2008 IEEE/RSJ International Conference on Intelligent Robots and Systems
Acropolis Convention Center
Nice, France, Sept, 22-26, 2008

Barge-in-able Robot Audition Based on ICA and Missing Feature Theory under Semi-Blind Situation

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

Abstract— **This paper describes a robot audition system that allows the user to barge-in; that is, the user can speak simultaneously when the robot is speaking. Our** *"barge-in-able"* **system consists of two stages: (1) cancellation of robot speech and (2) recognition of the separated user speech under the "semi-blind situation". The semi-blind situation is where a robot's speech signal is known but a user's speech signal is not. The first stage is achieved by using an adaptive filter based on time-frequency domain Independent Component Analysis, because that can separate robot speech more robustly against noise than conventional echo cancellers. To improve performance in on-line processing, we utilized known source normalization and the exponentially weighted stepsize method. The second stage is achieved by automatic speech recognition (ASR) based on the missing feature theory which provides robust recognition by exploiting the reliability of speech features distorted due to noise and/or separation. The semi-blind situation simplifies the estimation of such reliabilities. Experiments demonstrated that our system improved word correctness of ASR by 10.0 %.**

I. INTRODUCTION

A robot should recognize a target source from a mixture of sounds with the minimium amount of prior information because the robot has to work in unknown and/or dynamical environments. The mixture of sounds may include a robot's own speech because microphones are equipped on its body, not attached close to the mouth of a user. Therefore, the robot's own speech should be suppressed to enhance the user's speech. In human-robot or human-computer interaction, they may speak simultaneously when the robot is speaking. This situation is called "barge-in". A robot audition should be *"barge-in-able"* for smoother speech interaction so that the user does not necessarily need to wait until the robot finishes speaking.

Few research studies deal with barge-in-able systems from the viewpoint of robot audition because a conventional spoken dialogue system assumes a close-talking microphone. Miyabe *et al.* reported sound field control with many loudspeakers, and they applied independent component analysis (ICA) to semi-blind source separation to reduce the influence of echoes using in the speech of the system [1]. Creating silent zones around the microphones by placing many loudspeakers, they effectively separated the system speech by semi-blind source separation. Many loudspeakers are assumed to be installed in the environment, and their method is not suitable for robot audition.

R. Takeda, 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, komanita, ogata, okuno}@kuis.kyoto-u.ac.jp

K. Nakadai is with the Honda Research Institute Japan Co., Ltd., Wako, Saitama, 351-0114, Japan nakadai@jp.honda-ri.com

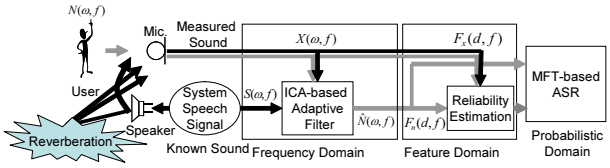

Fig. 1. Outline of our system: First, robot speech is separated by ICA-AF. Second, we estimate the reliability of features of separated sound. Finally, the separated sound is recognized by MFT-based ASR.

We solve the barge-in problem using two stages as shown in Fig. 1: (1) canceling the robot speech including its echoes and (2) recognizing the separated user's speech. Especially, we focused on **reverberations of the robot's speech** and designed these functions for real-time processing.

At the first stage, we use an adaptive filter based on ICA (ICA-AF) because

1) ICA-AF works well against noise, such as user's speech, unlike conventional echo canceling [1], [2], [3], and
2) ICA provides a natural interface to blind source separation (BSS), and beamforming method.

We have proposed time-frequency domain (TFD) ICA for echo cancellation (thereafter TFD-ICA-AF), and confirmed that it outperformed the time domain [2] and frequency domain (FD) ICA-AF [1] in terms of computational cost and performance for the case of batch processing [4].

This batch TFD-ICA-AF does not work well as itself when applied to on-line processing. This is because (A) the learning of the filter is insufficient, and (B) the performance depends on the stepsize (learning) parameter essentially, in on-line processing. These problems result in poor speech recognition performance.

We solved the problems by utilizing

I. known source (input vector) normalization for (B),
II. exponentially weighted stepsize method [5] for (B),
III. automatic speech recognition (ASR) based on missing feature theory (MFT) for (A).

The (I) and (II) methods can reduce the influence of the known source and the trasfer function to the stepsize, and improve the performace of cancellation with one fixed stepsize parameter. The effectiveness of (I) and (II) have already been confirmed experimentally in the adaptive filter, but neither in ICA-AF nor in terms of speech recognition as far as the authors know. In other words, the total performance of echo cancelling and speech recognition is essential in robot audition.

