Proceedings of the 2006 IEEE/RSJ
International Conference on Intelligent Robots and Systems
October 9 - 15, 2006, Beijing, China

Missing-Feature based Speech Recognition for Two Simultaneous Speech Signals Separated by ICA with a pair of Humanoid Ears

Ryu Takeda, Shun'ichi Yamamoto, Kazunori Komatani, Tetsuya Ogata, and Hiroshi G. Okuno
Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto 606-8501, Japan
{*rtakeda, shunichi, komatani, ogata, okuno*}*@kuis.kyoto-u.ac.jp*

***Abstract*— Robot audition is a critical technology in making robots symbiosis with people. Since we hear a mixture of sounds in our daily lives, sound source localization and separation, and recognition of separated sounds are three essential capabilities. Sound source localization has been recently studied well for robots, while the other capabilities still need extensive studies. This paper reports the robot audition system with a pair of omni-directional microphones embedded in a humanoid to recognize two simultaneous talkers. It first separates sound sources by Independent Component Analysis (ICA) with single-input multiple-output (SIMO) model. Then, spectral distortion for separated sounds is estimated to identify reliable and unreliable components of the spectrogram. This estimation generates the missing feature masks as spectrographic masks. These masks are then used to avoid influences caused by spectral distortion in automatic speech recognition based on missing-feature method.**

The novel ideas of our system reside in estimates of spectral distortion of temporal-frequency domain in terms of feature vectors. In addition, we point out that the voice-activity detection (VAD) is effective to overcome the weak point of ICA against the changing number of talkers. The resulting system outperformed the baseline robot audition system by 15 %.

***Index Terms*— Robot audition, Multiple Speakers, ICA, Missing-feature Methods, Automatic Speech Recognition**

I. Introduction

Many types of robots including humanoid robots have appeared recently, in particular, around Expo 2005 Aichi. They are expected to operate not in laboratory environments but in real-world environments in order to attain symbiosis between people and robots in our daily lives. Since verbal communication is the most important in our daily lives, hearing capabilities are essential for robots to attain symbiosis between people and robots. Current automatic speech recognition (ASR) systems work well in laboratory environments, while they do not in noisy environments. In the latter, we usually hear a mixture of sounds, in particular, a mixture of speech signals. Since speech signals are considered as non-quasi-stationary noises, normal noise reduction techniques are not applicable for recognizing a mixture of speech signals. For that purpose, three capabilities are mandatory; *sound source localization*, *sound source separation* (SSS), and *recognition of separated sounds*. Sound source localization has been recently studied well for robots, while the other capabilities still need extensive studies.

Since robots are usually deployed in real-world environments, robot audition should fulfill three requirements. First, they should work even in unknown and/or dynamically-changing environments. Second, they should listen to several speakers at the same time, and third, they should recognize what each speaker said. To fulfill the first two requirements, we use Independent Component Analysis (ICA) for source separation, because it is one of well-known methods of Blind Source Separation (BSS). ICA assumes only the mutual independence of component sound signals, and does not need *a priori* information about room transfer functions, head related transfer functions of the robot, or sound sources. The number of microphones needed by ICA is larger than or equal to that of sound sources. We use the SIMO-ICA (Single-Input Multiple-Output) [1], since our robot has only two microphones. In this paper, we assume that the number of sound sources are at most two.

To cope with the third requirement, we adopt the missing-feature theory (MFT) for ASR. Again, MFT-based ASRs usually use a clean acoustic model without requesting a priori information about sound sources or acoustic characteristics. MFT models the effects of interfering sounds on speech as the corruption of regions of time-frequency representations of the speech signal. Usually speech signal separated by ICA or any other technologies suffers from spectral distortion due to ill-posed inverse problems. Reliable and unreliable components are estimated to generate missing-feature masks. We use a binary mask, i.e., reliable or unreliable, in this study.

The main technical issues in using ICA and MFT-based ASR are (1) selecting which channel of separated signal to be recognized from SIMO signals, (2) estimating signal leakage from the other sound source, (3) detecting the number of sound sources, and last but not least (4) generating missing-feature masks (MFM) by estimating reliable or unreliable components of separated signals. In this paper, we use humanoid robot SIG2 which has a pair of microphones each of which is embedded in each ear.

The first issue is solved by using sound source localization with Interaural Intensity Difference to estimate the relative position between microphones and speakers. The second and third ones are solved partially by using voice activity detection (VAD) and sound source localization. The last problem, i.e.,

automatic generation of MFM is realized by taking into consideration the influence of the distortion estimated in spectral domain on the feature domain, and by deciding which features are reliable or not.

A. Related Work

Although a good deal of research on robot audition has been done in recent years, most efforts have focused on sound source localization and separation. Only a few researchers have focused on simultaneous speech signals, SSS, and the recognition of separated sounds. The humanoid, *HRP-2*, uses a microphone array to localize and separate a mixture of sounds, and can recognize speech commands for a robot in noisy environments [2]. *HRP-2*, however, only focused on a single speech signal. The humanoid robot *SIG* uses a pair of microphones to separate multiple speech signals with the *Adaptive Direction-Pass Filter (ADPF)* and recognizes each separated speech signal by ASR [3]. When three speakers uttered words, *SIG* recognized what each speaker had said. Yamamoto *et al.* recently developed a new interfacing scheme between SSS and ASR based on MFT [4]. They demonstrated that their interfacing scheme worked well for different types of humanoids, i.e., *SIG-2*, *Replie*, and *ASIMO* with manually created missing-feature masks, so-called *a priori* masks.

