The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

Improvement of Speech Recognition Performance for Spoken-Oriented Robot Dialog System Using End-fire Array

†Hiroshi Sawada, †Jani Even, †Hiroshi Saruwatari, †Kiyohiro Shikano and ‡Tomoya Takatani

***Abstract*— In this paper, we propose a microphone array structure for a spoken-oriented robot dialog system that is designed to discriminate the direction of arrival (DOA) of the target speech and that of the robot internal noise. First, we investigate the performance of the noise estimation conducted by semi-blind source separation (SBSS) in presence of both the diffuse background noise and the robot internal noise. The result indicates that the noise estimation of the SBSS is not good. Next, we analyze the DOA of the robot internal noise in order to determine the reason of the above result; we find out that the internal noise is always in-phase at the microphone array and overlap spacial with the target speech. Based on this fact, we propose to change the microphone array structure from the broadside array to the end-fire array in order to discriminate the DOAs of the target speech and the internal noise. Finally, we evaluate the word accuracy in a dictation task in presence of both diffuse background noise and robot internal noise to confirm the advantage of the proposed structure. Simulation results shows that the proposed microphone array structure results in approximately 10% improvement of the speech recognition performance.**

I. Introduction

In a hands-free dialog system, the user's voice is picked at a distance with a microphone array resulting in a more natural and stress-free interface for humans. In this system, however, it is difficult to achieve a high recognition accuracy because the noise generated by surrounding sound sources and the room reverberation always contaminate the target speech. For hands-free dialog systems mounted on a robot, the situation is even more difficult as the robot itself has several internal noise sources: fans, servo motors, and several mechanical parts. Moreover these internal noise sources are relatively close to microphone array and thus highly contaminate the acquired user's speech. But contrary to the noise created by the sources that are outside of the robot (referred to as background environmental noise), it is possible to install some sensors (referred to as internal noise sensors) ,e.g., a acceleration sensor and a Non-Audible Murmur (NAM) microphone, inside of the robot that collect additional information on the noise from inside on the robot (referred to as internal noise).

One approach using such internal noise sensors to address this problem is a combination of semi-blind source separation (SBSS) [1] and Wiener filter (WF) [2] (see Fig. 1). First, both background environmental and internal noises are estimated by SBSS (which is based on independent component analysis (ICA) [3]). Next, the target speech extraction is achieved by applying the WF on each of the microphone array signals. Finally, the outputs of these WF are merged together with a delay-and-sum (DS) [4] beamformer to obtain the enhanced speech fed to the speech recognizer. However, when the user stands in front of the robot and the microphone array installed outside of the robot is a broadside array, the absolute speech recognition score is not efficient due to the insufficient noise estimation performance in SBSS (this situation is depicted by Fig. 2 when $\theta_m = 0$).

In this paper, we focus our attention on the performance

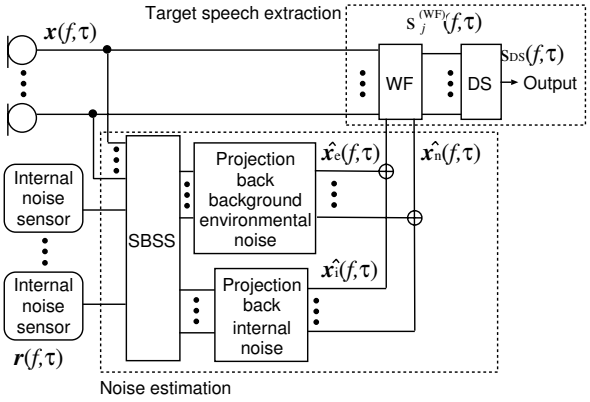

Fig. 1. Block diagram of the speech extraction method.

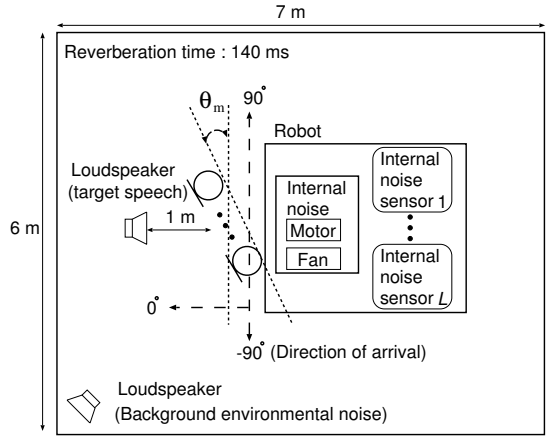

Fig. 2. Layout of the reverberant room used in our simulation.

This work was partly supported by the MIC Strategic Information and Communications R&D Promotion Programme in Japan.

