Proceedings of the 2007 IEEE/RSJ International
Conference on Intelligent Robots and Systems
San Diego, CA, USA, Oct 29 - Nov 2, 2007

Exploiting Known Sound Source Signals to Improve ICA-based Robot Audition in Speech Separation and Recognition

Ryu Takeda†, Kazuhiro Nakadai‡, Kazunori Komatani†, Tetsuya Ogata†, Hiroshi G. Okuno†
†*Graduate School of Informatics, Kyoto University, Sakyo, Kyoto 606-8501, Japan*
{*rtakeda, komatani, ogata, okuno*}*@kuis.kyoto-u.ac.jp*
‡*Honda Research Institute, Japan, 8-1 Honcho, Wako, Saitama, 351-0114, Japan nakadai@jp.honda-ri.com*

***Abstract*— This paper describes a new semi-blind source separation (semi-BSS) technique with independent component analysis (ICA) for enhancing a target source of interest and for suppressing other known interference sources. The semi-BSS technique is necessary for double-talk free robot audition systems in order to utilize known sound source signals such as self speech, music, or TV-sound, through a line-in or ubiquitous network. Unlike the conventional semi-BSS with ICA, we use the time-frequency domain convolution model to describe the reflection of the sound and a new mixing process of sounds for ICA. In other words, we consider that reflected sounds during some delay time are different from the original. ICA then separates the reflections as other interference sources. The model enables us to eliminate the frame size limitations of the frequency-domain ICA, and ICA can separate the known sources under a highly reverberative environment. Experimental results show that our method outperformed the conventional semi-BSS using ICA under simulated normal and highly reverberative environments.**

I. INTRODUCTION

Robot audition systems should be robust against unknown or known noises, because robots are expected to work even in unknown and/or dynamically-changing environments. For example, in real environments, people often talk at the same time. This situation is called "double-talk". In human-robot interactions, or in human-computer interactions, while the robot or the system speaks, the user speaks at the same time. This situation is called "barge-in". Robot audition in human-robot interactions should be double-talk free. Here, we use double-talk to include barge-in.

The idea in attaining double-talk free robot audition is to use the fact that the robot knows the original signals of its utterance, in either digital or analog format. Usually, the robot hears by its ears (microphones) sounds distorted by spatial transfer functions including the influence of reflection and echoes. If the system knows original sound source signals, the problems in sound source separation and recognition may be alleviated and thus robot audition performance is expected to improve.

In this paper, we present a new method called "semi-blind source separation" by exploiting known source signals to attain double-talk free robot audition. Barge-in is a well-known problem in the speech recognition community and many researchers have attacked it. We extend the concept of barge-in or double-talk to include the cases where the robot knows the original sound source signals such as line-in signals of TV or audio devices. In these cases, the robot need not separate such signals from observed ones, since it knows the original ones. The problem in the semi-blind source separation is to estimate signals affected by the spatial transfer functions and to improve the performance of sound source separation (SSS) and recognition of normal blind source separation. Such double-talk free robot audition system requires SSS with less *a priori* information on an environment, because the environment around robots is usually unstable.

Only a few researchers have focused on the speech separation and the recognition of separated sounds. The humanoid *HRP-2* can localize and separate a mixture of sounds and recognize speech commands in noisy environments [1] with adaptive beam former. Yamamoto *et al.* developed a new interfacing scheme between source separation and automatic speech recognition (ASR) based on the Missing Feature Theory (MFT) [2]. Takeda *et al.* succeeded in integrating independent component analysis (ICA) and MFT-based ASR and created a robot audition system with less *a priori* information on the environment [3]. However, these systems do not handle the cancellation of known noises or double-in free interaction.

For a double-talk free spoken dialogue system, Miyabe *et al.* proposed semi blind source separation (semi-BSS) with frequency domain (FD)-ICA, which uses the known sound waveform information [4]. Semi-BSS does not require double-talk detection, unlike many types of acoustic echo cancellers (AECs), e.g., single channel, stereophonic, wave synthesis, and beamformer-integrated [5], [6], [7]. The conventional AECs need a duration in which only known noise is emitted, because adaptation to double-talk duration is difficult. Semi-BSS with ICA is also advantageous when more than two noises and unknown noises exist. However, FD-ICA limits the performance derived from the relationship between the independency and the window size of the short-time frequency transformation (STFT) analysis [8], and it degrades the suppression of known noise under a more reverberative environments.

