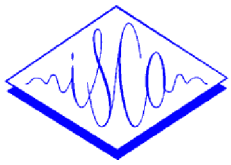


ITRW on Statistical and
Perceptual Audio Processing
Pittsburgh, PA, USA
September 16, 2006

Leak Energy Based Missing Feature Mask Generation for ICA and GSS and Its Evaluation with Simultaneous Speech Recognition

Shun'ichi Yamamoto[*], *Ryu Takeda*[*], *Kazuhiro Nakadai*[†], *Mikio Nakano*[†], *Hiroshi Tsujino*[†], *Jean-Marc Valin*[‡], *Kazunori Komatani*[*], *Tetsuya Ogata*[*], *and Hiroshi G. Okuno*[*]

[*]Graduate School of Informatics, Kyoto University, Kyoto, Japan
[†]Honda Research Institute Japan, Co., Ltd., Saitama, Japan
[‡]CSIRO ICT Center, Marsfield NSW, Australia

{shunichi,rtakeda,komatani,ogata,okuno}@kuis.kyoto-u.ac.jp,
{nakadai,nakano,tsujino}@jp.honda-ri.com, jean-marc.valin@csiro.au

Abstract

This paper addresses automatic speech recognition (ASR) for robots integrated with sound source separation (SSS) by using leak noise based missing feature mask generation. The missing feature theory (MFT) is a promising approach to improve noise-robustness of ASR. An issue in MFT-based ASR is automatic generation of the missing feature mask. To improve robot audition, we applied this theory to interface ASR and SSS which extracts a sound source originated from a specific direction by multiple microphones. In a robot audition system, it is a promising approach to use SSS as a pre-processor for ASR to be able to deal with any kind of noises. However, ASR usually assumes clean speech input, while speech extracted by SSS never fails to be distorted. MFT can be applied to cope with distortion in the extracted speech. In this case, we can assume that the noises included in extracted sounds are mainly leakages from other channels. Thus, we introduced leak noise based missing feature mask generation, which can generate a missing feature mask automatically by using information on leak noise obtained from other channels. To assess the effectiveness of the leak noise based missing feature mask generation, we used two methods for SSS: geometric source separation (GSS) and independent component analysis (ICA), and Multiband Julian for MFT based ASR. The two constructed systems, that is, GSS-based and ICA-based robot audition systems, were evaluated through recognition of simultaneous speech uttered by two speakers. As a result, we showed that the proposed leak noise based missing feature mask generation worked well in both systems.

1. Introduction

"*Listening to several things at once*" is people's dream and one goal of AI and robot audition, because psychophysical observations reveal that people can listen to at most two things at once [1]. Robot audition is an essential intelligent function for robots working with humans. Since robots encounter various sounds and noises, robot audition systems should be able to recognize a mixture of sounds and be noise-robust. Since robots are deployed in various environments, robot audition systems should require minimum *a priori* information about their acoustic environments and speakers [2, 3].

A robot audition system usually integrates sound source separation (SSS) and automatic speech recognition (ASR) subsystems. To minimize *a priori* information, we use blind source separation and a beamformer for SSS and missing feature theory (MFT) for ASR. The former literally separates sound signals from a mixture of sounds without assuming the characteristics of sound sources. The latter recognizes speech signals with a *clean* acoustic model by using missing feature masks (MFMs) that specify whether each spectral feature is reliable or not.

The most critical issue in missing-feature-mask generation is reliability estimation of spectral features in separated speech signals. The conventional studies on MFT focus only on cases where interfering sounds are quasi-stationary noises. This approach cannot handle two simultaneous speech signals. We assume that separated sounds are distorted mainly by signal leakage from other sound sources. If sound source separation is not perfect, separated sounds include sounds of non-target sources. We call the sounds leak noise. Therefore, in separating sounds, the system first estimates signal leakage and then identifies which spectral components are distorted. Finally, it creates MFMs that specify whether each spectral feature is reliable or not.

