Blind Noise Suppression for Non-Audible Murmur Recognition with Stereo Signal Processing

Shunta Ishii, Tomoki Toda, Hiroshi Saruwatari, Sakriani Sakti, Satoshi Nakamura

Graduate School of Information Science, Nara Institute of Science and Technology,
8916-5 Takayama-cho, Ikoma, Nara 630-0192, Japan
{ shunta-i, tomoki, sawatari, ssakti, s-nakamura }@is.naist.jp

***Abstract*—In this paper, we propose a blind noise suppression method for Non-Audible Murmur (NAM) recognition. NAM is a very soft whispered voice detected with NAM microphone, which is one of the body-conductive microphones. Due to its recording mechanism, the detected signal suffers from noise caused by speaker's movements. In the proposed method using a stereo signal detected with two NAM microphones, the noise is estimated with blind source separation, and then, spectral subtraction is performed in each channel to reduce the noise. Moreover, channel selection is performed frame by frame to generate less distorted monaural NAM signal. Experimental results show that 1) word accuracy in large vocabulary continuous NAM recognition is degraded from 69.2% to 53.6% by the noise and 2) it is significantly recovered to 63.3% in a simulated situation and 58.6% in a real situation with the proposed method.**

I. Introduction

The 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; and 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. New sensing devices as alternatives to the air-conductive microphone have been explored to detect *silent speech* signals, such as the throat microphone [2], electromyography (EMG) [3], ultrasound imaging [4], and so on. These sensing devices are also effective as noise robust speech interfaces; *e.g.*, Subramanya *et al.* [5] have reported that bone-conducted speech signals can be effectively used to enhance speech sounds under noisy conditions.

As one of the sensing devices to detect *silent speech* signals, Nakajima *et al.* [6] 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. It is capable of high-quality body-conductive recording and its usability is better than those of other devices such as EMG or ultrasound systems. There have been several attempts to develop a NAM recognition system by modeling acoustic characteristics of NAM [7], [8], [9], which are very different from those of normal speech.

In the conventional studies of 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 will 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 this paper, we propose a blind noise suppression method using stereo signal processing. A stereo signal is detected with two NAM microphones attached below the both ears. Blind Spatial Subtraction Array (BSSA) [10] is used to estimate the non-stationary noise signal by cancelling a target signal (*i.e.*, the NAM signal) with the beam forming process. In BSSA, unsupervised optimization of the beam former is performed with independent component analysis (ICA). And then, spectral subtraction (SS) [11] is performed with the estimated noise signal to reduce the noise in each channel. Moreover, to generate less distorted monaural NAM signal to be used in NAM recognition, frame-wise channel selection based on a time-varying signal-to-noise ratio (SNR) calculated with the estimated noise signal is implemented. We conduct experimental evaluations in large vocabulary continuous NAM recognition to show that the proposed method significantly improves NAM recognition performance by suppressing the noise caused by the speaker's movements.

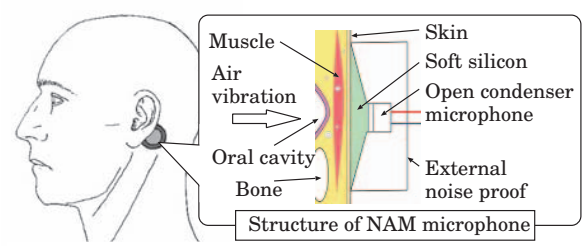

Fig. 1. Setting position and structure of NAM microphone.

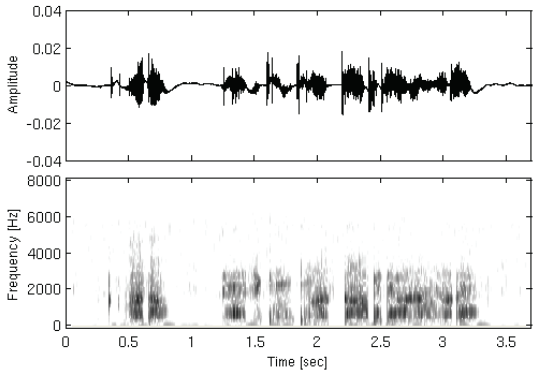

Fig. 2. Example of waveform and spectrogram of NAM signal.

