# Improving Speech Recognition of Two Simultaneous Speech Signals by Integrating ICA BSS and Automatic Missing Feature Mask Generation

*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

10.21437/Interspeech.2006-92

## Abstract

Robot audition systems require capabilities for sound source separation and the recognition of separated sounds, since we hear a mixture of sounds in our daily lives, especially mixed of speech. We report a robot audition system with a pair of omni-directional microphones embedded in a humanoid that recognizes two simultaneous talkers. It first separates the sound sources by Independent Component Analysis (ICA) with the single-input multiple-output (SIMO) model. Then, spectral distortion in the separated sounds is then estimated to generate missing feature masks. Finally, the separated sounds are recognized by missing-feature theory (MFT) for Automatic Speech Recognition (ASR). The novel aspects of our system involve estimates of spectral distortion in the temporal-frequency domain in terms of feature vectors and based on estimates error in SIMO-ICA signals. The resulting system outperformed the baseline robot audition system by 7 %.

**Index Terms**: Robot audition, ICA, missing-feature mask.

## 1. Introduction

Robots are expected to operate in real-world environments to attain symbiosis between people and robots in their daily lives. Since robots hear a mixture of sounds in the real world, robot audition systems require the essential capabilities of source localization and separation, and the recognition of separated sounds.

As robots are usually deployed in real-world environments, robot audition systems should fulfill three requirements. First, they should work even in unknown and/or dynamically-changing environments. Second, they should be able to listen to several speakers at the same time, and third, they should recognize what each speaker said. We used Independent Component Analysis (ICA) for source separation to fulfill the first two requirements, because it assumes 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 the sound sources. In this paper, we have assumed that the number of sound sources is two at most.

To cope with the third requirement, we adopted the missing-feature theory (MFT) for automatic speech recognition (ASR). Again, MFT-based ASRs usually use a clean acoustic model without requesting *a priori* information about clean 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, the speech signal separated by ICA or other technologies suffers from spectral distortion due to ill-posed inverse problems. Reliable and unreliable components are estimated to generate missing-feature masks. We used a binary mask, i.e., reliable or unreliable, in this study.

The main technical issues in using ICA and MFT-based ASR are (1) generating missing-feature masks (MFM) by estimating reliable or unreliable components for separated signals in speech period, and (2) estimating signal leakage from other sound sources in non-speech periods. We used a humanoid robot SIG2, which has a pair of microphones each of which is embedded in each ear.

The first problem, i.e., automatic generation of MFM was solved by taking the influence of distortion estimated in the spectral domain into consideration, and by determining which features were reliable by using separated SIMO-ICA signals. The second issues were solved by using *a priori* voice activity detection (VAD) information in this paper and we show VAD process is also effective to ICA signals.

The rest of the paper is organized as follows: Section 2 explains the ICA for speech signal. Section 3 presents how generates MFM automatically. Section 4 describes the experiments and evaluation, and Section 5 concludes the paper.

## 2. Sound source separation by ICA

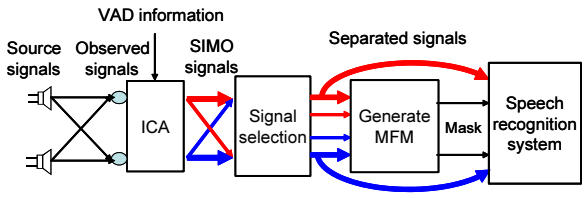


Figure 1: Overview of the system

The system consists of three components as outlined in Fig 1. These are (1) Independent Component Analysis (ICA) for BSS (2) MFT based ASR, and (3) MFM generation. The last one bridges the first and second components. This section focuses on the ICA. We first point out problems of sound source separation with ICA, and present our solutions using VAD technique.

