Blind Speech Extraction for Non-Audible Murmur Speech with Speaker's Movement Noise

† Miyuki Itoi, † Ryoichi Miyazaki, † Tomoki Toda, † Hiroshi Saruwatari, † Kiyohiro Shikano
† Nara Institute of Science and Technology, Nara, Japan (e-mail: miyuki-i@is.naist.jp)

***Abstract*—In this paper, we address an improved method of noise reduction used in multichannel Non-Audible Murmur (NAM) based on blind source separation. Recently, speech processing with NAM has been proposed for applying versatile speech interface into quiet environments where we hesitate to utter. NAM is a very soft whispered voice signal detected with the NAM microphone, which is one of the body-conductive microphone. The detected NAM signal always suffers from nonstationary noise caused by speaker's movement because it changes the setting condition of the NAM microphone. In order to reduce the noise signal, blind noise reduction using stereo NAM signals detected with two NAM microphones has been proposed by some of the authors. In this paper, we aim to achieve further improvement in the noise reduction ability by changing the noise estimation and postprocessing algorithms to enhance the target NAM signal. In addition, we evaluate the application of recording the NAM signals with various types of microphones.**

***Index Terms*—Non-Audible Murmur, blind spatial subtraction array, nonstationary noise**

I. Introduction

An explosive spread of portable devices with a lot of functions makes us realize importance of the development of natural interfaces to use them. A speech interface is one of the typical natural interfaces and speech recognition is a key technology to develop it. Although speech is a convenient medium, there are actually some situations where we face difficulties in using speech. For example, we would have trouble privately talking in a crowd; speaking itself would sometimes annoy others in quiet environments such as in a library. The development of technologies to overcome these inherent problems of speech is essential.

Recently, *silent speech interfaces* [1] have attracted attention as a technology to make speech interfaces more convenient. They enable speech input to take place without the necessity of emitting an audible acoustic signal. As one of the sensing devices to detect *silent speech* signals, Nakajima *et al.* [2] developed a Non-Audible Murmur (NAM) microphone. NAM is an extremely soft whispered voice, which is so quiet that people around the speaker hardly hear its emitted sound. Placed on the neck below the ear, the NAM microphone is capable of detecting extremely soft speech such as NAM from the skin through only the soft tissues of the head. There have been several attempts to develop a NAM recognition system by modeling acoustic characteristics of NAM [3], [4], [5], [6], which are very different from those of normal speech. In the past studies on NAM recognition, the speakers tried maintaining their positions as stably as possible during speaking to keep a setting condition of the NAM microphone as constantly as possible. However, this constraint should not be enforced in a real situation; the speaker often moves in speaking. Since the detected signal with NAM microphone is sensitive to the setting condition of the NAM microphone such as the pressure to attach the NAM microphone, noise is easily generated when the speaker moves. For example, when the speaker moves his/her head to look away, noticeable noise is generated if the attachment plane of the NAM microphone is rubbed by the skin. The NAM signal easily suffers from the generated noise and NAM recognition performance is significantly degraded. Since the generated noise is non-stationary and its frequency components widely overlap those of the NAM signal, it is not straightforward to suppress it. In order to resolve this problem, a blind noise suppression method using stereo signal processing has been proposed [7].

In this paper, we propose to apply blind spatial subtraction array (BSSA) to six-channel signals recorded simultaneously by a throat microphone and an adheresive NAM microphone, in addition to NAM microphone. In this part, we apply sparse signal extraction (SSE) to the noise estimation part, which is based on the sparseness between speech and diffuse noise, and we compare this method with the conventional method [7] in noise supression performance. Also, we introduce generalized spectral subtraction (GSS) and quasi-parametric Wiener filter (QPWF), comparing these two methods in noise supression performance.

II. Related works

A. NAM

NAM is defined as the articulated production of respiratory sounds without using the vocal-fold vibration, which can be conducted through only the soft tissues of the head without any obstruction such as bones [2]. NAM is recorded using the NAM microphone attached to the skin surface behind the ear, as shown in **Figure 1**. In this study, a neckband-type of NAM microphone [3], in which the necked presses the microphone against the skin, is used to stably attach it. Since NAM is a particularly soft whispered voice, the recorded sound by the NAM microphone is amplified with a special amplifier. **Figure 2** shows an example of the spectrogram of NAM. High-frequency components of NAM are usually not well observed owing to the mechanisms of body conduction, such as lack of radiation characteristics from lips and effect of low-pass characteristics of the soft tissues.

