ICA-BASED EFFICIENT BLIND DEREVERBERATION AND ECHO CANCELLATION METHOD FOR BARGE-IN-ABLE ROBOT AUDITION

Ryu Takeda[†]*,Kazuhiro Nakadai*[‡]*,Toru Takahashi*[†]*,Kazunori Komatani*[†]*,Tetsuya Ogata*[†]*,Hiroshi G. Okuno*[†]

[†]Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo, Kyoto 606-8501, Japan
[‡]Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako, Saitama, 351-0188, Japan
{rtakeda, tall, komatani, ogata, okuno}@kuis.kyoto-u.ac.jp, nakadai@jp.honda-ri.com

ABSTRACT

This paper describes a new method that allows "Barge-In" in various environments for robot audition. "Barge-in" means that a user begins to speak simultaneously while a robot is speaking. To achieve the function, we must deal with problems on blind dereverberation and echo cancellation at the same time. We adopt Independent Component Analysis (ICA) because it essentially provides a natural framework for these two problems. To deal with reverberation, we apply a Multiple Input/Output INverse-filtering Theorem-based model of observation to the frequency domain ICA. The main problem is its high-computational cost of ICA. We reduce the computational complexity to the linear order of reverberation time by using two techniques: 1) a separation model based on observed signal independence, and 2) enforced spatial sphering for preprocessing. The experimental results revealed that our method improved word correctness of reverberant speech by 10-20 points.

Index Terms— Barge-In, ICA, MINT, blind dereverberation, echo cancellation

1. INTRODUCTION

A robot should recognize an user's speech from a mixture of sounds with the least prior information, because the robot has to work in unknown and/or dynamical environments. A mixture of sounds may includes the robot's own speech and user's speech reverberations, because microphones are equipped on its body, not attached close to the mouth of a user. Therefore, these should be suppressed to enhance the user's speech (Fig. 1). In human-robot or in human-computer interaction, the user often interrupts and begins to speak while the robot or the system is speaking. This situation is called "barge-in". Robot audition systems should be *"barge-in-able"* for smoother speech interaction.

To achieve such a barge-in-able system, we must deal with the problems of echo cancellation (separation of robot's speech) and blind dereverberation (separation of user's speech reverberation) at the same time. We must especially focus on late-reverberation (late-reflection) because early-reverberation can be solved by cepstral mean normalization (CMN) or noise robust methods for automatic speech recognition (ASR). Nakatani *et al.* proposed a high-performance method of blind dereverberation based on Short-Time Fourier Transformation (STFT) representation [1]. Gomez *et al.* applied fast spectral subtraction for late reverberation by using a pre-recorded impulse response [2]. However, these and other familiar methods have not dealt with the echo-cancellation problem, or used a priori knowledge about the environment, such as room impulse response. This is similar to echo cancellation. Yang *et al.* recently proposed a noise-robust (user's speech-robust) method based on Independent Component Analysis (ICA) [3], and Miyabe *et al.* realizes a separation of the known and unknown sources efficiently with ICA framework [4]. However, these method also cannot deal with the target speech reverberation.

Thanks to the Global COE Program and the Grant-in-Aid for Scientific Research (S).

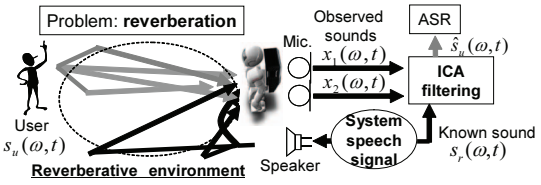

Fig. 1. Data flow and our problem

We focused on frequency domain ICA (FD-ICA) to deal with the two problems because 1) it provides a natural framework, such as blind source separation and adaptive filtering, 2) it is robust against Gaussian noise, such as fan noise, and 3) its convergence and computational cost are excellent compared with time domain ICA. However, Araki *et al.* reported a fundamental limitation of the performance of FD-ICA [5]. To overcome the limitation and deal with late-reverberation, we combined a Multiple Input/Output INverse-filtering Theorem (MINT)-based model of observation [6] and STFT representation [1] for FD-ICA. However, this naive application seriously increases the computational cost because this approach essentially separates all reflected sounds as other sources.