II. Non-Audible Murmur Recognition

A. Non-Audible Murmur (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 [6]. NAM is recorded using the NAM microphone attached to the skin surface behind the ear, as shown in **Figure 1**. In this work, a neckband-type of NAM microphone [7], which presses NAM microphone against the place, is used to stably attach it. Because NAM is a particularly soft whispered voice, it is amplified with a special amplifier. **Figure 2** shows an example of waveform and 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.

B. Conventional Work of NAM Recognition

The main difference between a normal speech recognition system and a NAM recognition system is an acoustic model. Since the amount of NAM data is still limited, most of the conventional work focused on the development of speaker-dependent NAM acoustic models. It has been reported that model adaptation techniques such as Maximum Likelihood Linear Regression (MLLR) [12] is effective to develop hidden Markov models (HMMs) for NAM from those for normal speech [8]. Moreover, the adapted speaker-dependent NAM acoustic models are further improved by refining the initial HMMs using NAM data from multiple different speakers [7] and also using many speakers' normal speech data transformed into NAM acoustic space [9]. Consequently, around 70% word accuracy has been achieved for various speakers in large vocabulary continuous NAM recognition. On the other hand,

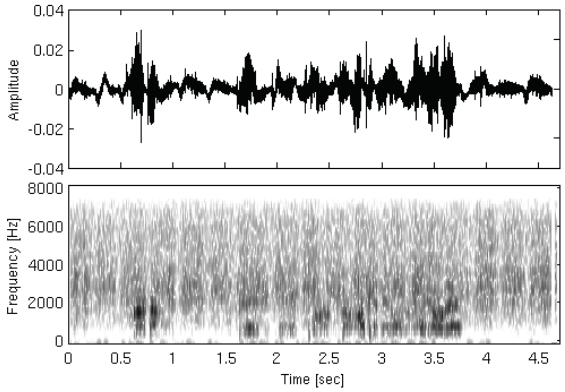

Fig. 3. Example of waveform and spectrogram of NAM signal when the speaker moves during speaking.

the effect of noise generated by the speaker's movements has been minimized in these studies by asking the speakers to maintain their positions as stably as possible during speaking.

C. Effect of Speaker's Movements on NAM Signal

The skin surface and muscles around the place of attachment of the NAM microphone usually move in conjunction with the speaker's movements in particular such as the movements of his/her head. These movements often change the setting condition of the NAM microphone. **Figure 3** shows an example of waveform and spectrogram of NAM when the speaker lightly shakes his head. We can see that the NAM signal is severely deteriorated by noise caused by the speaker's movements. The generated noise is non-stationary and causes substantially large acoustic differences compared with the NAM signal shown in **Figure 2**.

III. Blind Noise Suppression with Stereo Signals for NAM Recognition

In the proposed method, two NAM microphones are placed on the neck below the both ears to detect a stereo signal. The stereo signal allows us to use various effective noise suppression techniques such as the beam forming process. First, we represent the sound mixing model in the stereo signal detected with the NAM microphones. And then, the proposed blind noise suppression method is described.

A. Stereo Signal Modeling

When the speaker does not move, the stereo signal of NAM without suffering from any non-stationary noises is detected. A short-time analysis of the detected stereo signal is conducted by frame-by-frame discrete Fourier transform (DFT). The detected stereo NAM signal $\boldsymbol{s}(f,\tau) = [s_1(f,\tau), s_2(f,\tau)]^\top$ consisting of the first channel signal $s_1(f,\tau)$ and the second channel signal $s_2(f,\tau)$ is modeled by

$$\boldsymbol{s}(f,\tau) = \boldsymbol{a}(f)s_0(f,\tau), \quad (1)$$

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 $s_0(f,\tau)$, which is unobserved. It is linearly filtered