### 2.1. Inter-channel Signal Leakage and VAD

The model for mixtures of speech signals is assumed linear convolution in this paper. Since this linear convolution model does

not completely reflect actual acoustic environments, any method based on this model cannot decompose all signal components. The spectral distortion of separated signals is mainly caused by signal leakage from the other speech signals. Suppose that two speakers are talking and then one stops talking. It is often the case with ICA that signal leakage is observed during its silent period as is shown in Fig 2. The spectral parts in the red boxes are instances of signal leakage. If such leakage is very large, it is difficult to determine where speech ends. Inaccurate estimates of the period of speech would severely deteriorate the recognition accuracy.

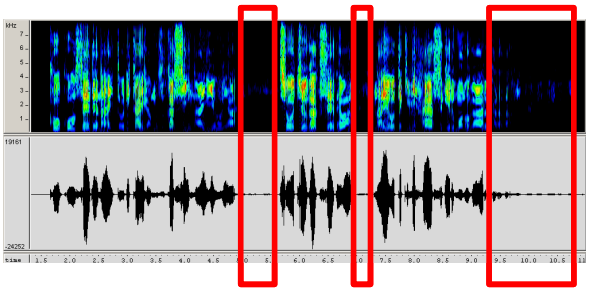

Figure 2: Leakage in spectrum for silent period

### 2.2. ICA for speech signals

We adopted frequency domain representation instead of that for the temporal domain. 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 the initial values.

#### 2.2.1. *Mixing process for speech signals*

We assumed that the signals would be observed by linearly mixing sound sources. This mixing process is expressed as:

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

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 a length of $N$, where $[X]_{ji}$ denotes a matrix that includes the element $X$ in the $i$-th row and the $j$-th column. In this paper, both the number of microphones, $J$, and that of sound sources, $L$, are 2.

#### 2.2.2. *Frequency-domain ICA*

We used frequency-domain ICA. First, short-time analysis of the observed signal was 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 $\omega$ and at each frame $t$. The unmixing process can be formulated in a frequency bin $\omega$ as

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

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 a frequency bin $\omega$.

An algorithm based on the minimization of Kullback-Leibler divergence is often used for speech signals to estimate the unmixing matrix $\boldsymbol{W}(\omega)$ in Eq. (2). Therefore, we used the following iterative equation with non-holonomic constraints [1]:

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

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

Problems specific to FD-ICA are ambiguities with scaling and permutation. We solved these with Murata's method, i.e., envelop of power spectrum [2]. This solution gives Single-Input Multiple-Output (SIMO) signals. In other words, one original speech signal is separated as a set of signals observed at each microphone.

### 2.3. Integration of VAD and ICA

ICA with the number of sound sources given by VAD is achieved by selecting signals to estimate the matrix:

$$\boldsymbol{Y}(\omega, t) = M(t)\hat{\boldsymbol{Y}}(\omega, t), and \tag{4}$$

$$M(t) = \begin{cases} 1 & I(t) = J \\ 0 & otherwise \end{cases}, \tag{5}$$

where $\hat{\boldsymbol{Y}}(\omega, t)$ is the observed signal vector, and $I(t)$ is the number of estimated sound sources at frame $t$ and $J$ is the number of microphones. The number of simultaneous sound sources is given in advance in this paper. After separation, the speech period is determined for recognition. This process improves the estimates for the matrix and reduces the computation time.

## 3. Speech Recognition with Automatic Generation of MFM

In this section, we explain how SIMO signals separated by ICA with VAD are recognized; i.e., by generating missing feature masks, estimating missing features with SIMO signals, and using MFT-based ASR.

### 3.1. MFT-based Speech Recognition

Due to interfering talkers, acoustic features of separated speech signal are severely distorted in the spectrum. By detecting distorted features, and identifying and masking unreliable features, MFT-based ASR improves its recognition accuracy [3]. Another merit of MFT-based ASR is that it only needs clean speech for training its acoustical model.

#### 3.1.1. *Features of MFT-based ASR*

Mel Frequency Cepstral Coefficients (MFCC) are not appropriate for recognizing separated sounds, because distorted features are identified in the spectrum. Mel Scale Log Spectrum (MSLS) is used instead, which is obtained by applying inverse Discrete Cosine Transform (DCT) to the MFCC features. We used 24 log spectral features and their 24 first-order time derivatives.