We demonstrated the performance of automatically generated MFMs by evaluating two robot audition systems:

GSS a Gometric Source Separation (GSS) with eight microphones, and automatic MFM generation for it, and

ICA an Independent Component Analysis (ICA) with two microphones and automatic MFM generation for it.

The separated speech signals and their associated MFMs were transmitted to MFT-based ASR (MFT-ASR) to recognize the speech.

In GSS, a multi-channel post-filter estimates signal leakage from different sources and quasi-stationary noises. In ICA, a SIMO (single-input multiple-output) model is used to obtain two channels (left and right) for each sound source. Then SIMO signals are used to estimate signal leakage.

This paper describes two systems, ICA and GSS, from the viewpoint of MFT. It presents MFT-ASR, explains benchmarks, and presents their results.

1.1. Related Work

Noise-robust ASRs have been studied extensively, for example in the AURORA project [4, 5]. One common method, in particular for in-car and telephony applications, is multi-condition training (training on a mixture of clean speech and noises) [6, 7]. Since an

acoustic model obtained by multi-condition training reflects all expected noises in specific conditions, ASR's use of such an acoustic model is effective only as long as speech including the expected noises is recognized. This assumption holds well for background noises in a car and on a telephone. However, multi-condition training may not be effective for robots, since they usually work in a dynamically changing noisy environment.

MFT-based ASR has been studied as a method of noise-robust ASR [8]. A spectrographic mask (also called MFM in this paper) is the set of tags that identify reliable and unreliable components of the spectrogram. MFT-based ASR uses this spectrographic mask to ignore corrupt signals during the decoding process. There are two main kinds of 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 recognize speech signals using reliable separated components and unreliable original input components [10, 11, 12, 13, 14].

Techniques of speech recognition in the presence of other speaker have been studied. McCowan *et al.* reported a combination of missing data speech recognition and microphone array [15]. Their system recognized speech mixed with stationary noise and a low level of background speech. Coy *et al.* reported a technique using speech fragment decoder based on missing data speech recognition [16]. Their systems divide spectral components into a set of spectro-temporal fragments, and recognized two simultaneous speech signals by using MFT. Brown *et al.* reported simultaneous speech recognition by using speech separation based on the statistics of binaural auditory features and missing data speech recognition. We present MFM generation method for a system based on sound source separation with multiple microphones and missing data speech recognition. We focus on simultaneous speech signals.

Although robot audition requires three essential functions, sound source localization, separation, and recognition of separated sounds, most researchers focus only on the first one. Nakadai *et al.* [17] have developed a robot audition system that can recognize three simultaneous speech signals for real-time and real-world processing using a pair of microphones installed in its ear position. Their system was developed by unifying four components: an active audition system to perceive auditory information better by controlling microphone parameters, a real-time multiple human tracking system that integrates an active audition system, face localization, face recognition and stereo vision, an active direction-pass filter (ADPF) to separate sound sources, and ASR using multiple direction- and speaker-dependent acoustic models. In other words, their system required a lot of information about acoustic environments and speakers.

2. General Recognition Architecture

A general architecture for recognizing several speech sources at once consists of three components:

1. Sound source separation,
2. MFT-based ASR, and
3. Automatic MFM generation.

The last is a bridge between the first and second components. In this section, we focus on the second component, MFT-based ASR, since it is used commonly by ICA and GSS systems. Overview of general recognition architecture is shown in Figure 1.