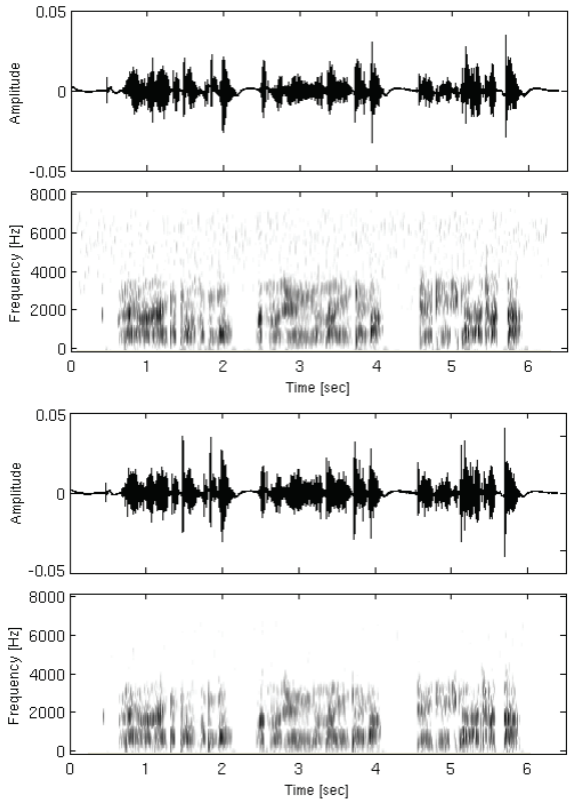

Fig. 4. Example of waveform and spectrogram of stereo NAM signal (top: the 1st channel, bottom: the 2nd channel).

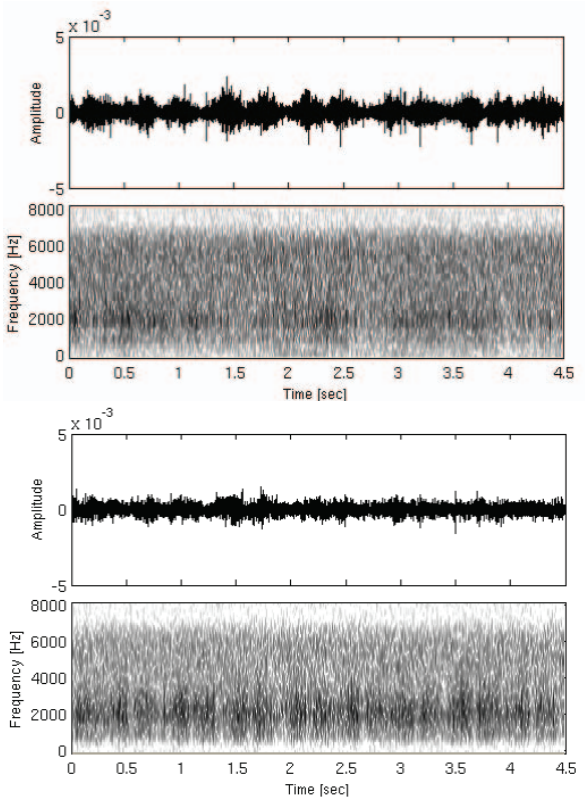

Fig. 5. Example of waveform and spectrogram of stereo noise signal caused by speaker's movement (top: the 1st channel, bottom: the 2nd channel).

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. **Figure 4** shows an example of the stereo NAM signal. We can see that there are some acoustic differences between NAM signals in different channels but they highly correlate to each other. We have confirmed from a result of our preliminary experiment that the NAM signal in one channel is effectively cancelled by the linearly filtered NAM signal in the other channel.

Non-stationary noise is generated depending on the speaker's movements. **Figure 5** shows an example of the stereo noise signal caused by a light shake of the head. The noise signals in individual channels seem to be synchronized at some parts but the correlation between them is actually weak. Therefore, the detected stereo noise signal is modeled by $\boldsymbol{n}(f,\tau) = [n_1(f,\tau), n_2(f,\tau)]^\top$ as diffuse noise signals. It may also be modeled by

$$\boldsymbol{n}(f,\tau) \quad = \quad \boldsymbol{b}(f,\tau) n_0(f,\tau), \tag{2}$$

where a component of an unobserved noise signal given by $n_0(f,\tau)$ is linearly filtered by channel-dependent and time-variant transfer functions $\boldsymbol{b}(f,\tau) = [b_1(f,\tau), b_2(f,\tau)]^\top$ affected by the changes of setting conditions of individual NAM microphones.

Assuming that the non-stationary noise is additive to the NAM signal, the detected stereo signal when the speaker moves during speaking in NAM is modeled by

$$\boldsymbol{x}(f,\tau) \quad \simeq \quad \boldsymbol{a}(f) s_0(f,\tau) + \boldsymbol{n}(f,\tau). \tag{3}$$

Note that to simplify the mixing process we also assume that the speaker's movements do not change the transfer function $\boldsymbol{a}(f)$. It is investigated in **Section IV-B** whether or not this assumption is valid.