978-1-4673-5604-6/12/$31.00 ©2012 IEEE

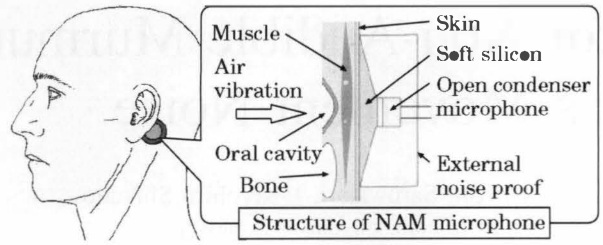

Fig. 1. Setting position and structure of NAM microphone.

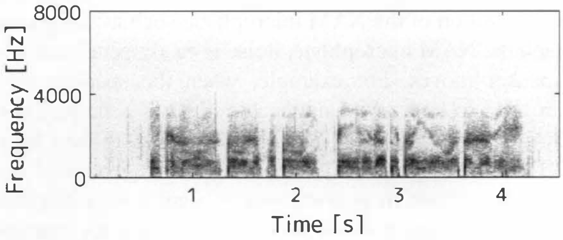

Fig. 2. Example of spectrogram of clean NAM signal.

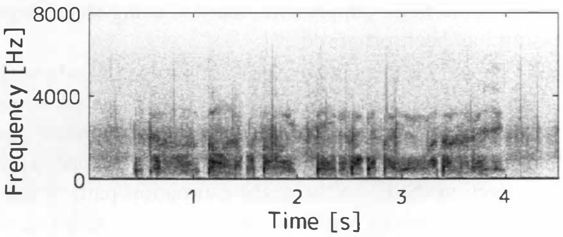

Fig. 3. Example of spectrogram of NAM signal when speaker moves during speaking.

B. *Effect of Speaker's Movements on NAM Signal*

In the past studies on NAM recognition ([2], [3], [8]), the speakers tried maintaining their positions as stably as possible during speaking to keep a setting condition of the NAM microphone as constantly as possible. However, this constraint should not be enforced in a real situation; the speaker often moves freely in speaking. The skin surface and muscles around the place of the NAM microphone attached usually move in conjunction with the speaker's movements, such as the movements of his/her head. These movements often change the acoustic condition of the NAM microphone. **Figure 3** shows an example of spectrogram of NAM when the speaker lightly shakes his head. We can confirm that the recorded NAM signal is severely deteriorated by noise caused by the speaker's movements. The generated noise is non-stationary and causes substantially large acoustic fluctuation compared with the NAM signal shown in **Figure 2**. This noise causes significant degradation in NAM recognition [7].

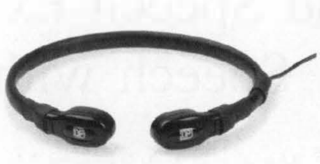

Fig. 4. Throat microphone.

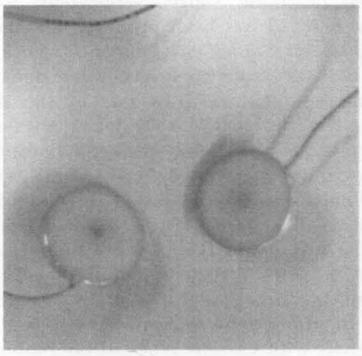

Fig. 5. Adhesive NAM microphone.

III. VARIATION OF MICROPHONES

In the past studies, NAM was recorded only by the NAM microphone, which is specialized for recording NAM. However, in order to make practical, it is worthwhile to test recording NAM by not only the conventional NAM microphone but also various kinds of other microphones. In this paper, we use the throat microphone and the adhesive NAM microphone, in addition to the conventional NAM microphone (hereafter "conventional NAM microphone" or simple "NAM microphone" is referred to as the neckband type microphone shown in Sect. II-A). The throat microphone used in our experiments consists of piezoelectric ceramics, shown in **Figure 4**. It is attached on talker's neck close to the vocal folds and receives uttered speech through the skin. It is a commercially available product for recording normal speech, not for NAM, and consequently it is necessary to investigate whether we can improve the recorded sound quality by using the microphone or not. The adhesive NAM microphone shown in **Figure 5** can receive uttered speech conducted through the soft tissues of body, similar to the conventional NAM microphone. Its surface is covered with adhesive material, fixing by sticking to body. By using this microphone, we can attach to the skin surface not only behind the ear but also to everywhere in body, and this makes it possible to record NAM without restricting attached places.

