euronoise

A Portable Robot Audition Software System for Multiple Simultaneous Speech Signals

Hiroshi Okuno[a], Shunichi Yamamoto[a], Kazuhiro Nakadai[b], Jean-Marc Valin[c], Tetsuya Ogata[a] and Kazunori Komatani[a]

[a]Kyoto University, Graduate School of Informatics, Yoshida-Honmachi, Sakyo, 606-8501 Kyoto, Japan
[b]Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako, 351-0114 Saitama, Japan
[c]CSIRO ICT Center, Cnr Vimiera & Pembroke Rds, NSW 2122 Marsfield, Australia
okuno@i.kyoto-u.ac.jp

Since a robot is deployed in various kinds of environments, the robot audition system should work with minimum prior information on environments to localize, separate and recognize utterances by multiple simultaneous talkers. For example, it should not assume either the number of speakers, the location of speakers for sound source separation (SSS), or specially tuned acoustic model for automatic speech recognition (ASR). We developed "HARK" portable robot audition that uses eight microphones installed on the surface of robot's body such as Honda ASIMO, and SIG-2 and Robovie-R2 at Kyoto University. HARK integrates SSS and ASR by using the Missing-Feature Theory. For SSS, we use Geometric Source Separation and multi-channel post-filter to separate each utterance. Since separated speech signals are distorted due to interfering talkers and sound source separation, multi-channel post-filter enhanced speech signals. At this process, we create a missing feature mask that specifies which acoustic features are reliable in time-frequency domain. Multi-band Julius, a missing-feature-theory based ASR, uses this mask to avoid the influence of unreliable features in recognizing such distorted speech signals. The system demonstrated a waitress robot that accepts meal orders placed by three actual human talkers.

1 Introduction

Listening to several things at once is a 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, 2]. *Robot audition*, or the robot's capability of listening by *its own ears* (microphones), is therefore an essential intelligent function for working in a daily environment in symbiosis with human. Since robots encounter various kinds of sounds and noises, robot audition should be able to recognize a mixture of sounds. The capability of listening to several things at once will increase the usability of robots, for instance, in assisting hearing-impaired or elder people. In addition, robots are deployed in various environments, in particular, dynamically changing ones, robot audition system should depend on minimum prior information about its deployment.

Robot audition system usually integrates various kinds of modules including sound source localization (SSL), sound source separation (SSS), and automatic speech recognition (ASR). The goals of robot audition system are summarized as follows:

1. *Extensible* for adding and replacing modules,
2. *Minimum prior information* for each modules,
3. *Portability* for various robot configurations, and
4. *Real-time processing.*

In other words, the technical issues in robot audition system focus on system-integration technology as well as individual technologies.

Related Work Research on robot audition has been active recently. IEEE Robotics and Automation Society and Robotics Society of Japan have provided organized sessions on robot audition since 2004. Robot audition community exploited a physical body of robot to improve the performance of sound source localization and separation. One good example of behavioral intelligence in robot audition is *active audition* [4, 3], which improved SSL and SSS by integrating these subsystems with active motion such as turning to a target sound source and/or visual processing.

Most robot audition research focused on SSL, and only a few on SSS and ASR. Nakadai et al. [5] used a pair of microphones embedded in ear parts of humanoid robot SIG. It succeeded in listening to three simultaneous utterances with a set of speaker- and direction-dependent acoustic models for ASR. Since their system needed a lot of prior information, it was difficult to deploy their robot to other environments. Valin et al. [6] have developed a microphone array system called "Manyears". It used steered beamformer to localize and used Geometric Source Separation (GSS) to separate sound sources. Eight microphones were placed on each vertex of a cubic. Valin and Yamamoto integrated Manyears and Missing Feature Theory (MFT) based ASR and reported preliminary results [7]. The microphones were embedded on the body of SIG2.

In signal processing community, a lot of methods in addition to GSS have been proposed to improves the SNR of the input speech signals before performing ASR [9, 10, 11]. GSS relaxes the limitation on the relationship between the number of sound sources and microphones. It can separate up to $N-1$ sound sources with N microphones, by introducing "geometric constraints" obtained from the locations of sound sources and the microphones. This means that GSS requires sound source directions as prior information. Given accurate sound source directions, GSS shows comparable performance with ICA (Independent Component Analysis). Usually ICA with more microphones costs more in computation, and thus is difficult for real-time processing. For near-field sound source localization, MUSIC (MUltiple SIgnal Classification) outperforms steered beamformer [9].

