Real-Time Semi-Blind Speech Extraction with Speaker Direction Tracking on Kinect

Yuji Onuma, Noriyoshi Kamado, Hiroshi Saruwatari, Kiyohiro Shikano
Nara Institute of Science and Technology,
Graduate School of Information Science,
Nara 630-0192, Japan
E-mail: yuji-o@is.naist.jp Tel: +81-743-72-5287

***Abstract*—In this paper, speech recognition accuracy improvement is addressed for ICA-based multichannel noise reduction in spoken-dialogue robot. First, to achieve high recognition accuracy for the early utterance of the target speaker, we introduce a new rapid ICA initialization method combining robot image information and a prestored initial separation filter bank. From this image information, an ICA initial filter fitted to the user's direction can be used to save the user's first utterance. Next, a new permutation solving method using a probability statistics model is proposed for realistic sound mixtures consisting of point-source speech and diffuse noise. We implement these methods using user tracking on Microsoft Kinect and evaluate it by speech recognition experiment in the real environment. The experimental results show that the proposed approaches can markedly improve the word recognition accuracy.**

I. Introduction

In a hands-free robot 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 accurate speech recognition because background environmental noises always degrade the target speech quality [1]. In this paper, speech recognition accuracy improvement is addressed for multichannel noise reduction in spoken-dialogue robot.

As a conventional noise reduction method, independent component analysis (ICA) [2] has been proposed. ICA is a well-known method that can be used to estimate environment noise components. Hence, one of the authors has proposed blind spatial subtraction array (BSSA) [3], which enhances target speech by subtracting an ICA-based noise estimate from noisy observations via spectral subtraction or Wiener filtering [4]. BSSA has been developed to be real-time processing [5], but there exist some inherent problems that (a) difficulty in removing the ambiguity of the source order, i.e., solve the permutation problem under diffuse noise conditions, and (b) BSSA's initial filter is always fixed to be, e.g., null beamformer (NBF) [6] steered to array normal. Therefore, the first utterance of the target user cannot be recognized.

In this paper, first, to achieve high recognition accuracy for the early utterance of the target speaker, we introduce a new rapid ICA initialization method combining image information and a prestored initial separation filter bank. To cope with the problem, we assume that the robot has its own video camera, and thus, the direction of the target user can be immediately detected. From this image information, the prestored ICA filter bank can be established with the user's direction tags, and an ICA initial filter fitted to the user's direction can be used to save the user's first utterance.

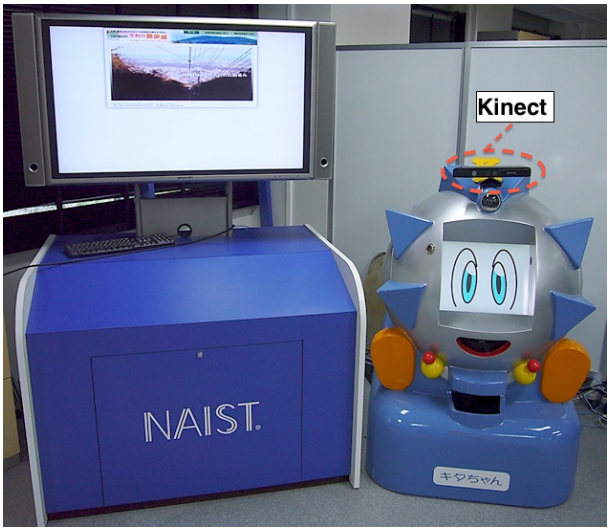

Fig. 1. Appearance of Kita-systems (Kita-chan and Kita-robo) with Kinect.

Next, a new permutation solving method using a probability statistics model is proposed. In this method, a shape difference between probability density functions of sources can cope with the permutation problem in realistic sound mixtures consisting of point-source speech and diffuse noise.

Finally, we implement the real-time target speech extraction system consisting of above-mentioned algorithms using user tracking ability of *Microsoft Kinect* (hereafter we call it Kinect) [7] and evaluate this system by speech recognition experiment in the real environment. This Kinect-based speech enhancement system is installed with our previously developed spoken dialog robot, *Kita-chan* [5], to construct a hands-free robot dialog system (see Fig.1). The experimental results in the real environment show that the proposed approaches can markedly improve the word recognition accuracy in the real-time ICA-based noise reduction developed in the robot dialogue system.

