2010 IEEE International Conference on Robotics and Automation
Anchorage Convention District
May 3-8, 2010, Anchorage, Alaska, USA

Upper-limit Evaluation of Robot Audition based on ICA-BSS in Multi-source, Barge-in and Highly Reverberant Conditions

Ryu Takeda, Kazuhiro Nakadai, Toru Takahashi, Kazunori Komatani, Tetsuya Ogata and Hiroshi G. Okuno

Abstract— **This paper presents the upper-limit evaluation of robot audition based on ICA-BSS in multi-source, barge-in and highly reverberant conditions. The goal is that the robot can automatically distinguish a target speech from its own speech and other sound sources in a reverberant environment. We focus on the multi-channel semi-blind ICA (MCSB-ICA), which is one of the sound source separation methods with a microphone array, to achieve such an audition system because it can separate sound source signals including reverberations with few assumptions on environments. The evaluation of MCSB-ICA has been limited to robot's speech separation and reverberation separation. In this paper, we evaluate MCSB-ICA extensively by applying it to multi-source separation problems under common reverberant environments. Experimental results prove that MCSB-ICA outperforms conventional ICA by 30 points in automatic speech recognition performance.**

I. INTRODUCTION

A. Background

Our goal is the development of a robot that can extract a user's speech from a mixture of sounds and can interact with humans naturally through speech in various environments. For example, a robot can talk with a target user a near loud television; a person may talk to it from far away; a user can interrupt a robot's utterance and begin speaking while the robot is speaking (called "barge-in"). Since speech is the most natural communication channel for human, such robots are useful and will help us in many situations, such as in housekeeping or rescue tasks. To achieve such a robot audition system, we must cope with the following three problems at the same time,

1) multi-source (speech and other noise) signals,
2) the robot's own speech signal, and
3) the reverberations of them.

These problems are posed by the microphones which are installed on its body, and not attached close to the user's mouth (Fig. 1). This degrades the performance of conventional automatic speech recognition (ASR) seriously because many ASRs or spoken dialogue systems work well in the laboratory but not in such noisy and reverberant environments. Additionally, robots must have **the least prior information** about the environment and should be adaptive to environments to work even in unknown environments. Therefore, we can say that robot audition is challenging research.

R. Takeda, T. Takahashi, K. Komatani, T. Ogata, and H. G. Okuno are with the Department of Intelligence Science and Technology. Graduate School of Informatics, Kyoto University, Kyoto, 606-8501, Japan. {rtakeda, tall, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp
K. Nakadai is with Honda Research Institute Japan Co., Ltd., Wako, Saitama, 351-0114, Japan. nakadai@jp.honda-ri.com

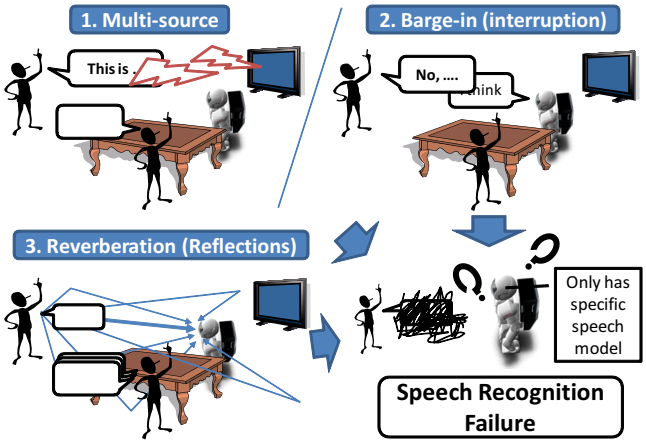

Fig. 1. Our target problems

B. Previous Works

1) Robot Audition Aspects: Most robot audition research does not tackle the reverberation and barge-in problems although they cover a broad range of topics, such as sound source separation (SSS), sound source localization (SSL), tracking (SST), integration SSS with ASR and real-time processing.

HARK is open source software for robot audition [1], and it can achieve a real-time SSS, SST and ASR. However, it also can not deal with reverberation and barge-in problems. The robot spoken dialogue system [2] guides at a railway station, and it accomplished a good ASR performance in a real station. This system also is not evaluated in a reverberant environment or in a situation where a speaker is far away from microphones.

2) Signal Processing Aspects: The problems that we tackle are categorized in blind source separation (BSS), blind dereverberation (BD, separation of reverberation) or multi-channel blind deconvolution (MBD), and echo cancellation (EC, separation of known source), respectively. Here, *Blind* means that a method only uses the observed and simple prior information. BSS or MBD problems under high reverberant environments are especially hot topics now. We pick up microphone array methods because it is easy to treat multi-source situations unlike missing data techniques [3] or others [4] that need some prior knowledge or models about sound sources and environment.