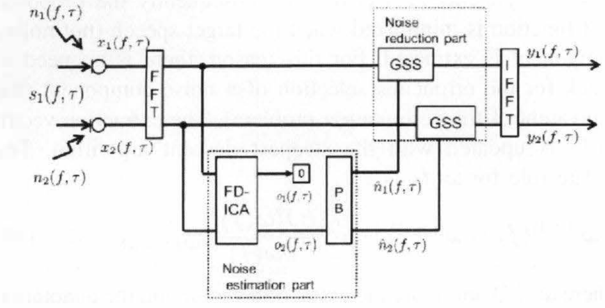

Fig. 6. Block diagram of blind noise reduction of conventional method for NAM.

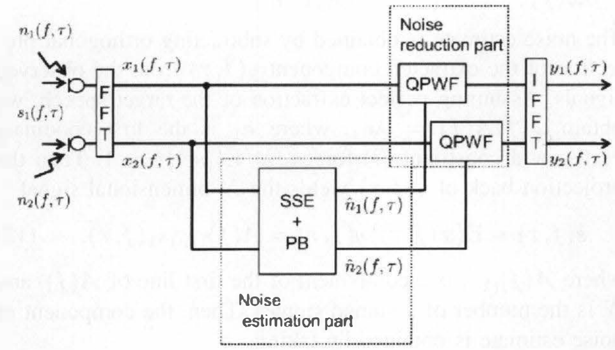

Fig. 7. Block diagram of blind noise reduction of proposed method for NAM.

IV. Blind Noise Suppression with Stereo NAM Signals

A. Overview

The conventional method [7] and proposed method use the NAM signals recorded via stereo channels of the NAM microphone, these of the throat microphone, and these of the adhesive microphone. Such stereo signals allow us to use various effective noise suppression techniques such as beamforming. First, we represent the sound mixing model in the stereo signals detected with those microphones. Then, the difference of blind noise suppression method between the conventional and proposed methods are described. These methods are based on BSSA, which consists of a noise estimation part and a noise suppression part. The block diagram of the conventional method is shown in **Figure 6**. We apply ICA based on infomax to the noise estimation part, and GSS to the noise suppression part. The block diagram of the proposed method is shown in **Figure 7**, where we apply SSE to the noise estimation part, and QPWF to the noise suppression part.

B. Mixing Model of NAM and Noise

The detected stereo NAM signals with speaker's movements, $\boldsymbol{x}(f,\tau) = [x_1(f,\tau), x_2(f,\tau)]^\top$ consisting of the first channel signal $x_1(f,\tau)$ and the second channel signal $x_2(f,\tau)$, are modeled by

$$\boldsymbol{x}(f,\tau) \quad \simeq \quad \boldsymbol{a}(f)s_1(f,\tau) + \boldsymbol{n}(f,\tau), \qquad (1)$$

where $\top$ denotes transposition of the vector, f is the frequency bin, and τ is the time index of DFT analysis. A component of the NAM signal before the body conduction is given by $s_1(f,\tau)$, which is unobserved. The signal $s_1(f,\tau)$ is linearly filtered with channel-dependent and time-invariant transfer functions $\boldsymbol{a}(f) = [a_1(f), a_2(f)]^\top$, which are affected by various factors such as a setting position of the NAM microphone, a setting of the amplifier, and so on. The detected stereo noise signals are modeled by $\boldsymbol{n}(f,\tau) = [n_1(f,\tau), n_2(f,\tau)]^\top$ as diffuse noise signals. Note that to simplify the mixing process we also assume that the speaker's movements do not change the transfer function $\boldsymbol{a}(f)$.

