The 2010 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 18-22, 2010, Taipei, Taiwan

An Improvement in Automatic Speech Recognition Using Soft Missing Feature Masks for Robot Audition

Toru Takahashi, Kazuhiro Nakadai, Kazunori Komatani, Tetsuya Ogata and Hiroshi G. Okuno

Abstract— **We describe integration of preprocessing and automatic speech recognition based on Missing-Feature-Theory (MFT) to recognize a highly interfered speech signal, such as the signal in a narrow angle between a desired and interfered speakers. As a speech signal separated from a mixture of speech signals includes the leakage from other speech signals, recognition performance of the separated speech degrades. An important problem is estimating the leakage in time-frequency components. Once the leakage is estimated, we can generate missing feature masks (MFM) automatically by using our method. A new weighted sigmoid function is introduced for our MFM generation method. An experiment shows that a word correct rate improves from 66 % to 74 % by using our MFM generation method tuned by a search base approach in the parameter space.**

I. INTRODUCTION

Human-robot interaction (HRI) is one of the most essential topics in humanoid robot research. HRI definitely of a humanoid robot with robot-embedded microphones improves by a function of a natural speech interaction because speech communication is usually used in the daily communication between humans.

A humanoid robot has to deal with multiple sound sources simultaneously because the robot might have to listen to a mixture of speech signals uttered by several users at the same time, which is called "simultaneous speech." As the robot receives the simultaneous speech by the robot-embedded microphones, each microphone receives a mixture of speech signals. Therefore, sound source separation is required before recognizing a desired speech signal. However, a conventional approach used in HRI was to use microphones near the speaker's mouth to collect only the desired speech.

"Robot Audition" was proposed to realize hearing capability that makes a robot listen to several things simultaneously by using robot-embedded microphones in [1]. In robot audition [2], sound source separation as preprocessing of automatic speech recognition (ASR) is an actively-studied research topic [3]. Valin *et al.* have developed sound source localization and separation by Geometric Source Separation (GSS) and a multi-channel post-filter with 8 microphones to perform speaker tracking [4], [5].

A problem in robot audition is an integration of preprocessing and ASR because there is a mismatch between preprocessing (e.g. sound source separation) and the conventional ASR systems. As sound source separation is an ill-posed, sound source separation produces separation errors. It is impossible to perfectly estimate effects of reverberations and environmental noises which change dynamically using microphones embedded in a mobile robot. Conventional ASR systems assume that the input speech is clean or contaminated with a known noise source, because their target is a mainly telephony application, which is able to assume a high signal-to-noise ratio (SNR).

To integrate them, Missing-feature-theory base ASR (MFT-ASR) is used [6], [7]. It is considered as a recognition system for dealing with weighted acoustic features. A problem is to generate missing-feature-masks (MFM), i.e. hard and soft masks.

In this paper, we design an automatic soft MFM generation method based on two weighted sigmoid functions. We implement the proposed soft MFM generation as a module of our open-sourced robot audition software HARK [8]. After that, to show validity of the proposed weighted soft MFM, we show effectiveness through simultaneous speech recognition.

The rest of this paper is organized as follows: Section II describes MFT. Section III describes the design of the soft MFM generation algorithm for robot audition. Section IV describes the implementation of robot audition with the proposed soft MFM method generation using our robot audition software HARK. Section V evaluates our proposed soft MFM generation method through recognition of three simultaneous speeches and a human-robot interaction scenario. The last section concludes this paper.

II. MISSING FEATURE THEORY

Missing-Feature-Theory (MFT) is a promising approach for such integration of the preprocessing and the ASR. MFT is known as a technique to improve noise-robustness of speech recognition by masking out unreliable acoustic features using a so-called *missing feature mask* (MFM) [9], [10], [11]. The effectiveness of MFT has been widely reported [6], [7].

Yamamoto *et al.* are the first research group which introduced a MFT to integrate ASR into preprocessing [12]. They showed the remarkable improvement of speech recognition of separated sounds although they use a priori knowledge to generate MFMs. First, the reliability of each time-frequency (TF) component was calculated by comparing separated speech with the corresponding clean speech (pre-mixed sound source signal). Then, a hard MFM consisting of 0 or 1 for each TF component was calculated from the reliability

T. Takahashi K. Komatani, T. Ogata and H. G. Okuno are with the Department of Intelligence andScience and Technology. Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan {tall, Komatani, ogata, okuno}@kuis.kyoto-u.ac.jp

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

978-1-4244-6676-4/10/$25.00 ©2010 IEEE

based on a manually-defined threshold. The generated MFM is called a priori MFM.