In the second stage, (III) MFT-ASR [6] copes with the remaining un-separated speech and the distorted parts caused

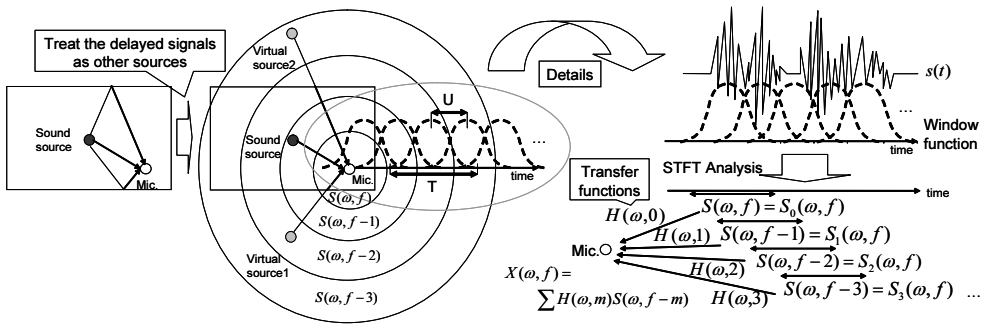

Fig. 2. Scheme of Time-Frequency Domain Convolution: $s(t)$ represents the original source signal, and $S(\omega, f)$ is the short-time Fourier Transform analysis of $s(t)$ with a window of size T and shift U. The observed signal $X(\omega, f)$ is the convolution of different frames and the transfer function $H(\omega, n)$. $S(w, f-1), S(w, f-2), \cdots$ are treated as **virtual sound sources**.

by insufficient separation. MFT-based ASR improves robustness against distortions caused by noise and/or the separation with the reliability of speech features. The performance of MFT-based ASR depends on the quality of reliability estimation. Common reliability-estimation methods are based on harmonics and pitch estimation, or time-frequency masking [7], [8]. We propose to use reliability estimation based on the result of TFD-ICA-AF, which reduces the computational cost of estimation. Thus, it is usefull for real-time processing.

II. Design of ICA-based Adaptive Filter

A. Modeling of Mixing and Unmixing Process

We modelled all processes about a source signal and an observed signal in the time-frequency (TF) domain. The merits are twofold; (1) it provides a natural interface to blind source separation (BSS), (2) features of speech for ASR can be extracted directly from separation result.

All signals in the time domain are analyzed by short-time Fourier transform (STFT) with a window of size T, and shift U. We assume that the original source spectrum $S(\omega, f)$ at time frame f and frequency ω affects the succeeding M frames of observed sound. Thus $S(\omega, f-1), S(\omega, f-2), \cdots, S(\omega, f-M)$ are treated as **virtual sound sources**. Fig. 2 depicts the scheme of the system. The observed spectrum $X(\omega, f)$ at a microphone is expressed as follows,

$$X(\omega, f) \quad = \quad N(\omega, f) + \sum_{m=0}^{M} H(\omega, m) S(\omega, f-m), \quad (1)$$

where $N(\omega, f)$ is the noise spectrum (user's speech) and $H(\omega, m)$ is the m th delay's transfer function in the TF domain. If $M = 0$, $X(\omega, f)$ represents the conventional instantaneous mixing model in the frequency domain. This TF model can be considered as multirate processing with an FFT filterbank.

The unmixing process for ICA is represented as:

$$\begin{pmatrix} \hat{N}(\omega, f) \\ \boldsymbol{S}(\omega, f) \end{pmatrix} = \begin{pmatrix} a(\omega) & -\boldsymbol{w}^T(\omega) \\ \mathbf{0} & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} X(\omega, f) \\ \boldsymbol{S}(\omega, f) \end{pmatrix}, \quad (2)$$

$$\boldsymbol{S}(\omega, f) = [S(\omega, f), S(\omega, f-1), \ldots, S(\omega, f-M)]^T \quad (3)$$

$$\boldsymbol{w}(\omega) = [w_0(\omega), w_1(\omega), \ldots, w_M(\omega)]^T, \quad (4)$$

where $\boldsymbol{S}$ and $\hat{N}(\omega, f)$ are a source spectrum vector and an estimated noise spectrum, respectively. $\boldsymbol{w}$ is an unmixing filter vector. $a(\omega)$ is a nonzero complex value. The unmixing process is described as a linear system with ICA and it is easy to integrate with FD-BSS.