Yamamoto, and Valin *et al.* further developed *Automatic-missing-feature Mask Generation (AMG)* by using a microphone array system consisting of eight microphones to separate sound sources. Their sound source separation system consisted of two components [5]: *Geometric Source Separation (GSS)* [6] and the *Multi-Channel Post-Filter (MCPF)* [5], [7]; GSS separated each sound source by using an adaptive beamformer with the geometric constraints of microphones, while MCPF refined each separated sound by taking channel-leakage and background noises into account. They developed AMG by using information obtained with MCPF to generate missing-feature masks for all separated sounds.

Missing-feature methods are usually adopted to improve the accuracy of recognition for ASR in noisy environments, in particular, with quasi-stationary noises [8]. A spectrographic mask is the set of tags that identify reliable and unreliable components of the spectrogram. MFT-based ASR uses a spectrographic mask to avoid corrupt signals during the decoding process. There are two main approaches for missing-feature methods; *feature-vector imputation* and *classifier modification*. The former estimates unreliable components to reconstruct a complete uncorrupted feature vector sequence and use it for recognition [9]. The latter modifies the classifier, or recognizer, to perform recognition using reliable separated components and unreliable original input components itself [10], [11], [8], [12], [13], [14]. Most studies have not focused on recognition of speech signals corrupted by interfering speech signals except [13], [14].

The rest of the paper is organized as follows: Section 2 explains the ICA and MFT-based ASR. Section 3 overviews our robot audition system. Section 4 describes the experiments we did for evaluation, and Section 5 discusses the results and observations. Section 6 concludes the paper.

II. Sound source separation by ICA

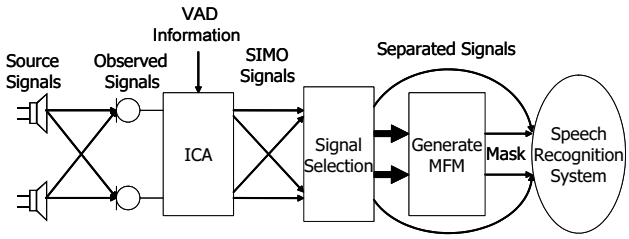

Fig. 1. Overview of the system

Our system consists of three components as is shown in Figure1. (1) Independent Component Analysis (ICA) as a blind source separation, (2) Missing-feature theory (MFT) based ASR, and (3) Automatic missing-feature mask (MFM) generation. The last one bridges between the first and second components. In this section, we focus on the first component, ICA. We first point out the problems of sound source separation with ICA, and show the improvement by using voice activity detection (VAD) technique.

A. Inter-channel Signal Leakage and Voice Activity Detection

We assume the model of mixtures of speech signals as convolution. Since this convolution model does not reflect actual acoustic environments, any methods based on this model cannot decompose each signal components. The spectral distortion of separated signals is mainly caused due to signal leakage in the desired speech signal. Suppose that two speakers are talking and one stops talking as is shown in Figure 2. It may often the case with ICA that signal leakage is observed during its silent period. The spectral parts enclosed in a red box are instances of signal leakage. If such leakage is very strong, it is difficult to determine the end of speech. A wrong estimation of speech period would deteriorate the recognition accuracy severely.

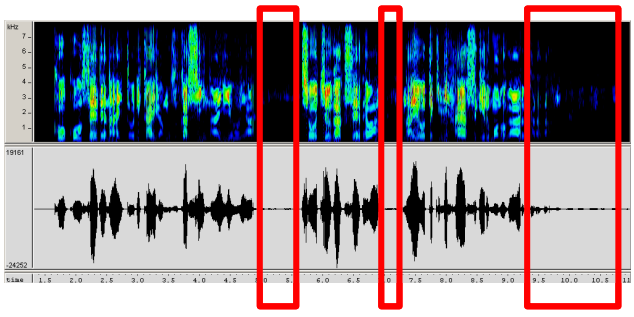

Fig. 2. Leakage in spectrum for silent period

The VAD information is useful for determining the period of utterance in order to improve the performance of separation. We can adopt a sound source localization and speaker tracking system, such as [15], to obtain VAD information. In this paper, the correct VAD information was manually given.

B. ICA for voiced signals

We adopt frequency domain representation instead of temporal domain one. The search space is smaller because the separating matrix is updated for each frequency bin, and thus its convergence is faster and less dependent on initial values.

1) Mixing process of speech signals: We assume that the signals are observed by mixing linearly sound sources. This mixing process is expressed as follows:

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

where $\boldsymbol{x}(t) = [x_1(t), \ldots, x_J(t)]^T$ is the observed signal vector, and $\boldsymbol{s}(t) = [s_1(t), \ldots, s_I(t)]^T$ is the source signal vector. In addition, $\boldsymbol{a}(n) = [a_{ji}(n)]_{ji}$ is the mixing filter matrix with the length of N, where $[X]_{ji}$ denotes the matrix which includes the element X in the i-th row and the j-th column. In this paper, the number of microphones, J is 2 and the number of multiple sound source, L, is 2.