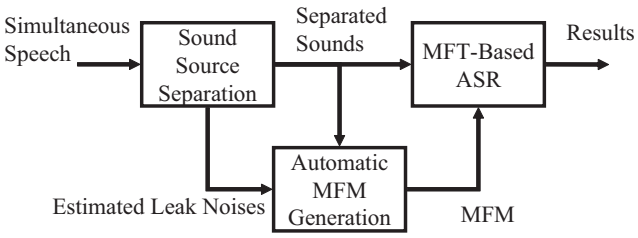


Figure 1: Overview of general recognition architecture

2.1. Acoustic Features of MFT-ASR

Since sound source separation is performed at the level of spectral representation, we adapt spectral features for MFT-ASR. Although Mel-Frequency Cepstrum Coefficient (MFCC) is a common acoustic feature for ASR, it is not appropriate for MFT-ASR, because a noise in each frequency band spreads to all coefficients in cepstral domain. We used the Mel Scale Log Spectrum (MSLS) obtained by applying Inverse Discrete Cosine Transformation to MFCCs. The calculation of MSLS is described by Yamamoto *et al.* [13]. The acoustic feature vector is composed of 48 spectral-related acoustic features: 24 spectral and 24 differential features.

2.2. Missing Feature Theory-based Automatic Speech Recognition

MFT-based ASR outputs a sequence of phonemes from acoustic features of separated speech and the corresponding MFMs. MFT-based ASR is an HMM-based recognizer, which is commonly used in conventional ASR systems. The only difference is in their decoding processes. In conventional ASR systems, estimation of a path with maximum likelihood is based on state transition probabilities and output probability in HMM. This process of estimating output probability is modified in MFT-ASR as follows: let $M(i)$ be an MFM vector which represents the reliability of the i-th acoustic feature. The output probability $b_j(x)$ is given by

$$b_j(x) = \sum_{l=1}^{L} P(l|S_j) \exp\left\{ \sum_{i=1}^{N} M(i) \log f(x(i)|l, S_j) \right\}, \quad (1)$$

where $P(\cdot)$ is a probability operator, $x(i)$ is an acoustic feature vector, N is the size of the acoustic feature vector, S_j is the j-th state, and $f(x|S_j)$ is a mixture of L multivariate Gaussians in j-th state. In marginalization approach [11], the output probability is calculated by using knowledge about unreliable features. If knowledge about any unreliable features is not available at all, the equation of output probability is equivalent to Equation 1.

We used hard mask (0-1 mask); i.e., 1 for reliable and 0 for unreliable. It is because performance of the hard masks were better than that of soft masks according to trying experiments.

For MFT-based ASR, we used Multiband Julius [18, 19], which is an extension of the Japanese real-time large vocabulary speech recognition engine Julius [20].

3. ICA-based Separation and MFM Generation

In the ICA system, sound source separation is ICA. In this section, we focus on ICA and MFM generation. Overview of ICA is shown in Figure 2.

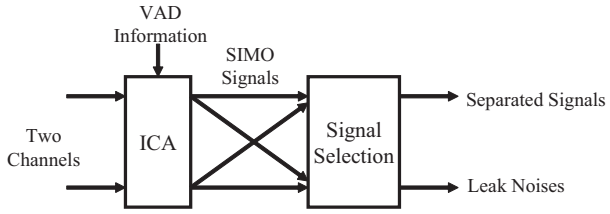


Figure 2: Overview of ICA

3.1. Frequency-domain ICA

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

The signals were assumed to be observed by linearly mixing sound sources, expressed as follows:

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

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 length N, where $[X]_{ji}$ denotes the matrix which includes element X in the i-th row and the j-th column. In our experiment, the number of microphones, J, was two and the number of multiple sound sources, L, was two.

The frequency-domain ICA works as follows. 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 for a frequency bin ω

$$\boldsymbol{Y}(\omega, t) = \boldsymbol{W}(\omega)\boldsymbol{X}(\omega, t), \tag{3}$$

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 (3), 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:

$$\boldsymbol{W}^{j+1}(\omega) = \boldsymbol{W}^{j}(\omega) - \alpha\{\text{off-diag}\langle\boldsymbol{\phi}(\boldsymbol{Y})\boldsymbol{Y}^h\rangle\}\boldsymbol{W}^{j}(\omega), \tag{4}$$

where α is a step size parameter that has effects on the speed of convergence, $[j]$ expresses the value of the jth step in the iterations, and $\langle\cdot\rangle$ denotes the time-averaging 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 $\phi(y_i) = \tanh(|y_i|)e^{j\theta(y_i)}$.