II. Related Work: Target Speech Extraction by BSSA

A. Noise Estimation by ICA

In this study, we assume a mixture of point-source target speech and diffuse noise, which typically arises in robot dialogue systems. It is known that ICA is proficient in noise estimation under the non-point-source noise condition [3]. The observed signal vector of the J-channel array in the time-frequency domain is given by

$$\boldsymbol{x}(f,\tau) = \boldsymbol{h}(f,\theta)s(f,\tau,\theta) + \boldsymbol{n}(f,\tau), \tag{1}$$

where f is the frequency bin, τ is the frame number, $\boldsymbol{x}(f,\tau) = [x_1(f,\tau),\ldots,x_J(f,\tau)]^{\mathrm{T}}$ is the observed signal vector, $\boldsymbol{h}(f,\theta) = [h_1(f,\theta),\ldots,h_J(f,\theta)]^{\mathrm{T}}$ is a column vector of transfer function from the target signal component to each microphone, $s(f,\tau,\theta)$ is a target speech signal component, and $\boldsymbol{n}(f,\tau) = [n_1(f,\tau),\ldots,n_J(f,\tau)]^{\mathrm{T}}$ is a column vector of the additive noise signal. In addition, θ is the direction-of-arrival (DOA) of the source.

In the ICA, we perform signal separation using a complex-valued matrix $\boldsymbol{W}_{\mathrm{ICA}}(f,\theta)$, so that the output signals $\boldsymbol{o}(f,\tau,\theta)$ become mutually independent; this procedure can be represented as

$$\begin{aligned}\boldsymbol{o}(f,\tau,\theta) &= [o_1(f,\tau,\theta),\ldots,o_K(f,\tau,\theta)]^{\mathrm{T}} \\ &= \boldsymbol{W}_{\mathrm{ICA}}(f,\theta)\boldsymbol{x}(f,\tau),\end{aligned} \tag{2}$$

where $\boldsymbol{o}(f,\tau,\theta)$ is the separated signal vector, K is the number of outputs, and $\boldsymbol{W}_{\mathrm{ICA}}(f,\theta)$ is the separation matrix for canceling out the signal coming from θ. The update rule of $\boldsymbol{W}_{\mathrm{ICA}}(f,\theta)$ is

$$\begin{aligned}\boldsymbol{W}_{\mathrm{ICA}}^{[p+1]}(f,\theta) = \mu\left[\boldsymbol{I} - \left\langle \boldsymbol{\varphi}\left(\boldsymbol{o}(f,\tau,\theta)\right)\boldsymbol{o}^{\mathrm{H}}(f,\tau,\theta)\right\rangle_{\tau}\right]\boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f,\theta) \\ + \boldsymbol{W}_{\mathrm{ICA}}^{[p]}(f,\theta),\end{aligned} \tag{3}$$

where μ is the step-size parameter, $[p]$ is used to express the value of the pth step in iterations, and $\boldsymbol{I}$ is the identity matrix. Moreover, $\langle\cdot\rangle_{\tau}$ denotes a time-averaging operator, $\boldsymbol{M}^{\mathrm{H}}$ denotes conjugate transpose of matrix $\boldsymbol{M}$, and $\boldsymbol{\varphi}(\cdot)$ is an appropriate nonlinear vector function.

Next, the estimated target speech signal is discarded as it is not required because we want to estimate only the noise component. Instead, assuming $o_U(f,\tau,\theta)$ as the speech component, we construct a *noise-only vector* $\boldsymbol{q}(f,\tau,\theta)$ as

$$\begin{aligned}\boldsymbol{q}(f,\tau,\theta) = [o_1(f,\tau,\theta),\ldots,o_{U-1}(f,\tau,\theta),0, \\ o_{U+1}(f,\tau,\theta),\ldots,o_K(f,\tau,\theta)]^{\mathrm{T}}.\end{aligned} \tag{4}$$

Following this, we apply the projection back operation to remove the ambiguity of amplitude, and we have

$$\begin{aligned}\hat{\boldsymbol{q}}(f,\tau,\theta) &= [\hat{q}_1(f,\tau,\theta),\ldots,\hat{q}_J(f,\tau,\theta)]^{\mathrm{T}} \\ &= \boldsymbol{W}_{\mathrm{ICA}}^{+}(f,\theta)\boldsymbol{q}(f,\tau,\theta),\end{aligned} \tag{5}$$

where $\boldsymbol{M}^{+}$ is the Moore-Penrose generalized inverse matrix of $\boldsymbol{M}$. Also, we should remove the ambiguity of the source order, i.e., solve the permutation problem. The solution of this permutation problem will be discussed in Sect. III-C.