C. Noise Estimation Based on Infomax

In this subsection, the conventional noise estimation method, which uses frequency domain ICA (FD-ICA) [9] based on higher-order statistics, is described. The detected stereo mixed-signals are separated with the complex valued demixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$ so that the output signals $\boldsymbol{o}(f,\tau) = [o_1(f,\tau), o_2(f,\tau)]^\top$ become mutually independent. The output signals are given by

$$\boldsymbol{o}(f,\tau) \quad = \quad \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau), \qquad (2)$$

where the demixing matrix $\boldsymbol{W}_{\mathrm{ICA}}(f)$ is determined by minimizing Kullback-Leibler divergence between the joint probability density function $p(\boldsymbol{o}(f,\tau))$ and the marginal probability density function $p(o_1(f,\tau))p(o_2(f,\tau))$ over a time sequence. The optimal $\boldsymbol{W}_{\mathrm{ICA}}(f)$ is obtained using the following iterative equation:

$$\begin{aligned}\boldsymbol{W}_{\mathrm{ICA}}^{[i+1]} &= \boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f) \\ &+\alpha\left[\boldsymbol{I} - \langle \boldsymbol{\Phi}(\boldsymbol{o}(f,\tau))\boldsymbol{o}^H(f,\tau)\rangle_\tau\right]\boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f), \quad (3)\end{aligned}$$

where α is the step-size parameter, $[i]$ indicates the value of the i-th step in iterations, $\boldsymbol{I}$ is the identity matrix, $\langle\cdot\rangle_\tau$ denotes the time-averaging operator, H denotes Hermitian transposition, and $\boldsymbol{\Phi}(\cdot)$ is the nonlinear vector function [10]. In this paper, we determine the demixing matrix utterance by utterance.

In the mixed signal modeled by (1), the separation process given by (2) is not obviously capable of suppressing the noise signal $\boldsymbol{n}(f,\tau)$. On the other hand, it is capable of suppressing a component of the NAM signal $s_1(f,\tau)$. In other words, ICA is proficient in well estimating a component related to the noise signal [11]. Therefore, only the noise component is useful in the output signals. To remove the NAM signal from the output signals, the following "noise-only" signal vector $\boldsymbol{o}^{(n)}(f,\tau)$ is constructed:

$$\boldsymbol{o}^{(n)}(f,\tau) = [\mathbf{0}, o_2(f,\tau)]^\top. \qquad (4)$$

To solve permutation problem, an initial matrix of $\boldsymbol{W}_{\mathrm{ICA}}$ is designed so that $o_2(f,\tau)$ becomes the noise component [10]. Following this, the projection back (PB) process [12] [13] is performed to remove the ambiguity of amplitude and estimate the non-stationary stereo noise signal $\hat{\boldsymbol{n}}(f,\tau) = [\hat{n}_1(f,\tau), \hat{n}_2(f,\tau)]^\top$ as follows:

$$\hat{\boldsymbol{n}}(f,\tau) \quad = \quad \boldsymbol{W}_{\mathrm{ICA}}^{+}(f)\boldsymbol{o}^{(n)}(f,\tau), \qquad (5)$$

where $\boldsymbol{M}^+$ denotes the Moore-Penrose pseudo inverse matrix of $\boldsymbol{M}$. It is obvious that this noise estimation is not perfect. But it is still useful to enhance the NAM signal with nonlinear noise reduction process using the noise spectral amplitude, as described later in Sect. IV-E.

D. Noise Estimation Based on SSE

In this subsection, the proposed noise estimation method is described. In noise estimation based on infomax described in Sect. IV-C, it is necessary to solve permutation problem. Thus we should discriminate which of $o_1(f,\tau)$ and $o_2(f,\tau)$ is speech or noise, but this is very difficult to solve because the conventional solution [14] is mainly based on direction-of-arrival (DOA) information, which is ambiguous in NAM. In addition, there is a problem that the nonlinear vector function such as $\tanh(x)$ is inappropriate for approximation of probability density of noise. To avoid these problems, SSE that exploits the sparsity of the modulus of the target speech signal is beneficial. In this method, it is not necessary to solve the permutation problem because we introduce statistical difference between speech and diffuse background noise. In addition, since there is no local minimum of the cost function, the convergence is stable. This results in higher quality of noise estimation than that of based on infomax.