3.2. ICA's Two Problems of Permutation and Scaling

Frequency-domain ICA suffers from two ambiguities: *scaling ambiguity*, i.e., the power of separated signals differs at each frequency bin, and *permutation ambiguity*, i.e., signal components are swapped among different channels. We solved these ambiguities in order to recover the spectral representation as completely as possible using Murata's method [21].

To cope with the scaling ambiguity, we applied the inverse filter $\boldsymbol{W}^{-1}$ to the estimated source signal vector $\boldsymbol{Y}$. Let the reconstructed observation assuming input from only source i be $\boldsymbol{v}_i$.

$$\boldsymbol{v}_i = \boldsymbol{W}^{-1}\boldsymbol{E}_i\boldsymbol{W}\boldsymbol{x} = \boldsymbol{W}^{-1}(\ 0\cdots\boldsymbol{u}_i\cdots 0\ )^t, \tag{5}$$

where $\boldsymbol{E}_i$ represents the matrix in which the i-th diagonal element is one, and the others are zero; i.e., $\sum_i \boldsymbol{E}_i = \boldsymbol{I}$. This solution thus produces single-input multiple-output (SIMO) signals. SIMO signals are used to generate MFMs.

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.

3.3. Improvement by Voice Activity Detection (VAD)

Since the convolution model does not reflect actual acoustic environments, no methods based on this model can completely decompose each signal component.

The spectral distortion of separated signals is mainly caused by signal leakage in the desired speech signal. Suppose that two speakers are talking and one stops talking, as shown in Figure 3. It may often be the case with ICA that signal leakage is observed during that speaker's silent period. The spectral parts enclosed in the red box are instances of signal leakage. If such leakage is very strong, it is difficult to determine the end of a speech signal. An incorrect estimation of a period of speech would degrade the recognition accuracy of ASR severely.

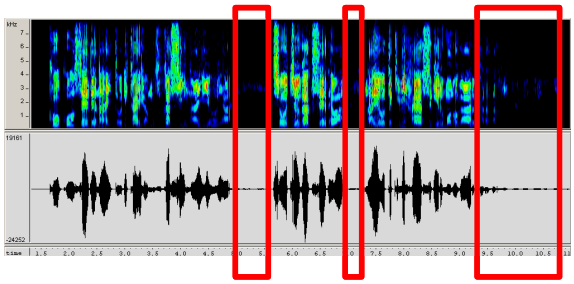


Figure 3: Leakage in spectrum for silent period in ICA

We used VAD that determines the period of utterance in order to improve the performance of separation and recognition. Since conventional VAD technologies assume quasi-stationary noises, they are usually not applicable for a mixture of simultaneous speech signals. The number of active speakers is used as VAD information, since ADPF provides such information stably [22]. The region of silent periods is filled with silent spectrum obtained in advance. If such a region is filled with 0 signals, it may not be treated as silence by ASR with an acoustic model that is trained with clean speech signals.

3.4. MFM Generation for an ICA System

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 $x_1(\omega, t)$ and $x_2(\omega, t)$ be SIMO signals of target source 1 and non-target source 2, respectively. These SIMO signals are selected from the elements of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ by using interaural intensity difference and interaural phase difference.