In this paper, we introduce two techniques to reduce the computational cost: 1) a separation model by assuming observed signal independence, and 2) enforced spatial sphering for preprocessing. In Section 2, we explain the MINT-based observation model, ICA and its problems. In Section 3, we explain the two techniques in detail and discuss our evaluations of the new method assessed by evaluation results obtained by speech recognition experiments.

2. CONVENTIONAL TECHNIQUES

We describe total models with STFT representation [1], which is a form of multi-rate processing because of a) its scalability with other methods and b) its low computational cost. We denote the spectrum after STFT as $s(\omega, t)$ at frequency ω and frame t. For the sake of readability, we skip denoting the frequency index, ω.

2.1. MINT-based observation model

Already mentioned in Section 1, MINT [6] confirms the existence of an inverse filter for the acoustic field if the number of microphones is larger than the number of sound sources.

We denote the observed spectra at microphones $1, \ldots, L$ as $x_1(t), \ldots, x_L(t)$ (L is the number of microphones). Then, we represent the observed vectors, $\boldsymbol{x}(t)$, $\boldsymbol{X}(t)$, and the user's and robot's (known) speech spectrum $\boldsymbol{S}_u(t)$ and $\boldsymbol{S}_r(t)$ as:

$$\boldsymbol{x}(t) = [x_1(t), x_2(t), \ldots, x_L(t)]^T, \quad (1)$$

$$\boldsymbol{X}(t) = [\boldsymbol{x}(t), \boldsymbol{x}(t-1), \ldots, \boldsymbol{x}(t-N)]^T, \quad (2)$$

$$\boldsymbol{S}_u(t) = [s_u(t), s_u(t-1), \ldots, s_u(t-M_u)]^T, \quad (3)$$

$$\boldsymbol{S}_r(t) = [s_r(t), s_r(t-1), \ldots, s_r(t-M_r)]^T, \quad (4)$$

where N, M_u and M_r mean the number of the delayed frames, and their proper sizes are decided by the following MINT condition.

We define the circulant "MINT transfer function matrix" $\boldsymbol{H}_x$ for source signal x with environment-dependent parameter K_x as:

$$\boldsymbol{h}_x(i) = [h_1^x(i), h_2^x, \ldots, h_L^x(i)]^T, \tag{5}$$

$$\boldsymbol{H}_x = \begin{pmatrix} \boldsymbol{h}_x(0) & \cdots & \cdots & \boldsymbol{h}_x(K_x) & \cdots & \boldsymbol{0} \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \boldsymbol{0} & \cdots & \boldsymbol{h}_x(0) & \cdots & \cdots & \boldsymbol{h}_x(K_x) \end{pmatrix}. \tag{6}$$

With the MINT transfer function matrices $\boldsymbol{H}_u$ and $\boldsymbol{H}_r$ of the user's and robot's speech, the observation model is represented as

$$\begin{pmatrix} \boldsymbol{X}(t) \\ \boldsymbol{S}_r(t) \end{pmatrix} = \begin{pmatrix} \boldsymbol{H}_u & \boldsymbol{H}_r \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{S}_u(t) \\ \boldsymbol{S}_r(t) \end{pmatrix}, \tag{7}$$

where $\boldsymbol{I}$ denotes a $(M_r+1) \times (M_r+1)$ unit matrix, and the size of $\boldsymbol{H}_r$ is $L(N+1) \times (M_r+1)$ and that of $\boldsymbol{H}_u$ is $L(N+1) \times (M_u+1)$. Here, if the MINT condition where $L(N+1) = (M_u+1)$ is satisfied, the whole mixing matrix becomes a holomorphic square matrix, i.e., its inverse system exists.

2.2. Independent Component Analysis

ICA for MINT-based separation model: Assuming that the source signals $\boldsymbol{S}_u$ and $\boldsymbol{S}_r$ are independent, respectively, this problem can be solved by instantaneous-model ICA [7]. With the separation matrices $\boldsymbol{W}_u$ and $\boldsymbol{W}_r$, the separation model is represented as

$$\begin{pmatrix} \hat{\boldsymbol{S}}_u(t) \\ \boldsymbol{S}_r(t) \end{pmatrix} = \begin{pmatrix} \boldsymbol{W}_u & \boldsymbol{W}_r \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{X}(t) \\ \boldsymbol{S}_r(t) \end{pmatrix}. \tag{8}$$