B. Estimation of the Unmixing Filter Vector

An algorithm based on minimizing the Kullback-Leibler divergence (KLD) with a natural (relative) gradient is commonly used to estimate the unmixing filter, $\boldsymbol{w}(\omega)$, in Eq.(2). Based on KLD, we applied the following iterative equations with non-holonomic constraints [9] to our model because of fast convergence,

$$\boldsymbol{w}(\omega, f+1) = \boldsymbol{w}(\omega, f) + \mu_1 \phi_{\hat{N}(\omega)} \left(\hat{N}(\omega, f) \right) \bar{\boldsymbol{S}}(\omega, f), \quad (5)$$

$$\phi_x(x) = -\frac{\mathrm{d} \log p_x(x)}{\mathrm{d}x}, \quad (6)$$

where μ_1 is a step-size parameter that controls the speed of convergence, and $\bar{y}$ represents the conjugate of y. $p_y(y)$ is defined as the probability distribution of y.

Here, because of the non-holonomic constraint, $a(\omega)$ is not updated and remains the initial value. This means that we can decide the value of $a(\omega)$ arbitrarily, and hence we set it to 1. We also should decide the mean and variance of $\hat{N}(\omega, f)$, because the algorithm uses probability distribution $p_{\hat{N}(\omega)}(\hat{N}(\omega))$. Since $p_{\hat{N}(\omega)}(\hat{N}(\omega))$ should be a variance-normalized distribution that satisfies $E[1 - \phi_x(x\alpha_x)\bar{x}\bar{\alpha_x}] = 1$, we have to estimate the normalizing factor of $\hat{N}(\omega, f)$.

According to the KLD minimization with a natural gradient, the normalizing factor γ_x of x at frame f+1 is generally calculated by the online learning algorithm as follows,

$$\gamma_x(f+1) = \gamma_x(f) + \mu_x \left[1 - \phi_x \left(x(f)\gamma_x(f) \right) \bar{x}(f)\bar{\gamma_x}(f) \right] \gamma_x(f). \quad (7)$$

Then, the $\hat{N}(f)$ is normalized with the estimated normalizing parameter α by Eq. (7),

$$\hat{N}_n(f) \quad = \quad \alpha(f)\hat{N}(f). \quad (8)$$

C. Multirate-Repeating Method

We applied the multirate repeating method [10] to our TFD-ICA-AF to improve the convergence speed and to make

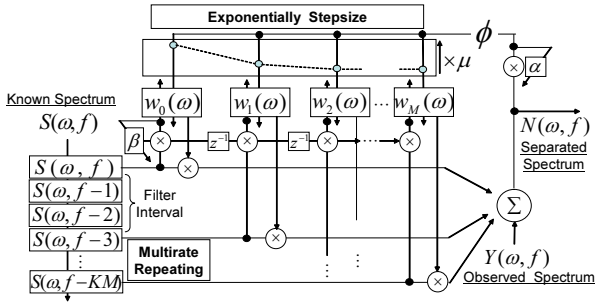

Fig. 3. Overview of ICA-AF part

the filter intervals independent of the shift size of STFT. In our previous work on TFD-ICA-AF, we found the shift size of STFT should be determined considering the performance and convergence of separation [4].

The multirate repeating method is equivalent to setting the interval of the filter K times longer than that of the input sample. Therefore, we changed the update equations as follows:

$$\hat{N}(f) = X(f) - \boldsymbol{S}_m^T(f)\boldsymbol{w}(f), \text{and} \quad (9)$$

$$\boldsymbol{S}_m(f) = [S(f), S(f-K), \cdots, S(f-MK)]^T, \quad (10)$$

where K is the filter-interval parameter.

D. Scaling Normalization of Input Vector

We normalized the power of the source spectrum $S(\omega, f)$ because the convergence speed also depends on it. For example, normalized-LMS method convergences faster than LMS for its input vector normalization [3]. This process is important for frequency-domain filtering because the power of the speech depends on the frequency. The Karhunen-Loeve Transformation (KLT) or ICA is ideal for normalization, but the computational cost is expensive [3].

This normalization is applied once per frame with the estimated scaling parameter β by Eq. (7):

$$S_n(f) = \beta(f)S(f), \text{and} \quad (11)$$

$$\boldsymbol{S}_n(f) = [S_n(f), S_n(f-K), \cdots, S_n(f-MK)]^T, \quad (12)$$

where $S_n(f)$ is the normalized input element and $\boldsymbol{S}_n(f)$ is the vector of $S_n(f)$.