2) Frequency-domain ICA: We use the frequency-domain ICA. First, the short-time analysis of observed signal is conducted by frame-by-frame discrete Fourier transform (DFT) to obtain the observed vector $\boldsymbol{X}(\omega, t) = [X_1(\omega, t), \ldots, X_J(\omega, t)]$ in each frequency bin ω and at each frame t. The unmixing process can be formulated in a frequency bin ω

$$\boldsymbol{Y}(\omega, t) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega, t) \tag{2}$$

where $\boldsymbol{Y}(\omega, t) = [Y_1(\omega, t), \ldots, Y_I(\omega, t)]$ is the estimated source signal vector, and $\boldsymbol{W}$ represents a (2 by 2) unmixing matrix in frequency bin ω.

For estimating the unmixing matrix $\boldsymbol{W}(\omega)$ in (2), an algorithm based on the minimization of the Kullback-Leibler divergence is often used. Therefore, we use the following iterative equation with non-holonomic constraints:

$$\boldsymbol{W}^{j+1}(\omega) = \boldsymbol{W}^{j}(\omega) - \alpha\{\text{off-diag}\langle\boldsymbol{\phi}(\boldsymbol{Y})\boldsymbol{Y}^{h}\rangle\}\boldsymbol{W}^{j}(\omega) \tag{3}$$

where α is a step size parameter that has effects on the speed of convergence, $[j]$ is used to express the value of the jth step in the eterations, and $\langle\cdot\rangle$ denotes the time-averging operator. The operation, off-diag$(\boldsymbol{X})$, replaces the diag-element of matrix X with zero. In this paper, the nonlinear function, $\boldsymbol{\phi}(\boldsymbol{y})$, is defined as $\phi(y_i) = \tanh(|y_i|)e^{j\theta(y_i)}$.

3) Solution of permutation and scaling problems in ICA: Frequency-Domain ICA suffers from two kinds of ambiguities; *scaling ambiguity*, i.e., the power of separated signals differs at each frequency bin, and *permutation ambiguity*, i.e., some signal components are swapped among different channels. These ambiguities occur, because ICA estimates both unmixing matrix $\boldsymbol{W}$ and source signal vector $\boldsymbol{Y}$ at the same time. The most important requirement in solving these ambiguities is to recover the spectral representation as complete as possible. In addition, the method to solve them should provide useful information to automatic missing-feature mask generatation. We solved these problems with the Murata's method in [16].

In order to cope with the scaling ambiguity, we apply the inverse filter $\boldsymbol{W}^{-1}$ to the estimated source signal vector $\boldsymbol{Y}$.

$$\boldsymbol{v}_i = \boldsymbol{W}^{-1}\boldsymbol{E}_i\boldsymbol{W}\boldsymbol{x} = \boldsymbol{W}^{-1}(\ 0 \cdots u_i \cdots 0\ )^t \tag{4}$$

where $\boldsymbol{x}$ is observed signal vector, and $\boldsymbol{W}$ is the estimated unmixing matrix, $\boldsymbol{E}_i$ represents the matrix in which the ith diagonal element is one, and the others are zero. Thus they satisfy the equation $\sum_i \boldsymbol{E}_i = \boldsymbol{I}$. This solution gives a Single-Input Multiple-Output (SIMO) signals. Here the term "SIMO" represents the outputs are the transmitted signals observed at multiple microphones.

The permutation ambiguity can be solved by taking into consideration correlation of envelopes of power spectrum among frequency bins. By calculating all correlations among frequency bins, the most highly correlated frequency bins are considered the spectrum of the same signal.

C. Integration of VAD and ICA

ICA with the number of sound sources given by VAD is realized by selecting signals. This selection is defined as follows:

$$\boldsymbol{Y}(\omega, t) = M\hat{\boldsymbol{Y}}(\omega, t) \tag{5}$$

$$M = \begin{cases} 1 & J = I \\ 0 & otherwise \end{cases} \tag{6}$$

where $\hat{\boldsymbol{Y}}(\omega)$ is the observed signal vector, and J is the number of microphones and I is the number of estimated sound sources. In this paper, the maximum number of simultaneous sound sources is given in advance.

Given the number of active speakers, the system must decide who stopped speaking. In case of two speakers, it is easy because the system focuses on only one speaker. It determines which one is speaking by using mean square error between the power spectrum of output signal of ICA and that of the observed signal spectrum for the estimated frames in which one speaker is speaking.

The region for silent periods are filled with silent spectrum that is obtained in advance. If such region is filled with 0 signal, it may not be treated as silence by ASR with acoustic model that is trained with clean speech signals.

Fig. 3. SIG2's ear

Fig. 4. Humanoid SIG2 with two ears

III. Speech Recognition with Automatic Generation of Missing Feature Mask

In this section, we explain how SIMO signals separated by ICA with VAD is recognized; that is, selecting speech signal out of SIMO signals, estimating missing features, generating missing feature masks, and MFT-based ASR.

A. Issues in applying MFT-based ASR to speech signals separated by ICA

In applying MFT-based ASR to SIMO signals separated by ICA, we have to solve the following three main issues:

1) Selecting speech signal out of SIMO signals for recognition
2) Designing acoustic features for MFT-based ASR, and
3) Estimating spectral distortion and generating MFM for MFT-based ASR.

We discuss these issues in the following subsections.

B. Selecting speech signal for recognition based on IID

As mentioned in section II, we solved the scaling ambiguity of ICA with inverse unmixing matrix. As a result, SIMO signals are obtained as outputs of ICA. Therefore, we must select speech signal out of SIMO signals for recognition for each sound source.