† are with the Graduate School of Information Science, Nara Institute of Science and Technology, Nara 630-0192, Japan (e-mail: {hiroshi-s, even, sawatari, shikano}@is.naist.jp). Tel: +81-743-72-5287, Fax: +81-743-72-5289.

‡ is with TOYOTA MOTOR CORPORATION, Aichi 471-8571, Japan.

of the SBSS based noise estimation and relate it to the direction of arrival (DOA) of the internal noise. An important finding is that the reason why the performance of the noise estimation is not good with a broadside array is because the internal noise is always in-phase at the microphone array and thus it's DOA is approximately the same as that of the target speech. Consequently, in order to improve the performance of SBSS for noise estimation, we modify the microphone array structure to discriminate DOAs of the target speech and the internal noise. In particular, we vary the angle of the microphone array to change the DOA of the internal noise (see angle θ_{m} in Fig. 2). We determine the optimal angle θ_{m} that results in the best estimation of the internal noise via a computer simulation. To illustrate the effectiveness of the proposed approach, we evaluate the word accuracy in a dictation task in presence of both diffuse background noise and robot internal noise, and show the improvement of the speech recognition performance.

II. Target speech extraction using Semi-blind source separation

A. Acoustic mixing model

We consider an acoustic mixing model where the number of microphones is J, and the number of internal noise sensors is L (see Fig. 3). Let f denotes the frequency bin number, and τ denotes the time-frame index number. The observed signal at the microphone array $\boldsymbol{x}(f,\tau) = [x_1(f,\tau),\ldots,x_J(f,\tau)]^{\mathrm{T}}$ is a mixture of one target speech signal $s(f,\tau)$, the background environmental noise signal $\boldsymbol{n}_{\mathrm{e}}(f,\tau) = [n_1^{(\mathrm{e})}(f,\tau),\ldots,n_J^{(\mathrm{e})}(f,\tau)]^{\mathrm{T}}$, and the internal noise signal $\boldsymbol{n}_{\mathrm{i}}(f,\tau) = [n_1^{(\mathrm{i})}(f,\tau),\ldots,n_K^{(\mathrm{i})}(f,\tau)]^{\mathrm{T}}$ (the number of internal noise signals is K). The observed signal vector at the internal noise sensors $\boldsymbol{r}(f,\tau) = [r_1(f,\tau),\ldots,r_L(f,\tau)]^{\mathrm{T}}$ depends only of the internal noise signal.

Then the observed signals at the microphone array and the internal noise sensors are given by,

$$\boldsymbol{x}(f,\tau) = \boldsymbol{h}_1(f)s(f,\tau) + \boldsymbol{n}_{\mathrm{e}}(f,\tau) + \boldsymbol{H}_2(f)\boldsymbol{n}_{\mathrm{i}}(f,\tau), \quad (1)$$

$$\boldsymbol{r}(f,\tau) = \boldsymbol{H}_3(f)\boldsymbol{n}_{\mathrm{i}}(f,\tau), \quad (2)$$

where $\boldsymbol{h}_1(f) = [h_1^{(1)}(f),\ldots,h_J^{(1)}(f)]^{\mathrm{T}}$ is the column vector containing the transfer functions from the target signal component to each microphone, $\boldsymbol{H}_2(f)$ $(J \times K)$ is the matrix containing the transfer functions from the internal noise components to each microphone, and $\boldsymbol{H}_3(f)$ $(L \times K)$ is the matrix containing the transfer functions from the internal noise components to each internal noise sensor.

The observed signal vector at the internal noise sensors depends only of the internal noise signal because the aerial vibration of the target speech is not recorded by the internal noise sensors. Thus, we can also assume that the internal noise sensors observe only the internal noise as vibrations transmitted through the chassis of the robot and not through the air. Figure 4 shows spectrograms of the observed signal at the internal noise sensors and the true internal noise signal at the microphone array. As showed in Fig. 4, the frequency characteristics of these signals differ (different types of sensors and propagation paths). Therefore, directly using the signal from the internal noise sensors to suppress the contribution of the internal noise at the microphone array results in a deep deterioration of the estimated target speech quality. Thus, we need to estimate the internal noise signal at the microphone array from the observed signals (microphone array and internal noise sensors).

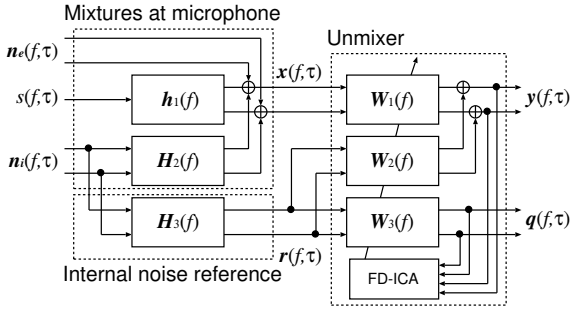

Fig. 3. Block structure of the mixing and the unmixing at the fth frequency bin.