E. Exponentially Weighted Stepsize

Makino *et al.* has proposed the exponetially weighted (EW) stepsize method for NLMS to improve the convergence speed by using the knowledge that the room impulse response decays exponentially [5]. We will reduce the influence of the room transfer function to the stepsize parameter with this method, because the filter coefficients estimated by TFD-ICA-AF also indicate exponential decay.

The stepsize of the i-th filter coefficient is decided as follows:

$$\mu_1(i) = \mu_1 \lambda^{-ic}, (i = 0, 1, ..., M), \quad (13)$$

where c is a decay-rate parameter that depends on the room reververation time.

F. On-Line Algorithms, Summary

The algorithms are summarized as follows and shown in Fig. 3 (ω is omitted for the sake of readability),

$$\boldsymbol{S}_m(f) = [S(f), S(f-K), \cdots, S(f-MK)]^T, \quad (14)$$

$$\hat{N}(f) = X(f) - \boldsymbol{S}_m(f)^T\boldsymbol{w}(f), \quad (15)$$

$$\hat{N}_n(f) = \alpha(f)\hat{N}(f), \quad S_n(f) = \beta(f)S(f), \quad (16)$$

$$\boldsymbol{S}_n(f) = [S_n(f), S_n(f-K), \cdots, S_n(f-MK)]^T, \quad (17)$$

$$\boldsymbol{w}(f+1) = \boldsymbol{w}(f) + \boldsymbol{\mu}_1\phi_{N_n}(\hat{N}_n(f))\bar{\boldsymbol{S}}_{\boldsymbol{n}}(f), \quad (18)$$

$$\alpha(f+1) = \alpha(f) + \mu_2[1 - \phi_{N_n}(\hat{N}_n(f))\bar{\hat{N}}_n(f)]\alpha(f), \quad (19)$$

$$\beta(f+1) = \beta(f) + \mu_2[1 - \phi_{S_n}(S_n(f))\bar{S}_n(f)]\beta(f), \text{and} \quad (20)$$

$$\boldsymbol{\mu}_1 = \text{diag}(\mu_1, \mu_1\lambda^{-c}, \cdots, \mu_1\lambda^{-cM}). \quad (21)$$

In particular, if the nonlinear function ϕ_x is in the form of $\phi_x(x) = r(|x|, \theta(x))e^{j\theta(x)}$, both $\alpha(f)$ and $\beta(f)$ become **real positive values**. For example, power-bounded function, $\phi(x) = \tanh(|x|)e^{j\theta(x)}$ is often used for super-gaussian distribution [11]. Since a speech signal is usually approximated by it, we use the nonlinear function.

Remark: If x follows a normalized gaussian distribution, $\phi_x(x)$ is converted to x. By applying $\phi_x(x) = x$ to Eq. (9), this simplification leads to the algorithm of LMS. LMS implicitly assumes that the distribution is known.

III. Integration of MFT-based ASR and ICA-based Adaptive Filter

A. Missing Feature Theory-based ASR

MFT-based ASR is similar to a hidden Markov model (HMM) -based recognizer, and the only difference is in their decoding processes. The output probability in the HMM is modified by the reliability $M(i)$ of the i-th acoustic features [12]. Here, we use a binary reliability as $M(i)$; 1 for reliable, and 0 for unreliable because of its low computational cost and good performance.

B. Reliability Estimation of the Separated Speech

We should estimate the reliability of the separated signal $\hat{N}(\omega, f)$ for MFT-based ASR. Our strategy is eliminating all suspicious features, because misestimation of unreliable features degrades recognition performance more severely than that of reliable features [7]. We just estimate the reliability of the separated speech that is influenced by a remained robot's speech which is caused by the insufficient learning of the separation.

We have two signals, the separated user's speech $\hat{N}(\omega, f)$ and the observed signal $X(\omega, f)$ from microphone. We assume that the difference between these signals is proportional to the robot's speech signal $S(\omega, f)$. Hence, the large difference between features of these two signals indicate that the features are affected by robot's speech $S(\omega, f)$ but not by other factors.

Let us denote $F_n(d, f)$ and $F_x(d, f)$ as the features of $\hat{N}(\omega, f)$ and $X(\omega, f)$ at frame f, and size d, respectively. The reliability $M(d, f)$ is calculated as follows;

$$M(d,f) = \begin{cases} 1, & |F_n(d,f) - F_x(d,f)| < T_{th} \\ 0, & otherwise \end{cases}, \quad (22)$$

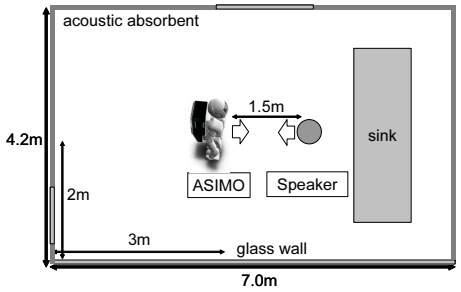

Fig. 4. The reverberation time (RT20) is 240 msec.