Saruwatari *et al.* [17] selected the strongest spectrum in order to apply binary mask. This selection method is not well suited to MFT-based ASR, partially because the binary mask also causes errors or distortion in spectrum, and partially because our system uses a pair of omni-directional microphones while they used a pair of directional microphones.

The selection based on power spectrum is usually good, but may not be for the speaker located in front of the robot. For example, if two speakers locate on the right and in front of the robots, the left channel of SIMO signals separated by ICA for the center speaker is less affected by the right speaker, although the power of the right channel may be larger. In summary, we should consider the relative location of microphones and speakers.

Since the positions of microphones are known, the relative position of sound sources is enough for selection in case of two speakers. One candidate for obtaining the relative position is information obtained by sound source localization. In solving the scaling ambiguity with inverse unmixing matrix, ICA generates SIMO signals consisting of left and right channels. Therefore, location may be obtained by using interaural intensity difference (IID) and interaural phase difference (IPD).

When the sounds are captured by the robot's ears (Figure 3), IID is emphasized because of the head-related transfer function (HRTF). IPD is usually not so stable due to permutation ambiguity. Even if the separation of ICA is not so accurate, the tendency of IID is usually recovered. Therefore, the relative position of speakers can be estimated by the normalized IID defined as follows:

$$I(f_p^L, f_p^R) \quad = \quad (f_p^L - f_p^R) / \max\left(f_p^L, f_p^R\right) \qquad (7)$$

where f_p is the intensity of signal f defined by the envelope of signal, or power spectrum. f_p^L and f_p^R are intensities of signal f observed at each microphones. The normalized IID, $I(f_p^L, f_p^R)$, is used to obtain the relative position of speakers by sorting the intensity of sound sources. Given the position of each microphone, we select the output of microphone closest to the speaker as speech signals for recognition.

C. Missing Feature Based Speech Recognition

When several people speak at the same time, each separated speech feature is severely distorted in spectrum from its original signal. By detecting and masking the distorted feature, the MFT-based ASR improves its recognition accuracy. As result, it needs only clean speech in training acoustical model of ASR.

1) Features for ASR: Since MFCC is not appropriate for recognizing separated sounds from simultaneous speeches by MFT-based ASR [8], we use Mel scale log spectrum (MSLS) that are obtained by applying Inverse Discrete Cosine Transform (DCT) to the MFCC features. The detailed flow of calculation is as follows:

1) FFT: 16 bit acoustic signals sampled by 16kHz are analyzed by FFT with 400 points of window and 160 frame shift to obtain spectrum.
2) Mel: Spectrum is analyzed by Mel-scale filter bank to obtain Mel-Scale spectrum of 24th order.
3) Log: Mel-scale spectrum of 24th order is converted to log-energies.
4) DCT: The log Mel-scale spectrum is converted by Discrete Cosine Tranform to the Cepstrum.
5) Lifter: Cepstral features 0 and 13-23 are set to zero so as to make the spectrum smoother.
6) CMS: Convolutive effects are removed using Cepstral Mean Substraction.
7) IDCT: The normalized Cepstrum is transformed back to the log Mel-scale spectral domain by means of an Inverse DCT.
8) Differentiation: The features are differentiated in the time domain. Thus, we obtain 24 log spectral features as well as their first-order time derivatives.

The [CMS] step is necessary in order to remove the influence of convoluted noise, such as reverberation and microphone frequency response.

2) Speech recognition based on missing feature theory : MFT-based ASR is a Hidden Markov Model (HMM) based recognizer which assumes that the input consists of reliable and unreliable spectral features. Most conventional ASRs are based on HMM, and estimate a path with maximum likelihood based on state transition probabilities and output probability in Viterbi algorithm. MFT-based ASRs differs from conventional ASRs in estimation of the output probability.

Let $f(x|s)$ be the output probability of feature vector x in state S. The output probability is defined by

$$f(x|S) = \sum_{k=1}^{M} P(k|S) f(x_r|k, S),$$

where M is the number of Gaussian mixture, and x_r is a reliable part in x.

This means that only reliable features are used in the probability calculation. Therefore, the recognizer can avoid severe degradation of performance caused by unreliable features.

D. Formulation of Generating Missing Feature Mask

Generating missing feature mask is formulated based on estimated error. By considering a function that converts spectrum to feature, we can reveal the relation between distortion of spectrum and that of feature. In addition, our method makes it possible to generate masks for the differential features.

1) A priori mask: As the result of ICA, a "true" vector, $\boldsymbol{s_0}$, is distorted by the "error" vector, $\Delta\hat{e}$. The distorted vector can be expressed as $\boldsymbol{x_0} = \boldsymbol{s_0} + \Delta\hat{\boldsymbol{e}}$. We define the smooth function $\boldsymbol{F}(\boldsymbol{s})$ as the mapping from spectrum $\boldsymbol{s}$ to feature.

Now, the error in feature space is expressed as follows:

$$\Delta_{\text{a priori}}\boldsymbol{F} \quad = \quad |\boldsymbol{F}(\boldsymbol{x}_0) - \boldsymbol{F}(\boldsymbol{s}_0)| \tag{8}$$

where the absolution operator is applied to each element of vector. Given the "ture" feature $\boldsymbol{F}(\boldsymbol{s}_0)$ of the vector $\boldsymbol{s}_0$, the MFM is defined as follows:

$$\boldsymbol{M} \quad = \quad \begin{cases} 1 & |\boldsymbol{F}(\boldsymbol{x}_0) - \boldsymbol{F}(\boldsymbol{s}_0)| < T \\ 0 & otherwise \end{cases} \tag{9}$$

where T is threshold parameter. We call the mask generated by (9) *a priori* mask.