$x_1(\omega, t)$ denotes the signal selected from SIMO signals by using interaural phase and level difference. They satisfy the following equation:

$$m(\omega, t) = x_1(\omega, t) + x_2(\omega, t) \tag{6}$$

$$x_1(\omega, t) = a_1(\omega) s_1^{'}(\omega, t) \tag{7}$$

$$x_2(\omega, t) = a_2(\omega) s_2^{'}(\omega, t) \tag{8}$$

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

$$m(\omega) = W_1(\omega) s_1(\omega) + W_2(\omega) s_2(\omega) \tag{9}$$

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

The errors of separated spectrum are expressed as

$$s_1^{'}(\omega, t) = \alpha_1(\omega) s_1(\omega, t) + \beta_1(\omega) s_2(\omega, t) \tag{10}$$

$$s_2^{'}(\omega, t) = \beta_2(\omega) s_1(\omega, t) + \alpha_2(\omega) s_2(\omega, t) \tag{11}$$

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

$$e_1(\omega, t) = \Big(\alpha_1(\omega) a_1(\omega) - W_1(\omega)\Big) s_1(\omega, t) + \beta_1(\omega) a_1(\omega) s_2(\omega, t) \tag{12}$$

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,

$$\Big(\alpha_1(\omega) a_1(\omega) - W_1(\omega)\Big) s_1(\omega, t) \simeq 0 \tag{13}$$

$$\beta_1(\omega) a_1(\omega) s_2(\omega, t) \simeq \gamma_1 x_2(\omega, t) \tag{14}$$

$$e_1(\omega, t) \simeq \gamma_1 x_2(\omega, t) \tag{15}$$

Thus, since the error can be regarded as a leak noise obtained from non-target source, we generate MFMs, $\boldsymbol{M}$, for the estimated observed spectrum, $\boldsymbol{x}$, with the estimated error spectrum, $\boldsymbol{e}$ as follows:

$$\boldsymbol{M} = \begin{cases} 1 & |\boldsymbol{F}(\boldsymbol{x}) - \boldsymbol{F}(\boldsymbol{x} - \boldsymbol{e})| < \theta \\ 0 & otherwise \end{cases} \tag{16}$$

In addition, the masks for time differential feature are generated as follows:

$$\boldsymbol{M}(k) = \begin{cases} 1 & |\Delta \boldsymbol{F}_k(\boldsymbol{x}) - \Delta \boldsymbol{F}_{k-1}(\boldsymbol{x} - \boldsymbol{e})| < \hat{\theta} \\ 0 & otherwise \end{cases} \tag{17}$$

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

These equations are based on the idea that if the error spectrum distorts the separated signal, there is a difference between x and $x - e$ in feature domain. Even if the error spectrum is large, small difference between x and $x - e$ in feature domain does not affect performance of speech recognition.

4. GSS-based Separation and MFM Generation

In the GSS system, sound source separation is GSS with multi-channel post-filter. The GSS system has been reported in the literature [13, 14, 23]. GSS with multi-channel post-filter is shown in Figure 4.

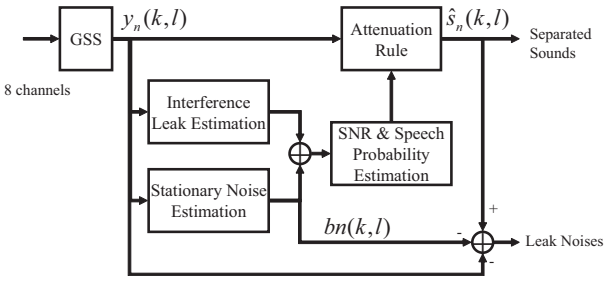


Figure 4: Overview of GSS with multi-channel post-filter

4.1. MFM Generation for GSS System

At first, we calculated leak noise by using input $y_n(k, l)$, output $\hat{s}_n(k, l)$, and the estimated background noise, $bn(k, l)$, of the multi-channel post-filter in frequency band k at frame l, where n is an index of a source. The variables filtered by the Mel filter bank are $Y_n(i, l)$, $\hat{S}_n(i, l)$, and $BN(i, l)$ in filter bank i, respectively. Leak noise $L_n(i, l)$ is defined by

$$L(i, l) = Y_n(i, l) - \hat{S}_n(i, l) - BN(i, l). \tag{18}$$

For each Mel-frequency band, the feature is considered reliable if the ratio of the leak noise over the input energy is greater than a threshold, T_{MFM}. This assumes that the more noise present in a certain frequency band, the lower the post-filter gain will be for that band.