Robot audition may be viewed as "noise-robust hands-free ASR" from signal processing community. Its common approach is the use of an acoustic model for ASR trained with noise adaptation techniques [8]. This approach won't work in dealing with unknown noises or in recognizing extremely noisy speech captured by a robot-embedded microphone.

Usually multi-channel sound source separation techniques such as GSS cause spectral distortion. Such a distortion affects acoustic feature extraction for ASR, especially the normalization processes of an acoustic feature vector, because the distortion causes fragmentation of the target speech in the spectro-temporal space, and produces a lot of sound fragments. To reduce the influence of spectral distortion for ASR, we employed two techniques; a multi-channel post-filter and white noise addition with missing-feature theory based ASR.

This paper reports the robot audition software system called "**HARK**" (HRI-JP Audition for Robots with Kyoto Univesity. The word "hark" stands for "*listen*".). It focuses on the refinements of each module and their integration and evaluates the performance of recognizing three simultaneous talkers.

Figure 1: Surround microphone

Figure 2: SIG2

Figure 3: RobovieIIs

Figure 4: ASIMO

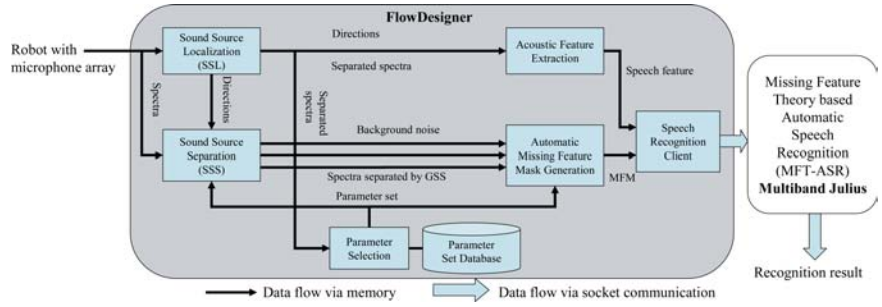


Figure 5: Overview of the real-time robot audition system

2 HARK Robot Audition System

The robot audition software system, HARK, consists of six modules as shown in Figure 5: Sound Source Localization (SSL), Sound Source Separation (SSS), Parameter Selection, Acoustic Feature Extraction, Automatic Missing Feature Mask Generation, and Missing Feature Theory based Automatic Speech Recognition (MFT-ASR). We used MUSIC for SSL, GSS for SSS, and Multi-band Julis for MFT-ASR.

HARK allows various kinds of microphone configuration. Figures 1– 4 show an 8-ch microphone array embedded in Humanoid SIG2, Robovie R2, and Honda ASIMO or a 7.1-ch surround microphone, H2Pro, of Holosound Inc. The positions of the microphones are bilaterally symmetric for all of them. This is because the longer the distance between microphones is, the better the performance of GSS is. We used GSS, which requires only the 3D position of each microphone.

The five modules except MFT-ASR are implemented as component blocks of *FlowDesigner* [12], a free data flow oriented development environment. The reason why MFT-ASR is treated separately is twofold; First, it needs a heavy CPU load in recognizing speech. Second, it uses a light-weighted data format in communication with the other modules. It uses acoustic features and MFM for communication with the other modules, while the other modules use raw signal data for their communication. FlowDesigner and Multiband Julian may run separately on different CPUs, since they can communicate with each other via a network.

Since the five modules communicate with each other a large amount of data, that is, raw signal data, the reduction of communication traffic is critical in real-time processing. FlowDesigner provides the mechanism of sharing data on a shared memory between modules. It also provides the re-usability of modules for rapid prototyping.

When two blocks have matching interfaces, they can be connected regardless of their internal processes. One-to-many and many-to-many connections are also possible. Thus, complex applications can be built simply by combining small reusable blocks. A block is coded in C++ and implemented as an inherited class of the fundamental block. It is compiled as a shared object on Linux. Since data communication is done by a function call with a pointer, it is faster than other middleware which use a communication buffer such as shared memory and a socket.

FlowDesigner, thus, achieves a well-balanced trade-off between independence and processing speed. Because a large amount of data is communicated in HARK, FlowDesigner is suitable for it. In fact, SSS in Figure 5 has the heaviest traffic, which requests a large bandwidth of 12.8 Mbps for input and 8 Mbps for output.

