9th IEEE-RAS International Conference on Humanoid Robots
December 7-10, 2009 Paris, France

Automatic Estimation of Reverberation Time with Robot Speech to Improve ICA-based Robot Audition

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

***Abstract*— This paper presents an ICA-based robot audition system which estimates the reverberation time of the environment automatically by using the robot's own speech. The system is based on multi-channel semi-blind independent component analysis (MCSB-ICA), a source separation method using a microphone array that can separate user and robot speech under reverberant environments. Perception of the reverberation time (RT) is critical, because an inappropriate RT degrades separation performance and increases processing time. Unlike most previous methods that assume the RT is given in advance, our method estimates an RT by using the echo's intensity of the robot's own speech. It has three steps: speaks a sentence in a new environment, calculates the relative powers of the echoes, and estimates the RT using linear regression of them. Experimental results show that this method sets an appropriate RT for MCSB-ICA for real-world environments and that word correctness is improved by up to 6 points and processing time is reduced by up to 60%.**

I. INTRODUCTION

Many humanoids, such as the ASIMO, have recently been developed to support and enrich human life. They are designed not for a particular environment but for various real-world environments. The ability of robots to hear is essential if symbiosis between people and robots is to be attained, because verbal communication is the most important tool in our daily lives. Many automatic speech recognition (ASR) or spoken dialogue systems work well in the laboratory but not in noisy and reverberant environments. Since the microphones used to pick up sound are installed on the robot and are thus not usually close to the mouth of the person speaking, a conventional dialogue system is not well suited for a robot audition. Moreover, the microphones also pick up the robot's own speech and its echoes, and it makes more difficult to recognize user's speech. The person and robot must thus interact *alternately* so as to avoid *"barge-in"* situations. "Barge-in" means that the person interferes and begins speaking while the robot is speaking, i.e., double-talk. This *"barge-in-able"* capability is also essential for robots to have smoother speech interactions as well as to distinguish human speech from noise and reverberation. While the benefits of the barge-in feature have been validated in the spoken dialogue research [1], [2], robot audition research has mainly focused on separation techniques, and little attention has been paid to the dialogue-related aspect [3]. Our goal is illustrated in Figure 1.

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

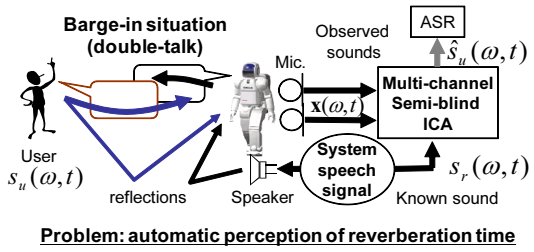

Fig. 1. Our approach

Since robots are usually deployed in real-world environments, robot audition should work even in unknown and/or dynamically changing environments. It should also distinguish the user's speech from noises, echoes, and reverberations. That is, it must deal with problems of echo cancellation, i.e., separation of the robot's/known speech, and blind dereverberation, i.e., separation of the user's speech reverberations, at the same time. The speech recognition with large vocabulary size in reverberant environment is very difficult and it remains hot topic in speech signal processing area, so many humanoids robot, including ASIMO, can not work actually. To satisfy these simultaneous requirements, we use multi-channel semi-blind independent component analysis (MCSB-ICA) [4], a statistical sound source separation method using a microphone array. This method has three particular advantages: 1) the only assumption is mutual independence of sound signals, i.e., *a priori* information about the environments and sound sources is not needed, 2) it is theoretically robust against Gaussian noise such as that from fans, 3) it can theoretically separate known speech, user's speech, and other sounds, including their reverberations. Other methods cannot deal with known-source signals [5], user speech signals [6] or reverberations [7], [8].

Two problems remain to be solved before MCSB -ICA can be implemented for robot audition. One is an stable incremental processing. MCSB-ICA needs 3 seconds duration data for filter estimation to achieve sufficient separation performance, which means a technique must be used to compensate for the latency caused by data buffering. The other is the filter length estimation of the dereverberation filter of MCSB-ICA, which is equivalent to estimation of the reverberation time (RT). The second one means that a dereverberation filter length appropriate for the environment must be set in advance because the length affects separation performance and computational cost, especially in the few available data case. We tackle the second problem because the first one has been discussed in several ICA-related papers [9] and many dereverberation techniques assume that the