In the fth frequency bin, we estimate $y(f,\tau)$ by applying extracting vector $\boldsymbol{\omega}(f)$ to the observed signals

$$y(f,\tau) = \boldsymbol{\omega}(f)\boldsymbol{x}(f,\tau) = \boldsymbol{\omega}(f)\boldsymbol{A}(f)\boldsymbol{s}(f,\tau) \qquad (6)$$

with constraint

$$\mathrm{E}\{|y(f,\tau)|^2\} = 1, \qquad (7)$$

where $\boldsymbol{A}(f)$ is a matrix whose entries represent the transfer functions, and $\boldsymbol{s}(f,\tau)$ is components of uttered signals. The first component of $\boldsymbol{s}(f,\tau)$, $s_1(f,\tau)$ is the target speech component. The vector $\boldsymbol{\omega}(f)$ is updated so as to minimize the cost function

$$J(\boldsymbol{\omega}(f)) = (\mathrm{E}\{|y(f,\tau)|\} - \gamma)^2, \qquad (8)$$

where $\gamma \geq 0$ is a parameter for controlling the sparsity of the extracted component. The constraint (7) can be written as (dropping frame and frequency indexes) $\mathrm{var}\{|y|\} + \mathrm{E}\{|y|\}^2 = 1$. Thus minimizing the cost function aims at extracting the component such that

$$\mathrm{E}\{|y|\} = \gamma \text{ and } \mathrm{var}\{|y|\} = 1 - \gamma^2. \qquad (9)$$

For a small γ, the extracted component has a modulus with a small mean $\mathrm{E}\{|y|\}$ and a large variance $\mathrm{var}\{|y|\}$ with respect to constraint imposed by (7). Namely, the modulus of the extracted component is sparse in the sense that most of the values are close to zero and only a few are significantly large. In the case of target speech in diffuse background noise, the speech modulus is sparser than that of the diffuse background noise components [15] [16], and consequently the proposed cost function is minimized when the target speech (not noise) component is extracted. For this reason, there is no need to check for the erroneous selection of a noise component (the equivarent of the permutation problem). The extraction vector $\boldsymbol{\omega}(f)$ is updated with the steepest descent algorithm. The update rule for $\boldsymbol{\omega}(f)$ is

$$\boldsymbol{\omega}^{[k+1]}(f) = \boldsymbol{\omega}^{[k]}(f) - \mu^{[k]}\frac{\partial J(\boldsymbol{\omega}(f))}{\partial \boldsymbol{\omega}(f)}|_{\boldsymbol{\omega}(f)=\boldsymbol{\omega}^{[k]}(f)}, \qquad (10)$$

where $\boldsymbol{\omega}_k(f)$ and μ are the extraction vector and the adaptation step at the kth iteration. The gradient of the cost function is given by

$$\frac{\partial J(\boldsymbol{\omega}(f))}{\partial \boldsymbol{\omega}(f)} = 2\mathrm{E}\left\{\boldsymbol{x}(f,t)\frac{y(f,\tau)^{\mathrm{H}}}{|y(f,\tau)|}\right\}(\mathrm{E}\{|y(f,\tau)|\} - \gamma). \qquad (11)$$

The noise estimate is obtained by subtracting orthogonal projection of the extracted component $y(f,\tau)$ from the observed signals. Assuming perfect extraction of the target speech, we obtain $\boldsymbol{\omega}(f)\boldsymbol{A}(f) = \lambda \boldsymbol{e}_1$, where $\boldsymbol{e}_1$ is the first coodinate vector, and constraint (7) forces $|\lambda|^2\mathrm{E}\{|s_1|\}^2 = 1$. Then the projection back of $y(f,\tau)$ yields the N-dimensional signal

$$\hat{\boldsymbol{s}}(f,\tau) = \mathrm{E}\{\boldsymbol{x}(f,\tau)\}y(f,\tau) = \boldsymbol{A}(f)_{(1,:)}s_1(f,\tau), \qquad (12)$$

where $\boldsymbol{A}(f)_{(1,:)}$ is a component of the first line of $\boldsymbol{A}(f)$, and N is the number of assumed signals. Then, the component of noise estimate is obtained by taking

$$\hat{\boldsymbol{n}}(f,\tau) = \boldsymbol{x}(f,\tau) - \hat{\boldsymbol{s}}(f,\tau) = \sum_{j=2}^{N}\boldsymbol{A}(f)_{(j,:)}s_j(f,\tau). \qquad (13)$$