2) Automatic generated mask: It is practically impossible to know the true vector $\boldsymbol{s}_0$ in advance. Therefore by using the separated speech signal vector $\boldsymbol{x}$ and estimated error vector $\boldsymbol{s}_0$, the error in feature space is expressed as follows on the assumption of the error, $\Delta\boldsymbol{e}$, is not so large.

$$\Delta F(\boldsymbol{x}) \quad \simeq \quad |\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x} - \Delta\boldsymbol{e})| \tag{10}$$

And MFM $\boldsymbol{M}'$ can be generated by

$$\boldsymbol{M}' \quad = \quad \begin{cases} 1 & |\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x} - \Delta\boldsymbol{e})| < T \\ 0 & otherwise \end{cases} \tag{11}$$

where T represent the threshold parameter.

Next we consider generating masks of the time differential feature. The time differential feature is defined

$$\Delta_t F(\boldsymbol{s}) = \boldsymbol{F}_t(\boldsymbol{s}) - \boldsymbol{F}_{t-1}(\boldsymbol{s}) \tag{12}$$

where the spectrum $\boldsymbol{s}$ includes all the time-frequency spectrum, and $\boldsymbol{F}_t(\boldsymbol{s})$ represents the tth frame feature of $\boldsymbol{F}(\boldsymbol{s})$. By using $\Delta\boldsymbol{F}$, the error vector of time differential feature is evaluated by the following equations:

$$\begin{aligned} &\Delta_t\boldsymbol{F}(\boldsymbol{x}) - \Delta_t\boldsymbol{F}(\boldsymbol{x} - \Delta\boldsymbol{e}) \\ &\quad = \quad \{\boldsymbol{F}_t(\boldsymbol{x}) - \boldsymbol{F}_{t-1}(\boldsymbol{x})\} \\ &\qquad\qquad -\{\boldsymbol{F}_t(\boldsymbol{x} - \Delta\boldsymbol{e}) - \boldsymbol{F}_{t-1}(\boldsymbol{x} - \Delta\boldsymbol{e})\} \\ &\quad = \quad \{\boldsymbol{F}_t(\boldsymbol{x}) - \boldsymbol{F}_t(\boldsymbol{x} - \Delta\boldsymbol{e})\} \\ &\qquad\qquad -\{\boldsymbol{F}_{t-1}(\boldsymbol{x}) - \boldsymbol{F}_{t-1}(\boldsymbol{x} - \Delta\boldsymbol{e})\} \\ &\quad = \quad \Delta\boldsymbol{F}_t(\boldsymbol{x}) - \Delta\boldsymbol{F}_{t-1}(\boldsymbol{x} - \Delta\boldsymbol{e}) \end{aligned} \tag{13}$$

With the threshold parameter T, we can generate the mask for time differential feature as follows:

$$\boldsymbol{M}_t \quad = \quad \begin{cases} 1 & |\Delta\boldsymbol{F}_t(\boldsymbol{x}) - \Delta\boldsymbol{F}_{t-1}(\boldsymbol{x} - \Delta\boldsymbol{e})| < T \\ 0 & otherwise \end{cases} \tag{14}$$

E. Generation of MFM with output of ICA

MFM is generated by estimating reliable and unreliable components of sounds separated by ICA. Since the influence of the signals leakage be weak, and we assume the error vector, $\Delta\boldsymbol{e}$, is not so large. In addition, the function, $\boldsymbol{F}$, can be assumed as smooth because our process of converting from spectrum to feature includes only filtering, log scaling and absolution operations.

Let $m(\omega, t)$ be the observed spectrum at a microphone, and $x_1(\omega, t), x_2(\omega, t)$ be the separated spectrum, then $x_1(\omega, t)$ denotes the signal selection as described in section III-B. Now they satisfy the following equation:

$$m(\omega, t) = x_1(\omega, t) + x_2(\omega, t) \tag{15}$$

$$x_1(\omega, t) = a_1(\omega)s_1^{'}(\omega, t) \tag{16}$$

$$x_2(\omega, t) = a_2(\omega)s_2^{'}(\omega, t) \tag{17}$$

where $a_1(\omega), a_2(\omega)$ and $s_1^{'}(\omega, t), s_2^{'}(\omega, t)$ are the estimated the elements of mixing matrix and separated spectrums. Ideally, $m(\omega, t)$ is separated as follows

$$m(\omega) = W_1(\omega)s_1(\omega) + W_2(\omega)s_2(\omega) \tag{18}$$

where $W_1(\omega), W_2(\omega)$ are transfer functions.

