PALADYN Journal of Behavioral Robotics

VERSITA

Research Article · DOI: 10.2478/s13230-010-0005-1 JBR · 1(1) · 2010 · 37-47

Soft missing-feature mask generation for Robot Audition

Toru Takahashi[1]*, Kazuhiro Nakadai[2 3]†, Kazunori Komatani[1], Tetsuya Ogata[1], Hiroshi G. Okuno[1]

1 Department of Intelligence and Science and Technology Graduate School of Informatics, Kyoto University, Kyoto 606-8501, Japan

2 Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako, Saitama 351-0114, Japan,

3 Mechanical and Environmental Informatics, Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Tokyo, 152-8552, Japan

Received 21 February 2010

Accepted 19 March 2010

Abstract

This paper describes an improvement in automatic speech recognition (ASR) for robot audition by introducing Missing Feature Theory (MFT) based on soft missing feature masks (MFM) to realize natural human-robot interaction. In an everyday environment, a robot's microphones capture various sounds besides the user's utterances. Although sound-source separation is an effective way to enhance the user's utterances, it inevitably produces errors due to reflection and reverberation. MFT is able to cope with these errors. First, MFMs are generated based on the reliability of time-frequency components. Then ASR weighs the time-frequency components according to the MFMs. We propose a new method to automatically generate soft MFMs, consisting of continuous values from 0 to 1 based on a sigmoid function. The proposed MFM generation was implemented for HRP-2 using HARK, our open-sourced robot audition software. Preliminary results show that the soft MFM outperformed a hard (binary) MFM in recognizing three simultaneous utterances. In a human-robot interaction task, the interval limitations between two adjacent loudspeakers were reduced from 60 degrees to 30 degrees by using soft MFMs.

Keywords

Robot Audition · HARK · missing-feature-theory · soft mask generation · simultaneous speech recognition · Automatic Speech Recognition · sound source separation · sound localization

1. Introduction

Human-robot interaction (HRI) is one of the most essential topics in behavioral robotics. HRI is improved by the inclusion of a natural speech communication function with robot-embedded microphones because we generally use speech in our daily communication. In an everyday environment a user may "barge in" or interrupt a robot while it is speaking, or several users may speak at the same time, which is termed "simultaneous speech." In addition, the robot itself generates sounds due to its fans and actuators, so the robot must be able to deal with multiple sound sources simultaneously. A conventional approach in human-robot interaction is to use microphones near the speaker's mouth to collect only the desired speech. Kismet of MIT has a pair of microphones with pinnae, but a human partner still used a microphone close to the speaker's mouth [4]. A group communication robot, Robita of Waseda University, assumes that each human participant uses a headset microphone [16]. Thus, "Robot Audition" was proposed to realize the hearing capability that allows a robot to listen to several things simultaneously by using the it's embedded microphones in [18].

Robot audition has now been actively studied for more than ten years, as typified by organized sessions on robot audition at the IEEE/RAS International Conferences on Intelligent Robots and Systems (IROS 2004–2009), and also a special session on robot audition at the IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP 2009) of the Signal Processing Society. Sound source separation as pre-processing of automatic speech recognition (ASR) is an actively-studied research topic in this field.

Hara et al. reported a humanoid robot, HRP-2, which uses a microphone array to localize and separate a mixture of sounds, and which is capable of recognizing speech commands in a noisy environment [12]. HRP-2 can recognize one speaker's utterance under noisy or interfering speakers. Nakadai et al. reported SIG, a humanoid robot which uses a pair of microphones to separate multiple speech signals through an active direction-pass filter, and recognizes each separated speech phrase using ASR [20]. They demonstrated that even when three speakers utter words at the same time, SIG was able to recognize what each speaker said. However, since their system used 51 acoustic models trained under different conditions at the same time, the system incurs a high computational cost, and performance deteriorates in an environment with unexpected and/or dynamically changing noises. Kim et al. have developed another binaural sound-source localization and separation method by integrating sound-source localization obtained by CSP (Cross-power Spectrum Phase) and that obtained by visual information with an EM algorithm [14]. This system assumes that only one predominant sound exists in each time frame. Valin et al. have developed sound-source localization and separation by Geometric Source Separation, and a multi-channel post-filter with 8 microphones to perform speaker tracking [32, 33].

Sound-source separation is an ill-posed problem, however, because it is impossible to perfectly estimate the effect of reverberation and environmental noises which change dynamically using microphones embedded in a mobile robot. Thus, sound-source separation produces separation errors. To remove such errors, a non-linear speech enhancement method such as Minima Controlled Recursive Average

*E-mail: {tall,komatani,ogata,okuno}@kuis.kyoto-u.ac.jp
†E-mail: nakadai@jp.honda-ri.com

(MCRA) [5] or Minimum Mean Square Error (MMSE) [10] is often used. Indeed, non-linear speech enhancement removes the separation errors, but it also generates some distortions like musical noise, which drops ASR performance.
ASR systems, on the other hand, assume that the input speech is clean or contaminated with a known noise source, because their target is mainly telephony applications, which generally involve a high signal-to-noise ratio (SNR).
There is, therefore, a mismatch between pre-processing sound source separation and ASR systems. It follows that one of the most important issues in robot audition is **integration between preprocessing and ASR**.