#### 3.1.2. *Speech recognition based on MFT*

MFT-based ASR is a Hidden Markov Model (HMM) based recognizer that assumes that 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 the output probability in the Viterbi algorithm. MFT-based ASRs differ from conventional ASRs in estimating 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 in performance caused by unreliable features.

### 3.2. Automatic Generation of MFM Using Estimated Error Spectrum — Theory

Generating the MFM is formulated based on the estimated error spectrum. Our method makes it possible to generate masks for differential features. Let $\boldsymbol{x}$ be the separated speech signal vector, $\Delta\boldsymbol{e}$ be the estimated error vector, and $\boldsymbol{F}(\boldsymbol{x})$ be the feature vector of spectrum $\boldsymbol{x}$. Assume that $\Delta\boldsymbol{e}$ is not too large. The errors in feature space are defined as follows:

$$\Delta\boldsymbol{F}(\boldsymbol{x}) = \left|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x} - \Delta\boldsymbol{e})\right|, \quad (6)$$

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

$$\boldsymbol{M}' = \begin{cases} 1 & \left|\frac{\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x}-\Delta\boldsymbol{e})}{E}\right| < T, \\ 0 & otherwise \end{cases} \quad (7)$$

where $T$ represents the threshold parameter, and $E$ is the maximum value of $|\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x} - \Delta\boldsymbol{e})|$.

We next explain how to generate the masks for the time differential features. The time differential feature is defined as

$$\Delta_t\boldsymbol{F}(\boldsymbol{s}) = \boldsymbol{F}_t(\boldsymbol{s}) - \boldsymbol{F}_{t-1}(\boldsymbol{s}), \quad (8)$$

where spectrum $\boldsymbol{s}$ includes all of the time-frequency spectrum, and $\boldsymbol{F}_t(\boldsymbol{s})$ represents the $t$-th frame feature of $\boldsymbol{F}(\boldsymbol{s})$. By using $\Delta\boldsymbol{F}$, the error vector for the time differential feature can be obtained with

$$\begin{aligned} &\Delta_t\boldsymbol{F}(\boldsymbol{x}) - \Delta_t\boldsymbol{F}(\boldsymbol{x} - \Delta\boldsymbol{e}) \\ &= \{\boldsymbol{F}_t(\boldsymbol{x}) - \boldsymbol{F}_{t-1}(\boldsymbol{x})\} - \{\boldsymbol{F}_t(\boldsymbol{x} - \Delta\boldsymbol{e}) - \boldsymbol{F}_{t-1}(\boldsymbol{x} - \Delta\boldsymbol{e})\} \\ &= \{\boldsymbol{F}_t(\boldsymbol{x}) - \boldsymbol{F}_t(\boldsymbol{x} - \Delta\boldsymbol{e})\} - \{\boldsymbol{F}_{t-1}(\boldsymbol{x}) - \boldsymbol{F}_{t-1}(\boldsymbol{x} - \Delta\boldsymbol{e})\} \\ &= \Delta\boldsymbol{F}_t(\boldsymbol{x}) - \Delta\boldsymbol{F}_{t-1}(\boldsymbol{x} - \Delta\boldsymbol{e}) \end{aligned} \quad (9)$$

With threshold parameter $T$, we can generate the mask for the time differential feature

$$\boldsymbol{M}_t = \begin{cases} 1 & |\Delta\boldsymbol{F}_t(\boldsymbol{x}) - \Delta\boldsymbol{F}_{t-1}(\boldsymbol{x} - \Delta\boldsymbol{e})| < T, \\ 0 & otherwise \end{cases} \quad (10)$$

### 3.3. Actual generation of MFM for output of ICA

The above theory is applied to the output of ICA. 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 for recognition. Error $E_1(\omega, t)$ of estimated spectrum $x_1(\omega, t)$ can now be expressed as

$$E_1(\omega, t) = e_{11}(\omega)s_1(\omega, t) + w_{11}(\omega)s_2(\omega, t), \quad (11)$$