The errors of separated spectrum are expressed as

$$s_1^{'}(\omega, t) = \alpha_1(\omega)s_1(\omega, t) + \beta_1(\omega)s_2(\omega, t) \tag{19}$$

$$s_2^{'}(\omega, t) = \beta_2(\omega)s_1(\omega, t) + \alpha_2(\omega)s_2(\omega, t) \tag{20}$$

where $\alpha_1(\omega), \alpha_2(\omega), \beta_1(\omega), \beta_2(\omega)$ are the error coefficients including scaling. Now the error of the estimated spectrum $x_1(\omega, t)$ is

$$\begin{aligned} e_1(\omega, t) = &\Big(\alpha_1(\omega)a_1(\omega) - W_1(\omega)\Big)s_1(\omega, t) \\ &+\beta_1(\omega)a_1(\omega)s_2(\omega, t) \end{aligned} \tag{21}$$

In this paper, we find that spectral distortion is caused by signal leakage and the distortion of original signal.

To estimate the error, we assume that the unmixing matrix approximates well to $W(\omega)$, and that the envelope of the power spectrum of leaked signal is similar to that of scaled $x_2(\omega, t)$. That is,

$$\Big(\alpha_1(\omega)a_1(\omega) - W_1(\omega)\Big)s_1(\omega, t) \simeq 0 \tag{22}$$

$$\beta_1(\omega)a_1(\omega)s_2(\omega, t) \simeq \gamma_1 x_2(\omega, t) \tag{23}$$

$$e_1(\omega, t) \simeq \gamma_1 x_2(\omega, t) \tag{24}$$

As discussed above, we generate MFMs, $\boldsymbol{M}$, for the estimated observed spectrum, $\boldsymbol{x}$, with the estimated error spectrum, $\boldsymbol{e}$, based on (11) as follows:

$$\boldsymbol{M} \quad = \quad \begin{cases} 1 & |\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x} - \boldsymbol{e})| < \theta \\ 0 & otherwise \end{cases} \tag{25}$$

In addition, the masks for time differential feature are generated based on (14)

$$\boldsymbol{M}(k) \quad = \quad \begin{cases} 1 & |\Delta\boldsymbol{F}_k(\boldsymbol{x}) - \Delta\boldsymbol{F}_{k-1}(\boldsymbol{x} - \boldsymbol{e})| < \hat{\theta} \\ 0 & otherwise \end{cases} \tag{26}$$

To simplify and thus speed up the estimate of the errors, we normalize the difference $\Delta \boldsymbol{F}$ with its maximum value.

IV. Experiments and Evaluation

A. Experiment Patterns

We use two omni-directional microphones placed in the ears of SIG2 humanoid robot (Figure 4) for evaluating the system. We compare speech recognition accuracy obtained in the following four different conditions:

1) ICA separation with utterances of different length,
2) ICA separation with VAD,
3) ICA separation with channel selection, and
4) ICA separation with VAD, channel selection, and missing feature masks.

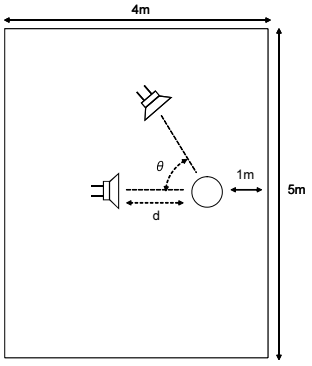

Fig. 5. Configuration 1: Asymmetric speakers

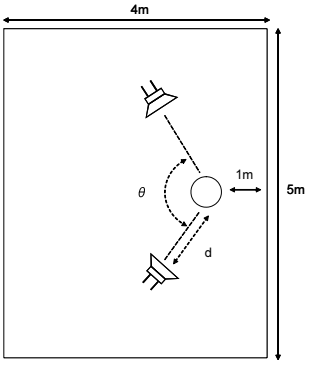

Fig. 6. Configuration 2: Symmetric speakers

1) Recording conditions: Two voices are recorded simultaneously from loudspeakers placed 0.5–1.5 m (d) away from the robot. The speech signals are assumed to arrive from different direction of sound sources, θ = 30°, 60°, and 90°. The female speaker is on the left side of the robot and the other is on its right side. The room size is 5 m×4 m, with a reverberation time of 0.2–0.3 sec. We use combinations of three different words selected from a set of 200 phonemically-balanced Japanese words.

2) Acoustic model for speech recognition: We use multi-band Julian as MFT-based ASR. It uses the triphone-based acoustic model trained by clean speech with utterances of 216 words by 25 male and female speakers. The acoustic model uses three states and four Gaussians per mixture.

3) Configurations for experiments: The SIG2 stands at 1 m from the wall with a glass window and female and male speakers are located in two configuration; one is asymmetric, female speaker is in the center and male is on the right (Figure 5). The other is symmetric (Figure 6). The main parameters for ICA are as follows; the sampling rate of data is 16 kHz, the frame length is 1,024 points, and the frame shift is 94. The initial values of unmixing matrix, $\boldsymbol{W}(\omega)$, are given at random. We tried all combinations of $\{\{0.88, 0.9, 0.92\}$ for the threshold, $\hat{\theta}$, and $\{0.04, 0.05, 0.06\}\}$ for the threshold, θ. These combinations are applied to each dataset, and finally obtained $\{0.92, 0.04\}$ as the best threshold.

B. Results of Experiments

1) Improvement of recognition accuracy by longer utterance period : Figure 7 shows the results of recognition accuracy of separate speech by ICA from two simultaneous utterances. The length of utterance varies from one word to one hundred words. As longer the speech period is, the recognition accuracy is improved. This is because the estimation of ICA becomes more accurate by using more samples. The recognition accuracy with twenty words is 20 % greater than that with one word. The recognition accuracy starts saturating around the length of twenty words.