1.1. Missing Feature Theory

" Missing Feature Theory (MFT)" is a promising approach for integration between pre-processing and ASR. MFT is a technique which is know to improve the noise-robustness of speech recognition by masking out unreliable acoustic features using a so-called " missing feature mask (MFM)" [6, 15, 27]. The effectiveness of MFT has been widely reported in connected digit recognition for telephony applications [1, 8], speaker verification [9, 23], de-reverberation [24], and recognition of separated speech in a binaural way [35].
Yamamoto and Nakadai et al. are the first research group to introduce a Missing Feature Theory (MFT) to integrate ASR with a binaural robot audition system [39]. First, the reliability of each time-frequency (TF) component was estimated by comparing separated speech with the corresponding clean speech. Then, a hard MFM consisting of 0 or 1 for each TF component was generated based on the reliability using a manually-defined threshold. Since this mask generation algorithm used reference speech signals to estimate the reliability, the generated MFM is called *a priori* MFM. Although they used *a priori* hard MFM, they showed a remarkable improvement in the speech recognition of separated sounds. This showed the effectiveness of MFT approaches.
Automatic MFM generation rises as an issue; actually, this is the primary issue in MFT approaches, and remains an open question despite numerous MFT studies. Although most works on automatic MFM generation focus on single-channel input, or on binaural input, Yamamoto and Valin et al. have developed an automatic MFM generation process based on microphone-array processing [37].
First, they showed that unreliable features generated by pre-processing are mainly caused by energy leakage from other sound sources. A microphone-array-based technique was developed to estimate the reliability of each time-frequency component from this energy leakage, by considering the properties of a multi-channel post-filter process and environmental noises. 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 of three voices was attained. However, they still used a hard binary MFM consisting of a value equal to 0 or 1, while the reliability of each TF component is estimated as a continuous value in the range 0 to 1.
This means that some useful information which is contained in the estimated reliability may be lost if hard MFM is used.

1.2. Soft Missing Feature Mask

A soft MFM with a continuous value from 0 to 1 was reported as a better masking approach[27] than hard MFM,both because soft masking can directly deal with the reliability of an input signal, and because probabilistic methods can be applied at the same time. Bayesian mask estimation algorithms were proposed in [28, 29], while Barker et al.[2] used a sigmoid function to estimate a soft MFM. We therefore believe that a soft MFM also improves the performance of robot audition in the recognition of pre-processed (separated) speeches. A hard MFM approach may work when a small number of time-frequency components are overlapped between the target speech and a noise, but in speech-noise cases such as barge-in or simultaneous speech, many time-frequency components are overlapped. Since a soft masking approach directly uses reliability, it can also deal with overlapped time-frequency components properly.
In this paper, we present an automatic, soft-MFM generation method based on a sigmoid function which is then implemented as a module of our open-sourced robot audition software, HARK [21]. To show the validity of the proposed soft-MFM method, we demonstrate its effectiveness through tasks including simultaneous speech recognition, and human-robot interaction involving a humanoid HRP-2 robot taking a meal order.
The rest of this paper is organized as follows: Section 2 describes the design of a soft-MFM-generation algorithm for robot audition. Section 3 describes the implementation of a robot-audition system with the proposed soft-MFM generation method, using HARK, our robot audition software. Section 4 illustrates how HRP-2 receives a meal order by means of robot audition functions. Section 5 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.

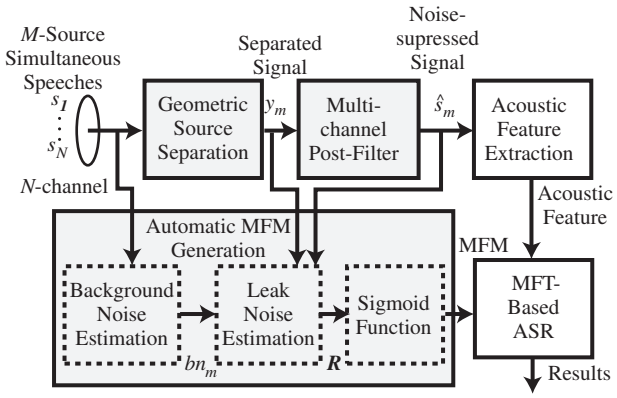

Figure 1. Geometric source separation with multi-channel post-filter.

2. The Design of Soft Missing Feature Mask

This section describes the design of our soft MFM which is based on reliability estimation for time-frequency components. First, the reliability of the time-frequency component is defined, then separated speeches are analyzed based on the measured reliability in order to model soft-MFM generation. Parameter optimization for the modeled soft-MFM generation is also shown.

2.1. Definition of reliability

Figure 1 shows the core steps of pre-processing in HARK, i.e. Geometric Source Separation (GSS) [25] and multi-channel post-filtering. GSS

is a hybrid sound-source separation method between beam forming and blind-source separation. Thus, an N-channel input signal 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.3.1). As mentioned in the previous section, however, 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 both of these types of noise and produces a noise-suppressed signal $\hat{s}_m$.
The reliability of $\hat{s}_m$ for each time-frequency component (frame and frequency indices are omitted for simplification) was defined by

$$R = \frac{\hat{s}_m + bn}{y_m}. \quad (1)$$

where bn is a background noise which is separately estimated using MCRA [5]. Note that R corresponds to leakage level because leakage is a dominant factor in making a time-frequency component unreliable, as mentioned in the previous section.

2.2. Analysis of separated speech based on reliability

We analyzed the characteristics of R and found that there are two peaks in the histogram for separated speeches when three speeches were uttered simultaneously.
One peak corresponds to the leakage components, and the other matches target-speech components. We checked several intervals from 10, 20, $\cdots$, 80, 90 degrees, and found the same tendency for every interval.

2.3. Modeling a soft mask

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

$$HM_m = \begin{cases} 1, & R > T_{MFM} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

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

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

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

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

where k now shows the frame index.

