COMPUTATIONAL AUDITORY SCENE ANALYSIS AND ITS APPLICATION TO ROBOT AUDITION

Hiroshi G. Okuno

Kyoto University
Graduate School of Informatics
Sakyo, Kyoto 606-8501, Japan
okuno@i.kyoto-u.ac.jp

Kazuhiro Nakadai

Honda Research Institute, Japan Co., Ltd.
8-1 Honcho, Wako, Saitama, 351-0188, Japan
nakadai@jp.honda-ri.com

ABSTRACT

Robot capability of hearing sounds, in particular, a mixture of sounds, by its own microphones, that is, *robot audition*, is important in improving human robot interaction. This paper presents the robot audition open-source software, called "HARK" (HRI-JP Audition for Robots with Kyoto University), which consists of primitive functions in computational auditory scene analysis; sound source localization, separation, and recognition of separated sounds. Since separated sounds suffer from spectral distortion due to separation, the HARK generates a time-spectral map of reliability, called "missing feature mask", for features of separated sounds. Then separated sounds are recognized by the Missing-Feature Theory (MFT) based ASR with missing feature masks. The HARK is implemented on the middleware called "FlowDesigner" to share intermediate audio data, which enables near real-time processing.

***Index Terms*—** robot audition, computational auditory scene analysis, Missing feature theory, simultaneous speakers

1. INTRODUCTION

Robot species have exploded at the time of AICHI EXPO 2005 like the Cambrian explosion of species starting about 542 million years ago. If the next step of evolution is species selection, what will make some species proliferate for symbiosis between human and robots? In human communication, speech signals and sounds are critical. Therefore, we believe that auditory capability is critical for such species selection for partner robots.

Robot audition is expected to facilitate capabilities similar to those of human. For example, people can attend one conversation and switch to another even in a noisy environment. This capability is known as the *cocktail party effect*. For this purpose, a robot should separate a speech stream from a mixture of sounds. It may realize the hearing capability of "*Prince Shotoku*" that, according to the Japanese legend, could listen to 10 people's petitions at once.

Since a robot produces various sounds and should be able to "understand" many kinds of sounds, auditory scene analysis is the process of simulating useful intelligent behavior, and even required when objects are invisible. While traditionally, auditory research has been focusing on human speech understanding, understanding auditory scenes in general is receiving increasing attention. Computational Auditory Scene Analysis (CASA) studies a general framework of sound processing and understanding [1]. Its goal is to understand an arbitrary sound mixture including speech, non-speech signals, and music in various acoustic environments.

The main research topics of CASA are sound source localization (SSL), sound stream separation (SSS), and its recognition including automatic speech recognition (ASR). In addition, CASA focuses on a general model and mechanism of separating various kinds of sounds, including voiced speech, music, and environmental sounds, from a mixture of sounds [1].

Based on primitive functions of CASA, A robot audition system usually integrates various kinds of CASA modules. The goal of design and implementation of the robot audition system is summarized as follows:

1. *Minimum prior information* for each module,
2. *Portability* for various robot configurations, and
3. *Real-time processing* by system integration.

In other words, the technical issues in robot audition systems focus on system-integration technology rather than individual technologies. This paper describes how various modules are integrated into the whole robot audition system and presents a portable robot audition open-source software called "HARK" (HRI-JP Audition for Robots with Kyoto University). It also demonstrates the feasibility of the resulting robot audition system.

2. COMPUTATIONAL AUDITORY SCENE ANALYSIS

This section describes CASA modules that the HARK uses.

2.1. Sound Source Localization (SSL)

We use two SSLs, a steered beamformer with geometrical refinement method [2] and a frequency-domain adaptive beamformer, MUSIC (MUltiple SIgnal Classification), with eight microphones. The former decomposes the whole 3D space into smaller subspaces gradually to obtain a peak power. MUSIC outperforms steered beamformer for near-field sound source localization [3], 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 to calculate a correlation matrix.

The selection of SSL depends on the configuration of microphone. Since most SSS requires the direction of sound source, SSS plays an important role in improving the performance of SSS. We use MUSIC for a configuration of microphones embedded in the head of a robot. We use a steered beamformer for a configuration of microphones embedded in the body near arms, because robot's arms affect impulse responses.