Previously, those problems, such as BSS and EC, have been treated as an isolated problem; some of them have not dealt with EC [5], others have not dealt with BD

[6], or have not been able to deal with BD or MBD [7]. Recently, integrated methods have been proposed, which can solve these problems at the same time. For example, Yoshioka integrates these techniques, such as BSS and BD, with maximum likelihood framework [8]. We also proposed a method that can solve BD and EC with Independent Component Analysis (ICA) framework [9]. Note that these integrated methods have not been discussed enough, and we can not conclude yet which one is the best way for robot audition.

C. Our Approach and Contribution

We adopted multi-channel semi-blind ICA (MCSB-ICA) [9], which we have proposed before for BD and EC, and which is categorized in a higher order statistics ICA. The reasons why we use it are that

1) it is theoretically robust against Gaussian noise, such as that from fans,
2) it can theoretically deal with MBD and EC with the linear order calculation cost of reverberation time, and
3) there is a lot of knowledge regarding real application, such as second order solution [10] and source model estimation [2].

The performance of MCSB-ICA has not been evaluated although its framework can deal with BSS or MBD as mentioned above. Because of the generality of the ICA framework, there is no need to extend the separation algorithm to deal with multi-source situation.

In this paper, we reveal the performance of MCSB-ICA as MBD and EC method

1) by using a large vocabulary and continuous ASR (not isolated word recognition with small size dictionary)
2) with many microphones embedded in a robot's head
3) under multi-source and reverberant situation.

And we compared it with the conventional frequency-domain ICA (FD-ICA) [7], which is used as a baseline method in [8]. We believe that these is no paper which evaluates a method under these conditions. Especially, ICA originally assumes that the number of microphones and sound sources are equal. In real applications, it is difficult to estimate the number of sound sources in advance, and we can use more microphones. Then, it is beneficial to show whether the MCSB-ICA works well or not with so many microphones. Note that we **ignore** the problems of *permutation* and *calculation cost* to evaluate the pure performance of MCSB-ICA. The latter problem will be solved by parallel processing or the development of a CPU in the future although the former problem is essential and we treat it as another problem as mentioned in the following section.

D. Paper organization

This paper consists of 7 sections. Section 2 explains the MCSB-ICA. Section 3 explains the conventional FD-ICA, and Section 4 shows other configuration of ICA. and discusses evaluations of our method in Section 5 and 6. The last section concludes the paper and discusses future work.

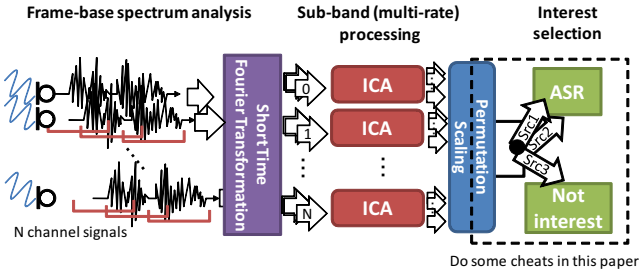

Fig. 2. STFT domain processing and overview of signal processing

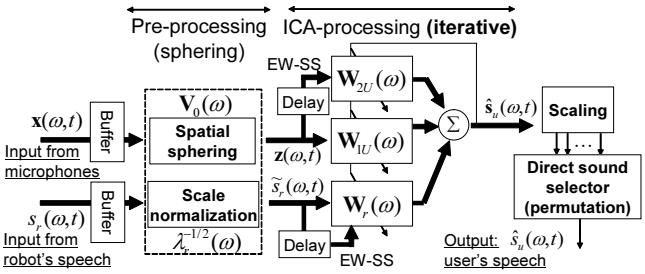

Fig. 3. Signal flow of MCSB-ICA

II. Multi-channel Semi-blind ICA

This section explains MCSB-ICA [9]. The MCSB-ICA model described here uses a short-time Fourier transformation (STFT) representation [5], which is a form of multi-rate processing (Fig. 2). We denote the spectrum after STFT as $s(\omega, t)$ at frequency ω and frame t. For the sake of simplicity, we have skipped denoting the frequency index, ω. The signal flow of MCSB-ICA is illustrated in Figure 3. We explain how the filter is estimated in this section.