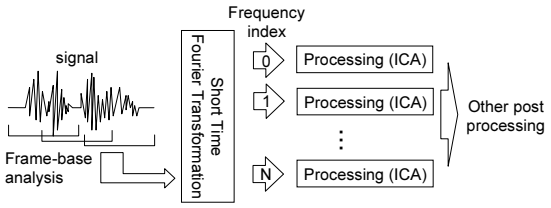

Fig. 2. STFT domain processing

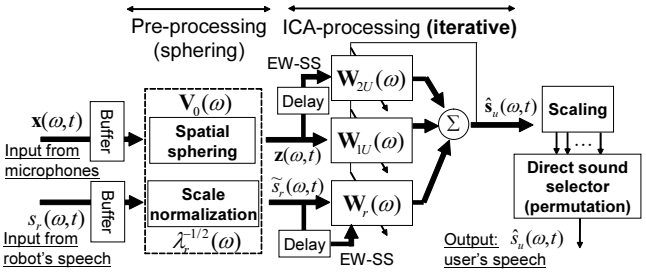

Fig. 3. Signal flow of MCSB-ICA

reverberation time is known. The ultimate goal is to enable a robot to automatically determine the reverberation time of the environment, i.e., estimate the dereverberation filter length.

Our approach is to use the typical situation of human-robot interaction. The idea is that the robot can perceive the reverberation time by speaking in the environment and by estimating the intensities of the echoes. This is considered to be a form of natural-active perception of reverberation time as people do. Since MCSB-ICA can also achieve echo cancellation, the dereverberation filter length can be set by using information from the echo cancellation filter. In our method, the decay rate of the relative power of the echo cancellation filter is used as a criterion of the reverberation time. While there are many methods for estimating the reverberation time accurately, many are unnatural for human-robot interaction because they use a noise-like sound [10] for the estimation. Moreover, self speech can be controlled by the robot while other sound sources cannot.

The rest of the paper is organized as follows: Section 2 explains basic MCSB-ICA. Section 3 explains the three techniques we use for active perception of the reverberation time, and Section 4 describes the implementation. We discuss our evaluations in Sections 5 and 6. The last section summarizes the key points and discusses future work.

II. Multi-channel Semi-blind ICA and Requirements for Robot Audition

This section explains MCSB-ICA [4] and the requirements related to robot audition. The MCSB-ICA model described here uses a short-time Fourier transformation (STFT) representation [11], 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 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 and some reflected signals of the user's speech.

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 [12].

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

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

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

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

where $\cdot^H$ denotes the conjugate transpose operation, and $\boldsymbol{\Lambda}$ is a non-holonomic constraint matrix, i.e.,

$\mathrm{diag}(\mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\hat{\boldsymbol{s}}^H(t)])$ [13]. The μ is a step-size parameter, and $\boldsymbol{\phi}(\boldsymbol{x})$ is a non-linear function vector, $[\phi(x_1), \cdots, \phi(x_L)]^H$. $\phi(x)$:

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

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 $\tilde{s}_r$.

C. Other considerations

1) Scaling: We used the projection back method [14] to solve the scaling problem. In this method, an element of the inverse separation matrix is multiplied by the corresponding separated signal. We used the i-th row and j-th column element c_j of $\hat{\boldsymbol{H}}_u = (\boldsymbol{W}_{1u}\boldsymbol{V}_0)^{-1}$, which satisfy 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) Permutation: We solved the permutation problem by using the average power of the separated signal. If the separated signals include direct and reflected sounds, the power of the direct sound is strongest in the separated signals. Hence, we selected the signal with the largest power.

3) 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 the frequency $\omega + 1$, $\boldsymbol{W}_{1u}(\omega + 1)$. We used the unit matrix for the initial value of the first separation matrix.

4) Step-size scheduling: To accelerate the convergence speed, we used an adaptive step size method [15].

D. Requirement for Robot Audition

For MCSB-ICA to be applied to robot audition, a method is needed for setting dereverberation filter length, N in equation (2), appropriate to the environment in advance. If an inappropriate value is set, the separation performance and the computational cost of MCSB-ICA will degrade and increase, respectively, especially in the few-data case. For example, if N is set shorter than the actual reverberation time, the reverberations cannot be suppressed sufficiently. On the other hand, if N is set longer than the actual reverberation time, the computational cost will only increase. Accurate estimation of the reverberation time is thus required for achieving an efficient robot audition system.