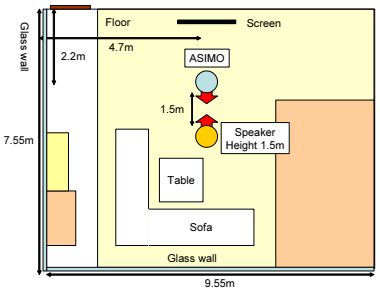

Fig. 5. The reverberation time (RT20) is 670 msec.

TABLE I
CONFIGURATION OF SEPARATION

Impulse Response	16kHz sampling
Reverberation time (RT20)	240 msec, 670 msec.
STFT setup	hanning:64 msec, overlap: 56 msec
Distance	1.5 m
Input data	[-1.0 1.0] normalized

TABLE II
CONFIGURATION OF ASR

TestSet	1 male and 1 female (each 200 words)
TrainingSet	11 males and 12 females (each 216 words)
Acoustic Model	Triphone: 3-state 4-mix. HMM
Language Model	Grammar
Feature	MFCC, 25 dimensions (12+Δ12+ΔPow)

where T_{th} is a threshold. If the feature includes the delta parameters, T_{th} should be changed accordingly. The computational cost is paid only in extracting feature of $X(\omega, f)$.

IV. EXPERIMENTS

A. Experimental Setups

The impulse responses for speech data were recorded at 16 kHz in a room shown in Figs. 3 and 4. The reverberation time (RT) in a normal room shown in Fig. 3 is short, 240 msec, and the RT in a hall-like room shown in Fig. 4 is long, 670 msec, where the speech recognition is typically very difficult because of the influence of reverberation. The sizes of the rooms were 4.2 × 7.0 m and 7.55 × 9.55 m, respectively. The speaker was 1.5 m away from a microphone mounted on the head of HONDA ASIMO. All data (16 bits, PCM) were normalized to $[-1.0 \;\; 1.0]$.

We selected 200 phonemically-balanced Japanese words for user speech, and they were convoluted with the recorded impluse responses. A male's speech was used for the robot's speech. The signal to noise ratios (SNRs) of user speech to robot speech were 0 dB and -10 dB.

Multi-band Julian [13] was used as the MFT-based ASR. The MFCC (12 + Δ12 + Δ Pow) was obtained after STFT with the window size 512 and shift size 160 for the speech features, and the cepstral mean normalization is applied to MFCC. A triphone-based acoustic model (3-state, 4-mixture) was trained with 216 words of clean speech uttered by 11 males and 12 females (word-closed). The training data sets do not include the data for the evaluation (speaker-open). These are summarized in Tabs. I, and II.

B. Experiment 1: TFD-ICA-AF on-line separation

We first examined the performance of TFD-ICA-AF with different frame lengths M in terms of word correctness (WC). In this experiment, we examined the effect of input-vector normalization and the exponentially weighted (EW) stepsize.

1) Without normalization and EW stepsize
2) With normalization and without EW stepsize
3) With normalization and EW stepsize

The parameters of this experiment were the step-size parameters (μ_1, μ_2), the window size T, shift size U, filter interval K, length of frame M, and the EW stepsize parameters λ, c. We chose values of 1024 (64 msec) for T, 128 (8 msec) for U, 3 for K, 0.75 for λ, and 1.0 for c. Since the FD-BSS has an optimal windows size [14], we select the appropriate $T = 1024$ considering the possibility of the integration with FD-BSS. $K = 3$ is suboptimal for the separation in an off-line experiment [4]. Since it is impossible to set the best learning step-size parameters (μ_1, μ_2), the they were set to the same value as μ, and tried as 1.0×10^k, and 5.0×10^k $(k = -3, -2, -1)$.

Before separating the mixture of sounds, TFD-ICA-AF separates about 3 seconds of impulse-response convoluted data for learning the initial value of the unmixing filter $\boldsymbol{w}(\omega)$ and the initial scaling parameters $(\alpha(\omega), \beta(\omega))$. After separation, we resynthesized a time-domain waveform from all the separated data.