A. Observation and Separation Model

We denote the spectra observed at microphones $M_1, \ldots, M_L$ as $x_1(t), \ldots, x_L(t)$ (L is the number of microphones) and its vector form as $\boldsymbol{x}(t) = [x_1(t), x_2(t), \ldots, x_L(t)]^T$. With the spectrum of the user's utterance, $s_u(t)$, and a known-source (robot's) spectrum, $s_r(t)$, the observed signals, $\boldsymbol{x}(t)$, can be described as a finite impulse response (FIR) filter model:

$$\boldsymbol{x}(t) = \sum_{n=0}^{N} \boldsymbol{h}_u(n) s_u(t-n) + \sum_{m=0}^{M} \boldsymbol{h}_r(m) s_r(t-n), \quad (1)$$

where $\boldsymbol{h}_u(n)$ and $\boldsymbol{h}_r(m)$ correspond to the N- and M-dimensional FIR coefficient vectors of the user's and known-source spectra.

Before explaining the MCSB-ICA separation model, let us define the observed vector, $\boldsymbol{X}(t)$, and the known-source vector, $\boldsymbol{S}_r(t)$:

$$\boldsymbol{X}(t) = [\boldsymbol{x}(t), \boldsymbol{x}(t-1), \ldots, \boldsymbol{x}(t-N)]^T \text{and} \quad (2)$$
$$\boldsymbol{S}_r(t) = [s_r(t), s_r(t-1), \ldots, s_r(t-M)]^T. \quad (3)$$

The separation model for MCSB-ICA is set so that the direct sound frame of a user's speech, $s_u(t)$, is independent

of the delayed-observed and known sound spectra, $\boldsymbol{X}(t-d)$ and $\boldsymbol{S}_r(t)$. Here, d (> 0) is an initial-reflection interval parameter, and we consider the dependence between the direct and adjacent frames of $s_u(t)$. The separation model is written as

$$\begin{pmatrix} \hat{\boldsymbol{s}}(t) \\ \boldsymbol{X}(t-d) \\ \boldsymbol{S}_r(t) \end{pmatrix} = \begin{pmatrix} \boldsymbol{W}_{1u} & \boldsymbol{W}_{2u} & \boldsymbol{W}_r \\ \boldsymbol{0} & \boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_r \end{pmatrix} \begin{pmatrix} \boldsymbol{x}(t) \\ \boldsymbol{X}(t-d) \\ \boldsymbol{S}_r(t) \end{pmatrix}, \quad (4)$$

where $\hat{\boldsymbol{s}}(t)$ is an estimated signal vector with an L dimension, and $\boldsymbol{W}_{1u}$ and $\boldsymbol{W}_{2u}$ correspond to $L \times L$ and $L \times L(N+1)$ blind separation and blind dereverberation matrices, respectively. $\boldsymbol{W}_r$ is the $L \times (M+1)$ echo cancellation separation matrix. $\boldsymbol{I}_2$ and $\boldsymbol{I}_r$ correspond to optimally-sized unit matrices.

Note that the estimated signal, $\hat{\boldsymbol{s}}(t)$, includes the direct sound signals, some reflected signals and uncertain independent signals because we assume the number of microphones is larger than that of sound sources. Then, we must select the target speech signals from the output signals by using some criteria. In this paper, **we select required signals by using reference signals** to evaluate the upper-limit of MCSB-ICA.

B. Estimation of Filter Parameters

The filter parameter set, $\boldsymbol{W} = \{\boldsymbol{W}_{1u}, \boldsymbol{W}_{2u}, \boldsymbol{W}_r\}$, is estimated by minimizing the Kullback-Leibler divergence between the joint probability density function (PDF) and the products of the marginal PDF of $\boldsymbol{s}(t)$, $\boldsymbol{X}(t-d)$ and $\boldsymbol{S}_r(t)$.

We obtain the following iterative update rules for $\boldsymbol{W}$ with a natural gradient method [11].

$$\boldsymbol{D} = \boldsymbol{\Lambda} - \mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\hat{\boldsymbol{s}}^H(t)], \quad (5)$$

$$\boldsymbol{W}_{1u}^{[j+1]} = \boldsymbol{W}_{1u}^{[j]} + \mu \boldsymbol{D}\boldsymbol{W}_{1u}^{[j]}, \quad (6)$$

$$\boldsymbol{W}_{2u}^{[j+1]} = \boldsymbol{W}_{2u}^{[j]} + \mu \left(\boldsymbol{D}\boldsymbol{W}_{2u}^{[j]} - \mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\boldsymbol{X}^H(t-d)] \right), \quad (7)$$