Such a linear discrimination with T_{MFM}, however, leads to misclassified time-frequency components; we therefore decided to introduce soft masking. We assume that these two groups follow Gaussian distributions. The distribution function for a 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

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

$$= \frac{1 + \mathrm{erf}\left(\frac{R-\mu_s}{\sigma_s\sqrt{2}}\right)}{2 + \mathrm{erf}\left(\frac{R-\mu_s}{\sigma_s\sqrt{2}}\right) - \mathrm{erf}\left(\frac{R-\mu_n}{\sigma_n\sqrt{2}}\right)} \quad (8)$$

This is a sigmoid-like function defined using error functions $\mathrm{erf}(\cdot)$. Since there is a high calculation cost for $B(R)$, 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 [30]:

$$SM_m = w_1 Q(R|a,b), \quad (9)$$

$$Q(x|a,b) = \begin{cases} \frac{1}{1+\exp(-a(x-b))}, & x > b \\ 0, & \text{otherwise} \end{cases}, \quad (10)$$

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 while b represents an x-offset. We also defined soft masks for Δ features as

$$\Delta SM_m(k) = w_2 \prod_{i=k-2, i\neq k}^{k+2} Q(R(i|a,b)). \quad (11)$$

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

2.4. Parameter optimization for soft masking

Figure 2 shows the relationship between soft and hard MFMs. When a is infinity and $w = 1.0$ in Equation (10), a soft MFM works as a hard MFM. In this case, b works as threshold, T_{MFM}. Parameters a and b can be derived from Eqs. (10) 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 measured the recognition performance of three simultaneous speech signals in order to optimize these parameters for a robot having eight omni-directional microphones as shown in Figure 8. Simultaneous speech signals were recorded in a room with RT_{20} = 0.35. Three different words were played simultaneously at the same volume from three loudspeakers located 2 m away

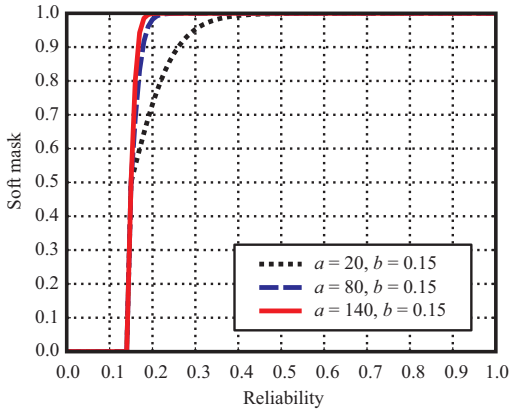

Figure 2. *Sigmoid function (Eq.(10) for soft mask generation when b = 0.15 and a = 20, 80, 140.*

Table 1. *A search space for soft MFM parameters*

parameter	range	step
a	40	–
b	0.1 – 1.5	0.1
w_1	0.1 – 1.5	0.1
w_2	0.1 – 1.5	0.1

from the robot. Each word was selected from the ATR phonetically balanced wordset consisting of 216 Japanese words. The direction of one loudspeaker was fixed in front of the robot, and the others were located at ±10, ±20, ···, ±80, ±90 degrees to the robot. For each configuration, 200 combinations of the three different words were played.

Table 1 shows a search space for a soft MFM parameter set $p = (a, b, w_1, w_2)$. Figure 3 shows an example of w_1-w_2 parameter optimization for the center speaker when the loud speakers were located at 0, 90, and -90 degrees. For 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 a similar result is obtained for every layout. We therefore obtained the optimized parameter set p_{opt} defined by

$$p_{opt} = \underset{p}{\text{argmax}} \frac{1}{9} \sum_{\theta=10}^{90} \frac{1}{3} (\text{WC}_\theta(a, b, w_1, w_2) + \\ + \text{WR}_\theta(a, b, w_1, w_2) + \text{WL}_\theta(a, b, w_1, w_2)) \quad (12)$$

where WC_θ, WR_θ, and WL_θ indicate the number of correct words for each of 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 $p_{opt} = (40, 0.5, 0.1, 0.2)$.

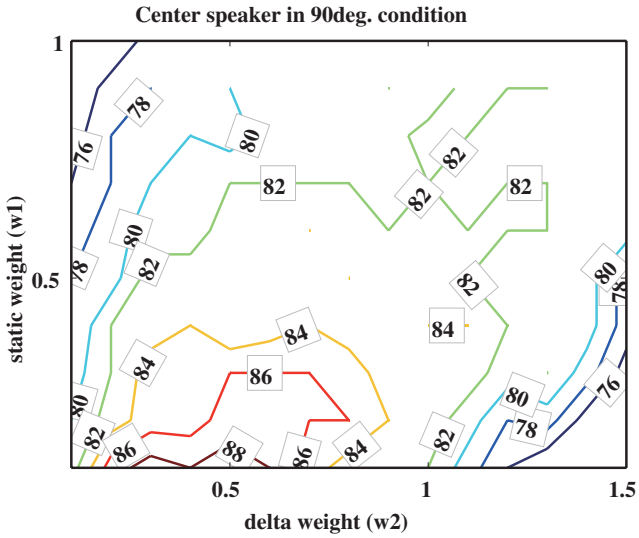

Figure 3. *ASR Performance for the center loudspeaker in a word correct rate. This is the case where three loudspeakers were located at (0. 90. -90). This shows the results for the parameters w_1 and w_2.*

3. A Robot Audition System

Our robot audition system consists of five major components shown in Figure 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,

D: Missing-Feature-Theory-based Automatic Speech Recognition (MFT-ASR).

Besides the five components, our robot audition system uses several techniques such as sound-source localization and tracking, which are described in [38].