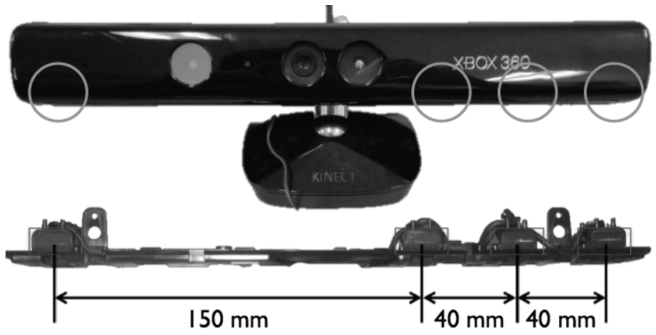

Fig. 2. Microphone array in Kinect.

B. Target Speech Extraction

In another path, we enhance the speech signal by delay-and-sum (DS) beamforming; the enhanced signal $y_{\mathrm{DS}}(f,\tau,\theta)$ is given by

$$y_{\mathrm{DS}}(f,\tau,\theta) = \boldsymbol{g}_{\mathrm{DS}}(f,\theta)^{\mathrm{T}}\boldsymbol{x}(f,\tau), \tag{6}$$

$$\boldsymbol{g}_{\mathrm{DS}}(f,\theta) = \left[g_1^{(\mathrm{DS})}(f,\theta),\ldots,g_J^{(\mathrm{DS})}(f,\theta)\right]^{\mathrm{T}}, \tag{7}$$

$$g_j^{(\mathrm{DS})}(f,\theta) = \frac{1}{J}\exp\left(-i2\pi\left(f/M\right)f_s d_j \sin\theta/c\right), \tag{8}$$

where $\boldsymbol{g}_{\mathrm{DS}}(f,\theta)$ is the coefficient of DS, f_{s} is the sampling frequency, d_j $(j = 1,\ldots,J)$ is the position of the microphones, M is the size of DFT, and c represents the sound velocity.

Finally, we apply generalized spectral subtraction (GSS) [8] as a post processing for target speech extraction, resulting in the target speech estimate $y_{\mathrm{BSSA}}(f,\tau,\theta)$ given by

$$\begin{aligned}&y_{\mathrm{BSSA}}(f,\tau,\theta) = \\ &\begin{cases}\sqrt[2n]{|y_{\mathrm{DS}}(f,\tau,\theta)|^{2n} - \beta\cdot|z_{\mathrm{ICA}}(f,\tau,\theta)|^{2n}} \\ \qquad\left(\text{if}\quad \left|y_{\mathrm{DS}}(f,\tau,\theta)\right|^{2n} - \beta\cdot\left|z_{\mathrm{ICA}}(f,\tau,\theta)\right|^{2n} \geq 0\right), \\ \gamma\cdot y_{\mathrm{DS}}(f,\tau,\theta) \qquad (\text{otherwise}) \quad ,\end{cases}\end{aligned} \tag{9}$$

where $2n$ is the exponent parameter in GSS, β is the subtraction parameter, γ is the flooring parameter and $z(f,\tau,\theta)$ is the estimated noise via DS, as

$$z(f,\tau,\theta) = \boldsymbol{g}_{\mathrm{DS}}^{\mathrm{T}}(f,\theta)\hat{\boldsymbol{q}}(f,\tau,\theta). \tag{10}$$

III. Real-Time Bssa with Speaker Tracking on Kinect

A. Outline of Kinect

In this paper, we introduce speaker position estimation by skeleton tracking on Kinect for DOA information of the real-time BSSA. Thus, we implement robot auditory interface to achieve high recognition accuracy for the early utterance of the target speaker.