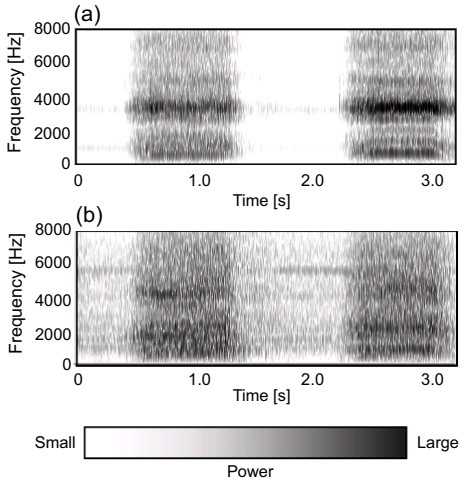

Fig. 4. Spectrograms of (a) the observed signal at internal noise sensors and (b) the true internal noise signal at the microphone array.

B. Target speech extraction

In ICA, the source separation is performed by applying the unmixing matrices $\boldsymbol{W}_i(f)$ $(i = 1,2,3)$ (of size $J \times J$, $J \times K$, and $K \times K$) to the observed signals

$$\boldsymbol{y}(f,\tau) = \boldsymbol{W}_1(f)\boldsymbol{x}(f,\tau) + \boldsymbol{W}_2(f)\boldsymbol{r}(f,\tau), \quad (3)$$

$$\boldsymbol{q}(f,\tau) = \boldsymbol{W}_3(f)\boldsymbol{r}(f,\tau), \quad (4)$$

and update these matrices such that the components of $\boldsymbol{y}(f,\tau) = [y_{\mathrm{s}}(f,\tau), y_{\mathrm{n}}(f,\tau)]^{\mathrm{T}}$ and $\boldsymbol{q}(f,\tau) = [q_1(f,\tau),\ldots,q_K(f,\tau)]^{\mathrm{T}}$ become mutually independent.

In this paper, we use an iterative update of the unmixing matrices. Using the superscript $\cdot^{[k]}$ to denotes a value at the kth iteration, we have the following update rules

$$\boldsymbol{W}_i^{[k+1]}(f) = \boldsymbol{W}_i^{[k]}(f) - \mu\Delta\boldsymbol{W}_i^{[k]}(f), \quad (5)$$

$$\Delta \boldsymbol{W}_1^{[k+1]}(f) = (\boldsymbol{I} - \langle \boldsymbol{\Phi}(\boldsymbol{y}(f,\tau)^{[k]})\boldsymbol{y}^{\mathrm{H}}(f,\tau)^{[k]}\rangle_\tau)\boldsymbol{W}_1^{[k]}(f), \quad (6)$$

$$\Delta \boldsymbol{W}_2^{[k+1]}(f) = (\boldsymbol{I} - \langle \boldsymbol{\Phi}(\boldsymbol{y}(f,\tau)^{[k]})\boldsymbol{y}^{\mathrm{H}}(f,\tau)^{[k]}\rangle_\tau)\boldsymbol{W}_2^{[k]}(f) - (\langle \boldsymbol{\Phi}(\boldsymbol{y}(f,\tau)^{[k]})\boldsymbol{q}^{\mathrm{H}}(f,\tau)^{[k]}\rangle_\tau)\boldsymbol{W}_3^{[k]}(f), \quad (7)$$

$$\Delta \boldsymbol{W}_3^{[k+1]}(f) = (\boldsymbol{I} - \langle \boldsymbol{\Phi}(\boldsymbol{q}(f,\tau)^{[k]})\boldsymbol{q}^{\mathrm{H}}(f,\tau)^{[k]}\rangle_\tau)\boldsymbol{W}_3^{[k]}(f), \quad (8)$$

where μ is the step size parameter, $\boldsymbol{I}$ is an identity matrix, $\langle\cdot\rangle_\tau$ denotes a time-averaging operator, and $\boldsymbol{M}^{\mathrm{H}}$ denotes hermitian transpose of matrix $\boldsymbol{M}$. The appropriate nonlinear vector function $\boldsymbol{\Phi}(\cdot)$ is estimated from the data using a kernel-based estimate of the score function [5]. After convergence, the permutation problem is resolved using the method combining DOA estimate and probability density distribution estimate [6].