To cope with this problem, we use the convolution in the time-frequency (TF)-domain to describe the reflection sounds and a new mixing process model for FD-ICA. This means we consider the reflections during some delay time as sounds originating from different sound sources, even if they originate from the same sound source. Therefore, FD-ICA separates the reflections as other sources. With this model,

1-4244-0912-8/07/$25.00 ©2007 IEEE.

we can effectively suppress the known sources in highly reverberative and noisy environments without the limitation of the FD-ICA.

The rest of the paper is organized as follows: Section Ⅱ explains the conventional semi-BSS with ICA. Section Ⅲ presents out semi-BSS with a time-frequency domain convolution model. Section Ⅳ describes the experiments by simulation, and Section V discusses the results. Section Ⅵ concludes the paper.

II. CONVENTIONAL SEMI-BLIND SOURCE SEPARATION WITH ICA

A. MIXING PROCESS

The signals observed by a process of linearly mixing sound sources in the time domain are expressed as

$$\hat{\boldsymbol{x}}(t) = \sum_{m=0}^{M-1} \hat{\boldsymbol{a}}(m)\boldsymbol{s}(t-m), \tag{1}$$

where $\hat{\boldsymbol{x}}(t) = [x_1(t), \ldots, x_J(t)]^T$ is the observed signal vector, and $\boldsymbol{s}(t) = [s_1(t), \cdots, s_I(t)]^T$ is the source signal vector. J and I are the number of microphones and sound sources respectively. In addition, $\hat{\boldsymbol{a}}(n) = [a_{ij}(n)]_{ij}$ is the mixing filter matrix of length M, where $[X]_{ij}$ denotes a matrix that includes the element X in the ith row and the jth column.

For the semi-BBS, the signals, $s_K(t), \cdots, s_I(t)$, are already known sound sources, and the mixing process is re-described as follows,

$$\boldsymbol{x}(t) = \sum_{n=0}^{N-1} \boldsymbol{a}(n)\boldsymbol{s}(t-n) \tag{2}$$

$$a_{ij}(n) = \begin{cases} 1 & (K \le j \le I \ and \ n = 0 \ and \ j = i) \\ 0 & (K \le j \le I \ and \ (n \ge 1 \ or \ j \ne i)) \end{cases}, \tag{3}$$

where $\boldsymbol{x}(t) = [x_1(t), \ldots, x_J(t), s_K(t), \ldots, s_I(t)]^T$ is the newly observed signal vector, and $\boldsymbol{a}(n) = [a_{ij}(n)]_{ij}$ is the new mixing matrix.

B. FD-ICA FOR SEMI-BLIND SOURCE SEPARATION

FD-ICA is often used to solve the unmixing problem because ICA converges faster in the frequency domain than in the time domain (TD)-ICA.

We expressed the original source signal, $s(t)$, as the signal, $S(\omega, f)$, at the fth frame and the ωth frequency bin in the time-frequency (TF) domain by a short-time analysis with the window frame size, T, and the shift size, U. The observed signal, $x(t)$, is also expressed as $X(\omega, f)$, and we obtain the observed vector $\boldsymbol{X}(\omega, f) = [X_1(\omega, f), \ldots, X_J(\omega, f), S_K(\omega, f), \ldots, S_I(\omega, f)]^T$. The unmixing process can be formulated in a frequency bin as

$$\boldsymbol{Y}(\omega, f) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega, f), \tag{4}$$

$$w_{ij}(\omega) = \begin{cases} 1 & (K \le j \le I \ and \ j = i) \\ 0 & (K \le j \le I \ and \ j \ne i), \end{cases} \tag{5}$$

where $\boldsymbol{Y}(\omega, f) = [Y_1(\omega, f), \ldots, Y_{K-1}(\omega, f), S_K(\omega, f), \ldots, S_I(\omega, f)]^T$ is the estimated source signal vector, and $\boldsymbol{W}(\omega) = [w_{ij}(\omega)]_{ij}$ represents an unmixing matrix in a frequency bin.

An algorithm based on the minimization of Kullback-Leibler divergence is often used on speech signals to estimate the unmixing matrix, $\boldsymbol{W}(\omega)$, in Eq. (4). Based on KLD, we used the following iterative equation with non-holonomic constraints [9].

$$\boldsymbol{W}^{[j+1]}(\omega) = \boldsymbol{W}^{[j]}(\omega) - \alpha\{\text{off-diag}\langle\boldsymbol{\phi}(\boldsymbol{Y})\boldsymbol{Y}^h\rangle\}\boldsymbol{W}^{[j]}(\omega), \tag{6}$$

where α is a step-size parameter that controls the speed of convergence, $[j]$ expresses the value of the jth step in the iteration, and $\langle\cdot\rangle$ denotes the time-averaging operator. The operation, off-diag($\boldsymbol{X}$), replaces each diagonal element of matrix X with zero. The nonlinear function, $\boldsymbol{\phi}(\boldsymbol{y})$, is defined as $\phi(y_i) = \tanh(|y_i|)e^{j\theta(y_i)}$ [10].