Kinect is a multi-modal interface with RGB camera, depth sensor and microphone array, which enable to add abilities of motion capture and voice recognition to Microsoft Xbox 360 game machines. By using Kinect for Windows SDK [9], we can utilize the microphone array with PC as a USB audio device of the four-ch input. Also, human skeleton tracking

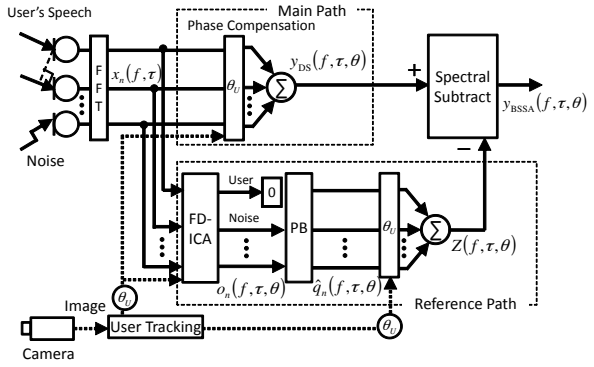

Fig. 3. Block diagram of BSSA with image information detection.

function has been provided to make possible to get information of the human head and each joint position instantly in millimeter accuracy by the SDK.

Figure 2 shows the microphone array of Kinect. It has four unidirectional microphones with an unequal interelement spacing. Signal (16kHz sampling frequency, 16bit quantization) from the microphone array is output to USB via a built-in preamplifier, an A/D converter and an audio stream controller.

B. Semi-Blind Speech Extraction with Video Information for Saving Early Utterance

Figure 3 shows a block diagram of the implemented system. In real-time BSSA, it is possible to process SS and DS in real-time. However, owing to the huge computational load, ICA cannot provide any effective separation matrix in real-time, especially for the first segment of user utterance. This causes a serious problem of poor recognition accuracy for the first word that often conveys important messages of the user, such as commands, user decision, and greetings.

To improve the speech recognition accuracy for the early utterances, we introduce a rapid ICA initialization approach combining video image information given by robot (Kinect's) eyes and a prestored initial separation filter bank. If tags of the prestored ICA filters in the filter bank do not coincide with the current user's DOA detected by image information, the system moves to the training phase; a new separation filter is updated by ICA and added to the filter bank with the detected-DOA tag. Otherwise, if the system can find the DOA tag that corresponds to the current user's DOA, the immediate detection of the user's DOA enables ICA to choose the *most appropriate* prestored separation matrix for saving the user's early utterances. Figure 4 and the following show the flow of processing.

Step 1: Construction of ICA filter bank

In this algorithm, we quantize the user's DOA for the ICA filter bank into five directions with 11.7°-wise, which divide angle of view of Kinect into five sections, where 0° is normal to the robot face. System holds a filter bank of ICA separation filters learned enough in all directions. If there is not the filter that has been learned, system set NBF of front direction to ICA initial value.

Step 2: User's DOA detection by Kinect user tracking

To estimate the direction of the user in the angle of view instantly, we use the skeleton tracking of Kinect. The skeleton tracking can detect the user's direction and our algorithm then quantize it into five DOAs (see Fig. 5). This calculation can be processed within about 50 ms, which is immediate enough to detect the user in advance of the utterance. The nearest DOA θ of the user's DOA estimated via Kinect is used as a tag of $\boldsymbol{W}_{\mathrm{ICA}}(f,\theta)$, which has already been iteratively optimized.

Step 3: Enhancement of early utterance

Based on the user direction θ obtained by STEP 2, load the ICA filter $\boldsymbol{W}_{\mathrm{ICA}}(f,\theta)$ from the filter bank. Using this, noise suppression processing is performed by BSSA for the current audio input, and outputs the target speech signal.

Step 4: Enhancement after 2nd-block utterances

After 2nd-block utterances of the user, we perform ICA and speech enhancement processing as described in Sect. II in a blockwise batch manner (see Fig. 6). When the learning of the filter is end, the current $\boldsymbol{W}_{\mathrm{ICA}}(f,\theta)$ is overwritten in the filter bank for preparing the next future user. Then, go back to Step 2.