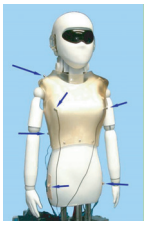


Fig. 1. SIG2

Fig. 2. Robovie R2

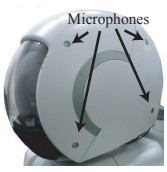


Fig. 3. ASIMO

Fig. 4. H2Pro

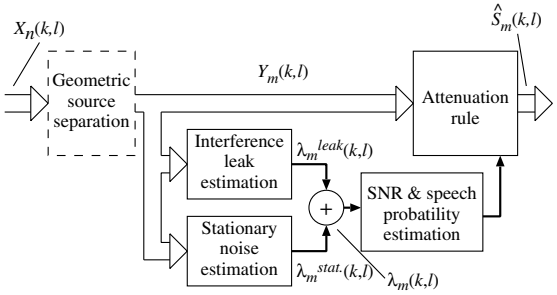


Fig. 5. Scheme of Multi-Channel Post-Filter

2.2. Sound Source Separation (SSS)

SSS consists of GSS and the multi-channel post-filter [4]. The original GSS proposed by [5] is modified in order to speed up adaptation by using a stochastic gradient and shorter time frame estimation. The multi-channel post-filter is used to enhance the output of GSS (Figure 5) based on the optimal estimator originally proposed by Ephraim [6]. This original method is extended to support multi-channel signals so that they can estimate both stationary and non-stationary noise. These estimations are used to estimate the reliability of acoustic features of separated sounds.

In spite of these simplifications, our implementation of SSS attained almost the same performance as the Valin's original implementation. SSS improved 10.3 dB in signal-to-noise ratio on average for separation of three simultaneous speech signals [4].

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 time-spectral 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.

2.3. Reliability Estimation of Separated Sounds

2.3.1. Acoustic features

To estimate reliability of acoustic features, we exploit the fact that noises and distortions are usually concentrated in some areas in the time-spectral space. As an acoustic feature, most conventional ASR systems use *Mel-Frequency Cepstral Coefficient (MFCC)* but noises and distortions are spread to all coefficients in MFCC. Thus, 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 [7]. Each of three normalization processes corresponds to one of the three normalization performed against MFCC; C0 normalization, liftering, and Cepstrum mean normalization.

2.3.2. Missing Feature Mask (MFM)

Based on the reliability estimation of separated sounds, a time-spectral map of reliability, called "Missing Feature Map" (MFM), is created. A MFM is created by comparing the input and the output of the multi-channel post-filter as shown in Figure fig:Post-Filter. Most studies on MFT have focused on a single channel input, but it is difficult to obtain enough information to estimate the reliability of acoustic features. We adopted a multi-channel approach using an 8-ch microphone array to alleviate this difficulty.

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:

$$m_k(i) = \frac{S_k^{out}(i) + N_k(i)}{S_k^{in}(i)}, \tag{1}$$

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. (1) 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. 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.

2.3.3. White noise addition

An additional method of recovering distortion is addition of a white noise to separated speech signals. This idea is motivated by the psychological evidence that noise may help human perception, which is known as *auditory induction*. This evidence is also useful for ASR, because an additive noise plays a role 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 [8]. They added office background noise after spectral subtraction, and showed the feasibility of this technique in noisy speech recognition.

2.4. Missing Feature Theory based ASR (MFT-ASR)

Missing Feature Theory (MFT) approaches use MFM of reliability to improve ASR. We adopted a classifier-modification method with marginalization, because other approaches such as cluster-based reconstruction of feature-vector imputation is not robust against mel-frequency based features [9]. 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. We use "Multi-band Julis" [7] as a MFT-ASR. 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, $x(i)$ be an acoustic feature vector, N be the size of the acoustic feature vector, and S_j be the j-th state. Let $P(\cdot)$ be a probability operator The output probability $b_j(x)$ is given below:

$$b_j(x) = \sum_{l=1}^{L} P(l|S_j) \exp\left\{ \sum_{i=1}^{N} M(i) \log f(x(i)|l, S_j) \right\}, \tag{2}$$

In accordance with the addition of white noises, we create a **white-noise-added (WNA) acoustic model** by training with both clean speech and white-noise-added speech. This WNA acoustic model is used as a single acoustic model for HARK's ASR, which reduces prior information for ASR.