When separating a point-source target speech and a non-point source noise, ICA estimates efficiently the noise by steering a directional null in the direction of the target, whereas the speech estimate is of poor quality [7]. Thus we utilize ICA as an estimator for both external and internal noises but not for the speech. These noise estimates are given by

$$\hat{\boldsymbol{x}}_{\mathrm{e}}(f,\tau) = \boldsymbol{W}_1^{+}(f)[0, y_{\mathrm{n}}(f,\tau)]^{\mathrm{T}}, \quad (9)$$

$$\hat{\boldsymbol{x}}_{\mathrm{i}}(f,\tau) = -\boldsymbol{W}_1^{+}(f)\boldsymbol{W}_2(f)\boldsymbol{W}_3^{+}(f)\boldsymbol{q}(f,\tau), \quad (10)$$

where $\hat{\boldsymbol{x}}_{\mathrm{e}}(f,\tau) = [\hat{x}_1^{(\mathrm{e})}(f,\tau),\ldots,\hat{x}_J^{(\mathrm{e})}(f,\tau)]^{\mathrm{T}}$ is the estimated background environmental noise signal vector, and $\hat{\boldsymbol{x}}_{\mathrm{i}}(f,\tau) = [\hat{x}_1^{(\mathrm{i})}(f,\tau),\ldots,\hat{x}_J^{(\mathrm{i})}(f,\tau)]^{\mathrm{T}}$ is the estimated internal noise signal vector (both estimated at the microphone array). Next, noise canceling is performed by applying a WF on each of the microphone array signals. The noise estimates used in the WF are obtained by adding the contributions of both external and internal noises at the microphones

$$\hat{\boldsymbol{x}}_{\mathrm{n}}(f,\tau) = \hat{\boldsymbol{x}}_{\mathrm{e}}(f,\tau) + \hat{\boldsymbol{x}}_{\mathrm{i}}(f,\tau), \quad (11)$$

where $\hat{\boldsymbol{x}}_{\mathrm{n}}(f,\tau) = [\hat{x}_1^{(\mathrm{n})}(f,\tau),\ldots,\hat{x}_J^{(\mathrm{n})}(f,\tau)]^{\mathrm{T}}$ contains all the components of the estimated noise signal vector. The WF gain is designed as follows:

$$g_j(f,\tau) = \frac{|x_j(f,\tau)|^2}{|x_j(f,\tau)|^2 + \beta|\hat{x}_j^{(\mathrm{n})}(f,\tau)|^2}, \quad (12)$$

where $g_j(f,\tau)$ is the WF gain at jth channel, and β is a gain factor. The J enhanced speech signals obtained by the Wiener filtering are

$$s_j^{(\mathrm{WF})}(f,\tau) = \sqrt{g_j(f,\tau)|x_j(f,\tau)|^2}\,\frac{x_j(f,\tau)}{|x_j(f,\tau)|}. \quad (13)$$

Finally, the J Wiener-filtered speech estimates are merged into a single-channel signal by applying a DS beamformer as follows:

$$s_{\mathrm{DS}}(f,\tau) = \boldsymbol{w}_{\mathrm{DS}}(f,\theta_{\mathrm{U}})^{\mathrm{T}}[s_1^{(\mathrm{WF})}(f,\tau),\ldots,s_J^{(\mathrm{WF})}(f,\tau)]^{\mathrm{T}}, \quad (14)$$

$$\boldsymbol{w}_{\mathrm{DS}}(f,\theta) = [w_1^{(\mathrm{DS})}(f,\theta),\ldots,w_J^{(\mathrm{DS})}(f,\theta)]^{\mathrm{T}}, \quad (15)$$

where $s_{\mathrm{DS}}(f,\tau)$ is the estimated target speech signal, θ_{U} is the look direction which is estimated from the unmixing matrix optimized by ICA [8], and $\boldsymbol{w}_{\mathrm{DS}}(f,\theta)$ is the coefficient vector of the DS array which is defined by

$$w_j^{(\mathrm{DS})}(f,\theta) = \frac{1}{J}\exp(-i2\pi(f/N)f_{\mathrm{s}}jd\sin\theta/c), \quad (16)$$

where f_{s} is the sampling frequency, d is the microphone inter spacing, N is the DFT size, and c is the sound velocity.

The internal noise frequency characteristics differ greatly depending on the robot actions. In particular the periods when the robot does not move that contain only fan noise are very different from the periods when the robot moves as mechanical and motor noises can be heard. It was experimentally reported in [9] that changing the gain factor β of the WF according to the type of the internal noise improves the speech recognition performance (The type of the internal noise being determined by using the control signals of the robot). In this paper, we also consider fixed and non fixed gain factor β.