The MFM $M_n(i, l), (i = 1, \cdots, N)$ for the spectral feature is defined as

$$M_n(i, l) = \begin{cases} 1, & \frac{L_n(i,l)}{Y_n(i,l)} < T_{MFM} \\ 0, & \text{otherwise} \end{cases}. \tag{19}$$

The MFM $M_n(i, l), (i = N + 1, \cdots, 2N)$ is defined as

$$M_n(i, l) = \prod_{t=l-2, t \neq l}^{l+2} M_n(i, t). \tag{20}$$

5. Experiments and Evaluation

To evaluate efficiency of automatic MFM generation based on leak estimation, we performed experiments on recognition of two simultaneous speech signals.

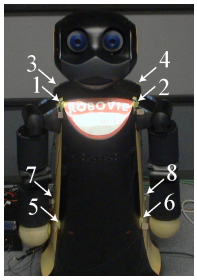


Figure 5: Robovie-R2

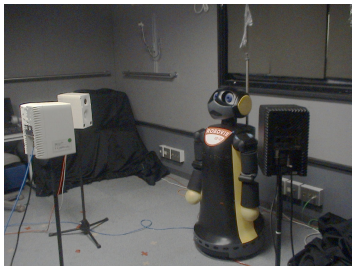


Figure 6: Robovie-R2 in the experiment room

5.1. Recording Conditions

We used Robovie-R2 for the experiments, with eight omnidirectional microphones on the body symmetrically. The transfer function of robot's body influences captured sound since microphones are not in the air. The positions of the microphones are shown in Figure 5. The distances between microphones 1 and 2, 1 and 3, and 1 and 5 are 25.6 cm, 18.8 cm, and 47.8 cm, respectively. For the ICA system, a pair of upper front microphones (1 and 2) was used. Simultaneous speech signals were recorded in a room, as shown in Figure 6. Their reverberation time was about 0.35 seconds (RT20). Japanese words were played simultaneously through loudspeakers at the same distance from the robot. Locations varied over five distances (at 50, 100, 150, 200, and 250 cm from the robot) and three directions. Because the waveform from thedistance of about 130cm can be treated as a plane wave for the most distant pair of microphones, we define 50 and 100 cm are near-field, and 150, 200 and 250 cm are far-field in this experiment. One loudspeaker was fixed in front of the robot, and the other was placed at 30, 60, or 90° left of the robot. The volume of the loudspeakers was set at the same level for all locations. 200 combinations of three different words were played for each configuration. The words were selected from 216 phonemically balanced words distributed by ATR. In other words, our systems recognize three simultaneous speech signals 200 times in each configuration.

5.2. Speech Recognition

Multiband Julius was used as the MFT-ASR. In the experiments, we used a triphone acoustic model and a grammar-based language model to recognize an isolated word. The triphone is an HMM which has 3 states and 4 mixtures in each state, and trained on 216 clean phonemically balanced words distributed by ATR. The size of the vocabulary was 200 words.

5.3. Configuration for Experiments

Parameters of our systems are determined experimentally. In ICA system, the threshold $\theta = 0.92$, and $\hat{\theta} = 0.05$ in Equations 16 and 17. In GSS system, the threshold $T_{MFM} = 0.75$.

5.4. Results

Figures 7 and 8 show word recognition rates for the ICA and GSS systems, respectively. The horizontal axis indicates speakers' positions, and the vertical one indicates word correct rates. For example, "30 deg., and 50 cm" on the horizontal axis means that one speaker is located 50 cm in front of the robot, and the other one is located 50 cm away at 30° left of center.