E. Noise Suppression Part

In the noise suppression part, GSS or QPWF is applied to each channel of the mixed signal. The GSS-applied NAM signal $\hat{\boldsymbol{s}}^{[\mathrm{GSS}]}(f,\tau) = [\hat{s}_1^{[\mathrm{GSS}]}(f,\tau), \hat{s}_2^{[\mathrm{GSS}]}(f,\tau)]^\top$ is obtained by

$$\hat{s}_c^{[\mathrm{GSS}]}(f,\tau) = \begin{cases} \sqrt[2\xi]{|x_c(f,\tau)|^{2\xi} - \beta|\hat{n}_c(f,\tau)|^{2\xi}}e^{j\arg(x_c(f,\tau))} \\ \qquad (\text{if } |x_c(f,\tau)|^{2\xi} > \beta|\hat{n}_c(f,\tau)|^{2\xi}), \\ \eta \cdot x_c(f,\tau) \qquad (\text{otherwise}), \end{cases} \qquad (14)$$

where c is the channel index, β is the processing strength parameter, η is the flooring parameter, and ξ is the exponent parameter. Also, the QPWF-applied NAM signal $\hat{\boldsymbol{s}}^{[\mathrm{QPWF}]}(f,\tau)$ is obtaind by

$$\hat{s}_c^{[\mathrm{QPWF}]} \quad = \quad \sqrt[2\xi]{\frac{|x_c(f,\tau)|^{2\xi}}{|x_c(f,\tau)|^{2\xi} + \beta|\hat{n}_c(f,\tau)|^{2\xi}}} \cdot|x_c(f,\tau)|e^{j\arg(x_c(f,\tau))}. \qquad (15)$$

V. Experimental Evaluations

A. Experimental Conditions

Target signals are six-channel NAM signals uttered by a Japanese female speaker. The target speech utterance was selected from Japanese Newspaper Corpus [17], where the length of the utterance is about 10 s. These NAM data were

TABLE I
EXPERIMENTAL CONDITIONS

ICA method	Infomax, SSE
NRR [dB]	3, 6, 9
Value of exponent	1.0, 0.7, 0.4, 0.1
Objective evaluation mesaure	Cepstral distortion (CD)

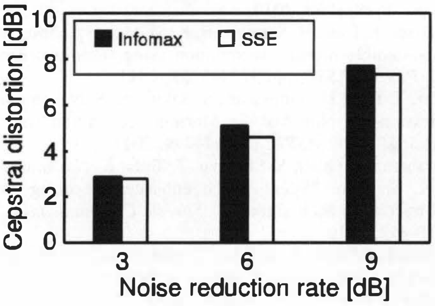

Fig. 8. Results of cepstral distortion for infomax and SSE methods with equivalent NRR.

recorded with two-channel NAM microphones, two-channel throat microphones, and two-channel adhesive NAM microphones simultaneously. The throat microphone is attached on speaker's neck close to the vocal cords, the NAM microphone is attached on the neck below the ear, and the adhesive NAM microphone is attached on the speaker's clavicle. The sampling frequency was set to 16 kHz. We used *simulated mixed-signals* generated by superimposing the non-stationary noise signals, recorded when the speaker moved without speaking in NAM, on the NAM signals recorded when the speaker did not move, with 0-dB SNR. In addition, we apply BSSA to the signals from the same kinds of microphone pairs. Then, we adjusted the processing strength parameter of GSS and QPWF so that noise reduction rate (NRR) [10] of each speech-enhanced output is identical. The NRR is defined as

$$\mathrm{NRR} = 10\log_{10}\frac{\mathrm{E}[s_{\mathrm{out}}^2]/\mathrm{E}[n_{\mathrm{out}}^2]}{\mathrm{E}[s_{\mathrm{in}}^2]/\mathrm{E}[n_{\mathrm{in}}^2]}, \quad (16)$$

where s_{in} and s_{out} are the input and output speech signals, respectively, and n_{in} and n_{out} are the input and output noise signals, respectively. Initial adaptation step of SSE was set to 0.001. The fast Fouriere transform (FFT) size was 1024, and the frame shift length was 256. The rest of the experimental conditions is listed in Table I.