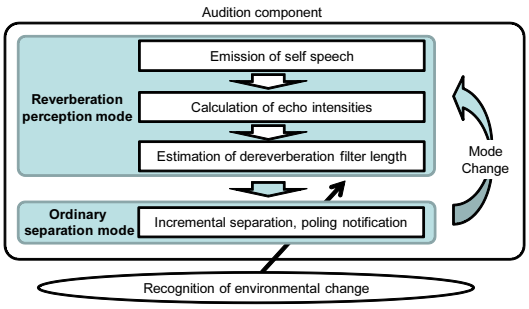

Fig. 4. Strategy for reverberation perception

III. Active Perception of Reverberation Time

This section explains the estimation of the reverberation time based on self speech emission. The objective is to set dereverberation filter length N automatically in accordance with the environment.

A. Strategy for Estimating Dereverberation Filter Length

We first clarify the strategy before explaining the reverberation time estimation. We define two modes for the audition system: reverberation perception, and ordinary separation. This enables us to design reverberation perception independently of ordinary separation. The flowchart is illustrated in Figure 4.

Our strategy

A robot enters a different environment (e.g., a room).

1) The robot recognizes the changes in environment.
2) The robot speaks a sentence.
3) The robot estimates the reverberation time using the echo's intensities.
4) The estimated reverberation time is used as long as the robot remains in the environment.

Since here we focus on robot audition, we only deal with 2-4). We assume that the robot can recognize environmental changes with the aid of a device, such as a camera.

Additionally, we make the following assumption.

> The robot does not recognize any sounds until the estimation of the reverberation time has finished.

This means that a robot concentrates on estimating the reverberation time for a few seconds (1–4 s.). Of course, the robot can recognize that there are other sound sources by using the separation results and can react to some response after reverberation estimation if other sound sources exist. Note that the filter length of MCSB-ICA does not require the accurate reverberation time while the accurate reverberation time of the room is affected by the object locations, that is, the rough estimation is enough for MCSB-ICA. We do not discuss this point as it is beyond our scope.

In the following subsections, we explain reverberation time estimation using robot speech echoes.

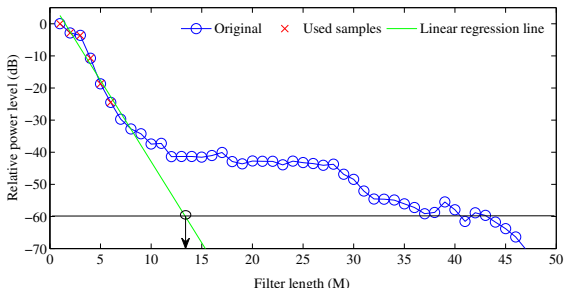

Fig. 5. Example of filter length estimation

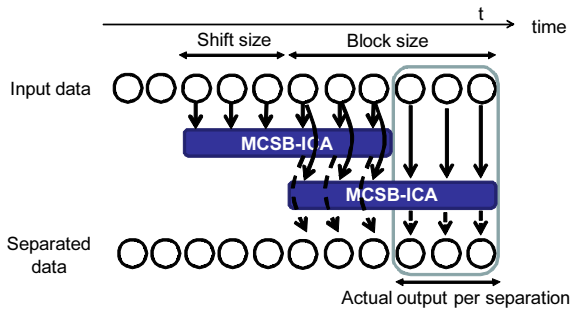

Fig. 6. Incremental processing of MCSB-ICA

B. Estimation of Echo Cancellation Matrix

We can get the echo cancellation matrix, $\boldsymbol{W}_r$, by using the equation (8). In reverberation perception mode, the dereverberation filter length is set to 1 to reduce the calculation cost because our interest is only the $\boldsymbol{W}_r$. We also must set the maximum each filter length, N_{max}, M_{max}, taking into account the resources of the robot.

C. Criteria and Estimation of the reverberation time

We rewrite the echo cancellation matrix as $\boldsymbol{W}_r = [\boldsymbol{w}_r(0)\boldsymbol{w}_r(1)\ldots\boldsymbol{w}_r(M)]$, where $\boldsymbol{w}_r(m) = [w_r^1(m)w_r^2(m)\ldots w_r^L(m)]^T$ is an $L \times 1$ vector. We no longer omit frequency index ω as we did in previous sections. We define the normalized power function of these filters at frequency ω as

$$p_r^i(\omega, m) = |w_r^i(\omega, m)|^2 / \max_m |w_r^i(\omega, m)|^2, \quad (14)$$

where i is a microphone index and m is a filter index.