C. Permutation Solver

In the context of the permutation problem in the ICA study, there exist many methods for solving permutation, such as source-DOA-based method [6] and its combination with subband correlation [10]. These approaches, however, often fail to solve the permutation problem in real environments because of the inherent problem that noise is sometimes diffuse and has no discriminable property on DOA. Therefore, in this paper, we introduce a new cue for the classification of sources, i.e., source statistics difference, instead of DOA information.

We assume that the separated signal by ICA, o_k, can be modeled using the gamma distribution as

$$P(o_k) = \frac{o_k^{\alpha-1}\exp\left(-\frac{o_k}{\sigma}\right)}{\sigma^{\alpha}\Gamma(\alpha)}, \tag{11}$$

where k is the index of channel of the separated signal by ICA, α is the shape parameter corresponding to the type of sources (e.g., $\alpha=1$ is Gaussian and $\alpha<1$ is super-Gaussian), σ is the scale parameter of the gamma distribution, and $\Gamma(\alpha)$ is the *gamma function*, defined as

$$\Gamma(\alpha) = \int_0^{\infty} t^{\alpha-1}\exp(-t)dt. \tag{12}$$

Modeling by the gamma distribution enables us to easily estimate the scale parameter σ and shape parameter α from observed raw data samples. These parameters can be estimated by the maximum likelihood estimation method as

$$\hat{\alpha}(o_k) = \frac{3-\omega+\sqrt{(\omega-3)^2+24\omega}}{12\omega}, \tag{13}$$

$$\hat{\sigma}(o_k) = \frac{\mathrm{E}[o_k]}{\hat{\alpha}(o_k)}, \tag{14}$$

$$\omega = \log\left(\mathrm{E}\left[o_k\right]\right) - \mathrm{E}\left[\log o_k\right]. \tag{15}$$