C. Problem of semi-blind source separation

We conducted a preliminary experiment to confirm the poor noise estimation performance of SBSS. Figure 2 ($\theta_{\mathrm{m}} = 0$) depicts the layout of the reverberation room used in this experiment. We used a four-element microphone array with an inter element spacing of 2.15 cm and three internal noise sensors were installed inside of the robot. We used 10 utterances (female speakers, 16kHz-sampled signals) convoluted with the impulse response that were recorded in this reverberant room to simulate a user standing in front of the microphone array at a distance of one meter. The background environmental noise signal is a noise recorded in an exhibition hall. The internal noise is a recording of the actual robot internal noise (fan noise, mechanical noise and motor noise). The input signal-to-noise ratio (SNR) between the target speech and the background environmental noise is set to 10 dB, and the input SNR between the target speech and the internal noise (referred to as internal SNR) is 16.6 dB.

We evaluate the noise estimation performance in SBSS on the basis of the spectral distortion (SD) $e(f)$ which is defined as follows:

$$e(f) = 10\log_{10}\left(\frac{1}{J}\sum_j\sum_\tau |x_j^{(\mathrm{n})}(f,\tau) - \hat{x}_j^{(\mathrm{n})}(f,\tau)|^2\right), \quad (17)$$

where $x_j^{(\mathrm{n})}$ denotes the true noise signal at the jth channel (the sum of the internal and the background environmental noises). The small SD indicates the high noise estimation performance. Thus, if we achieve the perfect noise estimation, the SD will be minus infinity.

Figure 5 shows the SD averaged on the ten target speaker utterances. We can see that at low frequencies the noise estimate given by SBSS is severely distorted.

III. Proposed method

A. Overview

In this section, we first analyze the DOA of the internal noise to clarify the cause of the poor noise estimation performance in SBSS. As a result of this analysis, we confirm that

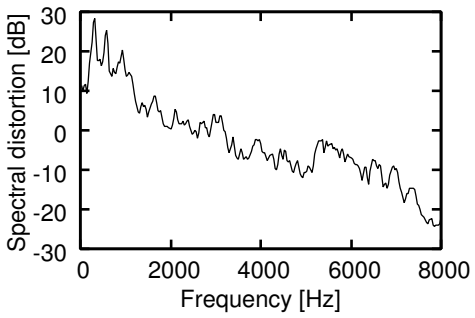

Fig. 5. Spectral distortion between the components of the true noise and the estimated noise (averaged on channels and utterances).

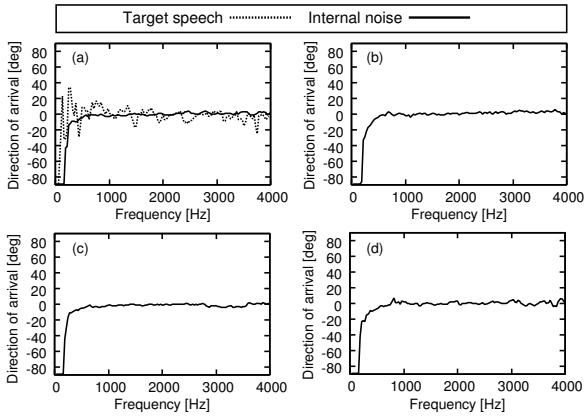

Fig. 6. DOAs of the internal noise (a) type 1 and the target speech, (b) type 2, (c) type 3, (d) type 4.

the internal noise is always in-phase at the microphone array. Based on this finding, we propose a solution to overcome the problem of SBSS.

B. DOA-based analysis

We first conduct an experiment to characterize the internal noise DOA. The conditions are the same as the previous experiment.

As showed in Fig. 2, the axis used for measuring the DOA is such that the target speech for a user standing in front of the broadside array ($\theta_{\mathrm{m}} = 0$) has a DOA of zero degree.

The robot actions and fan noise (internal noise) were recorded in four situations (type 1 to 4) while performing different movements that create additional mechanical and motor noises for short periods. The internal noise of type 1 designs the shaking of the robot's head and is a relatively calm noise. But the other internal noise types (2-4) design movements of the arms and these noises are loud to the point of masking the target speech.