FD-ICA for the semi-blind source separation is achieved by not updating the elements of $\boldsymbol{W}$ related to the known sound source $S_K(\omega, f), \ldots, S_I(\omega, f)$. Obviously, these elements are constant values of $1, 0$ and do not require learning for separation.

III. OUR SEMI-BLIND SOURCE SEPARATION

A. TIME-FREQUENCY DOMAIN CONVOLUTION MODEL AND NEW MIXING PROCESS

We considered all the processes from the source signal to the observed signal in the TF domain because the source separation and the extraction of the features for ASR are usually done in the domain. The key idea for dealing with the reflections is to consider the reflections as different sounds and separate them with BSS.

We assumed that the reflections of the sound affects the succeeding frames' observed sound, $X(\omega, f)$. With the number of assumed reflection filter (NRF), N, it is formulated as

$$X(\omega, f) = \sum_{n=0}^{N} A(\omega, n)\hat{S}(\omega, f-n), \tag{7}$$

where $A(\omega, n)$ is the nth delay's transfer function in the TF domain. This model describes the N frames' transfer function, and it can deal with highly dereverberated sound that includes many reflections. Fig. 1 illustrates them.

Then, the mixing process of sounds is redefined as

$$X_j(\omega, f) = \sum_{i=0}^{I}\sum_{n=0}^{N} A_{ij}(\omega, n)\hat{S}_i(\omega, f-n), \tag{8}$$

where $\hat{S}_i(\omega, f)$ is the ith sound source spectrum at frame n, and $A_{ij}(\omega, n)$ denotes the transfer function from the sound source, i, to the microphone, j, with delay, n. Suppose that $\hat{S}_i(\omega, k)$ and $\hat{S}_j(\omega, l)$ $(k \ne l)$ are different sounds, the mixing process assumes the $(I+1)(N+1)$ sound sources are mixed. We express it as

$$\boldsymbol{X}(\omega) = \boldsymbol{H}(\omega)\boldsymbol{S}(\omega) \tag{9}$$

$$\boldsymbol{X}(\omega) = [X_1(\omega, f), X_2(\omega, f), \ldots, X_J(\omega, f)]^T \tag{10}$$

$$\boldsymbol{S}(\omega) = [S_1(\omega, f), S_1(\omega, f-1), \ldots, S_I(\omega, f-N)]^T \tag{11}$$

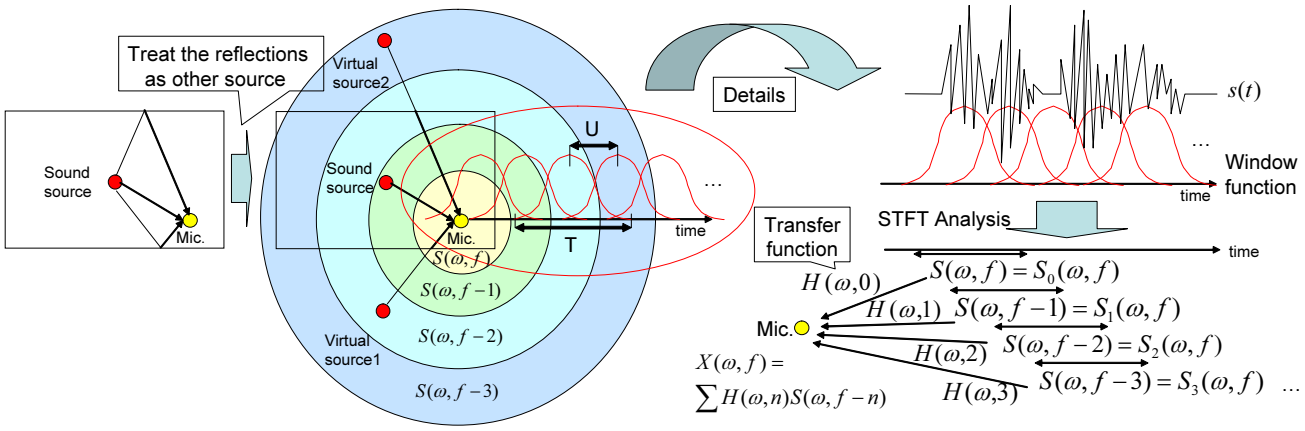

Fig. 1. Model of time-frequency convolution. $s(t)$ represents the original source signal, and $S(\omega, f)$ is the short-time Fourier analyzed $s(t)$ with window size T and shift size U. $S(\omega, t)$ with a different frame is treated as different (virtual) sources. Therefore, the observed signal, $X(\omega, f)$, is the convolution of these signals and the transfer function, $H(\omega, n)$, of each source.