C. Experiment 2: Reliability Estimation

This experiment confirmed the relationship between the threshold and the total improvement in WC with ICA-AF and MFT-based ASR. For threshold T_{th}, we tried 50 values, i.e., $n10^k$, $(n = 1, 2, ..., 9,\ k = -3, -2, -1, 1, 2)$, and checked the relationship between WC and the frame length M with the best threshold T_{th}.

We did not use the reliability of the delta features (Δ12 of MFCC) because that does not contribute to WC. We selected well separated user's speech from Experiment 1 for this experiment.

V. RESULTS

A. Experiment 1: TFD-ICA-AF on-line separation

The improvement in word correctness (WC) attained by the TFD-ICA-AF is summarized in Figs. 6 and 7. The reverberation time (RT) covered by filter in the graph indicates

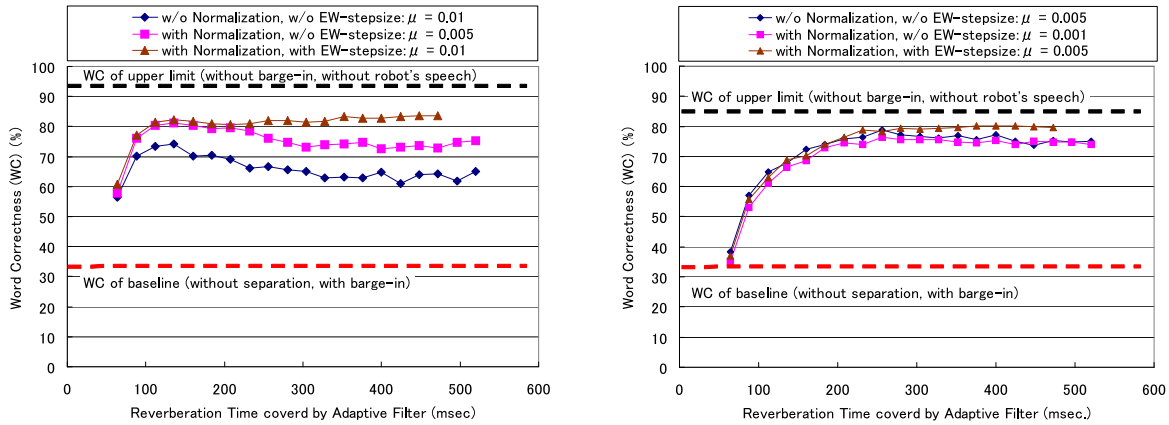

Fig. 6. Word Correctness of TFD-ICA-AF (SNR 0 dB). RT20s are 240 msec (left figure) and 670 msec (right figure).

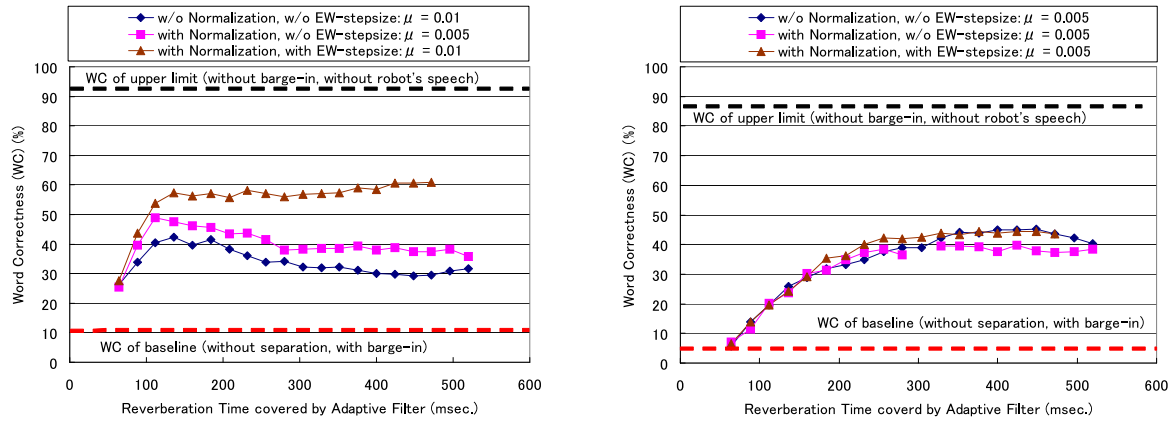

Fig. 7. Word Correctness of TFD-ICA-AF (SNR -10 dB). RT20s are 240 msec (left figure) and 670 msec (right figure).

the time-domain filter length of TFD-ICA-AF. For example, with $T = 1024$ (64 msec), $U = 128$ (8 msec), $K = 3$, and $M = 4$, the filter RT is $64 + 8 \times 3 \times 4 = 160$ msec. The best learning parameters μ are also shown in the figures.