We base the estimation of the reverberation time on this power function because it reflects the echo intensities and are related to the reverberation time of the environment. We define frequency- and microphone-averaged power function P and its logarithm L as a criteria for the reverberation time:

$$P(m) = \sum_i \sum_{\omega\in\Omega} p_r^i(\omega, m) / \max_m \sum_i \sum_{\omega\in\Omega} p_r^i(\omega, m), \quad (15)$$

$$L(m) = 20\log_{10} P(m), \quad (16)$$

where Ω is the frequency-band set to be considered.

We introduce a linear regression model for estimating the filter length:

$$y = am + b, \quad (17)$$

where a and b are the parameters, m is the filter index, and y is equivalent to $L(m)$. Using a few samples after the peak of $P(m)$, we estimate the a and b by the least mean squares method.

Finally, the dereverberation filter length is defined for the value of m that satisfies $L(m) = L_d$:

$$\hat{N} = \frac{L_d - b}{a}. \quad (18)$$

Figure 5 shows an example of this criterion and estimation. The *original* plot points show the $P(m)$ obtained from the echo cancellation filter for an environmental in which $\mathrm{RT}_{20} = 240$, where RT_{20} is reverberation time, and the *linear regression line* was estimated using Equation (17). In this case, the estimated filter length was about 13 according to Equation (18) with $L_d = -60$.

IV. Incremental Separation

For ordinary separation mode, we use block-wise incremental separation of MCSB-ICA. Since ICA buffers some duration data for stable estimation of the separation matrix, we use the previous I_b samples for time t separation. We introduce shift-size I_s to reduce latency caused by buffering. As illustrated in Figure 6, if I_s is increased, the latency also increases; if I_s is reduced, computational cost increases. This implementation is virtually the same as one used previously [9], except for the use of overlap parameter I_s here.

V. EXPERIMENTS

A. Settings

The impulse responses for speech data were recorded at 16 kHz in two different environment (rooms).

Env. I: a normal room (RT_{20} = 240 ms)
Env. II: a hall-like room (RT_{20} = 670 ms).

The first room was 4.2 × 7.0 m, and the second was 7.55 × 9.55 m. In both rooms, the speaker was 1.5 m from the microphone, which was mounted on the head of Honda ASIMO, and the angle between the speaker and the front of ASIMO was 0, 45, 90, -45, or -90 degrees. We recorded the impulse response of the robot's speech in each environment.

We used 200 Japanese sentences for the user's speech and 100 sentences for the robot's speech; they were convoluted in the corresponding recorded impulse responses. All synthesized data (16 bits, PCM) were normalized to [−1.0 1.0]. Julius[1] was used for HMM-based ASR with the statistical language model. Mel-frequency cepstral coefficients (12+Δ12+ΔPow) were obtained after STFT with a window size of 512 points (32 ms) and a shift size of 160 points (10 ms) for the speech features. We then applied cepstral mean normalization. A triphone-based acoustic model (three-state, four-mixture) was trained with 150 sentences of clean speech uttered by 200 male and female speakers (word-closed). The lengths of these utterances were from 1.5 to 15 s. The statistical language model consisted of **20,000**

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

TABLE I

CONFIGURATION FOR DATA AND SEPARATION

Impulse response	16-kHz sampling
Reverberation time (RT_{20})	240 and 670 ms
Distance and direction	1.5 m and 0°, 45°, 90°, −45°, −90°
Number of microphones	Two (embedded in ASIMO's head)
STFT analysis	Hanning: 32 ms and shift: 12 ms
Input wave data	[-1.0 1.0] normalized

TABLE II

CONFIGURATION FOR SPEECH RECOGNITION

Test set	200 sentences
Training set	200 persons (150 sentences each)
Acoustic model	PTM-triphone: 3-state, HMM
Language model	Statistical, vocabulary size 20 k
Speech analysis	Hanning: 32 ms and shift: 10 ms
Features	MFCC 25 dim. (12+Δ12+ΔPow)

words extracted from newspapers. The other experimental conditions are summarized in Tables I and II.

B. Evaluation

We carried out three experiments in each environment, using two microphones in each one.

1) Filter length estimation (Exp. A): Experiment A was used to evaluate the sensitivity of our method to noise and parameter M_{max}. The mean and standard derivation of the estimated filter length were evaluated for 100 robot's sentences. We used two type of data: one was only the robot's own speech (**without noise**), and the other was the robot's and people's speech (two sources in total), which is assumed to be noisy conditions (**with noise**). We tested several maximum filter lengths M_{max}, including 20, 30 and 50.