$$\boldsymbol{H} = \begin{pmatrix} H_{11}(\omega) & H_{12}(\omega) & \dots & H_{1(I+1)(N+1)}(\omega) \\ H_{21}(\omega) & H_{22}(\omega) & \dots & H_{2(I+1)(N+1)}(\omega) \\ \vdots & \vdots & \ddots & \vdots \\ H_{J1}(\omega) & H_{J2}(\omega) & \dots & H_{J(I+1)(N+1)}(\omega) \end{pmatrix}, \quad (12)$$

where $\boldsymbol{X}$ is the observed signal vector, $\boldsymbol{H}$ is the mixing matrix, and $\boldsymbol{S}$ is the new source signal vector with size $(I+1)(N+1)$. Therefore, the new mixing process can be described as linear mixing.

B. ADAPTATION OF SEMI-BSS FD-ICA TO THE MODEL

To simplify the problems, we assume one known sound source and one unknown sound source. In addition, we did not need the convolution of the reflection of the unknown sound, because solving the convolution of unknown sounds is equal to dereverberation. We focused on eliminating the known sound. In this case, we treated the reflected sounds as one sound,

$$X(\omega, f) = \sum_{n=0}^{N} H(\omega, n) S(\omega, f-n) \quad (13)$$

$$= \left(\sum_{n=0}^{N} H(\omega, n) \frac{S(\omega, f-n)}{S(\omega, f)} \right) S(\omega, f) \quad (14)$$

$$= \hat{H}(\omega) S(\omega, f). \quad (15)$$

Therefore, the model of the mixing process is reduced to,

$$\boldsymbol{X}(\omega) = \boldsymbol{H}(\omega)\boldsymbol{S}(\omega) \quad (16)$$

$$\boldsymbol{X}(\omega) = [X_1(\omega, f), S_2(\omega, f), \dots, S_2(\omega, f-N)]^t \quad (17)$$

$$\boldsymbol{S}(\omega) = [S_1(\omega, t), S_2(\omega, f), \dots, S_2(\omega, f-N)]^t \quad (18)$$

$$\boldsymbol{H} = \begin{pmatrix} H_{11}(\omega) & H_{12}(\omega) & \dots & H_{1(1+N)}(\omega) \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{pmatrix}. \quad (19)$$

Then, the conventional FD-ICA is applied to this model.

Before applying FD-ICA, the whitening process is done to speed the convergence of the learning unmixing matrix for practical use. The problems specific to FD-ICA are ambiguities with scaling and permutation. We solved the ambiguities of scaling using the projection back method proposed by Murata [11]. The permutation problem does not need to be solved if the number of unknown sounds is one.

IV. EXPERIMENTS BY SIMULATION

We evaluated our method and conventional semi-BSS with ICA using the simulation data. The following two experiments were conducted,

A) An evaluation of the relationship between the number of filters and the noise reduction rate (NRR) and the convergence speed in the highly reverberated environment

B) An evaluation by speech recognition and NRR in the normal reverberated environment .

Experiment A) required examining whether the TF domain convolution model is efficient for a reverberative environment. Experiment B) shows the contribution to speech recognition for robot audition.

A. EXPERIMENTAL SETUP IN THE HIGHLY REVERBERATIVE ENVIRONMENT

1) RECORDING CONDITIONS: The impulse response data were recorded at 16 kHz in the room shown in Fig. 2. Speaker A was 2.0 m away from the microphone, and speaker B was 3.0 m away from speaker A. The room and reverberation time we used was 7.55×9.55×3.2 m and 0.9–0.93 sec (RT60), respectively.