An automatic MFM generation rises as an issue without a priori knowledge. Actually, this is the primary issue in MFT approaches. Regardless of a lot of studies on MFT, this is still an open question. Although most studies on automatic MFM generation focused on single channel input or on binaural input, Yamamoto *et al.* have developed an automatic MFM generation based on microphone array processing [13]. First, they showed that unreliable features generated by preprocessing are mainly caused by leakage energy from other sound sources. They developed a microphone array based technique to estimate the reliability of each TF component from this leakage energy by taking the property of the multi-channel post-filter and environmental noises into consideration. Their automatic MFM generation was able to correctly estimate around 70% of unreliable TF components compared to a priori MFM. Thus, the ASR performance drastically improved, and simultaneous speech recognition by three speakers was attained. However, they still used a hard MFM consisting of 0 or 1, while the reliability of each TF component is estimated as a continuous value from 0 to 1. This means that some useful information included in the estimated reliability may be thrown away with hard MFM.

A soft MFM with a continuous value from 0 to 1 was reported as a better masking approach[11], because soft masking can directly deal with the reliability of an input signal and probabilistic methods can be applied at the same time. Actually, Bayesian mask estimation algorithms were proposed in [14], [15]. Barker *et al.*[7] used a sigmoid function to estimate a soft MFM. Therefore, we believe that a soft MFM also improves the performance of robot audition in recognition of preprocessed (separated) speeches. A hard MFM approach may work when a small number of TF components are overlapped between a target speech and a noise, but in speech noise cases like barge-in and simultaneous speech, many TF components are overlapped. Since a soft masking approach directly uses reliability, it can also deal with overlapped TF components properly.

III. The Design of Soft Missing Feature Mask

This section describes the design of our soft MFM. It is based on reliability estimation of TF components. First, the reliability of TF component is defined, and then, separated speeches are analyzed based on the reliability to model soft MFM generation, and parameter optimization for the modeled soft MFM generation is also shown.

A. Definition of reliability

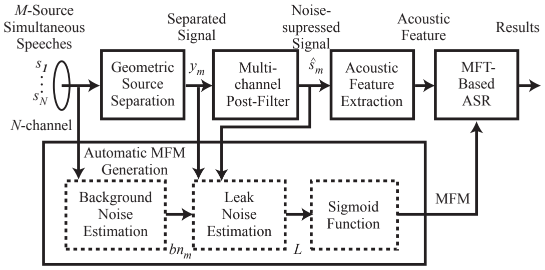

Fig. 1. Geometric source separation with multi-channel post-filter.

Figure 1 shows the components in HARK. An automatic MFM generation component integrates the preprocessing with the ASR components. Geometric Source Separation (GSS) and multi-channel post-filtering are the preprocessing components. Acoustic feature and ASR components are common components of speech recognition system. GSS is a hybrid sound source separation method between beamforming and blind source separation. Thus, an N-channel input signal $\boldsymbol{S}$ which consists of M sound sources s_m is separated into each sound source, y_m. We use an 8 channel microphone array (N = 8), and the number of sound sources, M, is decided in a sound localization module (see Sec.IV-A). However, as mentioned in the previous section, sound source separation is an ill-posed problem, and thus y_m still includes non-stationary cross-talk (leakage) and stationary background noises. Multi-channel post-filtering suppresses these two types of noises and produces a noise-suppressed signal $\hat{s}_m$.

The reliability of $\hat{s}_m$ for each TF component (frame and frequency indices are omitted for simplification) was defined by

$$L = \frac{\hat{s}_m + bn}{y_m}. \tag{1}$$

where bn is a background noise separately-estimated by using a minima controlled recursive algorithm (MCRA). Note that L corresponds to leakage level because leakage is a dominant factor to make a TF component unreliable. When a background noise level is zeros, L means the ratio estimation between leakage and source levels. Thus, $L = 0.5$ corresponds to 0 dB.