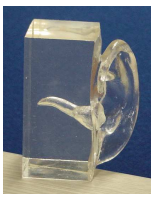

Figure 3: SIG2's ear

Figure 4: Humanoid SIG2 with two ears

where $e_{11}(\omega)$ is the filter error of $x_1(\omega, t)$, and $w_{11}$ is the scaled estimated filter of $x_1(\omega, t)$. $s_1(\omega, t)$ and $s_2(\omega, t)$ are an ideal separated signal spectrum.

To estimate an error $\Delta\boldsymbol{e}$, we assumed that the unmixing matrix would approximate well, and that the envelope for the power spectrum of the leaked signal would be similar to that for scaled $x_2(\omega, t)$. Therefore, the error spectrum of $x_1(\omega, t)$ can be expressed with scaling factor $\gamma_1$ as follows:

$$E_1(\omega, t) \simeq \gamma_1 x_2(\omega, t), \quad (12)$$

and we use $E_1(\omega, t)$ as the element of the error vector $\Delta\boldsymbol{e}$.

## 4. Experiments and Evaluation

### 4.1. Experiment Patterns

We used two omni-directional microphones installed as ears in the SIG2 humanoid robot (Figs.3 and 4) to evaluate the system. Three experiments are conducted and evaluation is performed in terms of SNR in MSLS and speech recognition accuracy.

1. Evaluating VAD for ICA in terms of SNR in MSLS,
2. Evaluating automatic MFM generation for original observed signals in terms of speech recognition rate, and
3. Evaluating automatic MFM generation for ICA output with/without VAD in terms of speech recognition rate.

Here, SNR in MSLS is defined as follows:

$$SNR = \frac{1}{F}\sum_{t=1}^{F} 10\log\sum_{i=0}^{24}\left(\frac{M_{ref}(i;t)^2}{(M_{out}(i;t) - M_{ref}(i;t))^2}\right), \quad (13)$$

where $F$ is the number of frames of the speech, $M_{out}(i;t)$ and $M_{ref}(i;t)$ represent $i$-th coefficient of frame $t$ of the target and the clean speech signal of MSLS, respectively. This indicates how the target speech signal is distorted compared with clean speech in the feature domain. We compare separated signal with clean signal, because the acoustic model of ASR is trained with clean speech signal in this paper.

### 4.2. Recording conditions

Two voices were recorded simultaneously from loudspeakers placed 1.0 m ($d$) from the robot shown in Figs.5 and 6. Asymmetric and symmetric configurations are used, because the performance of ICA separation is affected by the positions of speakers. The angle between two loudspeakers, $\theta$, is one of 30°, 60°, and 90°. The female speaker was to the left of the robot and the male was to its right. The room size was 4×5 m, with a reverberation time of 0.2–0.3 sec. We used combinations of two different words selected from a set of 200 phonemically balanced Japanese words.

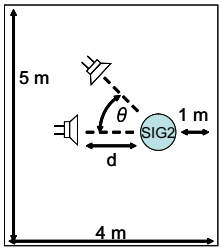


Figure 5: Configuration 1: Asymmetric speakers

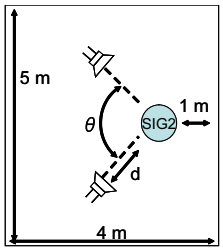


Figure 6: Configuration 2: Symmetric speakers

### 4.3. Configurations for experiments

We used multi-band Julian [4] as the MFT-based ASR. It uses a triphone-based acoustic model trained with clean speech of 216 words by 25 male and female speakers. These speakers were not used for evaluation (open test). The acoustic model uses three states and four Gaussians per mixture.