2) Separation performance (Exp. B, C): The other two experiments were used to evaluate the performance of MCSB-ICA in an ordinary separation mode with filter length estimation.

Exp. B: Dereverberation performance

Exp. C: Dereverberation and echo cancellation performance.

Word correctness (WC) was evaluated using **200 Japanese sentences**. The sounds included only the user's speech in Exp. B (**non-barge-in**); they included the user's and robot's speech in Exp. C (**barge-in**). In these experiments, we used 166 (2 s), 208 (2.5 s) and 255 (3 s) as block-size I_b; shift size I_s was set to half these values.

C. Separation Parameters Setting

The STFT parameters were set the same for all three experiments: the window size was 512 points (32 ms) and the shift-size was 192 points (12 ms). The frame interval parameter d was 2, and the filter lengths of echo cancellation and dereverberation in the ordinary separation mode was same, $N = M$.

The parameters for adaptive step-size control were set as a previous [15]. The filter estimation parameters were $\Omega = \{5, 6, \ldots, 200\}$ and $L_d = -60$, and the number of samples for the linear regression was 6.

VI. RESULTS

A. Filter Length Estimation (Exp. A)

Table III summarizes the means and standard deviations (std.) for Exp. A. The means were almost the same value with and without noise (14 for Env. I and 35 for Env. II). The standard deviations differed slightly with the situation (only by 0–2). The estimated filter lengths diffred by up to 5 according to the speech content.

TABLE III

ESTIMATED FILTER LENGTH

		Env. I (RT_{20} 240 ms)			Env. II (RT_{20} 670 ms)		
	M_{max}	20	30	50	30	40	50
w/o noise	Mean	14.0	13.7	13.2	35.0	35.3	35.4
	Std.	0.43	0.46	0.53	1.22	1.24	1.28
with noise	Mean	14.2	14.0	13.6	36.1	36.3	36.2
	Std.	1.25	1.17	1.05	2.38	2.41	2.30

These results show that our method is not affected by noise and M_{max}. Moreover, our method needs only a short filter length to estimate reverberation time. This means that we can set M_{max} to a small value in reverberation perception mode, which increases adaptation speed. The estimated reverberation time can be used for other methods, such as a criterion for acoustic model selection of ASR.

B. Separation Performance (Exp. B, C)

Figures 7 and 8 presents the position-averaged WC values for each dereverberation filter length N and each data duration for Exp. B, and Figures 9 and 10 show it for Exp. C. Table IV summarizes the WC values with the estimated filter length at Exp. A. The *no proc.* in the table means the WC without separation. WC for clean speech is about 93% in our configuration.

These figures show that the performance of MCSB-ICA in Env. I began to decrease at some filter length. This is most evident for the small-sample case (2 s) and Exp. B (barge-in situation). If we use an estimated filter length of 14, the performance degradation is reduced about 3 points in Exp. B and 5 points in Exp. C compared with the long filter length case, such as $N = 35$. The calculation cost with the estimated filter length in Env. I is 60% lower than with $N = 35$ because the computational cost of MCSB-ICA is proportional to the filter length, N. Of course, the performance in Env. II does not degrade with the estimated filter length.

The ability of our method to estimate a dereverberation filter length appropriate for the environment lead to the maintaining of separation performance and reducing a calculation cost.

VII. CONCLUSION AND FUTURE WORK

We proposed a multi-channel semi-blind ICA (MCSB-ICA) based robot audition. The reverberation time of the environment is perceived automatically using the robot's own speech and its echoes. The method consists of three step:

TABLE IV

WORD CORRECTNESS WITH ESTIMATED FILTER LENGTH (%)

	Exp. C (non-barge-in)				Exp.C (barge-in)			
	no proc.	2 s	2.5 s	3 s	no proc.	2 s	2.5 s	3 s
Env. I (RT_{20} 240 ms), $N = 15$	74.3	76.9	78.5	78.2	28.2	67.8	70.2	71.7
Env. II (RT_{20} 670 ms), $N = 36$	26.1	63.9	66.8	69.2	11.0	37.1	41.2	43.3