ICA estimates the separation matrices $\boldsymbol{W}_u$ and $\boldsymbol{W}_r$ and the user's speech $\boldsymbol{S}_u$ blindly at each frequency ω by minimizing Kullback-Leibler Divergence (KLD),

$$J(\boldsymbol{W}_u, \boldsymbol{W}_r) = \int p(\boldsymbol{S}_u, \boldsymbol{S}_r) \log \frac{p(\boldsymbol{S}_u, \boldsymbol{S}_r)}{q_1(\boldsymbol{S}_u) q_2(\boldsymbol{S}_r)} d\boldsymbol{S}_u d\boldsymbol{S}_r, \tag{9}$$

where p is the joint Probability Density Function (PDF) of $\boldsymbol{S}_u$ and $\boldsymbol{S}_r$. q_1 and q_2 correspond to the products of the marginal PDF of $\boldsymbol{S}_u$ and $\boldsymbol{S}_r$. Usually, these parameters are estimated by the iterative gradient-based method because of the non-linearity of J.

Sphering for Pre-processing: To achieve fast-convergence of ICA, sphering transformation works well as pre-processing [7]. This is a linear transformation, $\boldsymbol{V}$, which decorrelates the input signals and normalizes the variances. With the eigenvalue diagonal matrix, $\boldsymbol{\Lambda}$, and the eigenvector unitary matrix $\boldsymbol{E}$ of the temporal-spatial correlation matrix, $\boldsymbol{R}$, the transformation is usually done as,

$$\boldsymbol{R} = \begin{pmatrix} \mathrm{E}[\boldsymbol{X}(t)\boldsymbol{X}^H(t)] & \mathrm{E}[\boldsymbol{X}(t)\boldsymbol{S}_r^H(t)] \\ \mathrm{E}[\boldsymbol{S}_r(t)\boldsymbol{X}^H(t)] & \mathrm{E}[\boldsymbol{S}_r(t)\boldsymbol{S}_r^H(t)] \end{pmatrix}, \tag{10}$$

$$\begin{pmatrix} \boldsymbol{Z}_1(t) \\ \boldsymbol{Z}_2(t) \end{pmatrix} = \boldsymbol{V} \begin{pmatrix} \boldsymbol{X}(t) \\ \boldsymbol{S}_r(t) \end{pmatrix}, \boldsymbol{V} = \boldsymbol{E}\boldsymbol{\Lambda}^{-\frac{1}{2}}\boldsymbol{E}^H, \tag{11}$$

where $\mathrm{E}[\cdot]$ is a time-averaging operator, and $\boldsymbol{Z}_x$ denotes the transformed observed signal. The size of correlation matrix $\boldsymbol{R}$ is $(L(N+1)+M_r+1)^2$. $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ are used as the input of ICA instead of $\boldsymbol{S}_u$ and $\boldsymbol{S}_r$.

2.3. Problems with MINT-based separation model for ICA

1. Computational Cost: The calculation cost of the ICA is $O(L^2N^2)$, and that of the sphering is $O(L^3N^3)$. We need to reduce the computational cost to the linear order of the reverberation time, N, for practical use.

2. Permutation: ICA has ambiguity in permutation and scaling of the output signal, $\hat{\boldsymbol{S}}_u$. Because we applied ICA in the frequency domain, they have to be solved to re-syntheize the signal in the time domain. Hence, we must select the direct sound from $\hat{\boldsymbol{S}}_u$.

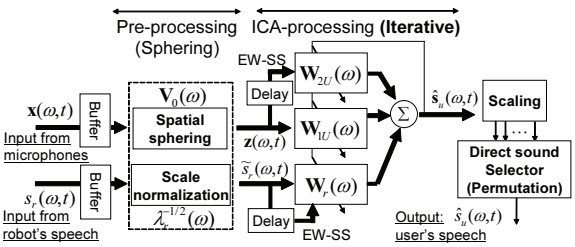

Fig. 2. Signal flow in our method for MINT-based model

3. EFFICIENT ICA-BASED SEPARATION METHOD

We explain the two techniques for the reduction of the computational cost of ICA in this section. As a result, the computational cost was reduced to $O(L^3)$ in the sphering, and $O(L^2N)$ in the ICA, i.e., it became the linear order of N. An overview of our method is illustrated in Fig. 2