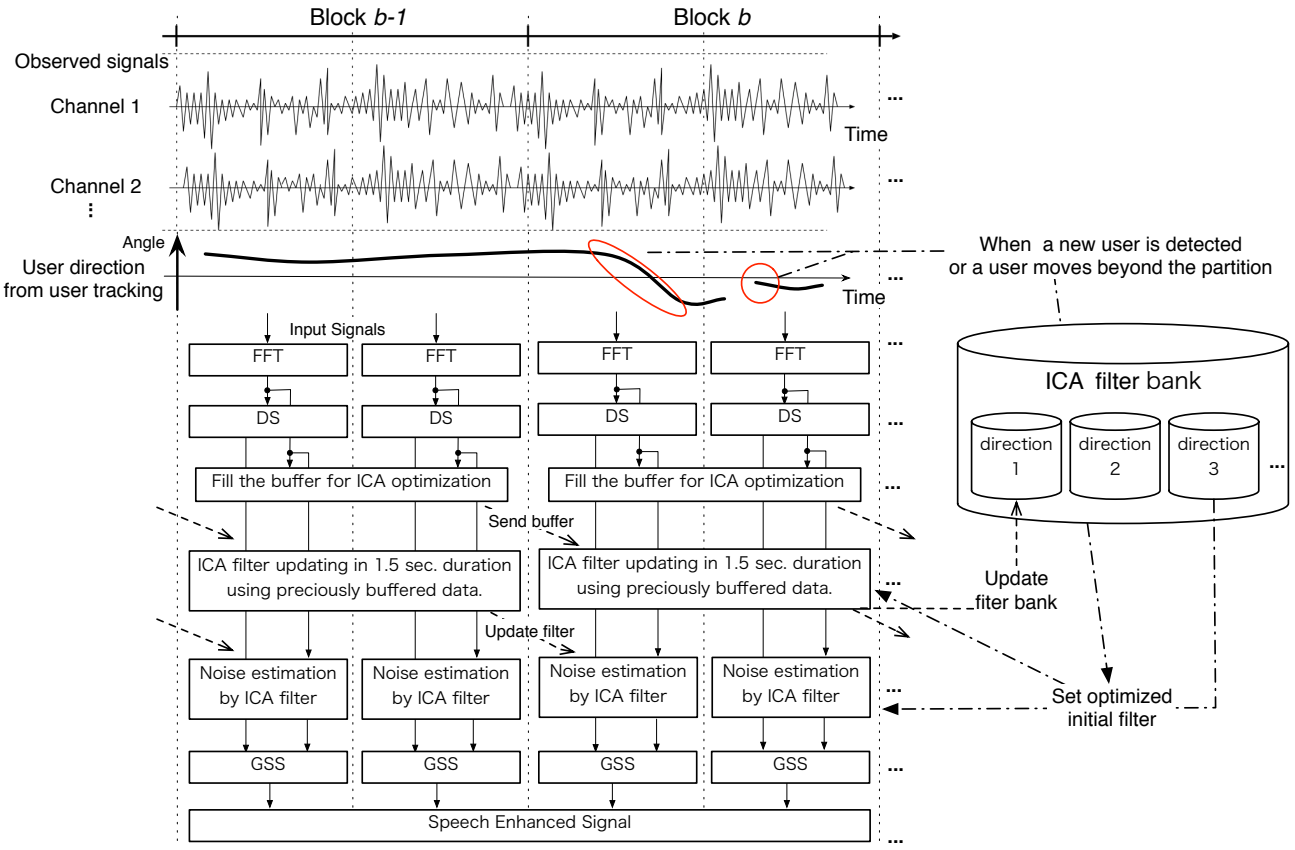

Fig. 4. Signal flow in real-time implementation of proposed system using user tracking.

For example, it is well known that the subband power spectral component of speech is well modeled by very spiky distribution as (11) with $\alpha = 0.1 \sim 0.2$, but that of diffuse noise is typically more Gaussian (i.e., $\alpha \simeq 1$) owing to the *center limit theorem*. Therefore, we can discriminate the sources and solve the permutation problem using this difference in statistical property.

Finally, in construction of noise-only vector in (4), we can set the user speech component index U as

$$U = \underset{k}{\operatorname{argmin}} \ \hat{\alpha}\left(o_k\left(f, \tau, \theta\right)\right). \tag{16}$$

Thus, the component with the smallest $\hat{\alpha}(o_k)$ is regarded as speech, which should be cancelled out in noise estimation.

D. Implementation

To process noise suppression with robot audiovisual information, Kinect is connected to the PC via USB, where the system was built on the PC using the Microsoft Visual C++ 2010 (8 GB of memory and CPU of 1.86 GHz Core i7). In this implementation, the STFT frame length is 512 points, the frame shift size is 128 points, and the length of the signal analysis window size for the separation filter update by ICA is set to 256 points.

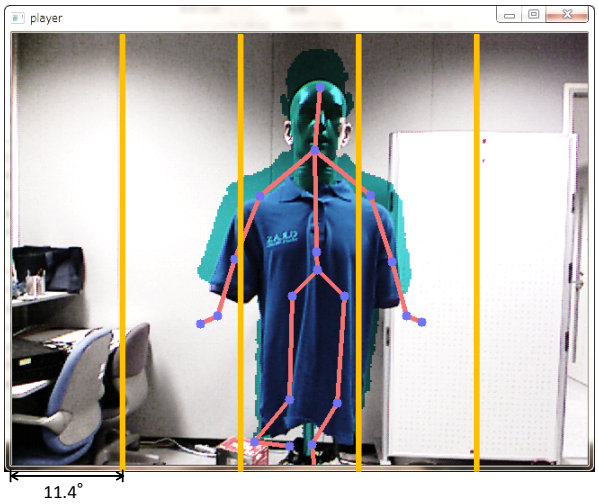

Fig. 5. DOA estimation by skeleton tracking on Kinect.

IV. Speech Recognition Experiment

A. Experimental Conditions

We conduct a speech recognition experiment to evaluate the effectiveness of the proposed method, where we compare the

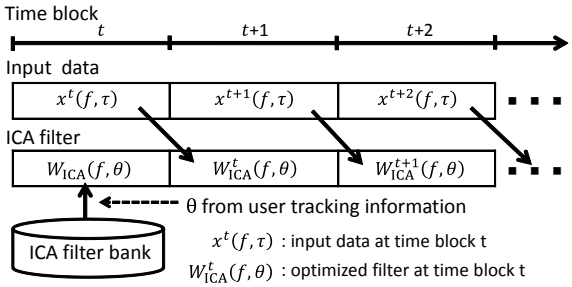

Fig. 6. Configuration of updating ICA separation filter in real-time BSSA with image information.