2) EXPERIMENTAL CONDITIONS: A simultaneous speech signal was generated with impulse responses, and the speeches were Japanese sentences of about 13 sec each. We set speaker A as the unknown signal and B as the known signal. We used NRR for the criteria on how the separation was effective. They were calculated by the formulation

$$SNR_{tar} = 10 \log_{10} \frac{\sum_t \sum_\omega |S_{ori}(\omega, t)|^2}{\sum_t \sum_\omega |S_{tar}(\omega, t) - S_{ori}(\omega, t)|^2} \quad (20)$$

$$NRR = SNR_{sep} - SNR_{obs}, \quad (21)$$

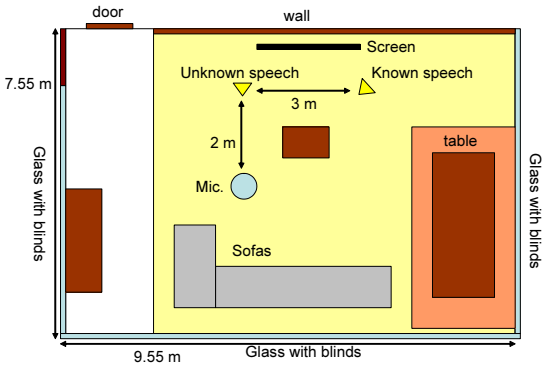

Fig. 2. Layout of room used for experiment A. Reverberation time (RT60) = 0.9–0.93 sec

where $S_{ori}(\omega, t)$ is the convoluted sound without any noise, and S_{obs} and S_{sep} are the observed mixture of sounds and separated sounds, respectively.

The parameters of this experiment are the shift size, U, and the number of assumed reflection filter (NRF), N. We obtained values from 160 to 500 with a difference of 40 for U. The window frame size, T, was 1,024 points (64 msec), and the learning parameter, α, was 0.45. The initial values for the unmixing matrix, $\boldsymbol{W}(\omega)$, were given at random. The SNR of the unknown signal in the observed signal was about -11 dB. We examined only NRR under these conditions because other techniques are needed for recognition under such high reverberative environment.

B. EXPERIMENTAL SETUP IN THE NORMAL REVERBERATIVE ENVIRONMENT

1) RECORDING CONDITIONS: In this experiment, we used two kinds of known sources, speech and music. The former involved the cancellation of self speech, while the latter involved cancellation of other known sound obtained through line-in or network. The impulse response data were also recorded at 16 kHz in this room, as shown in Fig. 3. The room we used was 7.0×9.0×3.2 m, and the reverberation time was 0.3–0.35 s (RT60). For the former, the angle, θ, from the front of the microphone was 0, 15, 30, 45, 60, 90, 270, 300, 315, 330, and 345 degrees, and for the latter, θ, was 0, 15, 30, 60, 90, 150, 180, 210, 270, 300, 330, and 345 degrees. The distance between the microphone and the speaker was 1.5 m in both cases.

2) EXPERIMENTAL CONDITIONS: We used combinations of two different words selected from a set of 200 phonemically balanced Japanese words for known and unknown speech sounds. For the music, 200 words were used as the unknown speech sounds. We used Julian [12] as the ASR, and the mel frequency cepstrum coefficients (MFCC: 12+Δ12+ΔPow) for the speech features. It uses a triphone-based acoustic model (3-state, 4-mixture) trained with 216 words of clean speech uttered by 22 male and female speakers. The training data sets do not include the data for the evaluation (open test).

The parameters of this experiment are the angle, θ, and

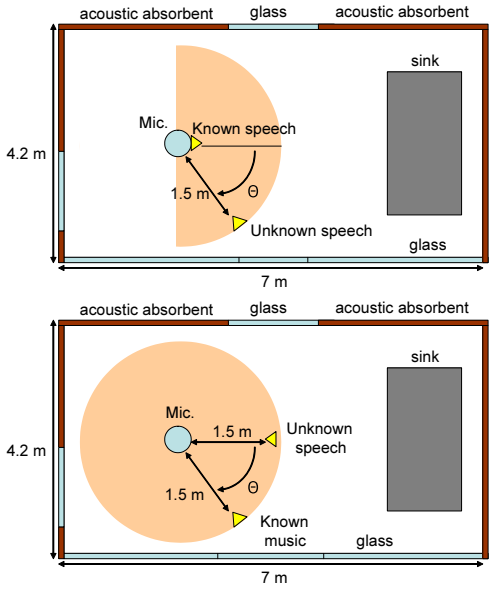

Fig. 3. Layout of room used for experiment B. The upper part is for the known speech situation, and the lower part is for the known music situation. Reverberation time (RT60) = 0.3–0.35 sec

NRR, N. The window frame size, T, and the shift size, U, were 1,024 points (64 msec) and 256 points, (16 msec) respectively, and the learning parameter α was 0.45. The initial values for the unmixing matrix, $\boldsymbol{W}(\omega)$, were given at random. The SNR of the unknown signal in the observed signal was set to about -11 dB. To examine the best performance, we estimated the unmixing matrix from data of 10 consecutive words of which length is 13 sec. The word correctness (WC) and the NRR were evaluated.