3.1. Separation model based on observed signal independence

We designed the following separation model to extract the direct sound frame of the target speech. The main idea is that we substituted the independence of the original source to that of the observed signal. Equation (8) uses the condition that "*Direct sound frame* $s_u(t)$ *is independent of* $\{\boldsymbol{S}_u(t-1), s_r(t), \boldsymbol{S}_r(t-1)\}$". Instead of this, we used the condition that "*Direct sound frame* $s_u(t)$ *is independent of* $\{\boldsymbol{X}(t-1), s_r(t), \boldsymbol{S}_r(t-1)\}$". Here, we add 1 frame redundant signal. Note that these two conditions represent the relationship between the sufficient and necessary conditions. If the $\boldsymbol{H}_u$ in Equation. (7) is holomorphic, the equivalence relationship is approximately true because the mapping from $\{\boldsymbol{S}_u(t-1), s_r(t), \boldsymbol{S}_r(t-1)\}$ to $\{\boldsymbol{X}(t-1), s_r(t), \boldsymbol{S}_r(t-1)\}$ is bijective and the relationship between their jouint PDFs is invertible. Whether this works well or not depends on the time-independence of the speech signal.

If we assume the time-independence of the speech signal, a new separation model is expressed by substituting the independence of the observed signal as follows:

$$\begin{pmatrix} \hat{\boldsymbol{s}}(t) \\ \boldsymbol{X}(t-d) \\ \boldsymbol{S}_r(t) \end{pmatrix} = \begin{pmatrix} \boldsymbol{W}_{1u} & \boldsymbol{W}_{2u} & \boldsymbol{W}_r \\ \boldsymbol{0} & \boldsymbol{I}_2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}_r \end{pmatrix} \begin{pmatrix} \boldsymbol{x}(t) \\ \boldsymbol{X}(t-d) \\ \boldsymbol{S}_r(t) \end{pmatrix}, \tag{12}$$

where $\hat{\boldsymbol{s}}$ is an estimated signal vector with L dimension, $\boldsymbol{W}_{1u}$ and $\boldsymbol{W}_{2u}$ correspond to $L \times L$ and $L \times L(N+1)$ separation matrices, $\boldsymbol{W}_r$ is the $L \times (M_r+1)$ separation matrix, and $\boldsymbol{I}_2$ and $\boldsymbol{I}_r$ are corresponding proper-sized unit matrices. $d > 0$ is an initial-reflection interval parameter and we can consider the dependence between the direct and adjacent frame of $s_u(t)$. Here, we assume $d = 2$.

By minimizing KLD based on a natural gradient [8], this leads to the the following iterative update rules for the separation matrices $\boldsymbol{W}_{1u}$, $\boldsymbol{W}_{2u}$, and $\boldsymbol{W}_r$.

$$\boldsymbol{D} = \boldsymbol{\Lambda} - \mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\hat{\boldsymbol{s}}^H(t)], \tag{13}$$

$$\boldsymbol{W}_{1u}^{[j+1]} = \boldsymbol{W}_{1u}^{[j]} + \mu \boldsymbol{D} \boldsymbol{W}_{1u}^{[j]}, \tag{14}$$

$$\boldsymbol{W}_{2u}^{[j+1]} = \boldsymbol{W}_{2u}^{[j]} + \mu \left(\boldsymbol{D} \boldsymbol{W}_{2u}^{[j]} - \mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\boldsymbol{X}^H(t-d)] \right), \tag{15}$$

$$\boldsymbol{W}_r^{[j+1]} = \boldsymbol{W}_r^{[j]} + \mu \left(\boldsymbol{D} \boldsymbol{W}_r^{[j]} - \mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\boldsymbol{S}_r^H(t)] \right), \tag{16}$$

where μ is a step-size parameter, $\boldsymbol{\phi}(\boldsymbol{x}) = [\phi(x_1), \cdots, \phi(x_L)]^T$ is a non-linear function vector, and $\boldsymbol{\Lambda}$ is a non-holonomic constraint matrix, $\boldsymbol{\Lambda} = \mathrm{diag}(\mathrm{E}[\boldsymbol{\phi}(\hat{\boldsymbol{s}}(t))\hat{\boldsymbol{s}}^H(t)])$ [9]. We used $\tanh(100|x|)e^{\theta(x)}$ as a non-linear function $\phi(x)$ [10]. Equation (14) is used to estimate the blind separation filter, $\boldsymbol{W}_{1u}$, the same as a standard FD-ICA. Equation (15) and (16) are used to estimate the dereverberant filter, $\boldsymbol{W}_{2u}$, and so-called adaptive filter, $\boldsymbol{W}_r$, respectively. Note that a fast-ICA algorithm [7] cannot be applied because the orthogonal condition of the separation matrix is not true with this formulation.