B. Comparison of ICA Method

We compare noise estimation based on infomax and SSE. We apply these methods to the noise estimation part of BSSA, and calculate CD when NRR is 3 dB, 6 dB, and 9 dB. The result of the experiment is shown in **Figure 8**. In the small NRR case, we cannot see the difference between two methods. However, in large NRR cases, infomax's CD becomes more larger than that of SSE. Thus, SSE is superior to infomax. This result is well consistent with the theoretical behavior of SSE as described in Sect. IV-D; thus SSE can automatically solve the permutation even when the noise DOA is ambiguous.

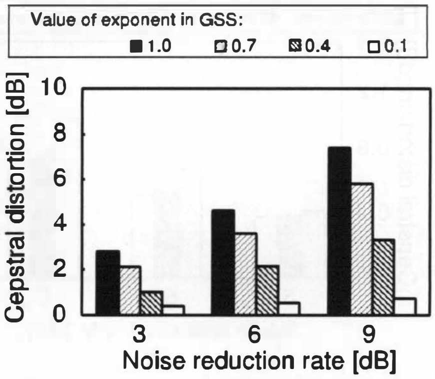

Fig. 9. Results of cepstral distortion with equivalent NRR when changing exponent parameter of GSS.

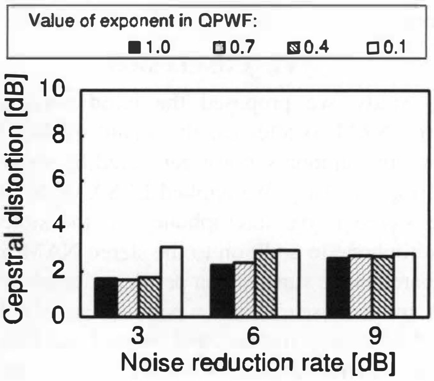

Fig. 10. Results of cepstral distortion with equivalent NRR when changing exponent parameter of QPWF.

C. Comparison of Postprocessing Method

In order to compare QPWF with GSS in postprocessing, the results of CD are shown in **Figure 9** and **Figure 10** with the internal parameter set like **Table I**. When we apply GSS to the noise suppression part, the smaller exponent parameter yields smaller CD. Thus, the small exponent parameter in GSS gives better performance for noise reduction. On the other hand, when we apply QPWF to the noise suppression part, we cannot confirm an apparent tendency in terms of the exponent parameter. In conclusion, GSS with ξ of 0.1 results in the best noise reduction performance.

D. Comparison between Microphone

Figure 11 shows the result of applying blind noise suppression to the NAM signal from various microphones. We use SSE in the noise estimation part and GSS in the noise reduction part with the exponent parameter of 0.1. From the result, the adhesive NAM microphone is better to use for improving noise reduction performance. Thus, we confirm the promising possibility that we can perform high quality noise suppression to NAM signal recorded not only by the

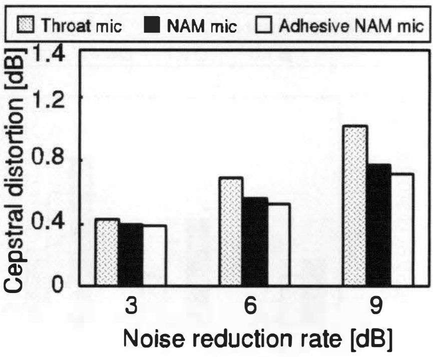

Fig. 11. Results of cepstral distortion for various microphones with equivalent NRR.

conventional NAM microphone, but also by the adhesive NAM microphone.