3. HARK: PORTABLE ROBOT AUDITION SOFTWARE

The HARK robot audition system integrates CASA modules described in the previous section in two ways; by Missing Feature Theory in the signal processing level and by data-flow oriented middleware, "FlowDesigner" [10] in the implementation level. Conventional robot audition systems considered SSS as preprocessing for

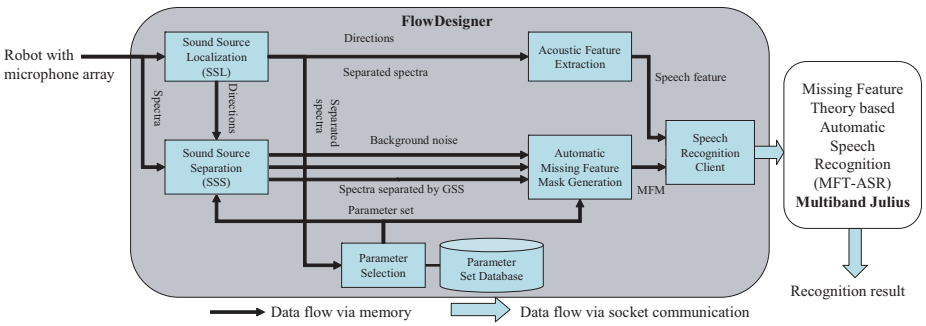


Fig. 6. Overview of the real-time robot audition system, HARK

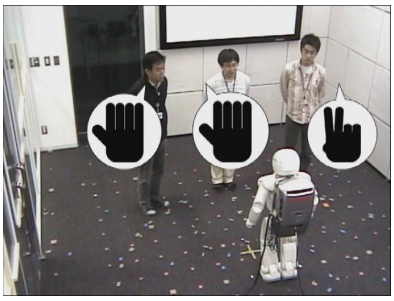


Fig. 7. Rock-paper-scissors sound game

ASR [11, 12]. They focused on the improvement of SNR and real-time processing without such integration.

The HARK consists of six modules as shown in Figure 6: 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).

The HARK allows various kinds of microphone configuration. Figure 1, 2 and 3 show an 8-ch microphone array embedded in Humanoid SIG2, Robovie R2, and Honda ASIMO, respectively. Figure 4 shows 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.

The five modules except MFT-ASR are implemented as component blocks of FlowDesigner. 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. This kind of communication is done by a function call with a pointer in FlowDesigner.

4. EVALUATION OF HARK

We evaluated the robot audition system in terms of the following three points; (1) recognition performance of simultaneous speech, (2) improvement of ASR by localization, and (3) processing speed.

4.1. Recognition of Simultaneous Speech Signals

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 $\times$ 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 of SIG2 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.

The recognition performance of three simultaneous speakers is evaluated with the following six 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** generated manually and the clean acoustic model . Since this mask is *ideal*, its result may indicate the potential upper limit of HARK.

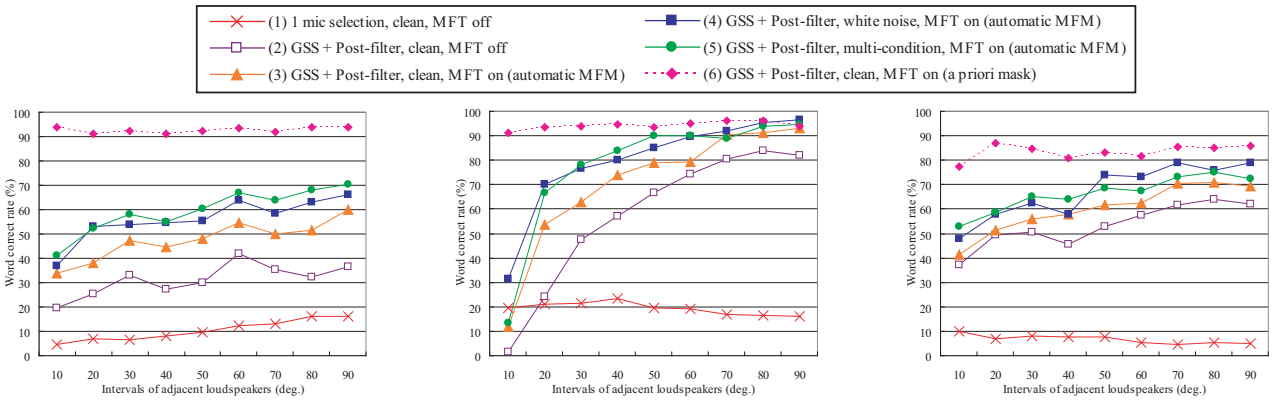
The **clean acoustic model** was trained with 10 male and 12 female ATR phonemically-balanced word-sets excluding the three word-sets (f101, m101, and m102) which were used for the recording. Thus, it was a speaker-open and word-closed acoustic model. The **MCT acoustic model** was trained with the same 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 word-sets, 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 8. 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-degrees was poor in particular for the center speaker, because any current sound source separation methods fails in separating such close three speakers. The fact that *A priori* mask showed quite a high performance may suggest some possibilities to improve the algorithms of MFM generation.