The ICA-based MFM generation (the ICA system) improved word correct rates by an average of 5.6%, and the GSS-based MFM generation (the GSS system) improved word correct rates by an average of 4.8%. The word correct rates of two simultaneous speech signals improved to an average of 67.8 and 88.0% for the ICA and GSS systems, respectively.

5.5. Discussion

The ICA system worked better in the near field than in the far field, because room transfer functions such as reverberation degraded the separation performance of ICA. The effect created by the intervals between the two speakers did not degrade the word correct rates much for the ICA system. The optimized unmixing matrix obtained by ICA is the reason for the system's robustness with intervals.

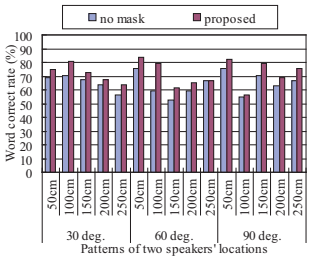
The GSS system worked better in the far field than in the near field, because a large difference in the time delay of arrival (TDOA) increases resolution of GSS. Narrow intervals between the speakers degraded the separation performance of GSS and the multi-channel post-filter, because the difference in TDOA decreased. GSS calculates the TDOA from locations of sound sources using the geometric constraints of microphones and does not take into consideration transfer functions of the body of the robot. The unmixing matrix obtained by ICA estimates such body transfer functions.

Some techniques which we used have limitations. In practical situations, the number and position of sources may vary. The GSS system with sound source localization can deal with the situation. The GSS system with eight microphones can separate up to eight sources, however, performance will decrease. On the other hand, it is difficult for the ICA system to deal with the situation. It is better that the number of microphones corresponds to the number of sources. Position of sources should be fixed while the ICA system adapts to training data during a few seconds. In this paper, we used simple missing data speech recognition with hard masks. However, there are more advanced MFT techniques, for example bounded marginalization, and the techniques may improve our system. Although hard masks was more effective than soft masks in our other experiments, there are possibilities of soft masks improving performance. MFT techniques also have the limitation. MFT cannot use orthogonal features since MFT should generally use spectral features. To cope with the limitation, we should improved spectral features or should develop MFM generation for mel-frequency cepstral coefficient which is commonly used for ASR.

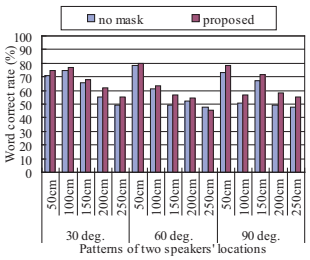
6. Conclusion

We presented two kinds of missing-feature approaches to separate and recognize two simultaneous speech signals. The ICA system uses two microphones for sound source separation. The GSS system uses GSS, a kind of beam-former, for sound source separation with eight microphones. Both separated sounds are recognized by MFT-ASR. These two systems were evaluated based on rates of recognition of simultaneous speech uttered by two speakers. We demonstrated that robot audition systems consisting of blind source separation and MFT-based ASR with automatic MFM generation recognized two simultaneous speech signals 5.6% and 4.8% better than conventional systems.

Since we focused on missing feature mask generation, we

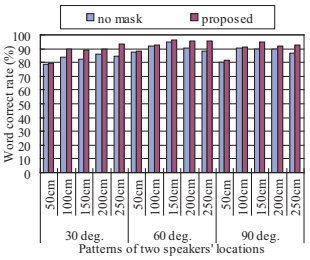


a) The speaker in the center direction

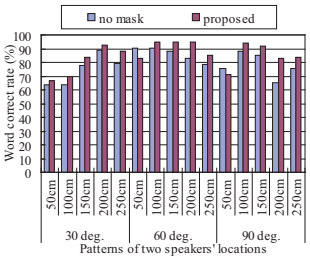


b) The speaker in the left direction

Figure 7: Recognition results with automatic MFM generation based on ICA (ICA system)



a) The speaker in the center direction



b) The speaker in the left direction

Figure 8: Recognition results with automatic MFM generation based on multi-channel post-filter (GSS system)

conducted the experiments using a recognition task as simple as possible. We are planning to conduct further experiments using more complicated tasks such as large-vocabulary continuous speech recognition.