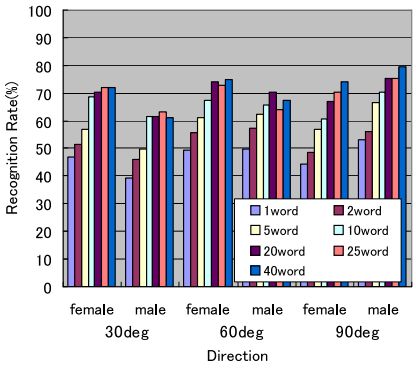

Fig. 7. Improvement of recognition accuracy by longer utterance period

2) Improvement of recognition accuracy by VAD: Figure 8 show the improvement of recognition accuracy by incorporating VAD into ICA for two configurations. Since in some benchmark, two speakers start and stop utterance at different times and thus in some period only one speaker utters. This caused signal leakage to a silent period, but VAD avoids such cases.

3) Improvement of recognition accuracy by channel selection : Figure 8 shows the improvement of recognition accuracy by selecting an appropriate channel of SIMO signals separated by ICA. This channel selection improves recognition accuracy either with VAD or without VAD.

4) Futher improvement of recognition accuracy by missing-feature masks : Figure 9 shows the improvement of recognition accuracy by *a priori* (ideal) mask and masks generated automatically (our masks). *A priori* mask attains the recognition accuracy of over 97 %. Auto generated mask improves by 5 % in average. It seems to depend on the location of speakers.

Finally, the improvements of recognition accuracy are summarized in Figure 10.

C. Discussions

Some observations about our system of listening to two things at once by integrating ICA, MFT-based ASR, and automatic missing feature mask generation are listed below:

1) ICA Separation with different utterance periods: Experiment (1) shows the longer utterance improves the recognition accuracy. The estimation of inverse unmixing matrix for ICA needs 30 seconds for stable estimation, which means about

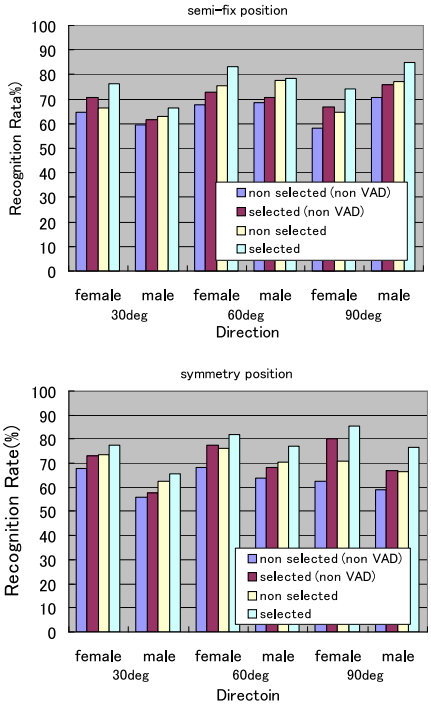

Fig. 8. Effects of signal selection and VAD on Recognition

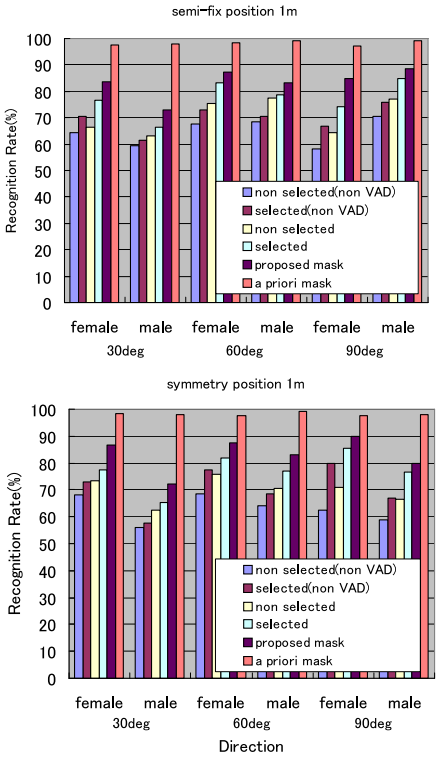

Fig. 10. Summary of improvements of recognition accuracy

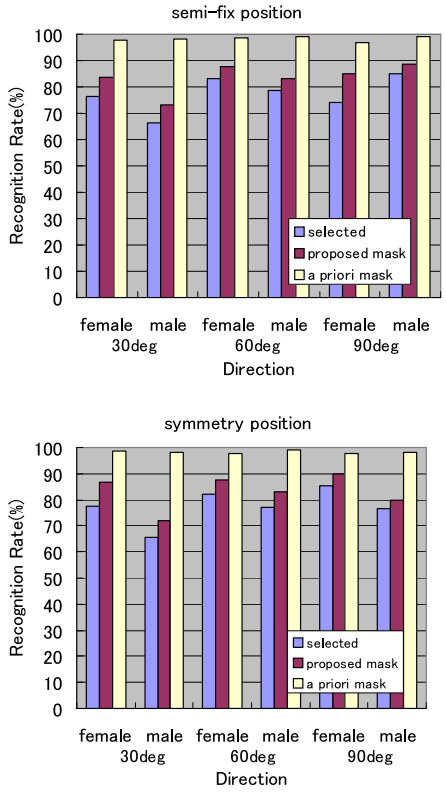

Fig. 9. Improvement of recognition accuracy by missing-feature masks

15 seconds of actual non-silent simultaneous speech signals. Needless to say, the period of utterance sufficient for stable separation depends on the implementation of ICA and other parameters of MFT-based ASR and automatic missing-feature mask generation.