2.1 Signal Processing Components

Sound Source Localization (SSL) HARK provides two SSLs: MUSIC and geometrical refinement method. The latter is a steered beamformer included in Manyears. MUSIC is a frequency-domain adaptive BF method. It outperforms the latter in the real world, in particular, in near fields, because a sharp local peak corresponding to a sound source direction is obtained from the MUSIC spectrum. Our implementation uses impulse responses measured every 5 degrees, which were used to calculate a correlation matrix.

Sound Source Separation (SSS) SSS consists of GSS and the multi-channel post-filter as [13]. We modified the original GSS proposed by Parra [11] in order to speed up adaptation by using stochastic gradient and shorter time frame estimation. Our implementation avoids divergence caused by numerical errors including division-by-zero and speeds up by simplifying some equations.

The multi-channel post-filter [13] is used to enhance the output of GSS (Figure 6). It is based on the optimal estimator originally proposed by Ephraim and Malah [14]. We extend the original method to support multi-channel signals so that they can estimate both stationary and non-stationary noise. These estimations are used to generate a missing feature mask.

In spite of these simplification, HARK implementation of SSS attained almost the same performance as Manyears' implementation. SSS improved 10.3 dB in signal-to-noise ratio on average for separation of three

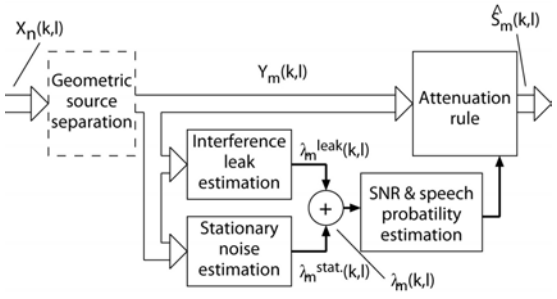


Figure 6: Scheme of Multi-Channel Post-Filter

simultaneous speech signals [13]. After separation, a white noise was added with a half power of the background noise.

White Noise Addition We exploit covering a distortion in any frequency band by adding a white noise, a kind of broad-band noises, to noise-suppressed speech signals. This idea is motivated by the psychological evidence that noise helps perception, which is known as *auditory induction.* It is known that in the human auditory system noises that pad temporal gaps between sound fragments help auditory perception organization.

This evidence is also useful for ASR, because an additive noise plays a roll to blur the distortions, that is, to avoid the fragmentation. Actually, the addition of a colored noise has been reported to be effective for noise-robust ASR [15]. They added office background noise after spectral subtraction, and showed the feasibility of this technique in noisy speech recognition.

In accordance with this addition of a white noise, we use an acoustic model trained with clean speech and white-noise-added speech. Thus, the system is able to assume only one type of noise included in speech, that is, white noise. It is easier for ASR to deal with one type of noise than various kinds of noises, and a white noise is suitable for ASR with a statistical model.

2.2 MFT Based Integration

Several robot audition systems with preprocessing and ASR have been reported so far [16, 17]. Those systems just combined preprocessing with ASR and focused on the improvement of SNR and real-time processing.

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 *missing feature masks (MFMs)* in a temporal-frequency map of reliability 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:

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

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.

MFT-based methods show high robustness against both stationary and non-stationary noises when the reliability of acoustic features is estimated correctly. The main issue in applying them to ASR is how to estimate the reliability of input acoustic features correctly. Because the distortion of input acoustic features are usually unknown, the reliability of the input acoustic features cannot be estimated. To estimate MFM, we used *Mel-Scale Log Spectrum (MSLS)* [18] as an acoustic feature and developed an automatic MFM generator based on the multi-channel post-filter.

Mel-Scale Log Spectrum Features 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 [18]. These three normalization processes correspond to three normalization performed against MFCC; C0 normalization, liftering, and Cepstrum mean normalization.

Automatic MFM generator Most reports on MFT have focused on a single channel input, so far. It is difficult to obtain information enough to estimate the reliability of acoustic features in a single channel approach. A multi-channel approach using a microphone array alleviates this difficulty. We developed an automatic MFM generator by using GSS and a multi-channel post-filter with an 8-ch microphone array.