3.1. Geometric source separation

GSS is a hybrid algorithm of Blind Source Separation (BSS) and beamforming [25]. BSS has a number of limitations such as permutation and scaling problems, which can be relaxed in GSS by the introduction of "geometric constraints". These are obtained from the locations of microphones and sound sources. Unlike the Linearly Constrained Minimum Variance (LCMV) beamformer which 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 approximations, 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). This is a frequency-domain adaptive beamforming method which

produces a sharp local peak corresponding to a sound-source direction, thus its noise robustness improves in the real world.
To formulate GSS, 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 $[s_1(\omega)s_2(\omega)\ldots s_M(\omega)]^T$, and a spectrum vector of signals captured by the N microphones at frequency ω, $\boldsymbol{x}(\omega)$, is denoted as $[x_1(\omega)x_2(\omega)\ldots x_N(\omega)]^T$, where T represents a transpose operator. $\boldsymbol{x}(\omega)$ is, then, calculated as

$$\boldsymbol{x}(\omega) = \boldsymbol{H}(\omega)\boldsymbol{s}(\omega), \tag{13}$$

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

$$\boldsymbol{y}(\omega) = \boldsymbol{W}(\omega)\boldsymbol{x}(\omega), \tag{14}$$

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

$$J_{SS}(\boldsymbol{W}) = \|E[\boldsymbol{y}\boldsymbol{y}^H - \mathrm{diag}[\boldsymbol{y}\boldsymbol{y}^H]]\|^2, \tag{15}$$
$$J_{GC}(\boldsymbol{W}) = \|\mathrm{diag}[\boldsymbol{W}\boldsymbol{D} - \boldsymbol{I}]\|^2, \tag{16}$$

where $\|\cdot\|^2$ indicates the Frobenius norm, $\mathrm{diag}[\cdot]$ 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

$$J(\boldsymbol{W}) = \alpha_S J_{SS}(\boldsymbol{W}) + J_{GC}(\boldsymbol{W}), \tag{17}$$

where α_S represents the weighting parameter which controls the weighting 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 [34]. In an online version of GSS, $\boldsymbol{W}$ is updated by minimizing $J(\boldsymbol{W})$

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \mu\boldsymbol{J}'(\boldsymbol{W}_t), \tag{18}$$

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.

3.2. Multi-channel post-filter

A multi-channel post-filter is used to enhance the output of the GSS algorithm[36]. It is a spectral filter using an optimal noise estimator described in [10]. This method is a kind of spectral subtraction [3], 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 only address the reduction of a specific type of noise: stationary background noise [17].
The output of GSS $\boldsymbol{y}$ forms an input of the multi-channel post-filter;

An output of the multi-channel post-filter $\hat{\boldsymbol{s}}$, is defined as

$$\hat{\boldsymbol{s}} = G\boldsymbol{y}, \tag{19}$$

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

$$\lambda_m = \lambda_m^{stat.} + \lambda_m^{leak}, \tag{20}$$

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 [5]. 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

$$\lambda_m^{leak} = \eta \sum_{i=0, i\neq m}^{M-1} Z_i, \tag{21}$$

where Z_i is the smoothed spectrum of the m-th source, Y_m and recursively defined (with $\alpha - 0.7$) [40]:

$$Z_m(f,t) = \alpha Z_m(f,t-1) + (1-\alpha)Y_m(f,t). \tag{22}$$

3.3. Acoustic feature extraction

To estimate the 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) [26] 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. Three normalization processes are then applied in order to obtain noise-robust acoustic features: mean-power normalization, spectrum-peak emphasis and spectrum-mean normalization. The details are described in [22]. 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.

3.4. Missing Feature Theory based ASR

Several robot audition systems with pre-processing and ASR have been reported so far [11, 19]. Such systems just combine preprocessing with ASR, and focus on the improvement of SNR and real-time processing.
Two critical issues remain: what kinds of pre-processing are required for ASR, and how does ASR use the characteristics of pre-processing 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 in 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:

$$b_j(x) = \sum_{l=1}^{L} P(l|S_j) \exp\left\{ \sum_{i=1}^{N} M(i) \log f(x(i)|l, S_j) \right\}, \qquad (23)$$

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 [13], which is based on the Japanese real-time large-vocabulary speech-recognition engine, Julian [31]. It supports various HMM types such as shared-state triphones and tied-mixture models. Network grammar is supported for a language model. It can work as a stand-alone 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.

4. System Implementation

This section introduces our open-sourced robot audition software, HARK, then we describe implementation of the proposed soft-MFM generation as a new module for HARK, and a robot audition system with the new module.

4.1. Open-Sourced Robot Audition Software HARK

We consider a software environment for robot audition research. Most studies focus on their own robot platforms, and their systems are unavailable for other researchers and research groups. It is pleasant and useful for robot audition researchers to share a common platform, because the researchers do not need to make their own robot platforms from scratch, and they can easily change a module to compare it with another. Thus, we implemented each technique described in the previous sections as a component for modular-based architecture called FlowDesigner [7][1]. FlowDesigner provides a flexible and efficient software development environment, which is achieved by flexible replacement of modules and fast data communication between modules. These advantages are achieved by the pull architecture of FlowDesigner.

We then released a set of components as HARK (Honda Research Institute Japan Audition for Robots with Kyoto University; the word also means of "listen" in old English).[2] [21]. HARK provides a user-customizable total robot audition system including multi-channel sound acquisition, sound localization, sound separation and ASR. HARK also provides opportunities to discuss general applicability, platform dependency, customization and tuning. Difficulties such as customization to another platform or tuning to another acoustic environment have not yet been discussed. In addition, HARK has a possibility to stimulate the robot audition research area, and to provide an effective tool for interdisciplinary research such as natural language processing, navigation, and HRI. According to user feedback, performance and stability of HARK will improve.

As related work, Valin released sound-source localization and separation software for robots called "ManyEars" as General Public License (GPL) open-source software (OSS). This is the first software which can provide generally-applicable and customizable robot audition systems. The only missing function is ASR. "ManyEars" is limited to sound-source localization and separation. ASR has yet not been included as it has a lot of parameters which affect the performance of a total robot audition system.