To estimate the DOA of the internal noise and the target speech we use a minimum variance (MV) method. First, we calculate the estimated power $P(f,\theta)$ given by

$$P(f,\theta) = \frac{1}{\boldsymbol{a}^{\mathrm{H}}(f,\theta)\boldsymbol{R}^{-1}(f)\boldsymbol{a}(f,\theta)}, \tag{18}$$

$$\boldsymbol{R}(f) = \mathrm{E}[\boldsymbol{z}(f,\tau)\boldsymbol{z}^{\mathrm{H}}(f,\tau)], \tag{19}$$

$$\boldsymbol{a}(f,\theta) = [a_1(f,\theta),\ldots,a_J(f,\theta)]^{\mathrm{T}}, \tag{20}$$

$$a_j(f,\theta) = \exp(i2\pi(f/N)f_{\mathrm{s}}jd\sin\theta/c), \tag{21}$$

where $\boldsymbol{a}(f,\theta)$ is the steering vector, $\boldsymbol{R}(f)$ is the correlation matrix, and $\boldsymbol{z}(f,\tau) = [z_1(f,\tau),\ldots,z_J(f,\tau)]^{\mathrm{T}}$ is the input signal vector (E[·] denotes an expectation operator). In this experiment, input signal vector is the true target speech signal or the internal noise signal at the microphone array. Next, we vary the angle θ from -90 to 90 using a unit increment and select the value giving the largest $|P(f,\theta)|$ as the DOA of the input signal.

The estimated DOAs for all the internal noise types and the target speech are plotted in Figs. 6(a), 6(b), 6(c) and 6(d). Since the wavelength in low frequencies is long, no DOA estimate methods can calculate the correct DOA. Thus, we do not consider the DOA result below 500 Hz. We can see that the DOAs of all the internal noise types are approximately zero degree. The four types of noises we consider are generated from various locations of the robot (e.g. neck, left arm, right arm) but all of these noises are in-phase at the microphone array. Meaning that no matter the location from which the noises are generated their apparent DOA at the microphone array is zero degree.

The reason of the above result is that when the robot makes a movement, the microphone array vibrates with the chassis. Thus the observed internal noise contains vibrations that propagate through the robot chassis. Generally speaking, the sound velocity in the solid is faster than that in the air. Therefore, the sound velocity of the internal noise is fast, and the time-difference-of-arrival of each microphone is smaller than that with propagation through the air. Since the internal noise image is always in-phase at the microphone array regardless of the moving part of the robot, when the user stands in front of the robot with the broadside array, the DOA of the internal noise and that of the target speech are the same see Fig. 6(a).

C. Proposed structure

We consider that while using a speech-oriented human-machine interface almost all users stand in front of the microphone array. Consequently, when using the broadside array, the target speech and the internal noise have approximately the same DOA (see Fig. 7(a)). In such situation, ICA cannot estimate properly the internal noise as steering a directional null in the speech direction also suppress the internal noise (ICA cannot separate the sources which are spatially adjacent). To overcome this limitation, we modify the microphone array structure to discriminate the DOAs of the target speech and the internal noise. In particular, the microphone array mounted outside of the robot varies from the broadside array ($\theta_{\mathrm{m}} \cong 0$) to the end-fire array ($\theta_{\mathrm{m}} \cong 90$). As a result, the DOA of the internal noise is shifted to a different direction of that of the target speech while keeping the face-to-face relationship of the user and the robot, see Fig.7(b). Therefore, the DOA of the target speech

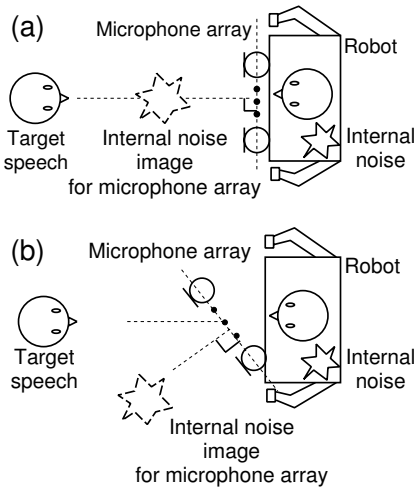

Fig. 7. (a) Conventional and (b) proposed microphone array structures and the internal noise image for the microphone array.

TABLE I
EXPERIMENTAL CONDITIONS FOR THE SPEECH RECOGNITION

test data	JNAS test set 100 utterances (female 23 speakers)
Speech recognition task	newspaper dictation (20 k word)
Acoustic model	phonetic tied mixture [12] based clean model with super-imposed noise (office noise 25 dB SNR)
Number of training speakers for acoustic model	260 speakers (150 utterances / 1speaker)
Decoder	Julius ver. 3.5.1

and that of the internal noise are no longer approximately the same. Consequently, it is expected that the quality of the noise estimate given by ICA improves. In the following, we determine the optimal angle of the microphone array for the internal noise of the robot and confirm the improvement of the speech recognition performance.