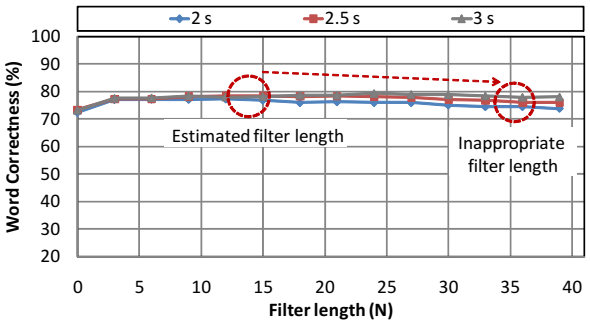

Fig. 7. Results of Exp. B in Env. I

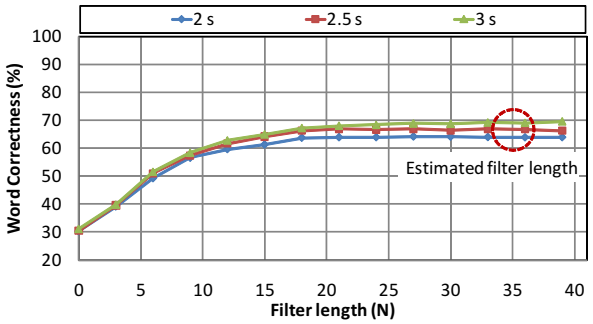

Fig. 8. Results of Exp. B in Env. II

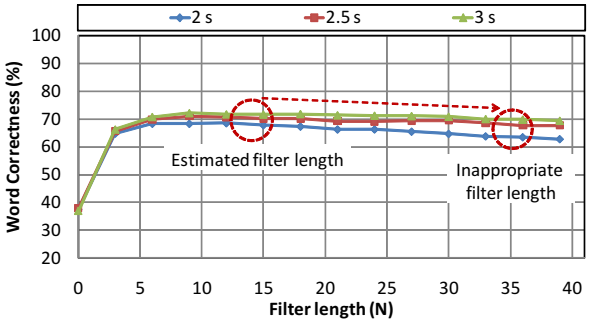

Fig. 9. Results of Exp. C in Env. I

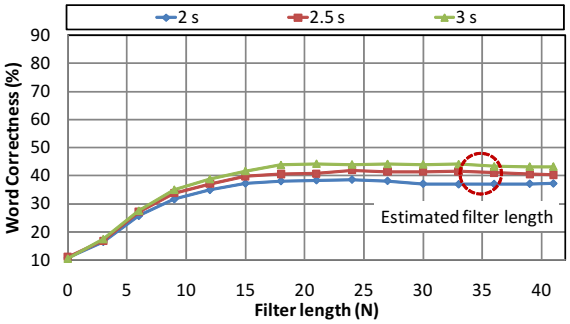

Fig. 10. Results of Exp. C in Env. II

emission of self speech, calculation of logarithmic relative power of echo cancellation filter, and filter length estimation based on linear regression model. Experimental results showed that our method can set a filter length appropriate to the environment.

In the future, we intend to work on the acoustic model selection based on the estimated filter length to improve word correctness more. And we will also work on improving the performance of incremental separation using MCSB-ICA by using a priori information, such as impulse responses. We also need to integrate MCSB-ICA with other methods 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).

REFERENCES

[1] K. Matsuyama, K. Komatani, T. Ogata, and Hiroshi G. Okuno, "Enabling a user to specify an item at any time during system enumeration," in *Proc. of Interspeech*, 2009 (to appear).

[2] K. Komatani, T. Kawahara, and Hiroshi G. Okuno, "Predicting asr errors by exploiting barge-in rate of individual users for spoken dialogue systems," in *Proc. of Interspeech*, 2008, pp. 183–186.

[3] 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.

[4] 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.

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

[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. Miyabe *et al.*, "Barge-in- and noise-free spoken dialogue interface based on sound field control and semi-blind source separation," in *EUSIPCO07*, 2007, pp. 232–236.

[8] 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.

[9] H. Saruwatari, Y. Mori, T. Takatani., S. Ukai, K. Shikano, T. Hiekata, and T. Morita, "Two-stage blind source separation based on ica and binary masking for real-time robot audition system," in *Proc. of IEEE/RSJ IROS05*. 2005, pp. 209–214, IEEE (2005).

[10] Y. Suzuki, F. Asano, H.-Y. Kim, and T. Sone, "An optimum computer-generated pulse signal suitable for the measurement of very long impulse responses," *J. Acoust. Soc. Am*, vol. 97, no. 2, pp. 1119–1123, 1995.

[11] 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.

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

[13] 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.

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

[15] 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.