4.2. Implementation of soft MFM generation for HARK

Figure 4 shows an example of a robot audition system constructed using HARK. A green rectangle represents a module, while a connection between modules is indicated by a black arrow. The top-left module (**AudioStreamFromMic**) captures sounds using a robot-embedded microphone array. After frequency analysis (**MultiFFT**), sound sources are localized using **LocalizeMUSIC**, **SourceTracker**, and **SourceIntervalExtender**. The localized sound sources are separated with **GSS**, and **PostFilter** enhances the separated sounds. MSLS features are calculated using **MSLSExtraction**, **Delta**, and **FeatureRemover** in a Mel-frequency domain (**MelFilterBank**). The MFM is estimated in **SMFMGeneration**. Finally, MSLS features and the corresponding MFMs are sent to MFT-ASR via a Socket interface using **SpeechRecognitionClient**. These modules are prepared in advance. Thus, users can easily construct their own robot audition systems by selecting modules and connecting them using a GUI interface.

The newly-developed module is shown as **SMFMGeneration** at the center of Figure 4. It has four terminals which are shown as black dots on the left and right edges of the box. Three terminals on the left edge correspond to the input signals such as $\hat{s}_m$, y_m, and bn in Equation (1). The last terminal on the right edge shows the output, that is, a soft-MFM vector for a frame. This module also has three parameters such as "FBANK", "THRESHOLD", and "TILT" shown in the property setting window in Figure 5, which appears by double-clicking the green box of **SMFMGeneration**. FBANK represents the number of dimensions for a static part of a MFM vector defined in Eq. (9), that is 13 in our setting. THRESHOLD and TILT correspond to a and b defined in Eq. (10), respectively.

5. 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 with the robot taking a meal order.

5.1. 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 Figure 8. Three loudspeakers

[1] *http://flowdesigner.sourceforge.net/*

[2] *It is available at http://winnie.kuis.kyoto-u.ac.jp/HARK/.*

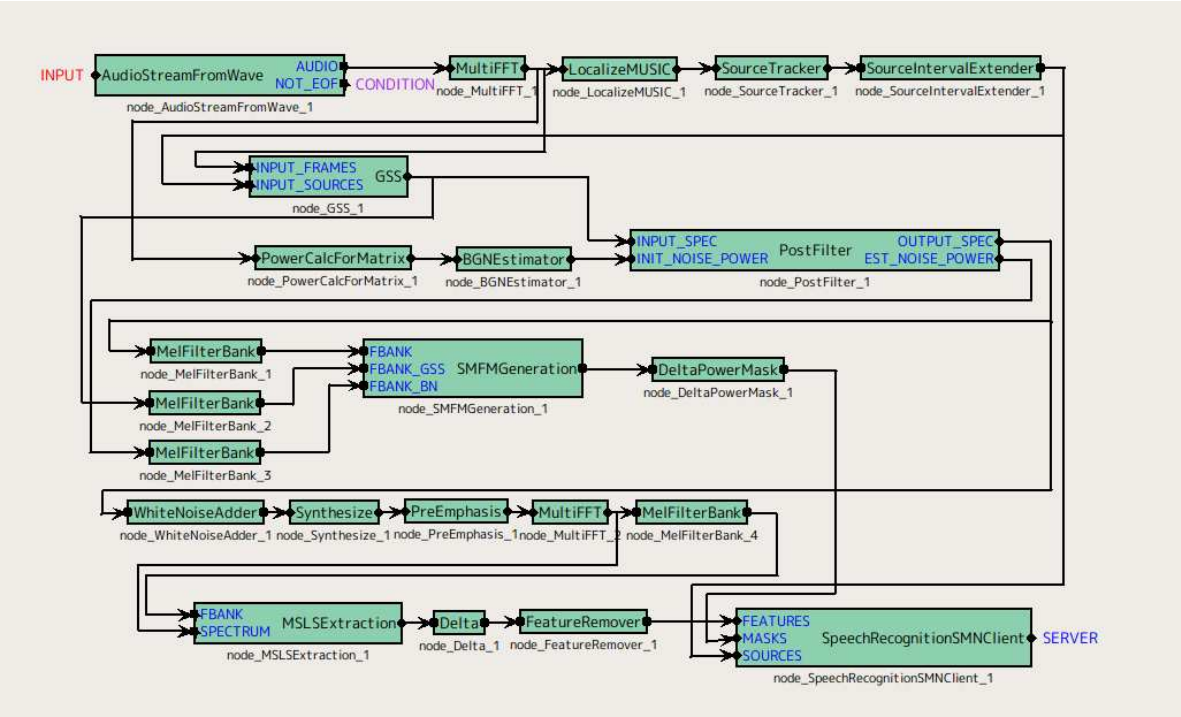

Figure 4. *A robot audition system constructed by HARK modules.*

Figure 5. *A parameter tab of a soft mask generation module.*

a) A humanoid robot HRP-2.

b) Layout of microphones.

Figure 8. *A humanoid robot HRP-2 with an 8 ch microphone array.*

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 shown in Table 2. The distance between the robot and each loudspeaker was 1 m. Four combinations of sound sources were used shown in Table 3. Thus, we generated 20 test data sets (3×4). 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 a Japanese News Article Speech Database containing 47,308 utterances by 300 speakers. After adding 20 dB of white noise to the speech data, the acoustic model was trained with the white-noise-added training data, which is a well-known technique for improving the noise-robustness of an acoustic model for ASR.