$$\boldsymbol{W}_{r}^{[j+1]} = \boldsymbol{W}_{r}^{[j]} + \mu \left(\boldsymbol{D}\boldsymbol{W}_{r}^{[j]} - \mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\boldsymbol{S}_r^H(t)] \right), \quad (8)$$

where $\cdot^H$ denotes the conjugate transpose operation, and $\boldsymbol{\Lambda}$ is a non-holonomic constraint matrix, i.e., $\mathrm{diag}(\mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\hat{\boldsymbol{s}}^H(t)])$ [12]. The μ is a step-size parameter, and $\boldsymbol{\phi}(\boldsymbol{x})$ is a non-linear function vector, $[\phi(x_1), \cdots, \phi(x_L)]^H$. $\phi(x)$ is defined as,

$$\phi(x) = -\frac{\mathrm{d} \log p(x)}{\mathrm{d}x}. \quad (9)$$

We assume that the source PDF is a noise-robust one $p(x) = \exp(-|x|/\sigma^2)/(2\sigma^2)$ with variance σ^2, and that $\phi(x)$ equals $x^*/(2\sigma^2|x|)$, where x^* denotes the conjugate of x. The two functions are defined in a continuous area, $|x| > \varepsilon$.

For pre-processing, we use enforced spatial sphering, which is an approximation of sphering. The observed signal, $\boldsymbol{X}(t)$, and the known signal, $\boldsymbol{S}_r(t)$, are transformed using two rules:

$$\boldsymbol{z}(t) = \boldsymbol{V}_u \boldsymbol{x}(t), \quad \boldsymbol{V}_u = \boldsymbol{E}_u \boldsymbol{\Lambda}_u^{-1/2} \boldsymbol{E}_u^H, \quad (10)$$

$$\tilde{s}_r(t) = \lambda_r^{-1/2} s_r(t), \quad (11)$$

where $\boldsymbol{E}_u$ and $\boldsymbol{\Lambda}_u$ are the eigenvector matrix and eigenvalue diagonal matrix of $\boldsymbol{R}_u = \mathrm{E}[\boldsymbol{x}(t)\boldsymbol{x}^H(t)]$. After sphering, $\boldsymbol{x}$ and s_r in Equations (4) – (8) are substituted into $\boldsymbol{z}$ and $\tilde{s}_r$.

III. FD-ICA: BASELINE METHOD

The FD-ICA is one of the most popular BSS method for real acoustic signal separation because of its fast convergence speed and good performance. The conventional FD-ICA, such as [7], is considered as a sub-set of MCSB-ICA in the formulation. If we set the filter length N and M to 0, the algorithm of MCSB-ICA reduces to that of FD-ICA.

The difference of FD-ICA and MCSB-ICA is whether the reverberation is considered or not. This extends the separation ability of FD-ICA, but it also increases the computational cost. Since the cost of MCSB-ICA is $O(L^2(M+N+2))$, the disadvantage will be overcome in the near future if L is small and N, M are not so huge. We do not discuss this disadvantage in this paper.

IV. COMMON CONSIDERATIONS

The FD-ICA and MCSB-ICA have problems called *scaling* and *permutation* problems. They are caused by the property of ICA which can not decide the amplitude and the permutation of output signals at all frequency bins. In this section, we explain these and other configurations of ICA, except the permutation problem. As mentioned in the previous section, we aligned output signals by using reference signals to solve the permutation problem.

1) Scaling: We used the projection back method [13]. We multiplied the i-th row and j-th column element c_j of $\hat{\boldsymbol{H}}_u = (\boldsymbol{W}_{1u}\boldsymbol{V}_0)^{-1}$, which satisfies the following equation for the scaling of the j-th element of $\hat{\boldsymbol{s}}_u(t)$.

$$l_j = \arg\max_l |\hat{\boldsymbol{H}}_u(l,j))| \quad (12)$$

$$c_j = \hat{\boldsymbol{H}}_u(l_j, j) \quad (13)$$

2) Initial value of separation matrix: The initial value of the separation matrix at the frequency ω, $\boldsymbol{W}_{1u}(\omega)$, was set to that of the estimated matrix at frequency $\omega+1$, $\boldsymbol{W}_{1u}(\omega+1)$. We used the unit matrix for the initial value of the first separation matrix. Empirically, the performance of MCSB-ICA degrades if we use a unit matrix as an initial value of the separation matrix for every frequency bin.

3) Step-size scheduling: To accelerate the convergence speed, we used a partial adaptive step-size method[14].

V. EXPERIMENTS

We evaluated MCSB-ICA by ASR performance with simulated data. The data is generated by using the impulse responses recorded in a real environment, and impulse responses can reconstruct the acoustic property, such as reflections. Note that robot's noises are not recorded and not used.