V. RESULTS

A. EXPERIMENT A

Fig. 4 is the waveforms of the observed sound, the convoluted sounds, the separated sound ($N = 1$) and the separated sound ($N = 15$). We can see how the separated sound is improved with large NRF. Fig. 5 shows the relationship between NRRs, NRF and shift size, U. As the number of filter increases, the NRR converges and improves about 17 points at shift size 280 and with 24 filters. About the shift size, U, the NRR peaks at $U = 280, 300$ different from NRF.

Next, we focus on the convergence speed. Fig. 6 reveals the relationship between the number of iteration, the number of assumed reflection filter (NRF), N, and the shift size U. The number of iteration increase as NRF increase at any shift size U. However, the increment speed is obviously different between small and large shift size.

From the results, our method outperformed the conventional method in the highly reverberative environment, and we moved beyond the limitations of the performance of the FD-ICA with the TF domain convolution model. And these two experiments about NRR and the number of iteration, we can say that a trade-off between the number of calculations and the performance of the separation apparently exists. Moreover, the shift size, U, affect the performance of our

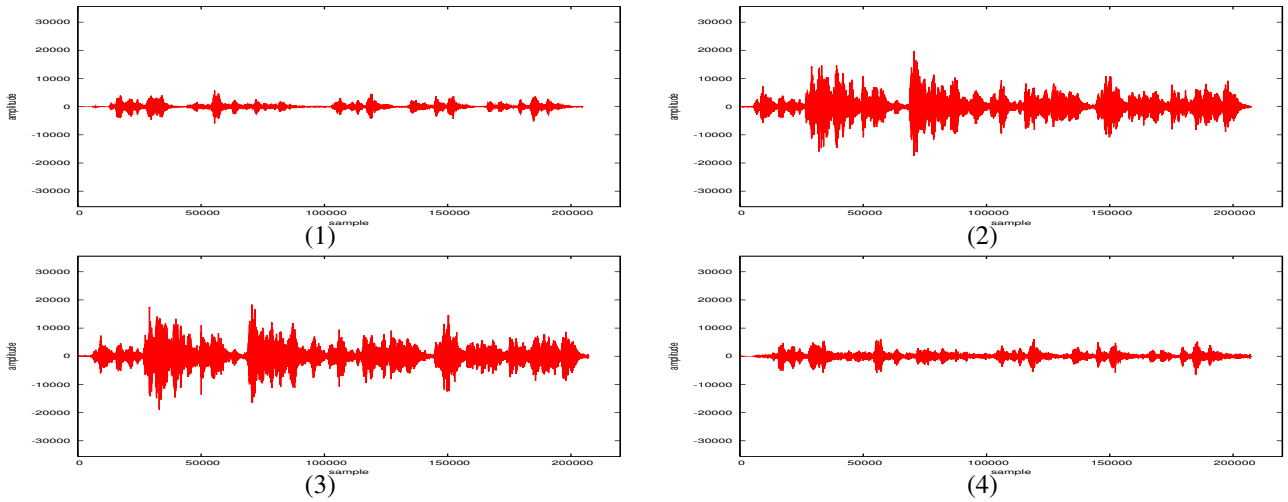

Fig. 4. Waveforms: (1) is the ideal signal of the unknown speech without noise. (2) is the observed signal. (3) is the separated signal with N = 1 (conventional semi-BSS ICA). (4) is the separated signal with N=15 (our method).

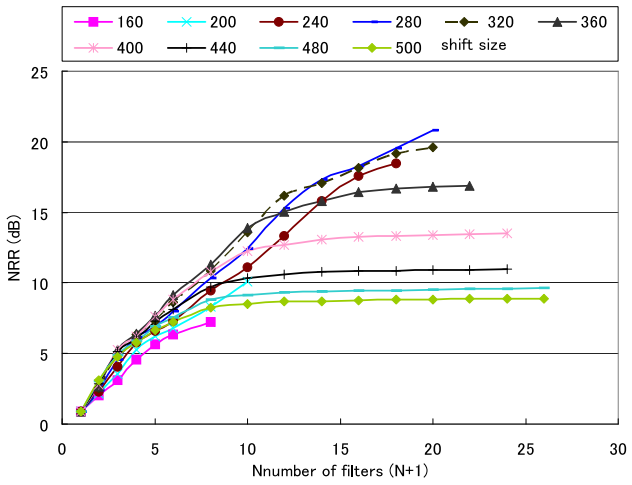

Fig. 5. Relationship between NRRs, shift size, and the number of assumed reflection filter, N.