B. Analysis of separated speech based on reliability

We analyzed the characteristics of L. We found that there are two peaks in a histogram of L for separated speeches when three speeches were uttered simultaneously. One corresponds to leakage components, and the other target speech components. We found the same tendency for some interval from 10 to 90 degrees.

C. Modeling a soft mask

In hard masking, a hard MFM is generated by thresholding as follows:

$$HM_m = \begin{cases} 1, & L > \theta \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where θ is a threshold. Dynamic acoustic features called Δ features are commonly used with static acoustic features to improve the ASR performance. Δ features are calculated by linear regression of five consecutive frames. Let static acoustic features be $m(k)$, Δ features are defined by

$$\Delta m(k) = \frac{1}{\sum_{i=-2}^{2} i^2} \sum_{i=-2}^{2} i \cdot m(k+i), \tag{3}$$

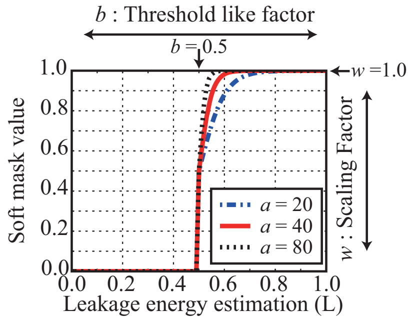

Fig. 2. Sigmoid function (Eq.(9) for soft mask generation.

where k is frequency indices. Thus, hard masks for Δ features are defined in the same way.

$$\Delta HM_m(n) = \prod_{i=n-2, i\neq n}^{n+2} HM_m(i). \quad (4)$$

where n shows the frame index. However, such a linear discrimination with θ makes misclassified TF components. Thus, we decided to introduce soft masking. We assume that these two groups follow Gaussian distributions. The distribution function for Gaussian is defined by

$$d(x) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right) \quad (5)$$

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}dt \quad (6)$$

Let the distribution functions for leakage and target speech be $d_n(R)$ and $d_s(R)$, respectively. A normalized speech reliability can be defined by

$$B(R) = \frac{d_s(R)}{d_s(R) + d_n(R)} \quad (7)$$

This is a sigmoid-like function defined using error functions $\mathrm{erf}(\cdot)$. Since the calculation cost of $B(R)$ is expensive, we decided to use a typical sigmoid function $Q(R)$ rather than to use this complicated function directly. We, then, defined a soft MFM based on $Q(R)$ as follows [16]:

$$SM_m = w_1 Q(R|a,b), \quad (8)$$

$$Q(x|a,b) = \begin{cases} \frac{1}{1+\exp(-a(x-b))}, & x > b \\ 0, & \text{otherwise} \end{cases}, \quad (9)$$

where w_1 is an weight factor for static features ($0.0 \leq w_1$). $Q(\cdot|a,b)$ is a modified sigmoid function which has two tunable parameters. a corresponds to a trend of the sigmoid function. b corresponds to an x-offset of the sigmoid function. We also defined soft masks for Δ features as

$$\Delta SM_m(k) = w_2 \prod_{i=k-2, i\neq k}^{k+2} Q(R(i|a,b)). \quad (10)$$

where w_2 is an weight factor for Δ features ($0.0 \leq w_2$).

TABLE I

A SEARCH SPACE FOR SOFT MFM PARAMETERS

parameter	range	step
a	20 – 80	20
b	0.5	–
w_1, w_2	0.1 – 1.5	0.1

D. Parameter optimization for soft masking

Figure 2 shows the relationship between soft and hard MFMs. When a is infinity and $w = 1.0$ in Eq. (9), a soft MFM works as a hard MFM. In this case, b works as threshold, θ. a and b can be derived from Eqs. (9) and (7), but it is difficult to attain analytical solutions for them. In addition, for w_1 and w_2, we have no theoretical evidence for parameter estimation. We, thus, the measured recognition performance of three simultaneous speech signals to optimize these parameters by using a robot having eight omni-directional microphones shown in Fig. 4. Simultaneous speech signals were recorded in a room with $RT_{20} = 0.35$. Three different words were played simultaneously with the same loudness from three loudspeakers located 1 m away from the robot. Each word was selected from the ATR phonetically balanced wordset consisting of 216 Japanese words. The direction of a loudspeaker was fixed in front of the robot, and the others were located at ±30, ±60, ±90, ±120, ±150 degrees to the robot. For each configuration, 200 combinations of the three different words were played.