The WCs without separation in the RT of 240 msec are 34.8% with 0 dB and 10.0% with -10 dB. And in the RT of 670 msec, the WCs are 36.8% at 0 dB and 6% at -10 dB. The upper limit indicates WC without barge-in and the WCs are 94.3% at 240 msec and 88.8% at 670 msec.

Adequate filter RT changes according to the RT of the room. In the RT of 670 msec, filter RT should also be set long. Therefore, adapting the filter RT to the RT of the environment dynamically is desirable.

With input vector normalization, WC is improved in the short RT environment compared to the WC without normalization. Nnormalization makes the stepsize independent of the input vector power. However, the normalization also decreases WC in the long RT because of the inadequacy of μ caused by other factors.

By applying the exponentially weighted (EW) stepsize, the WCs are improved well in both the short and long RT. In paticular, in the RT of 240 msec, performances with long filter RT are not worse than those with short filter RT. This is because the inadequacy of the stepsize caused by the transfer function is resolved by the knowledge of the filter coefficient. We can say EW stepsize is robust against the missselection of filter RT.

B. Experiment 2: Reliability Estimation

The relationship between WC and the threshold T_{th} is shown in Fig. 8. The WC with a larger threshold such as $T_{th} = 10^2$ is equal to the result obtained without using the missing feature technique. The separated data are the best results of experiment 1.

As T_{th} becomes smaller, the WC improves in the short RT condition. The WCs with different SNR are improved by 8.0 - 12.0 % with a threshold parameter $T_{th} = 10^{-2}$. However, in the RT of 670 msec, reliability estimation affects the WC. This is because our estimation method in this paper does not consider distortions caused by reverberation. This issue remains for future study.

The relationship between WC and filter RT in the short RT environment using the reliability threshold $T_{th} = 10^{-2}$ is shown in Fig. 9. WCs are improved with any filter RT. This result suggests our method works robustly against SNR and filter RT in the normal reverberation room.

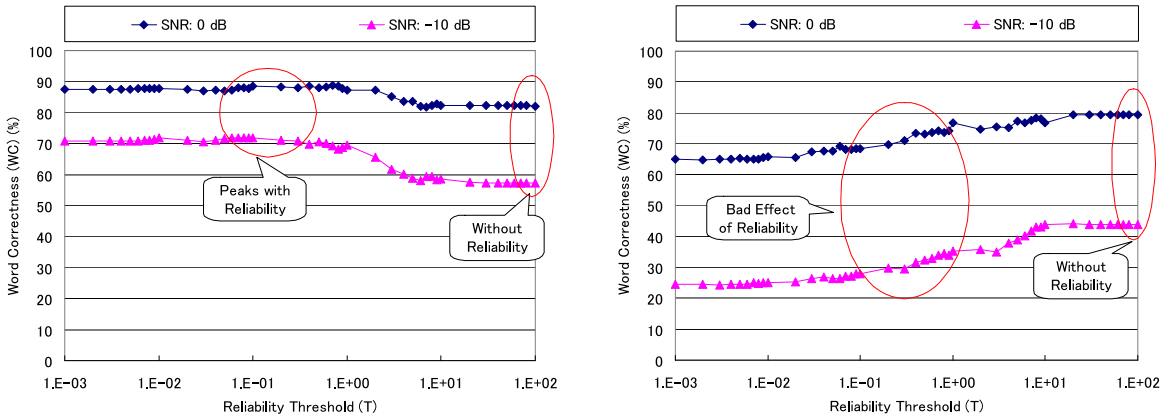

Fig. 8. WC with missing feature technique after TFD-ICA-AF. RT20s are 240 msec (left figure) and 670 msec (right figure).