A. Experimental Settings

1) Data for evaluation: The impulse responses for speech data were recorded at 16 kHz in a reverberant room, whose RT_{20} is about 940 [ms]. Here, RT_{20} means the reverberation time. The size of the room was 4.8 m × 5.55 m × about 3 m (depth x width x height). The target speaker was 1.0 m apart from a microphone mounted on the head of Honda ASIMO.

TABLE I

CONFIGURATION FOR DATA AND SEPARATION

Impulse response	16 kHz sampling
Reverberation time (RT_{20})	940 [ms]
Direction of speaker B	10°, 20°, 30°, 60°, 90°
Number of microphones	Eight (embedded in ASIMO's head)
STFT analysis	Hanning: 32 [ms] and shift: 10 [ms]
Input wave data	[-1.0 1.0] normalized

TABLE II

CONFIGURATION FOR SPEECH RECOGNITION

Test set	200 sentences
Training set	200 people (150 sentences each)
Acoustic model	PTM-Triphone: 3-state, HMM
Language model	Statistical, vocabulary size of 21 k
Speech analysis	Hanning: 32 [ms] and shift: 10 [ms]
Features	MFCC 25 dim. (12+Δ12+ΔPow)

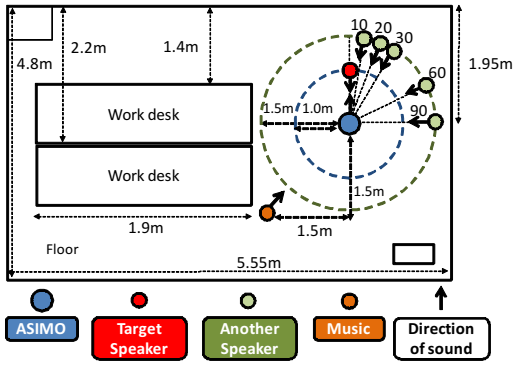

Fig. 4. The layout of room, robot, speaker, and music signal.

The noise speaker was located 1.5 m apart from ASIMO, and the angles between the noise speaker and the front of ASIMO were six patterns of 10, 20, 30, 60, 90 degrees. Music noise was at 2.12 m away from ASIMO coming from 315 degrees. We also recorded the impulse response from the robot's speech. The height of the microphones was 1.25 m and that of the other speakers were 1.4 m. These settings are illustrated in Figure 4. The relative amplitude among these impulse responses are saved. All data (16 bits, PCM) were normalized to $[-1.0 \;\; 1.0]$ for processing.

2) Separation parameters: The STFT parameters were set the same for all three experiments: the window size was 512 points (32 ms) and the shift-size was 160 points (8 [ms]). The frame interval parameter d was 2, and the filter lengths of echo cancellation and dereverberation was the same, that is, $N = M = 32$. The parameters for adaptive step-size control were set as a previous [14]. The number of iteration at the filter estimation was 25, and it is not a huge iteration number. To estimate the separation matrices, we used all observed data (off-line and batch processing). Note that the voice active section is given in our experiments.

3) ASR configuration: We used 200 Japanese sentences for the speaker's and robot's speech, and they were convoluted in the corresponding recorded impulse responses. Julius[1] was used for HMM-based ASR with the statistical language model. Mel-frequency cepstral coefficients (MFCC) (12+Δ12+ΔPow) were obtained after STFT with a window size of 512 points and a shift size of 160 points for the speech features, and we then applied Cepstral Mean Normalization. A triphone-based acoustic model (three-state and four-mixture) was trained with 150 sentences of clean speech uttered by 200 male and female speakers (word-closed). The statistical language model consisted of 21,000 words, which were extracted from newspapers. The other experimental conditions are summarized in Tables I and II.

[1]http://julius.sourceforge.jp/

Fig. 5. The microphone layout on ASIMO's head. The eight microphones forms like circle.

B. Combination of Noise Patterns and Evaluation Criteria

We compared the performance of MCSB-ICA with that of FD-ICA under the following noise combinations:

1) a target speech and a robot speech (and a music noise) for a barge-in situation,
2) a target speech and a noise speech (and a music noise) for a simultaneous-talk situation, and
3) a target speech, a noise speech and a robot speech (and a music noise) for a worst situation.

We can cluster the patterns with or without background music noise. Note that a maximum of four sources exists at the same time.