Table 1. Configuration of data and separation

Impulse response	16 kHz sampling
Reverberation time (RT_{20})	240 msec. and 670 msec.
Distance and direction	1.5 m and 0°, 45°, 90°, -45°, -90°
Number of microphones	two (embeded at ASIMO's head)
STFT analysis	Hanning: 64 msec. and shift: 24 msec.
Input wave data	[-1.0 1.0] normalized

Table 2. 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 20k
Speech analysis	Hanning: 32 msec. and shift: 10 msec.
Features	MFCC 25 dim.(12+Δ12+ΔPow)

3.2. Enforced Spatial Sphering

To reduce the calculation cost of sphering, we fix the spatial sphering for our separation model by substituting the values of the temporal and known-source correlation in Equation (10) to zero as,

$$\mathrm{E}[\boldsymbol{X}(t)\boldsymbol{X}^H(t)] = \mathrm{diag}\{\boldsymbol{R}_0, \boldsymbol{R}_0, \cdots, \boldsymbol{R}_0\}, \quad (17)$$

$$\mathrm{E}[\boldsymbol{S}_r(t)\boldsymbol{S}_r^H(t)] = \mathrm{diag}\{\lambda_r, \lambda_r, \cdots, \lambda_r\}, \quad (18)$$

$$\mathrm{E}[\boldsymbol{X}(t)\boldsymbol{S}_r^H(t)] = \boldsymbol{0}, \ \ \mathrm{E}[\boldsymbol{S}_r(t)\boldsymbol{X}^H(t)] = \boldsymbol{0}, \quad (19)$$

where $\boldsymbol{R}_0$ is a spatial correlation matrix, $\mathrm{E}[\boldsymbol{x}(t)\boldsymbol{x}^H(t)]$, and $\lambda_r = \mathrm{E}[s_r(t)s_r^H(t)]$ is a variance of known source $s_r(t)$.

Eventually, in enforced spatial sphering, the observed signal, $\boldsymbol{X}(t)$, and the known signal, $\boldsymbol{S}_r(t)$, are transformed as the following rules:

$$\boldsymbol{z}(t) = \boldsymbol{V}_0\boldsymbol{x}(t), \ \ \boldsymbol{V}_0 = \boldsymbol{E}_0\boldsymbol{\Lambda}_0^{-1/2}\boldsymbol{E}_0^H, \quad (20)$$

$$\tilde{s}_r(t) = \lambda_r^{-1/2}s_r(t), \quad (21)$$

where $\boldsymbol{E}_0$ and $\boldsymbol{\Lambda}_0$ are the eigenvector matrix and eigenvalue diagonal matrix of $\boldsymbol{R}_0$. After sphering, $\boldsymbol{x}$ and s_r in Equations. (12) – (16) are substituted into $\boldsymbol{z}$ and $\tilde{s}_r$.

3.3. Solution to scaling and permutation problems

Scaling: We used the projection back method [11] to solve the scaling problem. This method is achieved by multiplying the diagonal element of the inversed separation matrix by the corresponding separated signals. In our case, we used the diagonal element of $(\boldsymbol{W}_{1u}\boldsymbol{V}_0)^{-1}$.

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 max power.

3.4. Other configurations

Initial value of separation matrix: The initial value of the separation matrix, $\boldsymbol{W}_{1u}(\omega)$, is adjasted to the estimated matrix, $\boldsymbol{W}_{1u}(\omega+1)$. We used the unit matrix for the initial value of the first separation matrix.