2) ICA with VAD: By ICA with VAD, the recognition accuracy is improved. The rejection of leaked signals during non-speaking period contributed to the correct recognition. Actually, many people rarely speak for the completely same time, and thus this rejection is essential process for speech recognition.

3) SIMO-ICA with channel selection: The channel selection also proves valid for recognition. We consider that the localization with IID works well. The location or configuration of speakers affects the recognition accuracy, since the closer speakers makes sound source separation deteriorate. This is confirmed by comparing the results of asymmetrical and symmetrical configurations.

4) ICA with missing-feature masks: The recognition accuracy for speech signals separated by SIMO-ICA can be improved by channel selection and VAD (detection of the number of speakers). Further improvement can be attained by automatic missing-feature mask generation with MFT-based ASR. Missing-feature masks are generated by estimating reliable and unreliable components of separated signals. By employing all these improvements, the recognition accuracy of two simultaneous speech signals is more than 80%.

In contrast of two microphones, Yamamoto *et al.* uses eight

microphones to separate three simultaneous speech signals by Geometrical source separation with multi-channel post filter. They developed automatic missing-feature mask generation by taking into consideration the estimates of stationary noises and channel leakage to improve the recognition accuracy by 7 to 30%.

V. Conclusion

We constructed robot audition system for unknown and/or dynamically-changing environments without providing minimum *a priori* information. To fulfill such requirements, we employed ICA, MFT-based ASR and developed automatic missing feature mask generation.

A. Evaluation of system

In this paper, we use ICA for source separation and MFT-based ASR in order to construct robot audition system in real-world environment. Combination of VAD, MFM, and channel selection improves recognition accuracy by 20% . The comparison of our system with Yamamoto's system consisting of GSS, multi-channel post-filter, and automatic missing-feature mask generation may be interesting topics, which will be reported in a separate paper in a near future.

To improve the performance of individual subsystems is the first future task. Other remaining work includes more precise estimation of reliable and unreliable components of separated sounds.

B. Issues and Future Works

As said in discuss and evaluation section, there are rooms to improve this system. For example, there is ICA with information about speakers or accurate estimation of errors.

The issues include the limitation of the number of sound source. This is essential problem of ICA, and to realizing it leads to construct audition system for multi-environment. In this paper, we assumed the environment without noise, so a new method is desired to cope with noises. To operate in real world, we additionally consider the noise caused by robot itself, and the recognition of connected speech, and the real-time processing of this system.

By solving above problems, we will try connected speech recognition, or implement on the robot in the future.

References

[1] 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 *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*. IEEE, 2005, pp. 209–214.

[2] 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.

[3] K. Nakadai, H. G. Okuno, and H. Kitano, "Robot recognizes three simultaneous speech by active audition," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA-2003)*. IEEE, 2003, pp. 398–403.

[4] 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.

[5] J.-M. Valin, J. Rouat, and F. Michaud, "Enhanced robot audition based on microphone array source separation with post-filter," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004)*. IEEE, 2004, pp. 2123–2128.

[6] L. C. Parra and C. V. Alvino, "Geometric source separation: Mergin convolutive source separation with geometric beamforming," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 6, pp. 352–362, 2002.

[7] I. Cohen and B. Berdugo, "Microphone array post-filtering for non-stationary noise suppression," in *ICASSP-2002*, 2002, pp. 901–904.

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

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

[10] J. Barker, M. Cooke, and P. Green, "Robust asr based on clean speech models: An evaluation of missing data techniques for connected digit recognition in noise," in *Proc. of Eurospeech-2001*. ESCA, 2001, pp. 213–216.

[11] M. P. Cooke, P. Green, L. Josifovski, and A. Vizinho, "Robust automatic speech recognition with missing and unreliable acoustic data," *Speech Communication*, vol. 34, no. 3, pp. 267–285, 2000.

[12] P. Renevey, R. Vetter, and J. Kraus, "Robust speech recognition using missing feature theory and vector quantization," in *Proc. of 7th European Conference on Speech Communication Technology (Eurospeech-2001)*, vol. 2. ESCA, 2001, pp. 1107–1110.

[13] S. Yamamoto, J.-M. Valin, K. Nakadai, T. Ogata, and H. G. Okuno, "Enhanced robot speech recognition based on microphone array source separation and missing feature theory," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA 2005)*. IEEE, 2005, pp. 1489–1494.

[14] S. Yamamoto, K. Nakadai, J.-M. Valin, J. Rouat, F. Michaud, , K. Komatani, T. Ogata, and H. G. Okuno, "Making a robot recognize three simultaneous sentences in real-time," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2005)*. IEEE, 2005, pp. 897–902.

[15] H. G. Okuno, K. Nakadai, K. Hidai, H. Mizoguchi, and H. Kitano, "Human-robot interaction through real-time auditory and visual multiple-talker tracking," in *Proceedings of IEEE/RAS International Conference on Intelligent Robots and Systems (IROS-2001)*. IEEE, 2001, pp. 1402–1409.

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

[17] 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 *Proceedings of IEEE International Conference on Robots and Systems (IROS 2005)*. IEEE, 2005, pp. 209–214.