The performances are measured by word correctness (Cor.) and word accuracy (Acc.) of the target speech (Target sp.) and noise speech (Another sp.). The correctness and accuracy are defined by the following equation,

$$\text{Cor.} = \frac{\text{\# of correct words}}{\text{\# of all words}}, \tag{14}$$

$$\text{Acc.} = \frac{\text{\# of correct words} - \text{\# of inserted words}}{\text{\# of all words}}. \tag{15}$$

Cor. increases if the words in the sentence are recognized without missing words, and Acc. increases if the irrelevant words are not recognized.

VI. DISCUSSION

A. Results

Figure 6 shows the wave signal and spectrum examples of the clean speech signal, a reverberant one, an observed signal and two separated speech signals. The top left is a clean speech and the top right is a reverberant one. We can see the belt-like effect of reverberation from the spectra. The middle is an observed signal, the bottom left is the result of MCSB-ICA and the bottom right is that of FD-ICA. Obviously, MCSB-ICA separates the target speech from other sources better than FD-ICA.

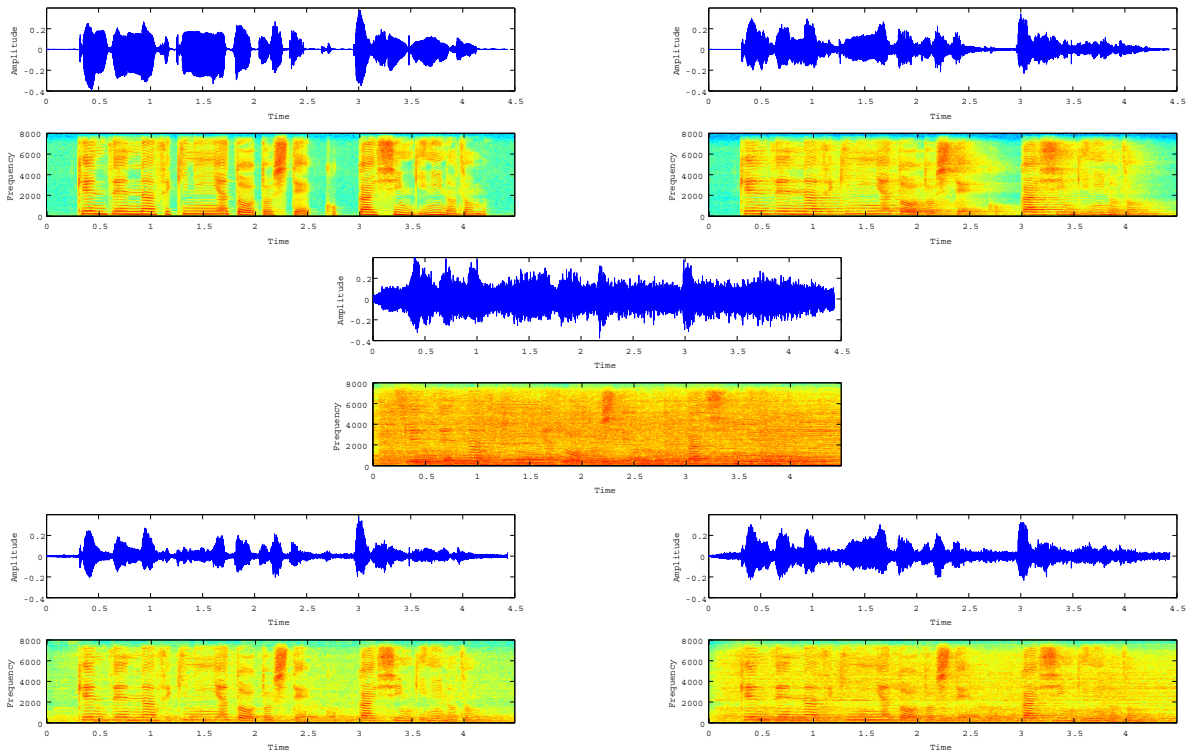

Fig. 6. Wave signals and spectra: clean (top left), reverberated (top right), observed signal (middle), separated by our method (bottom left) and separated by previous method (bottom right). The observed signal consists of two reverberated speech signals (30 degree interval) and one reverberated music signal.

Tables III and IV are the ASR result with music noise and without music noise, respectively. The ASR performance with clean speech is about 92%. In the case without music, the ASR results of MCSB-ICA are better than that of FD-ICA by an average of 30 points. Even in the case of 10 degree interval data, the performances are improved by about 30 points. In the music noise case, the results are similar to those without music. Since the spectrum of music is widespread to all the frequencies, ASR performance is degraded in all situations.

We concluded that MCSB-ICA accomplished MBD and EC at the same time, and outperforms conventional FD-ICA. Note that these results may be slightly affected by the location of microphones, but they will not change greatly.