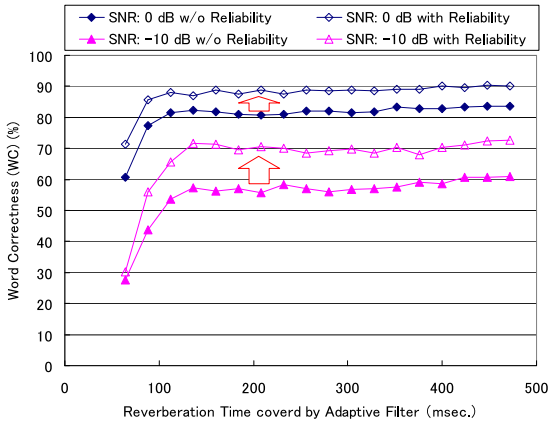

Fig. 9. WC of our system with $T_{th} = 10^{-2}$. RT20 of the room is 240msec

VI. CONCLUSION

We developed a barge-in-able robot audition system for smooth speech interaction. We designed the on-line TFD-ICA-AF and MFT-based ASR for a semi-blind situation. TFD-ICA-AF separates the known sound source even in the presence of a user's speech and reverberation. The input-vector normalization and exponentially weighted (EW) stepsize improved the performance in word correctness (WC) effectively. The reliablity of features is estimated by threshold processing according to the difference in features caused by TFD-ICA-AF with low computational cost. MFT-based ASR recognized the separated user speech and improved WC about 8.0 - 12.0 % after separation. With these two techniques, the total performance in the case of SNR 0dB almost reached the ideal performance in WC.

In the future, we will work on adaptation of the frame length M, stepsize μ, and λ, c in the EW stepsize, to the reverberation of the environment. For the reliability estimation, we must develop a hybrid method for the distortions caused by the separation, reverberation, and other noises. Eventually, we will try to develop a spoken dialogue system appropriate for the robot.

REFERENCES

[1] S. Miyabe, T. Takatani, Y. Mori, H. Saruwatari, K. Shikano, and Y. Tatekura, "Double-talk free spoken dialogue interface combining sound field control with semi-blind source separation," in *Proc. of IEEE ICASSP06*, 2006, vol. I, pp. 809–812.

[2] J.-M. Yang and H. Sakai, "A new adaptive filter algorithm for system identification using independent component analysis," in *Proc. of IEEE ICASSP07*, 2007, pp. 1341–1344.

[3] S. Haykin, *Adaptive Filter Theory*, Prentice-Hall, Upper Saddle River, NJ 07458, 4th edition, 1991.

[4] R. Takeda, K. Nakadai, K. Komatani, T. Ogata, and H. G. Okuno, "Exploiting known sound sources to improve ICA-based robot audition in speech separation and recognition," in *Proc. of IROS'07*, 2007.

[5] S. Makino, Y. Kaneda, and N. Koizumi, "Exponentially weighted stepsize nlms adaptive filter based on the statistics of a room impuse response," *IEEE Trans. on Speech And Audio Proc.*, vol. 1, no. 1, pp. 101–108, 1993.

[6] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," in *Speech Comm.*, 2000, vol. 34, pp. 267–285.

[7] M. L. Seltzer, B. Raj, and R. M. Stern, "A bayesian framework for spectrographic mask estimation for missing feature speech recognition," in *Speech Comm.*, 2004, vol. 43, pp. 379–393.

[8] D. Kolossa, H. Sawada, R. F. Astudill, R. Orglmeister, and S. Makino, "Recognition of convolutive speech mixtures by missing feature techniques for ICA," in *Proc. of ACSSC 06*, 2006, pp. 1397–1401.

[9] S. Choi, S. Amari, A. Cichocki, and R. Liu, "Natural gradient learning with a nonholonomic constraint for blind deconvolution of multiple channels," in *Proc. of Int'l Workshop on ICA and BBS*, 1999, pp. 371–376.

[10] H. Kiya, K. Nishikawa, and K. Ashihara, "Improvement of convergence speed for subband adaptive digital filter using the multirate repeating method," *Electronics and Communications in Japan, Part III*, vol. 78, no. 10, pp. 37–45, 1995.

[11] H. Sawada, R. Mukai, S. Araki, and S. Makino, "Polar coordinate based nonlinear function for frequency-domain blind source separation," *IEICE Trans. Fund.*, vol. E86-A, no. 3, pp. 505–510, 2003.

[12] S. Yamamoto, J.-M. Valin, K. Nakadai, J. Rouat, F. Michaud, T. Ogata, and H. G. Okuno, "Enhanced robot speech recognition based on microphone array source separation and missing feature theory," in *Proc. of IEEE ICRA05*, 2005, pp. 1489–1494.

[13] Y. Nishimura, T. Shinozaki, K. Iwano, and S. Furui, "Noise-robust speech recognition using multi-band spectral features," *Acoustical Society of America Journal*, vol. 116, pp. 2480–2480, 2004.

[14] S. Araki, R. Mukai, S. Makino, T. Nishikawa, and H. Saruwatari, "The fundamental limitation of frequency domain blind source separation for convolutive mixtures of speech," *IEEE Trans. on Speech and Audio Proc.*, vol. 11, pp. 109–116, 2003.