The missing feature mask is a matrix representing the reliability of each feature in the time-frequency plane. More specifically, this reliability is computed for each time frame and for each Mel-frequency band. This reliability can be either a continuous value from 0 to 1 (called "*soft mask*"), or a binary value of 0 or 1 (called "*hard mask*"). In this paper, hard masks were used.

We compute the missing feature mask by comparing the input and the output of the multi-channel post-filter. For each Mel-frequency band, the feature is considered reliable if the ratio of the output energy over the input energy is greater than threshold T. The reason for this choice is based on the assumption that the more noise present in a certain frequency band, the lower the post-filter gain will be for that band. The continuous missing feature mask $m_k(i)$ is thus computed as follows:

$$m_k(i) = \frac{S_k^{out}(i) + N_k(i)}{S_k^{in}(i)}, \quad (2)$$

where $S_k^{in}(i)$ and $S_k^{out}(i)$ are the post-filter input and output energy for frame k at Mel-frequency band i, and $N_k(i)$ is the background noise estimate for that band. The main reason for including the noise estimate $N_k(i)$ in the numerator of Eq. (2) is that it ensures that the missing feature mask equals 1 when no speech source is

present. Finally, we derive a hard mask $M_k(i)$ as follows:

$$M_k(i) = \begin{cases} 1 & \text{if } m_k(i) > T, \\ 0 & \text{otherwise} \end{cases}$$

where T is an appropriate threshold.

3 Evaluation of HARK

We evaluated the robot audition system in terms of the performance of three simultaneous speech recognition.

3.1 MFT and White Noise Addition

To evaluate how MFT and white noise addition improve the performance of automatic speech recognition, we conducted isolated word recognition of three simultaneous speech. In this experiment, Humanoid SIG2 with an 8-ch microphone array was used in a 4 m × 5 m room. Its reverberation time (RT_{20}) was 0.3–0.4 seconds.

Three simultaneous speech for test data were recorded with the 8-ch microphone array in the room by using three loudspeakers (Genelec 1029A). The distance between each loudspeaker and the center of the robot was 2 m. One loudspeaker was fixed to the front (center) direction of the robot. The locations of left and right loudspeakers from the center loudspeaker varied from ±10 to ±90 degrees at the intervals of 10 degrees. ATR phonemically-balanced word-sets were used as a speech dataset. A female (f101), a male (m101) and another male (m102) speech sources were used for the left, center and right loudspeakers, respectively. Three words for simultaneous speech were selected at random. In this recording, the power of robot was turned off.

By using the test data, the system recognized the three speakers with the following eight conditions:

(1) The raw input captured by the left-front microphone was recognized with the clean acoustic model.
(2) The sounds separated by SSS were recognized with the clean acoustic model.
(3) The sounds separated by SSS were recognized with MFM generated automatically and the clean acoustic model.
(4) The sounds separated by SSS were recognized with automatically generated MFM and the **WNA acoustic model**.
(5) The sounds separated by SSS were recognized with automatically generated MFM and the **MCT acoustic model**.
(6) The sounds separated by SSS were recognized with *a priori* **MFM** and the clean acoustic model . Since this mask is *ideal*, we consider its result as the potential upper limit of HARK.

The **clean acoustic model** was trained with 10 male and 12 female ATR phonemically-balanced word-sets excluding the three word-sets (f101, m101, and m102) which were used for the recording. Thus, it was a speaker-open and word-closed acoustic model. The **MCT acoustic model** was trained with the same ATR word-sets and separated speech datasets. The latter sets were generated by separating three-word combinations of f102-m103-m104 and f102-m105-m106, which were recorded in the same way as the test data. The **WNA acoustic model** was trained with the same ATR wordsets as mentioned above, and the clean speech to which white noise was added by 40 dB of peak power. Each of these acoustic models was trained as 3-state and 4-mixture triphone HMM, because 4-mixture HMM had the best performance among 1, 2, 4, 8, and 16-mixture HMMs.

The results were summarized in Figure 7. MFT-ASR with Automatic MFM Generation outperformed the normal ASR. The **MCT acoustic model** was the best for MFT-ASR, but the **WNA acoustic model** performed almost the same. Since the **WNA acoustic model** does not require prior training, it is the most appropriate acoustic model for robot audition. The performance at the interval of 10-degree was poor in particular for the center speaker, because any current sound source separation methods fails in seprating such close three speakers. The fact that *A priori* mask showed a quite high performance may suggest not a few possibilities to improve the algorithms of MFM generation.