Table 2. *Loudspeaker locations.*

Layout	center (deg.)	left (deg.)	right (deg.)
LL1	0	30	-30
LL2	0	60	-60
LL3	0	90	-900
LL4	0	120	-120
LL5	0	150	-150

Table 3. *Sound Source Combination.*

combination	center	left	right
SC1	female 1	female 2	female 3
SC2	female 4	female 5	female 6
SC3	male 1	male 2	male 3
SC4	male 4	male 5	male 6

5.2. Recognition of three simultaneous speeches

For comparison, we evaluated three kinds of MFMs as follows:

1. hard MFM : conventional hard MFM defined by Eqs. (2) and (4).
2. soft MFM (unweighted) : soft MFM defined by Eqs. (9) and (11) with $w_1 = w_2 = 1$.
3. soft MFM (proposed) : the proposed soft MFM defined by Eqs. (9) and (11) using the optimized MFM parameter set p_{opt}.

Word correct rates were measured with these MFMs for every test dataset described above.

Figsures 9–11 illustrate averaged word correct rates for the center, left and right speakers, respectively.

For the center speaker, we can say that our proposed soft MFM drastically improved ASR performance. For the left or right speaker, while the improvement was less than that for the center speaker, we still find improvements to some extent, especially, when the angle between loudspeakers is narrow. This difference is caused by the layout of the three speakers. The sound from the center speaker is affected by ***both*** the left and right speakers, while only the center speaker has a large effect on each of the side speakers. Thus, the number of overlapping TF components for the center speaker is larger than that of either the left or the right speaker individually. Also, their overlapping level for the center speaker is higher than the others. 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 10 points by averaging three-speaker cases.

When we focus on the difference between the unweighted soft MFM and the proposed soft MFM, we can find a similar tendency with respect to the difference between the soft and the hard MFMs; that is, the optimization of weighting factors is more effective when two speakers are closer together. This means that weighting factors work effectively to deal with highly overlapped TF components.

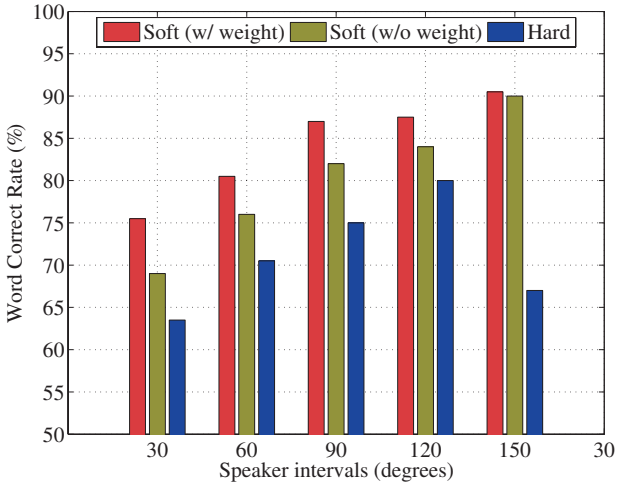

Figure 9. *Word correct rate for the center speaker.*

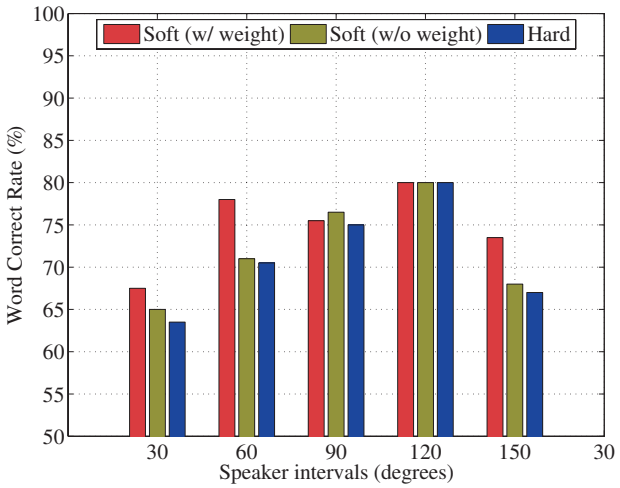

Figure 10. *Word correct rate for the left speaker.*

5.3. Human-robot interaction scenario

We applied a robot audition system including the proposed soft MFM generation to a human-robot interaction scenario. Figure 12 shows snapshots of the scenario. In this scenario, the robot receives a meal order, with three customers simultaneously asking the robot for what they want. The robot localizes and separates their speeches, and the recognizes the separated speeches. This demonstration was performed in the same room mentioned in Section 2.4. With the hard MFM we were previously using, speakers had to keep more than 60 degree intervals from the neighboring speaker for this kind of real situation, while benchmark tests show that the system maintains performance even when the interval between speakers is 30 degrees. This is caused by dynamically-changing noises in a real-world environment. On the

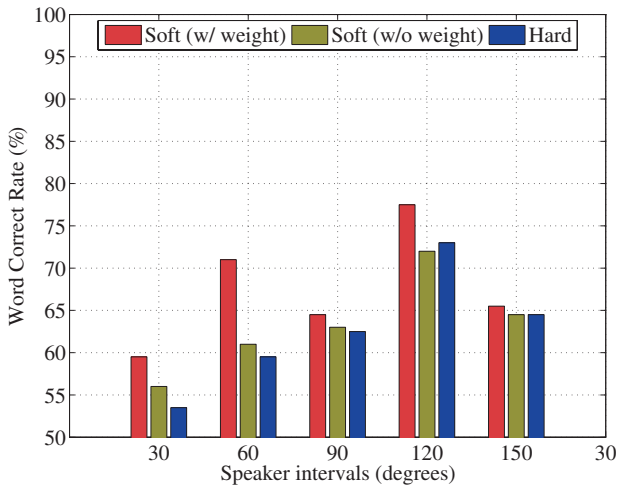

Figure 11. *Word correct rate for the right speaker.*

other hand, the robot with our proposed system maintained ASR performance even in the case of a 30 degree interval for this scenario. This means that the proposed soft MFM approach was effective in a real-world environment where dynamically-changing noises exist to some extent.