Table I shows a search space for a soft MFM parameter set $\boldsymbol{p} = (a, b, w_1, w_2)$. Figure 3 shows an example of the average case over loud speaker angles where three loudspeakers were localed at $(0.\ \theta.\ -\theta)$, $(\theta = 30, 60, 90, 120, 150)$. For the other conditions, we obtained a similar tendency for w_1-w_2 parameter optimization. We also performed parameter optimization for a and b, and found that the result with a common tendency is obtained for every layout. Therefore, we obtained the optimized parameter set $\boldsymbol{p}_{opt}$ defined by

$$\boldsymbol{p}_{opt} = \arg\max_{\boldsymbol{p}} \frac{1}{5\cdot 3} \sum_{\theta\in\{30,\ldots,150\}} \Big(\mathrm{WC}_\theta(a,b,w_1,w_2) + \mathrm{WR}_\theta(a,b,w_1,w_2) + \mathrm{WL}_\theta(a,b,w_1,w_2)\Big),$$

where WC_θ, WR_θ, and WL_θ show word correct rates for the center, right and left loudspeakers where their locations are $(0, \theta, -\theta)$ degrees, respectively.

Finally, we attained the optimal parameter set for the soft MFM as $\boldsymbol{p}_{opt} = (40, 0.5, 0.1, 0.1)$.

IV. A Robot Audition System

Our robot audition system consists of five major components shown in Fig. 1. Our proposed soft MFM generation was described in the previous section. This section explains the other four components, A: Geometric Source Separation, B: Multi-channel post-filter, C: Acoustic feature extraction, and D: MFT-ASR. Our robot audition system uses several techniques such as sound source localization and tracking, which are described in [17].

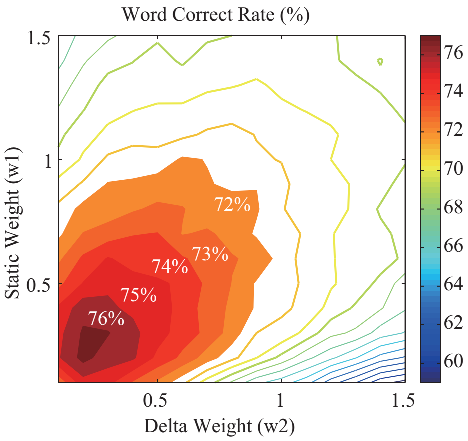

Fig. 3. ASR Performance for the center loudspeaker in a word correct rate. This is the average case over loudspeaker angles where three loudspeakers were located at (0. θ. $-\theta$), ($\theta = 30, 60, 90, 120, 150$). This shows the results for the parameters w_1 and w_2.

A. Geometric source separation

GSS is a hybrid algorithm of Blind Source Separation (BSS) and beamforming. It relaxes BSS's limitations such as permutation and scaling problems by introducing "geometric constraints" obtained from the locations of microphones and sound sources. Unlike the Linearly Constrained Minimum Variance (LCMV) beamformer that minimizes the output power subject to a distortion-less constraint, GSS explicitly minimizes cross-talk, leading to faster adaptation. The method is also interesting for use in the mobile robotics context because it allows easy addition and removal of sources. Using some approximation, it is also possible to implement separation with relatively low complexity.

Our GSS was modified so as to provide faster adaptation using stochastic gradient and shorter time frame estimation. The locations of sound sources are estimated with Multiple Signal Classification (MUSIC). It is a frequency-domain adaptive bearmforming method that produces a sharp local peak corresponding to a sound source direction, thus its noise robustness improves in the real world.