B. Blind Spatial Subtraction Array (BSSA)

BSSA [10] is a blind noise suppression method based on two key processes, 1) noise estimation based on the target signal canceller with an adaptive beam former designed by blind source separation (BSS) based on ICA and 2) SS with the estimated noise signal. This technique is capable of effectively enhancing the target signal as long as it is well suppressed by the beam former even if the additive noise signal is diffuse, which is essentially difficult to be suppressed by the beam former. These conditions are matched with the mixing process given by Eq. (3). Moreover, since BSSA is a blind process, we don't have to know a position of microphones, the direction of arrival (DOA) of the target signal, and so on. In our case, the position of NAM microphones and the transfer function in Eq. (1) substantially vary depending on individual speakers and the setting conditions of NAM microphones. Moreover, we don't know the DOA of the NAM signal. Therefore, the blind process is essential in NAM recognition. The proposed noise reduction process based on BSSA for NAM recognition is shown in **Figure 6**.

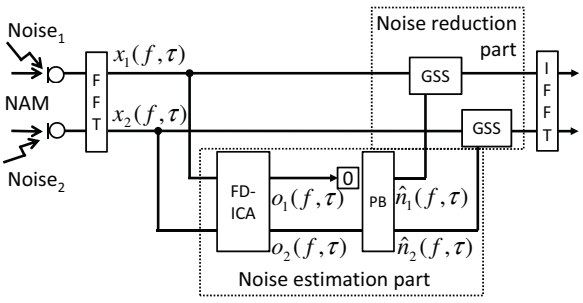

Fig. 6. Block diagram of BSSA for NAM recognition.

1) Noise Estimation Process: Frequency domain ICA (FD-ICA) based on higher-order statistics is used to estimate the noise signal. The detected stereo mixed-signal is separated with the complex valued unmixing 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

$$\boldsymbol{o}(f,\tau) \quad = \quad \boldsymbol{W}_{\mathrm{ICA}}(f)\boldsymbol{x}(f,\tau), \tag{4}$$

where the unmixing 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 iterative equation:

$$\begin{aligned}\boldsymbol{W}_{\mathrm{ICA}}^{[i+1]} \;=\; & \boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f) \\ & +\alpha\left[\boldsymbol{I} - \langle \boldsymbol{\Phi}(\boldsymbol{o}(f,\tau))\boldsymbol{o}^H(f,\tau)\rangle_\tau\right]\boldsymbol{W}_{\mathrm{ICA}}^{[i]}(f),\end{aligned} \tag{5}$$

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 [13]. In this paper, we determine the unmixing matrix utterance by utterance.

In the mixed signal modeled by Eq. (3), the separation process given by Eq. (4) 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 $s_0(f,\tau)$. In other words, it is capable of well estimating a component related to the noise signal $n_0(f,\tau)$. 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:

$$\boldsymbol{o}^{(n)}(f,\tau) = [0, o_2(f,\tau)]^\top. \tag{6}$$

An initial matrix of $\boldsymbol{W}_{\mathrm{ICA}}$ is designed so that $o_2(f,\tau)$ becomes the noise component. The DOA analysis may also be used to determine which is the noise component, $o_1(f,\tau)$ or $o_2(f,\tau)$ [13]. And then, the projection back (PB) process [14] 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:

$$\hat{\boldsymbol{n}}(f,\tau) \quad = \quad \boldsymbol{W}_{\mathrm{ICA}}^{+}(f)\boldsymbol{o}^{(n)}(f,\tau) \tag{7}$$

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.

2) Noise Reduction Process: In the noise reduction process, SS using power spectrum of the estimated noise signals is performed to effectively reduce the non-stationary noise components. In the original BSSA, the delay-and-sum (DS) process is performed to combine the multi-channel target and noise signals into monaural target and noise signals before the SS. On the other hand, it is not straightforward to combine the stereo NAM signal into the monaural NAM signal with the DS process since the transfer function of the NAM signal in each channel (*i.e.*, each component of $\boldsymbol{a}(f)$ in Eq. (1)) is quite different from each other and highly depends on not only the DOA but also various factors. Therefore, in the proposed method, SS is performed in each channel. Moreover, to improve the noise estimation accuracy, power spectrum of the estimated noise signal in each channel is compensated with frequency-dependent and time-invariant weights. These weights are determined so that the time-averaged power spectrum of the estimated noise signal in a noise-only segment (*e.g.*, a silence part at the beggining of an utterance) is close to that of the detected mixed-signal in the corresponding noise-only segment, which is regarded as the target reference. The generalized SS (GSS) [15] is performed to reduce the artificial distortion usually caused by an oversubtraction as much as possible. The enhanced stereo NAM signal $\hat{\boldsymbol{s}}(f,\tau) = [\hat{s}_1(f,\tau), \hat{s}_2(f,\tau)]^\top$ extracted from the detected stereo mixed-signal $\boldsymbol{x}(f,\tau)$ is given by