6. Conclusion and future work

We have presented an improvement in automatic speech recognition of robot audition to allow it to realize natural human-robot interaction which is a topic essential to behavioral robotics. The missing-feature theory is adopted to integrate microphone-array-based pre-processing of sound-source localization and separation. For the missing feature mask, we used a soft missing feature mask taking a continuous value between 0 and 1, instead of a conventional hard missing feature mask taking a binary value, 0 or 1. The soft missing feature mask is generated automatically by estimating the reliability of a time-frequency component based on a sigmoid function. The automatic soft mask generation is incorporated as a set of modules into the HARK open-sourced robot audition system.

The resulting HARK-based robot audition system with automatic soft mask generation improves the performance of automatic speech recognition in the case of three simultaneous speeches, in particular for narrower intervals of two adjacent speakers up to 30 degrees. The conventional system worked for speaker-sepearations greater than 60 degrees. Therefore, the soft mask system provides opportunities to deploy a robot audition system in more realistic multi-party interaction. As a proof of concept, a humanoid HRP-2 demonstrated the role of taking three meal orders at the same time.

Future work includes detailed analysis and more applications (for example, extensive benchmarking to analyse the performance of automatic speech recognition with wide variations of speaker configuration under various acoustic environments); and application of the HARK-based system to actual multi-party interactions. Typical scenarios will include barge-in utterances, utterances of moving talkers, and recognition while the robot is in motion. These applications are expected

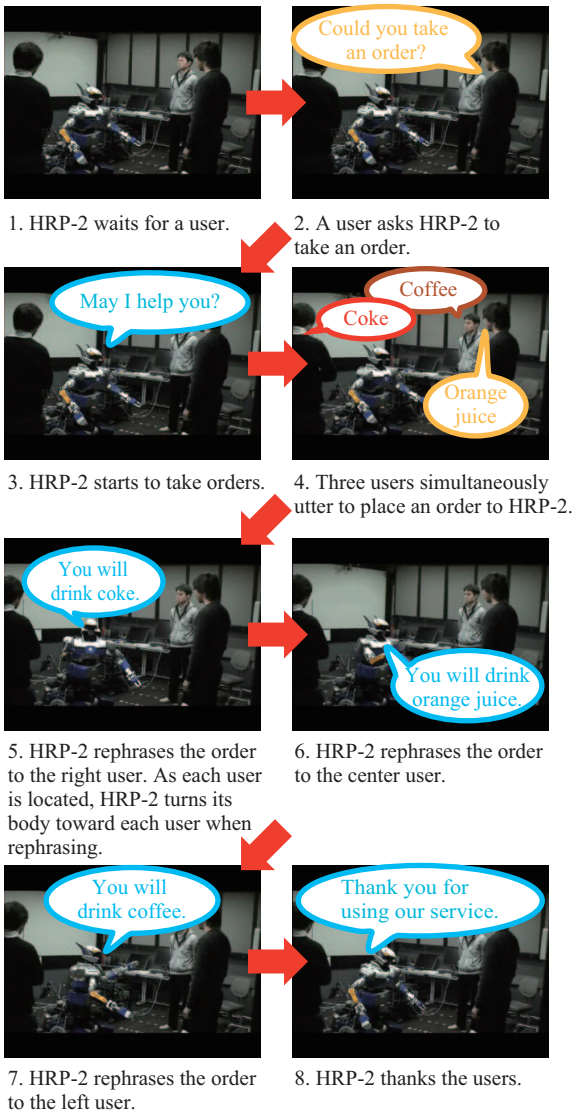

Figure 12. *Snapshots of a meal order taking task.*

to gather more experience in coping with a mixture of sounds, and to guide new research towards symbiosis of human and robots through verbal communication and auditory scene analysis.

Acknowledgments

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

References

[1] 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 Procedings of Eurospeech-2001, pages 213–216. ESCA, 2001.

[2] J. Barker, L. Josifovski, M. Cooke, and P. Green. Soft decisions in missing data techniques for robust automatic speech recognition. In Proc. of 6th International Conference on Spoken Language Processing (ICSLP-2000), volume I, pages 373–376, 2000.

[3] S. F. Boll. A spectral subtraction algorithm for suppression of acoustic noise in speech. In roceedings of 1979 International Conference on Acoustics, Speech, and Signal Processing (ICASSP-79), pages 200–203. IEEE, 1979.

[4] C. Breazeal. Emotive qualities in robot speech. In Proceedings of 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2001), pages 1389–1394, 2001.

[5] I. Cohen and B. Berdugo. Speech enhancement for non-stationary noise environments. Signal Processing, 81(2):2403–2418, 2001.

[6] M. Cooke, P. Green, L. Josifovski, and A. Vizinho. Robust automatic speech recognition with missing and unreliable acoustic data. Speech Communication, 34(3):267–285, May 2000.

[7] C. Côté, D. Létourneau, F. Michaud, J. M. Valin, Y. Brosseau, C. Räievsky, M. Lemay, and V. Tran. Reusability tools for programming mobile robots. In Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004), pages 1820–1825. IEEE, 2004.

[8] J. de Veth, F. de Wet, B. Cranen, and L. Boves. Missing feature theory in asr: Make sure you miss the right type of features. In Proceedings of Workshop on Robust Methods for ASR in Adverse Conditions, Tampere, pages 231–234, 1999.

[9] A. Drygajlo and M. El-Maliki. Speaker verification in noisy environments with combined spectral subtraction and missing feature theory. In Proceedings of 1998 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 1998), pages 121–124, 1998.

[10] Y. Ephraim and D. Malah. Speech enhancement using minimum mean-square error short-time spectral amplitude estimator. IEEE Transactions on Acoustics, Speech and Signal Processing, ASSP-32(6):1109–1121, 1984.