The formulation of GSS is described. Suppose that there are M sources and N ($\geq M$) microphones. A spectrum vector of M sources at frequency ω, $\boldsymbol{s}(\omega)$, is denoted as $[s_1(\omega)s_2(\omega)\dots s_M(\omega)]^T$, and a spectrum vector of signals captured by the N microphones at frequency ω, $\boldsymbol{x}(\omega)$, is denoted as $[x_1(\omega)x_2(\omega)\dots x_N(\omega)]^T$, where T represents a transpose operator. $\boldsymbol{x}(\omega)$ is, then, calculated as

$$\boldsymbol{x}(\omega) = \boldsymbol{H}(\omega)\boldsymbol{s}(\omega), \tag{11}$$

where $\boldsymbol{H}(\omega)$ is a transfer function matrix. Each component H_{nm} of the transfer function matrix represents the transfer function from the m-th source to the n-th microphone. The source separation is generally formulated as

$$\boldsymbol{y}(\omega) = \boldsymbol{W}(\omega)\boldsymbol{x}(\omega), \tag{12}$$

where $\boldsymbol{W}(\omega)$ is called a *separation matrix*. The separation is defined as finding $\boldsymbol{W}(\omega)$ which satisfies the condition that output signal $\boldsymbol{y}(\omega)$ is the same as $\boldsymbol{s}(\omega)$. In order to estimate $\boldsymbol{W}(\omega)$, GSS introduces two cost functions, that is, separation sharpness (J_{SS}) and geometric constraints (J_{GC}) defined by

$$\begin{aligned} J_{SS}(\boldsymbol{W}) &= \|E[\boldsymbol{y}\boldsymbol{y}^H - \text{diag}[\boldsymbol{y}\boldsymbol{y}^H]]\|^2, & (13)\\ J_{GC}(\boldsymbol{W}) &= \|\text{diag}[\boldsymbol{W}\boldsymbol{D} - \boldsymbol{I}]\|^2, & (14)\end{aligned}$$

where $\|\cdot\|^2$ indicates the Frobenius norm, $\text{diag}[\cdot]$ is the diagonal operator, $E[\cdot]$ represents the expectation operator and H represents the conjugate transpose operator. $\boldsymbol{D}$ shows a transfer function matrix based on a direct sound path between a sound source and each microphone. The total cost function $J(\boldsymbol{W})$ is represented as

$$J(\boldsymbol{W}) = \alpha_S J_{SS}(\boldsymbol{W}) + J_{GC}(\boldsymbol{W}), \tag{15}$$

where α_S means the weight parameter that controls the weight between the separation cost and the cost of the geometric constraint. This parameter is usually set to $\|\boldsymbol{x}^H\boldsymbol{x}\|^{-2}$ according to [4]. In an online version of GSS, $\boldsymbol{W}$ is updated by minimizing $J(\boldsymbol{W})$

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \mu\boldsymbol{J}'(\boldsymbol{W}_t), \tag{16}$$

where $\boldsymbol{W}_t$ denotes $\boldsymbol{W}$ at the current time step t, $\boldsymbol{J}'(\boldsymbol{W})$ is defined as an update direction of $\boldsymbol{W}$, and μ means a step-size parameter.

B. Multi-channel post-filter

A multi-channel post-filter is used to enhance the output of the GSS algorithm. It is a spectral filter using an optimal noise estimator. This method is a kind of spectral subtraction, but it generates less musical noises and distortion, because it takes temporal and spectral continuities into account. We extended the original noise estimator to estimate both stationary and non-stationary noise by using multi-channel information, while most post-filters address the reduction of a type of noise, stationary background noise.An input of the multi-channel post-filter is the output of GSS; $\boldsymbol{y}$. An output of the multi-channel post-filter is $\hat{\boldsymbol{s}}$, which is defined as

$$\hat{\boldsymbol{s}} = G\boldsymbol{y}, \tag{17}$$

where G is a spectral gain. The estimation of G is based on minimum mean-square error estimation of spectral amplitude. To estimate G, noise variance is estimated.

The noise variance estimation λ_m is expressed as

$$\lambda_m = \lambda_m^{stat.} + \lambda_m^{leak}, \tag{18}$$

where $\lambda_m^{stat.}$ is the estimate of the stationary component of the noise for source m at frame t for frequency f, and λ_m^{leak} is the estimate of source leakage.