B. Unsolved Problems

There remains several problems for MCSB-ICA to apply it to robot audition. The first one is computational cost, the second one is data buffering time and the last one is a permutation problem. The computational cost will be solved by the development of hardware and parallel-processing techniques. However, the latter two should be discussed seriously.

The data buffering time is very important for ICA because the separation performance is directly affected by it. Even if we adopt block-wise processing, the separation matrices must be re-estimated to perform best because the degree of the time-independency is different among the blocks and the estimated matrices of them are different to some extent. Then, the multi-layer recognition mechanism will be required, such as the combination of fast-processing middle-quality recognition, and slow-processing high-quality recognition to cope with any situation.

The permutation and interest problem can be solved if the likelihood of sound source is defined. For example, permutation should be determined to maximize the likelihood, and we can select the speech signals from many separated signals according to the likelihood. However, we must consider which abstraction level of the likelihood is appropriate for required processing. The ASR likelihood result can be a criteria of it, but it must evaluate likelihood many times and this results in over-spec processing. It will be enough to have a simpler model only to distinguish noise or speech. The cost function estimation [2] is the lowest and abstract level model estimation, that is, wave signal level. Therefore, the moderate likelihood construction is required to solve the permutation and interest problem efficiently.

VII. CONCLUSION

Our goal is the development of a robot that can distinguish a user's speech from a mixture of sounds and can interact with humans naturally through speech in various environments. The MCSB-ICA is a useful MBD method to satisfy such requirements. However, MCSB-ICA has not been applied to the problems actually while is has been applied to robot speech and reverberation separation. We evaluated the performance limitation of MCSB-ICA under multi-source, barge-in and reverberant environments. The experimental results demonstrated the effectiveness of our methods compared with conventional FD-ICA by about 30

TABLE III

ASR PERFORMANCES (%) WITHOUT MUSIC NOISE.

Noise type	angle	method	Target sp.		Another sp.	
			Cor.	Acc.	Cor.	Acc.
reverberated speech w/o noise			26.0	19.7	16.9	12.4
Robot		no proc.	15.9	8.4	–	–
		fd-ica	51.7	44.6	–	–
		mcsb-ica	86.6	85.0	–	–
Speech	10	no proc.	10.9	5.0	–	–
		fd-ica	27.9	18.4	11.7	7.2
		mcsb-ica	64.7	54.2	48.7	37.9
	20	no proc.	11.1	3.4	–	–
		fd-ica	37.1	28.6	17.1	11.2
		mcsb-ica	73.6	66.2	65.2	58.3
	30	no proc.	10.5	3.6	–	–
		fd-ica	39.7	31.8	23.8	17.1
		mcsb-ica	81.3	78.0	72.7	67.9
	60	no proc.	12.3	5.1	–	–
		fd-ica	38.4	30.2	20.9	14.1
		mcsb-ica	82.5	79.2	73.4	67.9
	90	no proc.	11.7	5.5	–	–
		fd-ica	41.4	33.7	22.3	14.3
		mcsb-ica	81.0	78.0	67.9	60.3
Speech & Robot	10	no proc.	9.9	3.9	–	–
		fd-ica	22.2	13.8	10.3	6.6
		mcsb-ica	57.8	47.8	37.6	28.3
	20	no proc.	9.9	4.1	–	–
		fd-ica	28	20.8	13.4	8.6
		mcsb-ica	69.2	61.8	57.0	49.5
	30	no proc.	9.6	3.8	–	–
		fd-ica	30.9	23.7	17.0	11.3
		mcsb-ica	75.8	72.2	66.7	61.7
	60	no proc.	9.8	5.2	–	–
		fd-ica	31.6	23.5	15.2	8.9
		mcsb-ica	77.6	74.2	67.7	63.1
	90	no proc.	9.1	5.1	–	–
		fd-ica	32.8	25.3	15.9	9.7
		mcsb-ica	77.2	73.2	64.7	58.2

TABLE IV

ASR PERFORMANCES (%) WITH MUSIC NOISE.