$$\hat{s}_c(f,\tau) = \begin{cases} \sqrt[2\xi]{|x_c(f,\tau)|^{2\xi} - \beta|\hat{n}_c(f,\tau)|^{2\xi}}e^{j\arg(x_c(f,\tau))} \\ \quad (\text{if } \; |x_c(f,\tau)|^{2\xi} > \beta|\hat{n}_c(f,\tau)|^{2\xi}) \\ 0 \quad (\text{otherwise}) \end{cases}, \tag{8}$$

where c is a channel index ($c = 1, 2$), β is an oversubtraction parameter, and ξ is an exponential domain parameter.

The estimated stereo NAM signal includes the residual noise components and is still artificially distorted even if carefully tuning the GSS parameters. Consequently, there are negligible acoustic differences between the enhanced stereo NAM signal $\hat{\boldsymbol{s}}(f,\tau)$ and the original stereo NAM signal $\boldsymbol{s}(f,\tau)$. To reduce the effect of these differences on NAM recognition, noise superimposition process [16] is performed. A pre-defined stationary noise signal is superimposed on the detected NAM signals with a constant SNR, and then the acoustic model for NAM recognition is developed using them as the training data. In the recognition process, the same noise signal is superimposed on the enhanced stereo NAM signal.

C. Frame-Wise Channel Selection Process

There are several choices to recognize the estimated stereo NAM signal; *e.g.*, a recognition process is performed in only one pre-decided channel, or recognition processes are separately performed in individual channels and then a recognition result in either channel is selected using some measures. Since the noise signal in each channel varies differently from each other, the channel causing larger distortion of the enhanced NAM signal also changes frame by frame. Therefore, it is expected that a frame-wise channel selection process is effective.

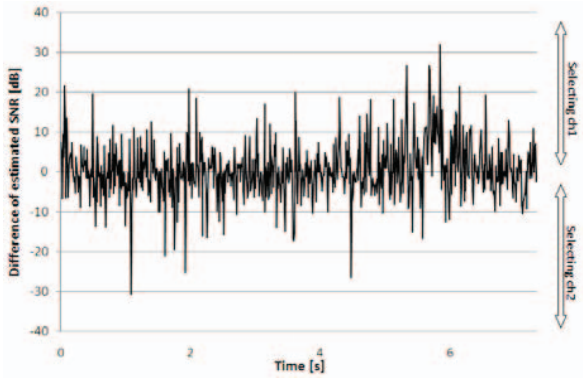

Fig. 7. Difference of SNRs estimated frame by frame in individual channels (*i.e.*, $\mathrm{SNR}_{1,\tau} - \mathrm{SNR}_{2,\tau}$ in Eq. (9)).

As the selection measure to detect the enhanced NAM signal less distorted by GSS, a SNR is estimated frame by frame based on the detected stereo mixed-signal $\boldsymbol{x}(f,\tau)$ and the estimated stereo noise signal $\hat{\boldsymbol{n}}(f,\tau)$ as follows:

$$\mathrm{SNR}_{c,\tau} = 10\log_{10}\frac{\sum_f |x_c(f,\tau)|^2 - \sum_f |\hat{n}_c(f,\tau)|^2}{\sum_f |\hat{n}_c(f,\tau)|^2}. \quad (9)$$

The channel with a higher estimated SNR is selected frame by frame. **Figure 7** shows an example of the difference of estimated SNR trajectories of individual channels. Based on the selection result, the acoustic feature sequences of the enhanced NAM signals in individual channels are combined into the single acoustic feature sequence.

There are other choices as selection measures, such as a residual SNR using the noise suprressed signal, but a result of our preliminary experiment showed that the SNR shown in Eq. (9) yields the best recognition accuracy.