We computed the stationary noise estimate, $\lambda_m^{stat.}$, using MCRA technique To estimate λ_m^{leak}, we assumed that the interference from other sources is reduced by factor η (typically -10dB $\leq \eta \leq$ -5 dB) by LSS. The leakage estimate is thus expressed as

$$\lambda_m^{leak} = \eta \sum_{i=0, i\neq m}^{M-1} Z_i, \tag{19}$$

where Z_i is the smoothed spectrum of the m-th source, Y_m and recursively defined (with $\alpha - 0.7$) [18]:

$$Z_m(f,t) = \alpha Z_m(f,t-1) + (1-\alpha) Y_m(f,t). \quad (20)$$

C. Acoustic feature extraction

To estimate reliability of acoustic features, we have to exploit the fact that noises and distortions are usually concentrated in some areas in the spectro-temporal space. Most conventional ASR systems use *Mel-Frequency Cepstral Coefficient* (MFCC) as an acoustic feature, but noises and distortions are spread to all coefficients in MFCC. In general, Cepstrum based acoustic features like MFCC are not suitable for MFT-ASR, Therefore, we use *Mel-Scale Log Spectrum* (MSLS) as an acoustic feature.

MSLS is obtained by applying inverse discrete cosine transformation to MFCCs Then three normalization processes are applied to obtain noise-robust acoustic features; mean power normalization, spectrum peak emphasis and spectrum mean normalization. The details are described in [19]. These three normalization processes correspond to three normalization performed against MFCC; C0 normalization, liftering, and Cepstrum mean normalization. The acoustic feature vector composes 13 MSLS features, their derivatives and Δ log power, i.e., a 27-dimensional MSLS-based acoustic vector was used.

D. Missing Feature Theory based ASR

Two critical issues remain; what kinds of preprocessing are required for ASR, and how does ASR use the characteristics of preprocessing besides using an acoustic model with multi-condition training. We exploited an interfacing scheme between preprocessing and ASR based on MFT.

MFT uses MFMs in a temporal-frequency map to improve ASR. Each MFM specifies whether a spectral value for a frequency bin at a specific time frame is reliable or not. Unreliable acoustic features caused by errors in preprocessing are masked using MFMs, and only reliable ones are used for a likelihood calculation in the ASR decoder. The decoder is an HMM-based recognizer, which is commonly used in conventional ASR systems. The estimation process of output probability in the decoder is modified in MFT-ASR.

Let $M(i)$ be a MFM vector that represents the reliability of the i-th acoustic feature. The output probability $b_j(x)$ is given by the following equation:

$$b_j(x) = \sum_{l=1}^{L} P(l|S_j) \exp\left\{ \sum_{i=1}^{N} M(i) \log f(x(i)|l, S_j) \right\}, \quad (21)$$

where $P(\cdot)$ is a probability operator, $x(i)$ is an acoustic feature vector, N is the size of the acoustic feature vector, and S_j is the j-th state.

For implementation, we used Multiband Julian, which is based on the Japanese real-time large vocabulary speech recognition engine Julian. It supports various HMM types such as shared-state triphones and tied-mixture models. Network grammar is supported for a language model. It works as a standalone or client-server application. To run as a server, we modified the system to be able to communicate acoustic features and MFM via a network.

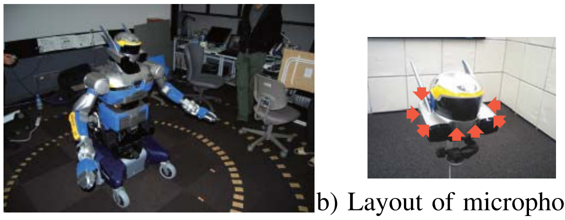
b) Layout of microphones.
a) A humanoid robot HRP-2.

Fig. 4. A humanoid robot HRP-2 with an 8 ch microphone array.

V. Evaluation

To evaluate the proposed robot audition system with soft MFM generation, simultaneous speech recognition was performed in a manner of isolated word recognition. Also, the system was introduced to a human-robot interaction scenario, that is, a meal order taking task.

A. Experimental setup

We used a humanoid robot HRP-2 with eight microphones around the top of the head for an experiment of simultaneous speech recognition. It was placed at the center of a circle in Fig. 4. Three loudspeakers were used to play three speeches simultaneously. A loudspeaker was fixed in front of the robot, and two other loudspeakers were located at ±30, ±60, ±90, ±120, or ±150. The distance between the robot and each loudspeaker was 1 m. Three males are used as sound sources. Each test dataset consists of 200 combinations of three different words randomly-selected from ATR phonetically balanced 216 Japanese words.