Noise type	angle	method	Target sp.		Another sp.	
			Cor.	Acc.	Cor.	Acc.
reverberated speech w/o noise			26.0	19.7	16.9	12.4
Robot & Music		no proc.	0.7	0.7	–	–
		fd-ica	20.5	18.8	–	–
		mcsb-ica	77.3	75.6	–	–
Speech & Music	10	no proc.	1.0	1.0	–	–
		fd-ica	9.0	8.5	3.0	2.8
		mcsb-ica	40.7	31.3	18.7	13.9
	20	no proc.	0.7	0.7	–	–
		fd-ica	13.9	13.0	3.4	3.4
		mcsb-ica	59.4	55.2	36.7	33.0
	30	no proc.	0.7	0.6	–	–
		fd-ica	14.6	13.8	5.0	4.8
		mcsb-ica	68.7	65.6	49.0	46.3
	60	no proc.	0.4	0.4	–	–
		fd-ica	14.2	13.1	4.9	4.6
		mcsb-ica	70.4	68.4	51.3	48.4
	90	no proc.	0.7	0.7	–	–
		fd-ica	14.9	13.8	4.8	4.5
		mcsb-ica	68.6	66.2	43.4	40.1
Speech & Robot & Music	10	no proc.	0.7	0.7	–	–
		fd-ica	7.3	6.7	2.5	2.4
		mcsb-ica	33.9	25.5	14.8	10.8
	20	no proc.	0.5	0.5	–	–
		fd-ica	9.8	9.1	3.3	3.1
		mcsb-ica	54.0	49.4	30.1	27.4
	30	no proc.	0.5	0.5	–	–
		fd-ica	12.8	12.1	4.8	4.6
		mcsb-ica	62.0	59.1	42.8	40.6
	60	no proc.	0.4	0.4	–	–
		fd-ica	12.3	11.7	3.6	3.5
		mcsb-ica	64.6	62.3	43.0	40.7
	90	no proc.	0.6	0.6	–	–
		fd-ica	12.3	11.7	4.1	3.8
		mcsb-ica	64.3	61.7	36.5	33.3

points in ASR performance. And we revealed the actual problems of BSS for robot audition.

In the future, we intend to work on the permutation and buffering problems as mentioned in the discussion section. We also need to integrate MCSB-ICA with other methods, such as sound source localization, to enable real-time processing for robot audition.

VIII. ACKNOWLEDGMENTS

This research was partially supported by the Global COE Program and a Grant-in-Aid for Scientific Research (S) and JSPS Fellows.

REFERENCES

[1] K. Nakadai, Hiroshi G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, "An open source software system for robot audition hark and its evaluation," in *Proc. of Humanoids*, 2008, pp. 561–566.

[2] Y. Takahashi *et al.*, "Source adaptive blind signal extraction using closed-form ica for hands-free robot spoken dialogue system," in *Proc. of ICASSP09*, 2009, pp. 3681–3684.

[3] B. Raj and R. M. Sterm, "Missing-feature approaches in speech recognition," *IEEE Signal Processing Magazine*, vol. 22, no. 5, pp. 101–116, 2005.

[4] R. Gomez *et al.*, "Distant-talking robust speech recognition using late reflection components of room impulse response," in *ICASSP08*. 2008, pp. 4581–4584, IEEE.

[5] T. Nakatani *et al.*, "Blind speech dereverberation with multi-channel linear prediction based on short time fourier transform representation," in *ICASSP08*. 2008, pp. 85–88, IEEE.

[6] J.-M. Yang *et al.*, "A new adaptive filter algorithm for system identification using independent component analysis," in *ICASSP07*. 2007, pp. 1341–1344, IEEE.

[7] S. Araki *et al.*, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. on Speech & Audio Proc.*, vol. 11, pp. 109–116, 2003.

[8] T. Yoshioka *et al.*, "An integrated method for blind separation and dereverberation of convolutive audio mixtures," in *EUSIPCO08*, 2008.

[9] R. Takeda *et al.*, "ICA-based efficient blind dereverberation and echo cancellation method for barge-in-able robot audition," in *Proc. of ICASSP09*, 2009, pp. 3677–3680.

[10] K. Tachibana, H. Saruwatari, Y. Mori, S. Miyabe, K. Shikano, and A. Tanaka, "Efficient blind source separation of combining closed-form second-order ica and nonclosed-form higher-order ica," in *Proc. of IEEE ICASSP07*, 2007, vol. I, pp. 45–48.

[11] S. Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[12] S. Choi *et al.*, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," in *Int'l Workshop on ICA and BBS*, 1999, pp. 371–376.

[13] N. Murata *et al.*, "An approach to blind source separation based on temporal structure of speech signals," in *Neurocomputing*, 2001, pp. 1–24.

[14] R. Takeda, K. Nakadai, T. Takahashi, K. Komatani, T. Ogata, and Hiroshi G. Okuno, "Step-size parameter adaptation of multi-channel semi-blind ica with piecewise linear model for barge-in-able robot audition," in *Proc. of IROS09 (to appear)*, 2009.