IV. Experimental Evaluations

To investigate how much recognition accuracy is degraded by the speaker's movements and evaluate the effectiveness of the proposed method, large vocabulary continuous NAM recognition experiments were conducted.

A. Experimental Conditions

Stereo NAM data from a single Japanese male speaker were recorded with two NAM microphones. The sampling frequency was set to 16 kHz. To check how reasonable the assumption in the mixing process given by Eq. (3) was, we conducted two types of experimental evaluations. In one evaluation, we used *simulated mixed-signals* generated by superimposing the non-stationary noise signals, which were detected when the speaker moved without speaking in NAM, on the NAM signals, which were detected when the speaker did not move. In the other evaluation, we used *real mixed-signals* detected in a real situation where the speaker moved while speaking in NAM. In generating the *simulated mixed-signals*, power of the non-stationary noise signals was adjusted so that the recognition performance for the *simulated mixed-signal* was nearly equal to that for the *real mixed-signal*.

We adopted 12 MFCCs, 12 ΔMFCCs and Δ power as the acoustic features. The DFT size, window size, and shift size were set to 1024, 512, and 256, respectively. Left-to-right 3 state triphone HMMs with no skip were used as an acoustic model. The number of shared states was 2189 and the state output probability distribution was modeled with 16 mixture components of GMMs. This acoustic model was initially developed with normal speech database designed for training speaker-independent model, which included voices of several hundreds of speakers. Then, it was adapted to the NAM signals without non-stationary noise using iterative MLLR. The number of adaptation sentences was 208 and the number of test sentences was 143 sentences. They were selected from Japanese newspaper articles. We simultaneously used the NAM signals in both channels (*i.e.*, 416 utterances in total) to develop a common acoustic model for both channels. We used 60 k word trigram language model trained with Japanese newspaper articles.

The following five settings were evaluated:

- *Unprocessed* : The mixed-signals without any noise suppression processes were directly used.
- *GSS* : GSS using the time-averaged noise spectrum in the noise-only segment was performed in each channel.
- *Proposed BSSA* : The proposed method using only BSSA was performed.
- *Proposed BSSA and selection* : The proposed method using both BSSA and frame-wise channel selection was performed.
- *Clean* : NAM signal without suffering from non-stationary noise was used.

In the methods using GSS (*GSS*, *Proposed BSSA*, and *Proposed BSSA and selection*), a SNR of the superimposed noise mentioned in **Section III-B2** was set to 30 dB in both model adaptation process and recognition process. The oversubtraction parameter and the exponential domain parameter in Eq. (8) were set to 0.1 and 1/3, respectively. These settings were experimentally determined so that the recognition performance was maximized.

B. Experimental result

Figure 8 shows the result for the *simulated mixed-signals*. The non-stationary noise causes substantial degradation in word accuracy from 69.2% to 53.6% in the first channel (ch1) and from 67.3% to 52.1% in the second channel (ch2). The improvement in word accuracy by the conventional noise reduction with a monaural signal (*GSS*) is observed but it is limited (55.5% in ch1 and 52.9% in ch2) since this process cannot suppress the non-stationary noise components. *Proposed BSSA* yields significantly larger improvements in word accuracy (61.4% in ch1 and 61.6% in ch2) compared with *GSS* since the non-stationary noise estimation is performed by the stereo signal processing. Moreover, frame-wise channel selection yields further improvement (63.3% of *Proposed BSSA and selection*). There still remain the noticeable difference in word accuracy between *Proposed BSSA and selection* and *Clean*

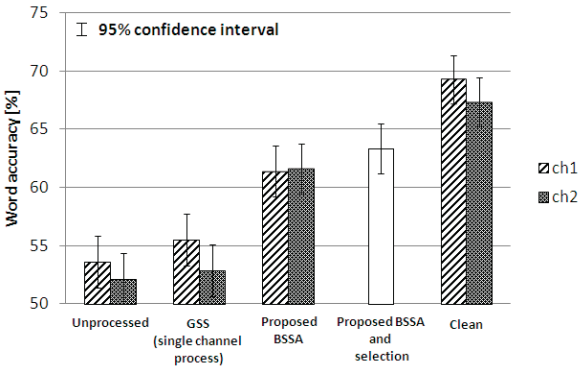

Fig. 8. Result for *simulated mixed-signals*.