7. References

[1] M. Kashino and T. Hirahara, "One, two, many – judging the number of concurrent talkers," *Journal of Acoustic Society of America*, vol. 99, no. 4, pp. Pt.2, 2596, 1996.

[2] H. G. Okuno, T. Nakatani, and T. Kawabata, "Interfacing sound stream segregation to speech recognition systems — preliminary results of listening to several things at the same time," in *Proc. of AAAI-96*. pp. 1082–1089, AAAI.

[3] H. G. Okuno, T. Nakatani, and T. Kawabata, "Understanding three simultaneous speakers," in *Proc. of IJCAI-1997*, pp. 30–35.

[4] AURORA, "http://www.elda.fr/proj/aurora1.html" "http://www.elda.fr/proj/aurora2.html," .

[5] D. Pearce, "Developing the ETSI AURORA advanced distributed speech recognition front-end & what next," in *Proc. of Eurospeech-2001*. 2001, ESCA.

[6] R. P. Lippmann, E. A. Martin, and D. B. Paul, "Multi-styletraining for robust isolated-word speech recognition," in *Proc. of ICASSP-87*. 1987, pp. 705–708, IEEE.

[7] M. Blanchet, J. Boudy, and P. Lockwood, "Environmentadaptation for speech recognition in noise," in *Proc. of EUSIPCO-92*, 1992, vol. VI, pp. 391–394.

[8] Bhiksha Raj and Richard M. Stern, "Missing-feature approaches in speech recognition," *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*. 2001, pp. 213–216, ESCA.

[11] M. 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, May 2000.

[12] P. Renevey, R. Vetter, and J. Kraus, "Robust speech recognition using missing feature theory and vector quantization," in *Proc. of Eurospeech-2001*. vol. 2, pp. 1107–1110, ESCA.

[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 *Proc. of ICRA 2005*. pp. 1489–1494, IEEE.

[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 *Proc. of IROS 2005*. pp. 897–902, IEEE.

[15] I. McCowan, A. Morris, and H. Bourlard, "Robust speech recognition with small microphone arrays using the missing data approach," in *Proc. of ICSLP-2002*, Martigny, Switzerland, pp. 2181–2184.

[16] André Coy and J. Barker, "Soft harmonic masks for recognising speech in the presence of a competing speaker," in *Proc. of INTERSPEECH-2005*. pp. 2641–2644, ISCA.

[17] K. Nakadai, H. G. Okuno, and H. Kitano, "Robot recognizes three simultaneous speech by active audition," in *Proc. of ICRA-2003*. pp. 398–403, IEEE.

[18] Y. Nishimura, T. Shinozaki, K. Iwano, and S. Furui, "Noise-robust speech recognition using multi-band spectral features," in *Proceedings of 148th Acoustical Society of America Meetings*, 2004, number 1aSC7.

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

[20] T. Kawahara and A. Lee, "Free software toolkit for Japanese large vocabulary continuous speech recognition," in *Proc. of ICSLP-2000*, vol. 4, pp. 476–479.

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

[22] K. Nakadai, K. Hidai, H. Mizoguchi, H. G. Okuno, and H. Kitano, "Real-time auditory and visual multiple-object tracking for robots," in *Proc. of IJCAI-2001*, pp. 1424–1432.

[23] S. Yamamoto, K. Nakadai, J.-M. Valin, J. Rouat, F. Michaud, K. Komatani, T. Ogata, and H. G. Okuno, "Genetic algorithm-based improvement of robot hearing capabilities in separating and recognizing simultaneous speech signals," in *Proc. of IEA/AIE'06*. vol. LNAI 4031, pp. 207–217, Springer-Verlag.