For an acoustic model in ASR, we trained a 3-state and 16-mixture triphone model based on Hidden Markov Model (HMM) using 27 dimensional MSLS features. To make evaluation fair, we performed an open test, that is, the acoustic model was trained with a different speech corpus from test data. For training data, we used Japanese News Article Speech Database containing 47,308 utterances by 300 speakers [20]. After adding 20 dB of white noise to the speech data, the acoustic model was trained, which is a well-known technique to improve noise-robustness of an acoustic.

B. Recognition of three simultaneous speeches

For comparison, we evaluated two kinds of MFMs, the hard and soft masks. The conventional hard MFM is defined by Eqs. (2) and (4). The proposed soft MFM is defined by Eqs. (8) and (10). The parameters for the mask generation are optimized. Word correct rates (WCR) were measured with these MFMs for a test dataset. Figures 5,6 illustrate averaged word correct rates for the center and the left speakers.

For the center, left and right speakers, we can say that our proposed soft MFM drastically improved the ASR performance. The improvement is better, especially, when the angle between loudspeakers. When the angle is wider, the number of overlapping TF components is smaller. Thus the

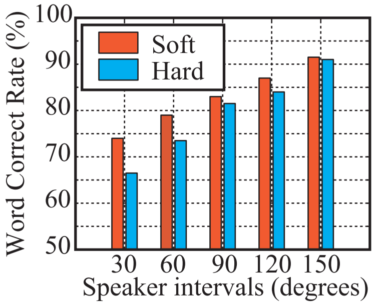

Fig. 5. Word correct rate for the center speaker.

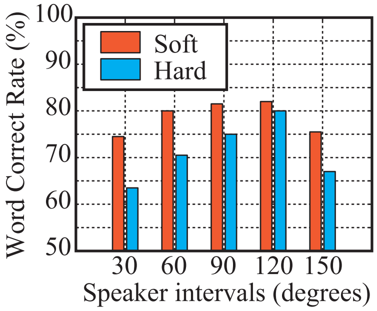

Fig. 6. Word correct rate for the left speaker.

difference of the ASR performance between hard and soft masks is less.

In case of 150 degrees, the angle between the left and right speakers is 60 degrees. Thus, the WCR of the left speaker degrades because this is a kind of the high interfered speech signal recognition. In this condition, our proposed soft MFM also improved the ASR performance, drastically. This proves that the proposed soft MFM is able to cope with the large number of overlapping TF components even in the highly-overlapped cases. The improvement of the proposed soft MFM reached around 8 points by averaging three speaker cases. For the right speakers, its trend is similar to the left speaker. The average WCR is depends on a speaker. For some speakers, WCR is less than the other speakers, but average WCR over speakers is almost same as the result showed in Figs. 5, 6.

VI. CONCLUSION AND FUTURE WORK

We presented the integration method between the preprocessing and the MFT-ASR to recognize a highly interfered speech signal. The MFT is adopted to integrate microphone-array-based preprocessing with sound source localization, and separation into ASR. For generating missing feature masks, we used a weighted soft missing feature mask taking a continuous value between 0 and w instead of a conventional hard missing feature mask taking a binary value, 0 or 1.

The resulting HARK-based robot audition system with automatic soft mask generation improves the performance of ASR in three simultaneous speeches, in particular for narrower intervals of two adjacent speakers up to 30 degrees. In 30 degrees condition, we improved word correct rate from 66% to 74% in the simultaneous speech recognition of robot audition to realize natural human-robot interaction. The conventional system worked up to 60 degrees. Therefore, the soft mask system provides opportunities to deploy a robot audition system to more realistic multi-party interaction.

Future work includes simultaneous speech recognition experiment with different distance values, the development of the automatic tuning method, the detailed analysis, and more applications. For example, extensive benchmarking to analysis the performance of ASR with a wide variation of speaker configurations under various acoustic environments.

Our research is partially supported by the Grant-in-Aid for Scientific Research and Global COE Program.