VI. Conclusion

In this study, we proposed the blind noise suppression method for NAM to alleviate the sound quality degradation caused by non-stationary noise generated by speaker's movements during speaking. We applied BSSA to the NAM signal from the stereo throat microphone and the stereo adhesive NAM microphone in addition to the stereo NAM microphone, and compared noise suppression performance of various methods. First, in the noise estimation part, we compared SSE to the conventional infomax method, and showed that SSE is superior to the conventional method. Secondly, in the noise reduction part, we applied QPWF and GSS to BSSA. When we used GSS, the setting if exponent parameter is an important issue in noise suppression and the small exponent parameter gives high quality noise suppression. On the other hand, when we used QPWF, the exponent parameter is not so sensitive to the performance. Finally, we compared noise suppression quality of the conventional NAM microphone, the throat microphone, and the adhesive NAM microphone. In conclusion, the adhesive NAM microphone performed best quality in noise suppression. Thus, it was shown that the investigation on how to record the NAM signals by various microphones and searching the best attached place are profitable.

Acknowledgment

This work was supported by the MIC SCOPE, and JST Core Research of Evolution Science and Technology (CREST), Japan.

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 Trans. Information and Systems*, vol.E89-D, no.1, pp.1–8, 2006.

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

[4] P. Heracleous, T. Kaino, H. Saruwatari, and K. Shikano, "Unvoiced Speech Recognition Using Tissue-conductive Acoustic Sensor," *EURASIP Journal on Advances in Signal Processing*, vol.2007, Article ID 94068, 11 pages, 2007.

[5] P. Heracleous, V.-A. Tran, T. Nagai, and K. Shikano, "Analysis and recognition of NAM speech using HMM distances and visual information," *IEEE Trans. Audio, Speech, and Language Processing*, vol.18, no.6, pp. 1528–1538, 2010.

[6] D. Babani, T. Toda, H. Saruwatari, K. Shikano, "Acoustic model training for non-audible murmur recognition using transformed normal speech data," *Proc. ICASSP*, pp. 5224–5227, 2011.

[7] S. Ishii, T. Toda, H. Saruwatari, S. Sakti, and S. Nakamura, "Blind noise suppression for Non-Audible Murmur recognition with stereo signal processing," *Proc. ASRU*, pp.494--499, 2011.

[8] 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.

[9] S. Araki, R. Mukaii, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech", *IEEE Transactions on Speech and Audio Processing*, vol.11, no.2, pp.109-116, 2003.

[10] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano, "Blind source separation combining independent component analysisand beamfoming," *EURASIP Journal on Applied Signal Processing*, vol.2003, no.11, pp.1135–1146, 2003.

[11] Y. Takahashi, T. Takatani, K. Osako, H. Saruwatari, and K. Shikano, "Blind spatial subtraction array for speech enhancement in noisy environment," *IEEE Trans. on Audio, Speech and Language Processing*, vol.17, no.4, pp.650–664, 2009.

[12] S. Ikeda and N. Murata, "A method of ICA in time-frequency domain," *Proc. ICA*, pp.365–360, 1999.

[13] Y. Mori, H. Saruwatari, T. Takatani, S. Ukai, K. Shikano, T. Hiekata, Y. Ikeda, H. Hashimoto, and T. Morita, "Blind Separation of Acoustic Signals Combining SIMO-Model-Based Independent Component Analysis and Binary Masking," *EURASIP Journal on Applied Signal Processing*, vol. 2006, Article ID 34970, 17 pages, 2006.

[14] H. Saruwatari, T. Kawamura, T. Nishikawa, A. Lee, and K. Shikano, "Blind source separation based on a fast-convergence algorithm combining ICA and beamforming," *IEEE Transactions on Speech and Audio Processing*, vol.14, no.2, pp.666-678, 2006.

[15] R. Prasad, H. Saruwatari, and K. Shikano, "Probability Distribution of Time-Series of Speech Spectral Components," *IEICE Trans. Fundamentals*, vol.E87-A, no.3, pp.584-597, 2004.

[16] R. Prasad, H. Saruwatari, and K. Shikano, "Estimation of shape parameter of GGD function by negentropy matching," *Neural Processing Letters*, vol.22, pp.377-389, 2005.

[17] A. Lee, T. Kawahara, and K. Shikano, "Julius -An open source realtime large vocabulary recognition engine, " *Proc. European Conference on Speech Communication Technology 2001*, pp.1691–1694, 2001.