[11] 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), pages 2404–2410. IEEE, 2004.

[12] H. Isao, A. Futoshi, K. Yoshihiro, K. Fumio, and Y. Kiyoshi. Robust speech interface based on audio and video information fusion for humanoid hrp-2. In Proceedings of 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2004), pages 2404–2410, 2004.

[13] Multiband Julius. http://www.furui.cs.titech.ac.jp/mbandjulius/.

[14] H. D. Kim, K. Komatani, T. Ogata, and H. G. Okuno. Human tracking system integrating sound and face localization using em algorithm in real environments. Advanced Robotics, 23(6):629–653, 2007.

[15] R. P. Lippmann and B. A. Carlson. Robust speech recognition with time-varying filtering, interruptions, and noise. In Proceedings of 1997 ISCA 5th European Conference on Speech Communication and Technology (EuroSpeech 1997), pages 365–372, 1997.

[16] Y. Matsusaka, T. Tojo, S. Kuota, K. Furukawa, D. Tamiya, K. Hayata, Y. Nakano, and T. Kobayashi. Multi-person conversation via multi-modal interface – a robot who communicates with multi-user. In Proceedings of 6th European Conference on Speech Communication Technology (Eurospeech 1999), pages 1723–1726, 1999.

[17] I. A. McCowan and H. Bourlard. Microphone array post-filter for diffuse noise field. In ICASSP-2002, volume 1, pages 905–908, 2002.

[18] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano. Active audition for humanoid. In Proc. of 17th National Conference on Artificial Intelligence (AAAI-2000), pages 832–839. AAAI, 2000.

[19] K. Nakadai, D. Matasuura, H. G. Okuno, and H. Tsujino. Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots. Speech Communication, 44(1-4):97–112, October 2004.

[20] K. Nakadai, D. Matsuura, H. G. Okuno, and H. Tsujino. Improvement of recognition of simultaneous speech signals using av integration and scattering theory for humanoid robots. Speech Communication, 44(1-4):97–112, 2004.

[21] K. Nakadai, H. G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino. An open source software system for robot audition hark and its evaluation. In Proceedings of 2008 IEEE/RAS International Conference on Humanoid Robots (HUMANOIDS 2008), pages 561–566, 2008.

[22] 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, number 1aSC7, 2004.

[23] M. T. Padilla, T. F. Quantieri, and D. A. Reynolds. Missing feature theory with soft spectral subtraction for speaker verification. In Proceedings of the 8th International Congress on Spoken Language Processing (InterSpeech 2006), pages 913–916, 2006.

[24] H.M. Park and R. M. Stern. Missing feature speech recognition using dereverberation and echo suppression in reerberation environments. In Proceedings of 2007 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2007), volume IV, pages 381–384, 2007.

[25] L. C. Parra and C. V. Alvino. Geometric source separation: Mergin convolutive source separation with geometric beamforming. IEEE Transactions on Speech and Audio Processing, 10(6):352–362, 2002.

[26] R. Plomp, L. C. W. Pols, and J. P. van de Geer. Dimensional analysis of vowel spectra. Acoustical Society of America, 41(3):707–712, 1967.

[27] B. Raj and R. M. Stern. Missing-feature approaches in speech recognition. Signal Processing Magazine, 22(5):101–116, 2005.

[28] P. Renevey and A. Drygajlo. Missing feature theory and probabilistic estimation of clean speech components for robust speech recognition. In Proceedings of European Conference on Speech Communication Technology (Eurospeech-1999), pages 2627–2630, 1999.

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

[30] T. Takahashi, K. Nakadai, K. Komatani, T. Ogata, and H. G. Okuno. Missing-feature-theory-based robust simultaneous speech recognition system with non-clean speech acoustic model. In Proceedings of 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2009), pages 2730–2735, 2009.

[31] K. Tatsuya and L. Akinobu. Free software toolkit for Japanese large vocabulary continuous speech recognition. In International Conference on Spoken Language Processing (ICSLP), volume 4, pages 476–479, 2000.

[32] J. M. Valin, F. Michaud, and J. Rouat. Robust localization and tracking of simultaneous moving sound sources using beamform-

ing and particle filtering. Robotics and Autonomous Systems Journal, 55(3):216–228, 2007.
[33] J. M. Valin, J. Rouat, and F. Michaud. Enhanced robot audition based on microphone array source separation with post-filter. In Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pages 2133–2128, 2004.
[34] 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), pages 2123–2128. IEEE, 2004.
[35] F. Wang, Y. Takeuchi, N. Ohnishi, and N. Sugie. A mobile robot with active localization and discrimination of a sound source. Journal of Robotic Society of Japan, 15(2):61–67, 1997.
[36] 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), pages 897–902. IEEE, 2005.
[37] 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/RSJ International Conference on Intelligent Robots and Systems (IROS 2005), pages 1489–1494. IEEE, 2005.
[38] S. Yamamoto, K. Nakadai, M. Nakano, H. Tsujino, J. M. Valin, K. Komatani, T. Ogata, and H. G. Okuno. Design and implementation of a robot audition system for automatic speech recognition of simultaneous speech. In Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU-2007), pages 111–116. IEEE, 2007.
[39] S. Yamamoto, K. Nakadai, H. Tsujino, T. Yokoyama, and H. G. Okuno. Improvement of robot audition by interfacing sound source separation and automatic speech recognition with missing feature theory. In Proceedings of IEEE International Conference on Robotics and Automation (ICRA 2004), pages 1517–1523. IEEE, 2004.
[40] 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 inseparating and recognizing simultaneous speech signals. In Proceedings of 19th International Conference on Industrial, Engineering, and Other Applications of Applied Intelligent Systems (IEA/AIE'06), volume LNAI 4031, pages 207–217. Springer-Verlag, 2006.