REFERENCES

[1] K. Nakadai, *et al.*, "Active Audition for Humanoid," *Proc. of AAAI-2000*, pp.832–839, 2000.
[2] I. Hara, *et al.*, "Robust Speech Interface Based on Audio and Video Information Fusion for Humanoid (HRP-2)," *Proc. of IROS 2004*, pp.2404–2410, 2004.
[3] K. Nakadai, *et al.*, "Improvement of Recognition of Simultaneous Speech Signals Using AV Integration and Scattering Theory for Humanoid Robots," *Speech Communication*, Vol.44, No.1–4, pp.97–112, 2004.
[4] J.-M. Valin, *et al.*, "Enhanced Robot Audition Based on Microphone Array Source Separation with Post-Filter," *Proc. of IROS 2004*, pp.2133–2128, 2004.
[5] J.-M. Valin, *et al.*, "Robust Localization and Tracking of Simultaneous Moving Sound Sources Using Beamforming and Particle Filtering," *Robotics and Autonomous Systems Journal*, Vol.55, No.3, pp.216–228, 2007.
[6] J. Veth, *et al.*, "Missing Feature Theory in ASR: Make Sure You Miss the Right Type of Features," *Proc. of Workshop on Robust Methods for ASR in Adverse Conditions, Tampere*, pp.231–234, 1999.
[7] J. Barker, *et al.*, "Soft Decisions in Missing Data Techniques For Robust Automatic Speech Recognition, " *Proc. of ICSLP 2000*, Vol.I, pp.373–376, 2000.
[8] K. Nakadai, *et al.*, "An Open Source Software System For Robot Audition HARK and Its Evaluation," *Proc. of HUMANOIDS 2008*, pp.561–566, 2008.
[9] R. P. Lippmann, *et al.*, "Robust Speech Recognition with Time-Varying Filtering, Interuptions, aned noise," *Proc. of Eurospeech 1997*, pp.365–372, 1997.
[10] M. Cooke, *et al.*, "Robust Automatic Speech Recognition with Missing and Unreliable Acoustic Data," *Speech Communication*, pp.267–285, Vol.34, No.3, May, 2000.
[11] B. Raj, *et al.*, "Missing-Feature Approaches in Speech Recognition," *Signal Processing Magazine*, Vol.22, No.5, pp.101–116, 2005.
[12] S. Yamamoto, *et al.*, "Improvement of Robot Audition by Interfacing Sound Source Separation and Automatic Speech Recognition with Missing Feature Theory," *Proc. of ICRA 2004*, pp.1517–1523, 2004.
[13] S. Yamamoto, *et al.*, "Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory," *Proc. of ICRA 2005*, pp.1489–1494, 2005.
[14] M. L. Seltzer, *et al.*, "A Bayesian classifier for spectrographic mask estimation for missing feature speech recognition," *Speech Communication*, Vol.43, pp.379–393, 2004.
[15] P. Renevey, *et al.*, "Missing feature theory and probabilistic estimation of clean speech components for robust speech recognition," *Proc. of Eurospeech 1999*, pp.2627–2630, 1999.
[16] T. Takahashi, *et al.*, "Missing-Feature-Theory-Based Robust Simultaneous Speech Recognition System with Non-clean Speech Acoustic Model," *Proc. of IROS 2009*, pp.2730–2735, 2009.
[17] S. Yamamoto, *et al.*, "Design and Implementation of A Robot Audition System for Automatic Speech Recognition of Simultaneous Speech, " *Proc.of ASRU 2007*, pp.111–116, 2007.
[18] S. Yamamoto, *et al.*, "Genetic Algorithm-Based Improvement of Robot Hearing Capabilities in separating and Recognizing Simultaneous Speech Signals," *Proc. of IEA/AIE'06*, LNAI 4031, pp.207–217, 2006.
[19] Y. Nishimura, *et al.*, "Noise-Robust Speech Recognition Using Multi-Band Spectral Features," *Proc. of 148th ASA Meetings*, 1aSC7, 2004.
[20] K. Itou, *et al.*, "JNAS: Japanese speech courpus for large vocavulary continuous speech recognition research", *J. of Acoustical Society Japan*, (E) 30(3), pp.199–206, 1999.