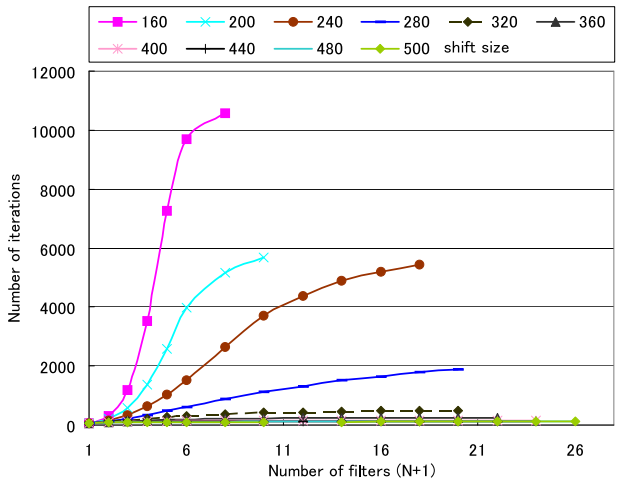

Fig. 6. Relationship between the number of iterations, shift size, and the number of assumed reflection filter, N.

method. If U is too large, the reflections cannot be separated well because of the effect of the sides of the window function. If the main power of the known source exits at the sides of the window, the model cannot deal with the reflection well. If U is too small, the independency among the original source $S(\omega, f - N)$ decreases, because the transfer function is almost the same. It is a factor of the slow convergence of the matrix.

B. EXPERIMENT B

Fig. 7 and 8 show the relationship between NRR and WC and NRF, respectively. The upper limit in Fig. 8 indicates the limitation of the WC at each angle, θ, that is, without noise. Obviously, both the NRR and WC are improved as the number of filters increases. Six NRF is enough to suppress the known sound in the normal reverberative environment. In particular, if the known sound is music, the performance with six filters is almost equal to that of the upper limit. And the WC of the music is better than that of the speech, because the noise is also words for recognition in the case of the known speech. Our method also outperformed the conventional method in this case.

The performance of the separation seems to saturate around $N = 6$. It is small number compared with the Experiment A. This indicates the validity of the TF-domain convolution model, because the optimal size differs with the reverberation time. In the two experiments, the highly reverberative environment requires a large NRF, and the normal reverberative environment requires a small NRF.

VI. CONCLUSION

We created a double-talk free robot audition system that can be used in unknown and/or dynamically-changing environments and deal with *a priori* information. To fulfill such requirements, we used FD-ICA with the TF domain convolution model. We achieved semi-BSS with the FD-ICA and the model, and it outperformed the conventional semi-BSS with ICA in two experiments under different conditions.

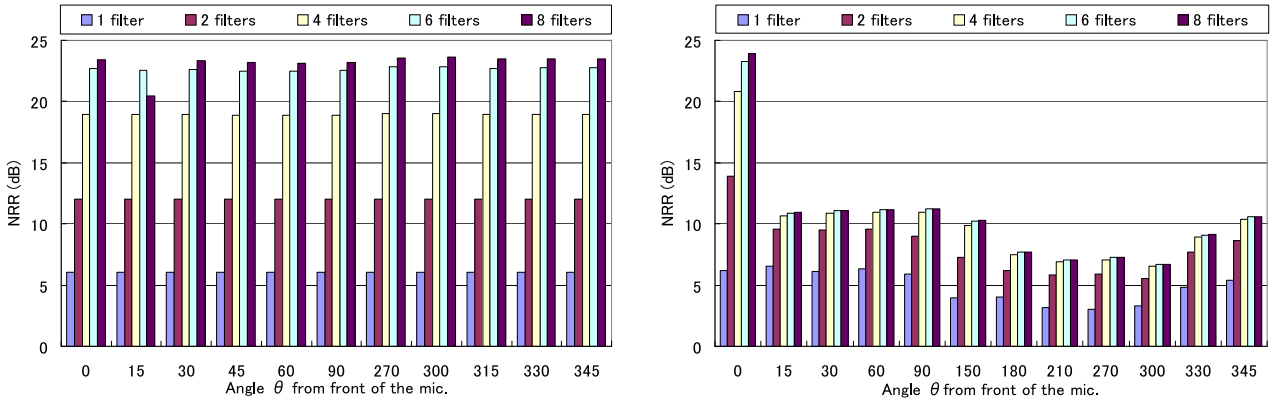

Fig. 7. Relationship between NRRs and NRF, N. The left part is the case of known speech, and the right part is that of known music.