IV. EXPERIMENT AND RESULT

A. Experiment 1

1) Experimental setup: To confirm the effectiveness of varying the angle of the microphone array, we conducted a computer-simulation-based experiment. We conform the conditions of the reverberant room to the experiment of Sect. 2.3. However, the number of utterances is 100 (female speakers), and we use four kinds of internal noises. The internal SNR of type 1 internal noise is 16.6 dB, type 2 is 4.5 dB, type 3 is 0.4 dB, and type 4 is 5.4 dB. Also, the gain factor β of the WF is fixed at 5 for all internal noise types. The experiment is repeated with eleven different angle θ_{m}: ±90, ±60, ±45 ,±30, ±10, 0. The experimental conditions for the speech recognition show the Table 1.

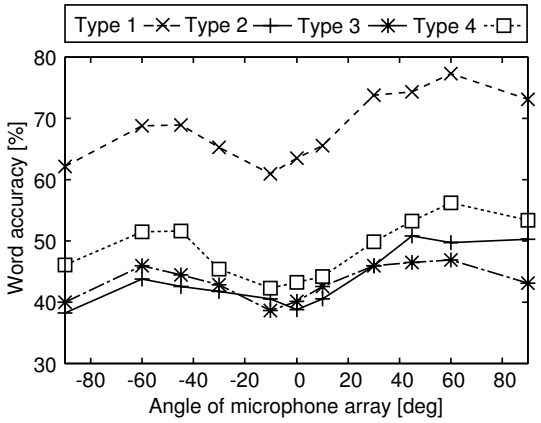

Fig. 8. Word accuracy for all the internal noise types with different angles of the microphone array.

2) Experimental result: We compared the different angles of the microphone array on the basis of word accuracy, noise reduction rate (NRR) which is defined as the output SNR in dB minus the input SNR in dB, cepstral distortion (CD) which is a measure of the degree of spectral envelope distortion in the cepstrum domain [10] and SD. Figure 8 shows the result of speech recognition test. We can see that word accuracy can be improved by varying the angle of the microphone array. In particular, word accuracy at θ_{m} = 60 degrees is obviously superior to that at θ_{m} = 0 degree for all internal noise types. We can achieve 14% (type 1), 11% (type 2), 7% (type 3) and 13% (type 4) improvements of the speech recognition result. NRR and CD (averaged on all target speaker utterances) are given in Figs. 9. We also show the result of SD in Fig.10. We can see that CD and SD at θ_{m} = 60 degrees is smaller than that at θ_{m} = 0 degree for all internal noise types, and that the NRRs for the cases of θ_{m} = 60 degrees and θ_{m} = 0 degree are almost the same except for the type 3 internal noise. This may be a clue to explain why the improvement is the least for type 3 noise. From these results, we can see that improving the SBSS noise estimation performance results in an improvement of the target speech extraction performance. This result also indicates that the improvement of the speech recognition performance is mainly due to the improvement of the CD.

B. Experiment 2

1) Experimental setup: We investigate the speech recognition performance at the optimal angle of the microphone array θ_{m} = 60 with optimized β. We conform the conditions of the reverberant room and speech recognition to the experiment 1 except parameter β. As mentioned in Sec.2.3, We change the β according to the noise type while the robot is moving. Values for stationary (β_1) and non stationary (β_2) parts are given are Table 2. These values are optimized based on word accuracy.

2) Experimental result: Figure 11 shows the result of the speech recognition test. We can see that word accuracy at θ_{m} = 60 and optimized β (Proposed 2) is superior to that

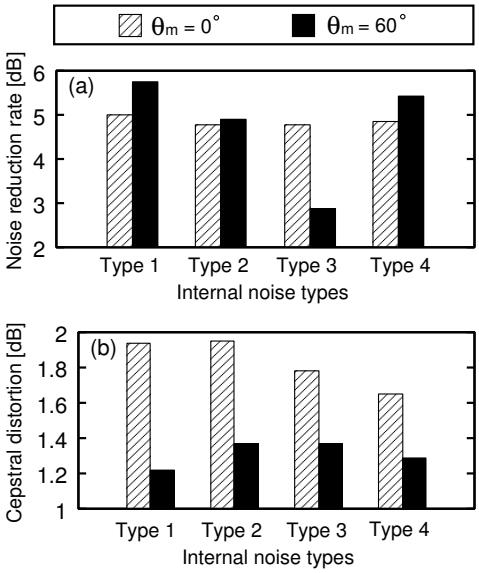

Fig. 9. Experimental results of (a) noise reduction rate (b) cepstral distortion for the cases of $\theta_m = 0$ degree and $\theta_m = 60$ degrees.