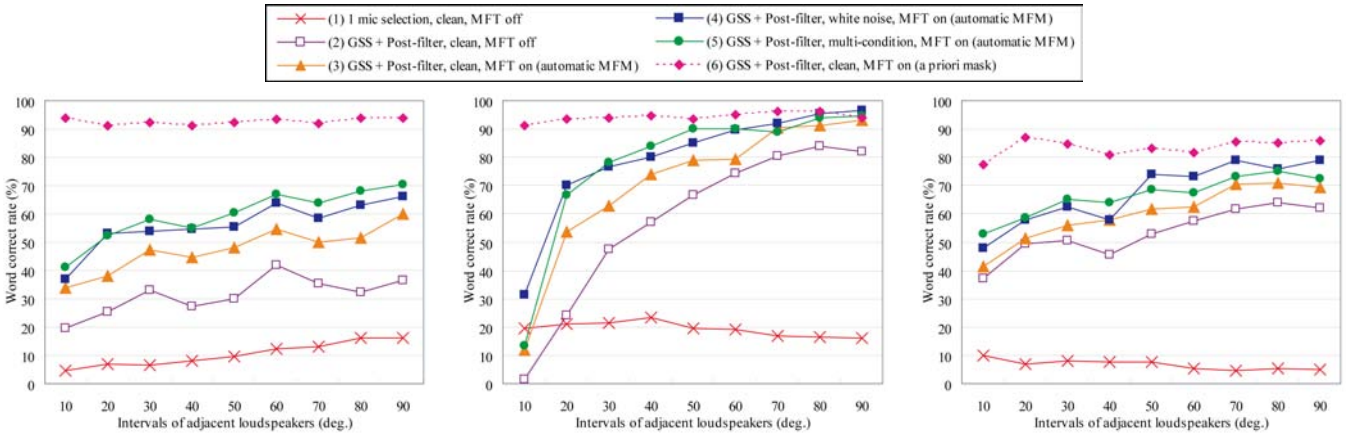
Table 1: Word correct rate (WCR in %) of the center speaker according to each localization method

Acoustic model	*White noise addition*			*Clean model*		
Interval	30°	60°	90°	30°	60°	90°
given	90.0	88.5	91.0	85.0	84.5	87.0
steered BF	82.3	90.5	89.0	65.5	70.6	72.4
MUSIC	86.0	83.3	86.7	57.0	74.0	64.5

3.2 Sound Source Localization Effects

This section evaluates how the quality of sound source localization methods including manually given localization, steered Beamformer and MUSIC affects the performance of ASR. SIG2 used steered BF. Since the performance MUSIC depends on the number of microphones on the same plane, we used Honda ASIMO shown in Figure 4, which was installed in a 7 m × 4 m room. Its three walls were covered with sound absorbing materials, while the other wall was made of glass which makes strong echoes. The reverberation time (RT_{20}) of the room is about 0.2 seconds. We used the condition (6) in Section 3.1, and used three methods of sound source localization with clean and WNA acoustic models.

The results of word correct rates were summarized in Table 1. With the **clean acoustc model**, MUSIC outperformed steered BF, while with the **WNA acoustic mode**, the both performances were comparable. In case of **given localization**, improvement by white noise addition training was small. On the other hand, training with white noise addition improved word correct rates greatly for both steered beamformer and MUSIC. We think that the ambiguity in sound source localization caused voice activity detection more ambiguous, which degraded the recognition performance with the clean acoustic model. On the other hand, white noise addition to separated sound with the WNA acoustic model reduced such degradation.



(a) The left speaker (b) The center speaker (c) The right speaker

Figure 7: Word correct rates of three simultaneous speakers with our system

Figure 8: Meal orders

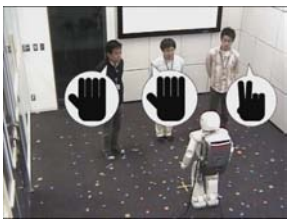


Figure 9: Rock-paper-scissors sound game

3.3 Three Simultaneous Talkers

Figure 8 demonstrates that when three actual human talkers place a meal order at the same time, the robot audition system recognizes each meal order and confirms their orders one by one and tells the total amount of the orders. The real-time implementation reduces the response time from 8.0 sec to 1.9 sec. The main factor of delay of 1.9 sec is caused by the detection of end of utterance. If the same input is given by an audio file, the response time is about 0.4 sec.