TABLE I
EXPERIMENTAL CONDITIONS FOR SPEECH RECOGNITION

Database	JNAS [11], 46 speakers (200 sentences)
Speech recognition task	20k words newspaper dictation
Acoustic model	Phonetic tied mixture (PTM), clean model
Number of training speakers for acoustic model	JNAS, 260 speakers (150 sentences / 1 speaker)
Decoder	Julius ver. 4.2 [11]

following four speech enhancement systems.

- Non noise suppression **(Unprocessed)**.
- DS beamforming **(DS)**.
- Conventional BSSA initialized by 0° steered NBF with conventional DOA-based permutation solver [6] **(Conventional)**.
- Proposed method **(Proposed)**.

Experimental environment is shown in Figure 7. Reverberation time is 200 ms. We emitted JNAS [11] speech database from a dummy head as a target user's voice and real railway-station noise from eight loudspeakers surrounding the Kinect. SNR of speech and noise is set to 10 dB and 15 dB. Speaker direction θ is changed in three conditions, namely, 0, 10, and 20 degrees. The rest of the experimental conditions is summarized in Table I. In GSS, the exponent parameter is 2.0, subtraction parameter is 1.4, and flooring parameter is 0.2; these values are determined experimentally. Note that this experiment was conducted on the real-time processing basis (not batch processing). The total latency of this processing is about 50 ms.

B. Results

Figures 8–10 show the results of speech recognition experiments. From these figures, we can confirm that the proposed method significantly outperforms the conventional methods in all cases of user directions. In particular, although the proposed method is aimed to only save the user's early utterance, total word accuracy for the full sentence can be considerably improved; this indicates the importance of the recognition of the first utterance. In the conventional method, recognition rate is markedly low if the user direction is shifted from 0° because the filter did not converge enough. In contrast, the recognition rate of the proposed method is maintained to be high because

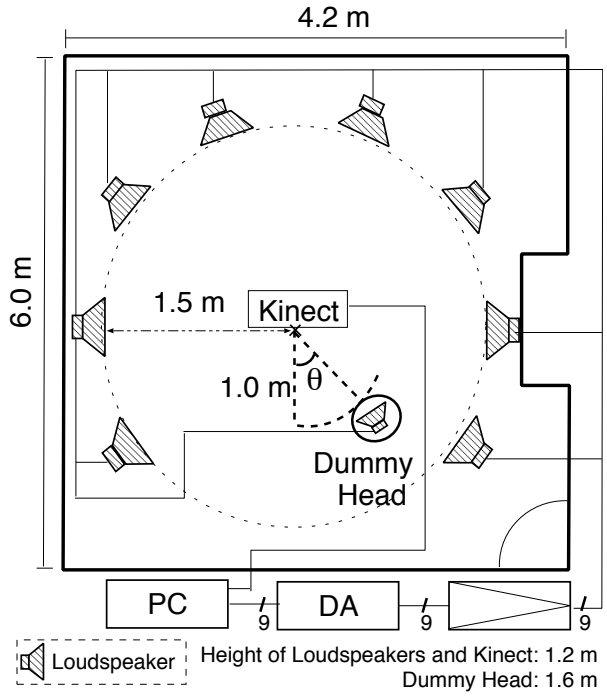

Fig. 7. Acoustical environment used in real-world experiment.

the optimal filter can be rapidly available with the assistance of image information.

V. CONCLUSIONS

In this paper, first, to save the user's first utterance, we introduce a new rapid ICA initialization method combining robot video information and a prestored initial separation filter bank. Next, a signal-statistics-based permutation solving approach is proposed. The experimental results show that the proposed real-time ICA-based noise reduction developed in Kinect can markedly improve the word recognition accuracy in the robot dialogue system.

ACKNOWLEDGEMENT

This work was partly supported by JST Core Research of Evolutional Science and Technology (CREST), Japan.