Step-size scheduling: The step-size parameter is adjasted by a combination of annealing and the exponentially weighted step-size (EWSS) [12] because they reduce the influence of the ignored temporal correlation when sphering. The step-size, μ_k, of the separation matrix at the j-th iteration and k-th delayed frame is defined by

$$\mu_k^{[j]} = \frac{\alpha}{j}\lambda^k + \beta, \quad (22)$$

where α, β and λ are constant values.

4. EXPERIMENTS

4.1. Speech Data and Experimental Setup for ASR

The impulse responses for speech data were recorded at 16 kHz in two different rooms,

Env. I) a normal room (RT_{20}=240 msec), and

Env. II) a hall-like room (RT_{20}=670 msec).

Here, RT_{20} means a reverberation time. The speaker was 1.5 m from a microphone mounted to the head of Honda's ASIMO, and the angles between the speaker and the front of the ASIMO were 5 patterns of 0, 45, 90, -45, -90 degrees. We also recorded the impulse response about the robot's speech at each environment.

We used 200 Japanese sentences for the user's and robot's speech, and they were convoluted the corresponding recorded impulse responses. Julius [13] was used for HMM-based ASR with statistical language model. Mel-frequncy 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 (CMN). A triphone-based acoustic model (3-state and 4-mixture) was trained with 150 sentences of clean speech uttered by 200 male and female speakers (word-closed). The other experimental conditions are summarized in Tables 1 and 2.

4.2. Evaluation

We carried out two experiments in two environments, Env.I and Env.II in terms of word correctness (WC):

Exp. A) dereverberation performance, and

Exp. B) dereverberation and echo cancellation performance.

Note that the observed sounds include only user's speech in Exp.A and includes user's and robot's speech in Exp.B. We changed the length of observed signal N (denotes number of frames) and the size of the data to estimate the matrices $\boldsymbol{W}_{1u}$, $\boldsymbol{W}_{2u}$ and $\boldsymbol{W}_r$, i.e., with 1, 2, 3 sec block-separated data, and with all data (batch). We used two microphones in these experiments.

We also evaluated the real-time factor (RTF) in the case of batch processing at each experiments. RTF is calculated by P/I, where P is a process time and I is a data time (duration). The CPU of the machine is Intel Pentium D 3.20GHz. In this experiment, we compared our method to the naive method which is based on the simple combination of MINT-based model and ICA with sphering.

4.3. Separation parameters

The same parameters for STFT were chosen, and the window size was $1,024$ points (64 msec.) which is suboptimal size [5], and its shift size was 384 points (24 msec.). The step-size parameters were $\lambda = 0.8$, $\alpha = 5.0e^{-1}$, and $\beta = 5.0e^{-3}$ for block-wise processing, and $\lambda = 0.9$ for batch processing. **We fixed the maximum number of the iterations for estimating matrices to 20** because the time for separation is usually restricted in practical use. With more iteration, the performance will improve slightly. In block-wise processing, the estimated $\boldsymbol{W}_{1u}$ in a certain block is used as an initial value for the $\boldsymbol{W}_{1u}$ of the next block.

5. RESULTS

Figures 4 and 5 present the position-averaged results of Exp.A, Figures 7 and 8 present those of Exp.B, and Table 3 summarizes the average improvement in WC. Figures 3 and 6 show the RTF in Exp.A and Exp.B, respectively.

In Exp.A, our method improved WC by 6 points at Env.I, and improved it by 45 points at Env.II. In Exp.B, it improved 40 and 30

Table 3. Best average word correctness (%)

	User's Speech (no proc.)	Exp.A (Dereverb.)				User's and Robot's speech (no proc.)	Exp.B (Dereverb. + echo cancel)			
		1 sec.	2 sec.	3 sec.	Batch		1 sec.	2 sec.	3 sec.	Batch
Env.I (RT_{20}=240 [ms])	74.3	77.7	81.4	83.3	84.2	28.2	60.9	69.0	72.0	73.2
Env.II (RT_{20}=670 [ms])	26.1	64.1	68.0	70.6	72.9	11.0	33.2	40.8	41.5	50.0