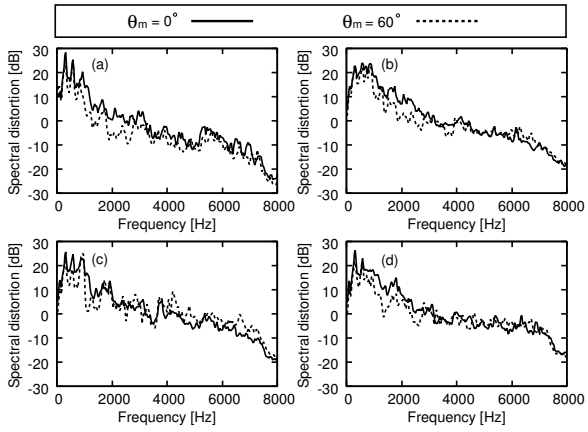

Fig. 10. Spectral distortion of internal noise (a) type 1, (b) type 2, (c) type 3, (d) type 4 for the cases of $\theta_m = 0$ degree and $\theta_m = 60$ degrees.

at $\theta_m = 60$ and β fixed at 5 for all internal noise types (Proposed 1). Compared with the conventional microphone structure (Conventional), we can achieve 15% (type 1), 15% (type 2), 8% (type 3) and 16% (type 4) improvements of the speech recognition result at $\theta_m = 60$ and optimized β (Proposed 2).

V. Conclusion

In this paper, we showed that the internal noise is always in-phase at the microphone array because they are transmitted through a solid (the robot chassis here). Then we proposed to replace the broadside array by an end-fire array for improving both noise estimation and speech recognition performances. This proposed approach is not limited to robot application and can be easily extended to car applications [13] because the road noise in car application is also transmitted through a solid (the car chassis).

TABLE II
GAIN FACTOR β OF WF

internal noise type number	$\{\beta_1, \beta_2\}$
1	{ 20, 30}
2	{ 30, 50}
3	{ 10, 20}
4	{ 10, 50}

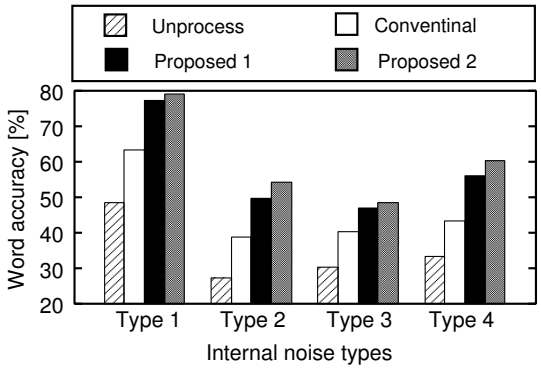

Fig. 11. Experimental results of word accuracy for the cases of the observed signal, $\theta_m = 0$ degree, $\theta_m = 60$ degrees with fixed $\beta = 5$ and $\theta_m = 60$ degrees with optimized β.

References

[1] J. Even, et al., "Frequency domain semi-blind signal separation: application to the rejection of internal noises," *Proc. International Conference on Acoustic Speech and Signal Processing,* pp. 157–160, 2008.

[2] P. Loizou, *Speech Enhancement: Theory and Practice,* CRC Press, 2007.

[3] P. Comon, "Independent component analysis, a new concept?," *Signal Processing,* vol.36, pp.287–314, 1994.

[4] M. Brandstein, et al., *Microphone Arrays Signal Processing Techniques and Applications,* Springer-Verlag, 2001.

[5] N. Vlassis, et al., "Efficient source adaptivity in independent analysis," *IEEE Trans. Neural Networks,* vol.12, no.3, pp. 559–566, 2001.

[6] J. Even, et al., "An improved permutation solver for blind signal separation based front-ends in robot audition," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems,* pp. 2172–2177, 2008.

[7] Y. Takahashi, et al., "Blind spatial subtraction array for noisy environment," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 17, no. 4, pp. 650–664, 2009.

[8] H. Saruwatari, et al., "Blind source separation combining independent component analysis and beamforming," *EURASIP J. Applied Signal Porc.,* vol.2003, no.11, pp. 1135–1146, 2003.

[9] J. Even, et al., "Semi-blind suppression of internal noise for hands-free robot spoken dialog system," *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems,* pp. 658–663, 2009.

[10] L. Rabiner, et al., *Fundamentals of speech recognition,* Upper Saddle River, NJ: Prentice Hall PTR, 1993.

[11] A. Lee, et al., "Julius? An open source realtime large vocabulary recognition engine," *Proc. Eur. Conf. Speech Commun. Technol.*, pp. 1691–1694, 2001.

[12] A. Lee, et al., "A new phonetic tied-mixture model for efficient decoding," *Proc. International Conference on Acoustics, Speech and Signal Processing* , pp. 1269–1272, 2000 .

[13] H. Saruwatari, et al., "Speech enhancement in car environment using blind source separation," *Proc. International Conference on Spoken Language Processing,* pp. 1781–1784, 2002.