The main parameters for ICA were a sampling rate for data of 16 kHz, a frame length of 1,024 points, and a frame shift of 94. The initial values for the unmixing matrix, $\boldsymbol{W}(\omega)$, were given at random. We used 0.92 and 0.04 as the best threshold for the normal feature and time-difference feature, respectively. And we used 0.2 as the scaling factor, $\gamma_1$. These values are obtained empirically.

### 4.4. Results of Experiment

Figure 7 plots the distortions of observed signals, ICA signals and with VAD. Figure 8 plots the improvements in recognition accuracy with ICA and with or without masks generated automatically as indeicated by "our masking". ICA improved recognition accuracy by 25.5 %, and the auto generated mask improved recognition accuracy by 6-7 % on average. *A priori* (ideal) mask was prepared by using a clean speech signal and attained a recognition accuracy of over 97 %.

Figure 9 indicates VAD information works effectively separated ICA signals. MFM was effective in the speech period, but less effective in the non-speech period because of the difficulty of accurate masking. Instead of VAD, a binary-mask [5] in time-frequency domain may work well to reduce spectral distortions in non-speech period not in speech period, because there is a possibility of increasing distortions in speech period. The MFM also worked well by not using VAD. Finally, the VAD and MFM improve recognition accuracy by 12 % on average, and the recognition rates were over 80 %.

## 5. Conclusion

We constructed a robot audition system for unknown and/or dynamically-changing environments with providing minimum *a priori* information. To fulfill such requirements, we employed ICA and MFT-based ASR, and developed automatic MFM generation for ICA output signals.

For improvement of the speech recognition rate, we proposed the combination of ICA and VAD in order to reduce the signal leakage during the silent period. Throught the experiment, we demonstrated the utility of MFM in ICA. A combination of ICA and MFM improved recognition accuracy well.

One important work remaining is more precise estimates of reliable and unreliable components of separated sounds. Other future work includes frame-wise VAD, stationary noise reduction, moving talkers, and non-speech sound sources.

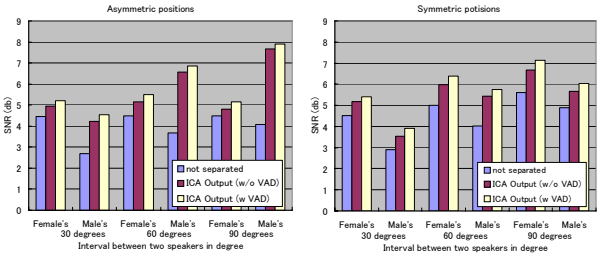


Figure 7: Improved SNR of MSLS with ICA and VAD

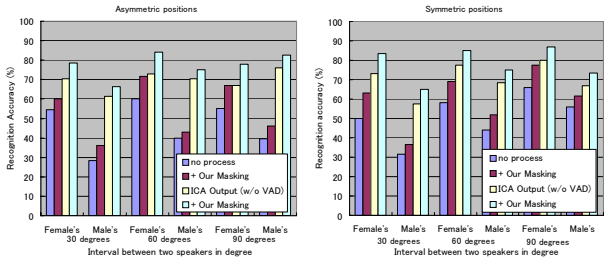


Figure 8: Improved recognition accuracy with ICA and MFM

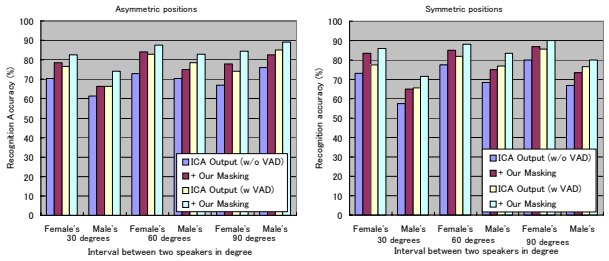


Figure 9: Improved recognition accuracy with MFM and VAD

## 6. References

[1] 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.

[2] 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.

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

[4] Multiband Julius, "http://www.furui.cs.titech.ac.jp/mband_julius/," .

[5] 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)*. 2005, pp. 209–214, IEEE.