We measured processing time by recognizing speech signals of 800 seconds. The total processing time on Pentium 2.4GHz CPU was 499 sec (130 sec for ASR and 369 sec for others) with 0.446 sec of output delay. As a whole, HARK runs almost in real time.

Another application demonstrates a rock-paper-scissors sound game Figure 9 where ASIMO recognizes three simultaneous utterances and judged who won the game by using only speech information.

4 Conclusion

This paper described the portable real-time robot audition software, HARK. The key technology is MFT-based integration of sound source separation and MFT-based ASR by automatically generating missing feature masks. We showed the effectiveness of HARK through several experiments, and the conventional noise-robust ASR approaches such as only the use of a multi-condition trained acoustic model, and/or a single channel preprocessing have difficulty in realizing robot audition. HARK installed on three robots demonstrates potential capabilities such as interactions with multiple people and auditory scene visualizer. HARK is available at the URL of http://winnie.kuis.kyoto-u.ac.jp/HARK/.

Several detail experiments are still missing. The robustness against speech contaminated non-speech directional noise sources like music, and reverberation should be evaluated.

References

[1] H.G. Okuno, T. Nakatani, & T. Kawabata. Understanding three simultaneous speakers. *IJCAI-1997*, 30–35.

[2] D. Rosenthal & H.G. Okuno (Eds.). *Computational Auditory Scene Analysis*. Lawrence Erlbaum Associates, 1998.

[3] H.G. Okuno, & K. Nakadai. Active audition for humanoid robots that can listen to three simultaneous talkers. *JASA*, 113(4):2230. 2003. 145th Meeting, 2pSC4.

[4] K. Nakadai, K. Hidai, H. Mizoguchi, H.G. Okuno, & H. Kitano. Real-time auditory and visual multiple-object tracking for robots. *IJCAI-2001*, 1424–1432.

[5] K. Nakadai, H.G. Okuno, & H. Kitano. Exploiting auditory fovea in humanoid-human interaction. *AAAI-2002*, 431–438.

[6] J.-M. Valin, J. Rouat, & F. Michaud. Enhanced robot audition based on microphone array source separation with post-filter. *IEEE/RSJ IROS-2004*, 2123–2128.

[7] J.-M. Valin, S. Yamamoto, J. Rouat, F. Michaud, K. Nakadai, & H.G. Okuno. Robust recognition of simultaneous speech by a mobile robot. *IEEE Tr. on Robotics*, 23(4):742–752, Aug. 2007.

[8] R. P. Lippmann, E. A. Martin, & D. B. Paul. Multi-styletraining for robust isolated-word speech recognition. *ICASSP-1987*, 705–708.

[9] F. Asano, H. Asoh, & T. Matsui. Sound source localization and signal separation for office robot "Jijo-2". *IEEE MFI-1999*, 243–248.

[10] C. Jutten & J. Herault. Blind separation and sources. *Signal Processing*, 24(1):1–10, 1995.

[11] L. C. Parra & C. V. Alvino. Geometric source separation: Mergin convolutive source separation with geometric beamforming. *IEEE Tr. on SAP*, 10(6):352–362, 2002.

[12] C. Côté, *et al.* Code reusability tools for programming mobile robots. *IEEE/RSJ IROS 2004*, 1820–1825.

[13] S. Yamamoto, J.-M. Valin, K. Nakadai, T. Ogata, & H.G. Okuno. Enhanced robot speech recognition based on microphone array source separation and missing feature theory. *IEEE ICRA-2005*, 1489–1494.

[14] Y. Ephraim & D. Malah. Speech Enhancement Using Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator. *IEEE Tr. on ASSP*, 32(6):1109–1121, 1984.

[15] S. Yamada, A. Lee, H. Saruwatari, & K. Shikano. Unsupervided speaker adaptation based on HMM sufficient statistics in various noisy environments. *Eurospeech-2003*, 1493–1496.

[16] I. Hara, *et al.* Robust speech interface based on audio and video information fusion for humanoid HRP-2. *IEEE/RSJ IROS-2004*, 2404–2410.

[17] K. Nakadai, D. Matasuura, H.G. Okuno, & 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.

[18] Y. Nishimura, T. Shinozaki, K. Iwano, & S. Furui. Noise-robust speech recognition using multi-band spectral features. *148th ASA Meetings*, 1aSC7, 2004.