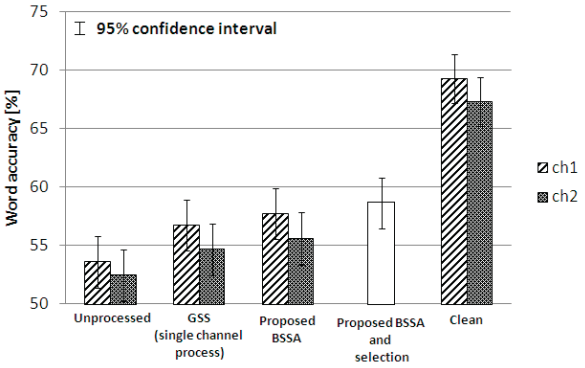

Fig. 9. Result for *real mixed-signals*.

because the diffuse noise components are essentially difficult to be estimated.

Figure 9 shows the result for the *real mixed-signals*. Compared with the result for the *simulated mixed-signals*, word accuracy in *Proposed BSSA* significantly decreases (57.7% in ch1 and 55.6% in ch2) and becomes close to that in *GSS* (56.7% in ch1 and 54.6% in ch2). We have noticed that a directivity pattern shaped by the beam former to suppress the NAM signal, which is given by a part of the unmixing matrix, has the deep null at around zero degree of the DOA in the *simulated mixed-signals* but the null becomes much shallower and its degree is slipped in the *real mixed-signals*. This is because the assumption in the mixing process given in Eq. (3) is not valid in a real situation; *i.e.*, the speaker's movements also cause the change of the transfer function $\boldsymbol{a}(f)$. Consequently, the estimation accuracy of the non-stationary noise becomes low. Nevertheless, the estimated stereo noise signal is still useful to perform the frame-wise channel selection that yields the best word accuracy (58.6%), which is statistically significantly better than the results of *Proposed BSSA* and *GSS* in the second channel. In a real situation, we don't know which channel yields better word accuracy and it will vary depending on the setting conditions of the NAM microphones. Therefore, *Proposed BSSA and selection* capable of blindly selecting a better channel is more effective than *Proposed BSSA* and *GSS*.

V. Conclusions

In this paper, we proposed the blind noise suppression method for Non-Audible Murmur (NAM) recognition to alleviate the word accuracy degradation caused by non-stationary noise generated by speaker's movements during speaking. Using a stereo signal detected with two NAM microphones, the non-stationary noise was estimated with blind source separation and independent component analysis, and spectral subtraction was performed to enhance the stereo NAM signal. Moreover, frame-wise channel selection was performed to construct less distorted monaural NAM signal. The experimental results demonstrated that the proposed method is capable of significantly reducing the degradation in NAM recognition accuracy caused by the speaker's movements.

Acknowledgment: This research was supported in part by MEXT Grant-in-Aid for Scientific Research (A).

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

[4] 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.

[5] 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.

[6] 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.

[7] 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, Brighton, UK, Sep. 2009.

[8] 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.

[9] 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, Prague, Czech Republic, May. 2011.

[10] 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.

[11] S. Boll. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, Vol. 27, No. 2, pp. 113–120, 1979.

[12] M.J.F. Gales. Maximum likelihood linear transformations for HMM-based speech recognition. *Computer Speech and Language*, Vol. 12, No. 2, pp. 75–98, 1998.

[13] H. Saruwatari, S. Kurita, K. Takeda, F. Itakura, T. Nishikawa, and K. Shikano. Blind source separation combining independent component analysis and beamforming. *EURASIP Journal on Applied Signal Processing*, Vol. 2003, No. 11, pp. 1135–1146, 2003.

[14] S. Ikeda and N. Murata. A method of ICA in time-frequency domain. *Proc. ICA*, pp. 365–370, Aussions, France, Jan. 1999.

[15] B.L. Sim, Y.C. Tong, J.S. Chang, and C.T. Tan. A parametric formulation of the generalized spectral subtraction method. *IEEE Trans. on Speech and Audio Processing*, Vol. 6, No. 4, pp. 328–337, 1998.

[16] S. Yamade, A. Lee, H. Saruwatari, and K. Shikano. Unsupervised speaker adaptation based on HMM sufficient statistics in various noisy environments. *Proc. INTERSPEECH*, pp. 1493–1496, Geneva, Switzerland, Sep. 2003.