REFERENCES

[1] R. Prasad, et al., "Robots that can hear, understand and talk," *A*dvanced Robotics, vol.18, pp.533–564, 2004.
[2] P. Comon, "Independent component analysis, a new concept," *Signal processing,* vol.36, pp.287–314, 1994.
[3] Y. Takahashi, H. Saruwatari, et al., "Blind spatial subtraction array for noisy environment," *IEEE Trans. Audio, Speech, and Language Processing*, vol.17, no.4, pp.650–664, 2009.
[4] P. Loizou, *Speech Enhancement: Theory and Practice,* CRC Press, 2007.
[5] H. Saruwatari, et al., "Hands-free speech recognition challenge for real-world speech dialogue systems," *Proc. ICASSP*, pp.3729–3782, 2009.
[6] H. Saruwatari, et al., "Blind source separation combining independent component analysis and beamforming," *EURASIP Journal on Applied Signal Processing*, vol.2003, no.11, pp.1135–1146, 2003.
[7] Microsoft, "Kinect - Xbox.com," http://www.xbox.com/ja-JP/kinect
[8] B. L. Sim, et al., "A parametric formulation of the generalized spectral subtraction method," *IEEE Trans. Speech and Audio Processing*, vol.6, no.4, pp.328–337, 1998.

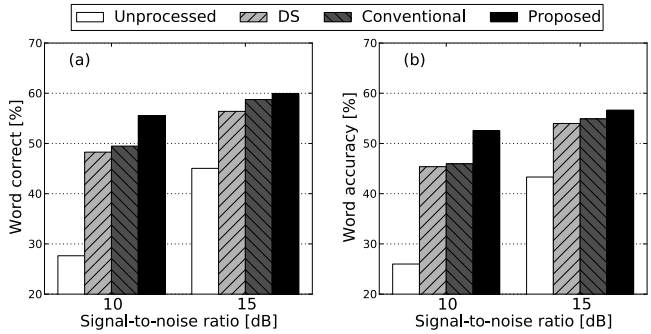

Fig. 8. Result of speech recognition test in real-world experiment, where speaker direction is 0 degree. (a) word correct, and (b) word accuracy.

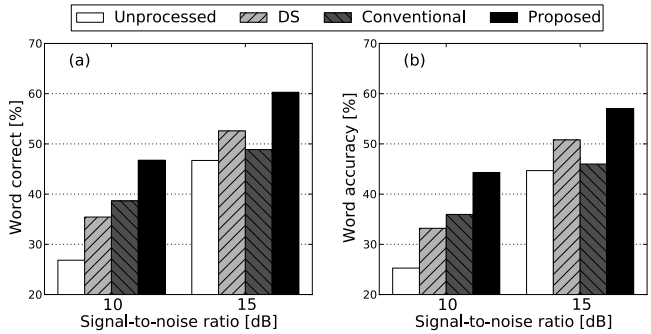

Fig. 9. Result of speech recognition test in real-world experiment, where speaker direction is 10 degree. (a) word correct, and (b) word accuracy.

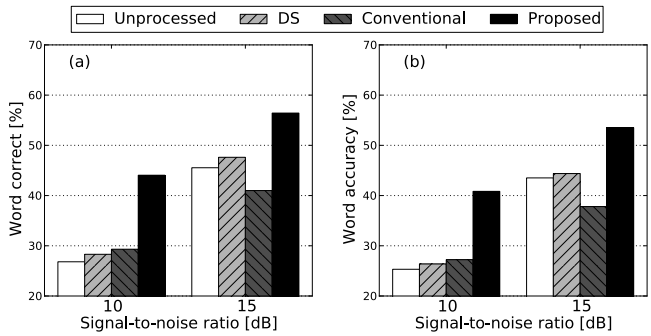

Fig. 10. Result of speech recognition test in real-world experiment, where speaker direction is 20 degree. (a) word correct, and (b) word accuracy.

[9] Microsoft, "Microsoft Kinect SDK for Developers| Develop for the Kinect | Kinect for Windows," http://kinectforwindows.org/

[10] H. Sawada, et al., "A robust and precise method for solving the permutation problem of frequency-domain blind source separation," *IEEE Transactions on Speech and Audio Processing*, vol.12, no.5, pp.530–538, 2004.

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