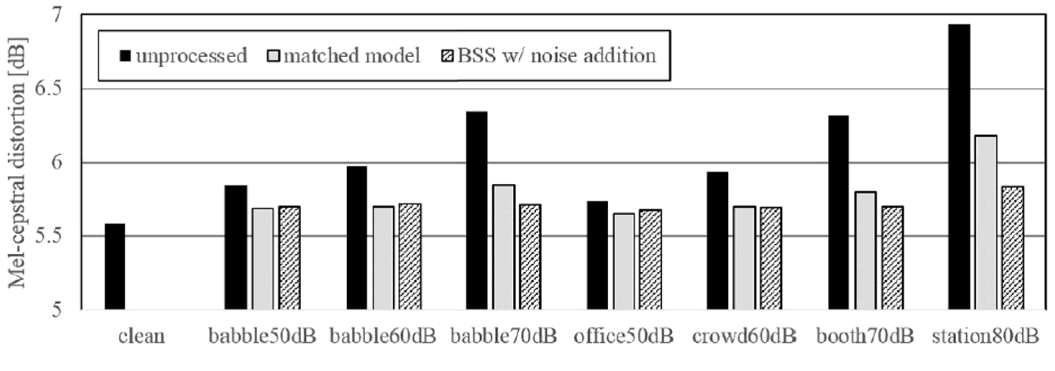


(a) *Result in NAM2SP*

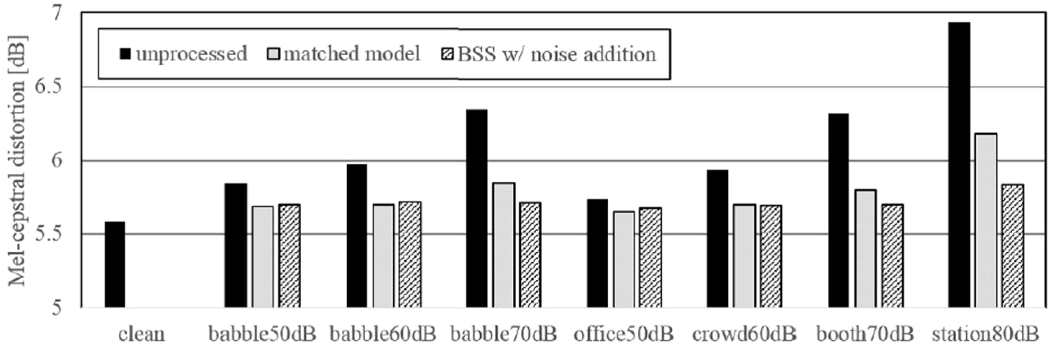


(b) *Result in NAM2WH*

Figure 6: *Comparison of different SNR settings in known noise superimposition.*



(a) *Result in NAM2SP*



(b) *Result in NAM2WH*

Figure 7: *Comparison of different enhancement methods.*

These results have demonstrated that the proposed method is very effective for improving robustness against external noise in the statistical NAM enhancement.

6. Conclusion

In this paper, we have proposed a method for improving noise robustness of the nonaudible murmur (NAM) enhancement processing based on statistical voice conversion. To make it possible to handle arbitrary noise conditions, the external noise suppression method based on external noise monitoring and the known noise superimposition method have been successfully implemented as the front-end processing for the statistical NAM enhancement processing. The experimental results have demonstrated that the proposed methods are capable of significantly improving the conversion accuracy under noisy conditions. We plan to conduct subjective evaluations to further examine the effectiveness of the proposed method.

Acknowledgements: This work was supported in part of JSPS KAKENHI Grant Numbers: 15K12064, 26280060, and 16J08977.

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] 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, Sep. 2004.

[3] 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.

[4] T. Schultz and M. Wand. Modeling coarticulation in EMG-based continuous speech recognition. *Speech Communication*, Vol. 52, No. 4, pp. 341–353, 2010.

[5] T. Hueber, E.-L. Benaroya, G. Chollet, B. Denby, G. Dreyfus, and M. Stone. Development of a silent speech interface driven by ultrasound and optical images of the tongue and lips. *Speech Communication*, Vol. 52, No. 4, pp. 288–300, 2010.

[6] T. Hirahara, M. Otani, S. Shimizu, T. Toda, K. Nakamura, Y. Nakajima, and K. Shikano. Silent-speech enhancement using body-conducted vocal-tract resonance signals. *Speech Communication*, Vol. 52, No. 4, pp. 301–313, 2010.

[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. Audio, Speech and Language Processing*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[9] T. Toda, M. Nakagiri, and K. Shikano. Statistical voice conversion techniques for body-conducted unvoiced speech enhancement. *IEEE Trans. Audio, Speech and Language Processing*, Vol. 20, No. 9, pp. 2505–2517, 2012.

[10] Y. Tajiri, K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. Non-audible murmur enhancement based on statistical conversion using air- and body-conductive microphones in noisy environments. *Proc. INTERSPEECH*, pp. 2769–2773, Dresden, Germany, Sep. 2015.

[11] Z. Liu, Z. Zhang, A. Acero, J. Droppo, and H. Huang. Direct filtering for air- and body-conductive microphones. *Proc. MMSP* pp. 363–366, 2004.

[12] 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.

[13] Y. Tajiri, T. Toda, and S. Nakamura. Noise suppression method for body-conducted soft speech enhancement based on external noise monitoring. *Proc. ICASSP*, pp. 5935–5939, Shanghai, China, Mar. 2016.

[14] S. Yamade, A. Baba, S. Yoshikawa, A. Lee, H. Saruwatari, and K. Shikano. Unsupervised speaker adaptation for robust speech recognition in real environments. *Electronics and Communications in Japan (Part II)* Vol. 88, No. 8, pp. 30–41, 2005.

[15] S. Tsuruta, K. Tanaka, T. Toda, G. Neubig, S. Sakti, and S. Nakamura. An evaluation of target speech for a nonaudible murmur enhancement system in noisy environments. *Proc. APSIPA ASC*, 4 pages, Siem Reap, Cambodia, Dec. 2014.

[16] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura. Speech parameter generation algorithms for HMM-based speech synthesis. *Proc. ICASSP*, pp. 1315–1318, Istanbul, Turkey, June 2000.

[17] A. Gilloire and M Vetterli. Adaptive filtering in subbands with critical sampling: analysis, experiments, and application to acoustic echo cancellation. *IEEE Tran. Signal Processing*, Vol. 2, No. 8, pp. 148–151, 1995.

[18] S. Amari. Natural gradient works efficiently in learning. *Neural Computation*, Vol. 10, No. 2, pp. 251–276, 1998.

[19] 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.

[20] K. Tokuda, T. Kobayashi, T. Masuko, and S. Imai. Mel-generalized cepstral analysis – a unified approach to speech spectral estimation. *Proc. ICSLP*, pp. 1043–1045, Yokohama, Japan, Sep. 1994.

[21] 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.

[22] 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.

[23] 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.