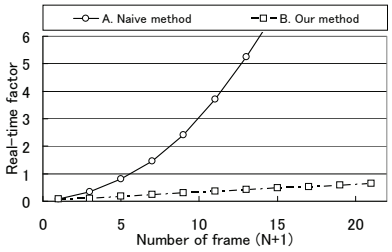

Fig. 3. Real Time Factor in Exp.A

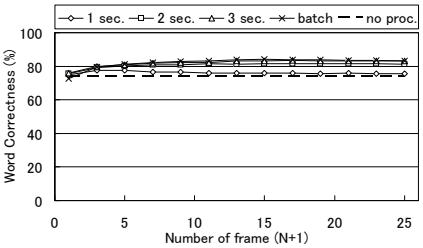

Fig. 4. Results of Exp.A in Env.I

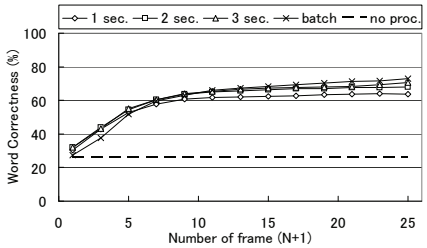

Fig. 5. Results of Exp.A in Env.II

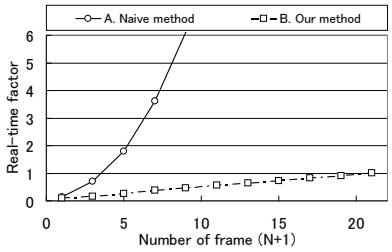

Fig. 6. Real Time Factor in Exp.B

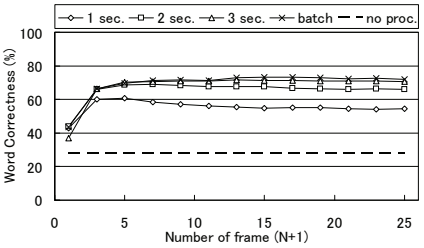

Fig. 7. Results of Exp.B in Env.I

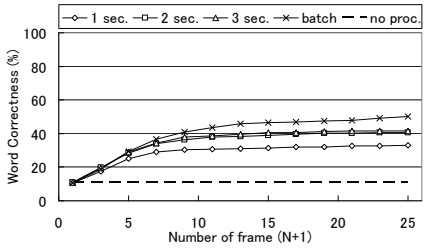

Fig. 8. Results of Exp.B in Env.II

points in the two environments, respectively. We can see that the RTF and the number of frame N is proportional relationship unlike naive method. We concluded that our method works well in almost all situations.

The number of samples seriously affects the performance, but around 2–3 sec. data seem to be sufficient. To improve the performance more, we must improve or develop adaptive step-size scheduling and adaptive frame-length N estimation for our method. In fact, performance can still be improved by 1 – 5 points by changing parameter λ according to all conditions. These parameter optimization is future work.

6. CONCLUSION

We developed a robot audition system that enabled barge-in for smooth speech interaction. To suppress reverberation and robot's speech, we introduced a MINT-based model of observation to FD-ICA. We reduced the calculation cost by using two techniques; 1) a separation model based on the independence between a direct sound signal and observed signals, and 2) enforced spatial sphering. The experimental results demonstrated the effectiveness of our methods.

In the future, we intend to work on step-size scheduling and adapting N for real-time implementation. We also intend to analyze the estimated filter and evaluate it when there are other sound sources in parallel. To accomplish more efficient processing, we need to integrate it with other methods according to the existing conditions.

7. REFERENCES

[1] 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.

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

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

[4] 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.

[5] 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.

[6] M. Miyoshi *et al.*, "Inverse filtering of room acoustics," *IEEE Trans. Acoust., Speech & Signal Process*, vol. 36, no. 2, pp. 145–152, 1988.

[7] A. Hyvarinen *et al.*, *Independent Component Analysis*, Wiley-Interscience, 2001.

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

[9] 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.

[10] H. Sawada *et al.*, "Polar coordinate based nonlinear function for frequency-domain blind source separation," *IEICE Tr. Fundamentals*, vol. E86-A, no. 3, pp. 505–510, 2003.

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

[12] S. Makino *et al.*, "Exponentially weighted stepsize nlms adaptive filter based on the statistics of a room impuse response," *IEEE Trans. on Speech & Audio Proc.*, vol. 1, no. 1, pp. 101–108, 1993.

[13] A. Lee *et al.*, "Julius – an open source real-time large vocabulary recognition engine," in *EUROSPEECH*, 2001, pp. 1691–1694.