4.2. Evaluation of 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 3, which was installed in a 7 m $\times$ 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 (4), 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 acoustic model**, MUSIC outperformed steered BF,



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

Fig. 8. Word correct rates of three simultaneous speakers with our system

Table 1. Word correct rate of the center speaker (in %)

Acoustic model	White noise addition			Clean model		
direction \ *Interval*	*30°*	*60°*	*90°*	*30°*	*60°*	*90°*
manually 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

while with the **WNA acoustic mode**, both the 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 to be more ambiguous, which degraded recognition performance with the clean acoustic model. On the other hand, white noise addition to separated sounds with the WNA acoustic model reduced such degradation.

4.3. Processing Speed

The processing time when HARK separated and recognized speech signals of 800 seconds on a Pentium 4 2.4 GHz CPU is 499 second consisting of 369 sec for FlowDesigner and 130 sec for MFT-ASR. The output delay is 0.446 second. As a whole, our robot audition system ran in real time.

5. LISTEN TO THREE SIMULTANEOUS SPEAKERS

We presents two applications of the HARK for recognizing actual human speakers. The HARK used the WNA acoustic model trained with Japanese Newspaper Article Sentences (JNAS) corpus of 306 speakers. Thus, these applications were *speaker-* and *word-open*.

One application is a referee for a rock-paper-scissors sound game that includes a recognition task of two or three simultaneous utterances. ASIMO was located at the center of the room, and three speakers stood 1.5 m away from ASIMO at 30 degree intervals (Figure 7). A speech dialog system specialized for this task was connected with the HARK. ASIMO judged who won the game by using only speech information, i.e., without using visual information. Because they said rock, paper, or scissors simultaneously in an environment where robot noises exist, the SNR input sound was less than -3 dB. All of the three utterances had to be recognized successfully to complete the task. The completing rate of referee task is around 60% and 80% in the cases of three and two speakers, respectively.

Another application is that Robovie accepts simultaneous meal orders that three actual human speakers place. The HARK recognizes each meal order and confirms their orders one by one and tells the total amount of the orders. The FlowDesigner implementation reduces the response time from 8.0 sec to 1.9 sec. If the same input is given by an audio file, the response time is about 0.4 sec.

6. CONCLUSION

This paper described the HARK portable real-time robot audition system. 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. Since the HARK is open source free software, we hope that it would contribute human robot interaction and "hands-free" ASR. Several detailed experiments remains; The robustness against speech contaminated non-speech directional noise sources like music, and reverberation should be evaluated.

7. REFERENCES

[1] D. Rosenthal and H. G. Okuno, Eds., *Computational Auditory Scene Analysis*, Lawrence Erlbaum Associates, Mahwah, New Jersey, 1998.

[2] J.-M. Valin, F. Michaud, B. Hadjou, and J. Rouat, "Localization of simultaneous moving sound sources for mobile robot using a frequency-domain steered beamformer approach," *ICRA-2004*, 1033–1038, IEEE.

[3] F. Asano, H. Asoh, and T. Matsui, "Sound source localization and signal separation for office robot "Jijo-2"," *MFI-99*, 243–248, IEEE.

[4] 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," *ICRA-2005*, 1489–1494, IEEE.

[5] L. C. Parra and C. V. Alvino, "Geometric source separation: Merging convolutive source separation with geometric beamforming," *IEEE Trans. on SAP*, vol.10, no.6 (2002) 352–362.

[6] Y. Ephraim and D. Malah, "Speech enhancement using minimum mean-square error short-time spectral amplitude estimator," *IEEE Trans. on ASSP*, vol.32, no.6 (1984) 1109–1121.

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

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

[9] H. Raj and R. M. Sterm, "Missing-feature approaches in speech recognition," *IEEE Signal Proc. Mag.*, vol.22, no.5 (2005) 101–116.

[10] C. Côté, D. Létourneau, F. Michaud, J.-M. Valin, Y. Brosseau, C. Raievsky, M. Lemay, and V. Tran, "Code reusability tools for programming mobile robots," *IROS-2004*, 1820–1825, IEEE,

[11] I. Hara, F. Asano, H. Asoh, J. Ogata, N. Ichimura, Y. Kawai, *et al.*, "Robust speech interface based on audio and video information fusion for humanoid HRP-2," *IROS-2004*, 2404–2410, IEEE.

[12] 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 Comm.*, vol.44, no.1-4 (Oct. 2004) 97–112.