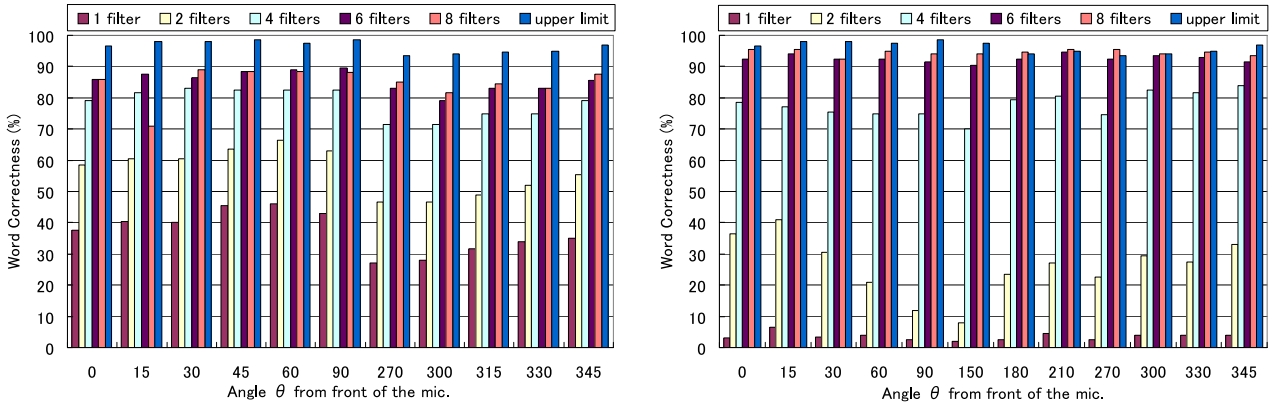

Fig. 8. Relationship between word correctness (WC) and number of filters. The left part is the case of known speech, and the right part is that of known music.

For improving the speech recognition rate, we have created a combination of ICA and MFT-based ASR to compensate for the remains of the separation. This technique improves the recognition rate by generating a more accurate missing feature mask because the interference sound source is already known.

In future work, we will first work on real-time processing of this technique and the optimal parameters for the shift size, the window frame size, and the initial values of the matrix. Integrating MFT-based ASR is the next challenge. After that, we will try to develop a better BSS with this model or a semi-BSS that has fewer constraint than this method.

REFERENCES

[1] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, F. Kanehiro, H. Hirukawa, and K. Yamamoto, "Robust speech interface based on audio and video information fusion for humanoid HRP-2," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*. IEEE, 2004, pp. 2404–2410.

[2] S. Yamamoto, K. Nakadai, H. Tsujino, T. Yokoyama, and H. G. Okuno, "Assessment of general applicability of robot audition system by recognizing three simultaneous speeches," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*. IEEE, 2004, pp. 2111–2116.

[3] R. Takeda, S. Yamamoto, K. Komatani, T. Ogata, and H. G. Okuno, "Missing-feature based speech recognition for two simultaneous speech signals separated by ica with a pair of humanoid ears," in *Proceedings of IEEE International Conference on Intelligent Robots and Systems (IROS 2006)*, 2006.

[4] 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. 2006 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP2006)*, vol. I, 2006, pp. 809–812.

[5] E. Hansler, "Acoustic echo and noise control: where do we come from –where do we go?" in *Proc. 7th International Workshop on Acoustic Echo and Noise Control*, 2001, pp. 1–4.

[6] S. Makino and S. Shimauchi, "Stereophonic acoustic echo cancellation – an overview and recent solutions," in *Proc. The 1999 IEEE Workshop on Acoustic Echo and Noise Control*, 1999, pp. 12–19.

[7] W. Herbordt, J. Ying, H. Buchner, and W. Kellermann, "A realtime acoustic human-machine front-end for multimedia applications integrating robust adaptive beamforming and stereophonic acoustic echo cancellation," in *Proc. 7th International Conf. on Spoken Language Processing*, vol. 2, 2002, pp. 773–776.

[8] 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.

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

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

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

[12] A. Lee, T. Kawahara, and K. Shikano, "Julius – an open source real-time large vocabulary recognition engine," in *Proc. of EUROSPEECH*, 2001, pp. 1319–1322.