Introduction to Open Source Robot Audition Software HARK

Kazuhiro Nakadai[†,‡], Hiroshi G. Okuno[*], Toru Takahashi[*], Keisuke Nakamura[†],
Takeshi Mizumoto[*], Takami Yoshida[‡], Takuma Otsuka[*], and Gökhan Ince[†]
† Honda Research Institute Japan Co., Ltd. ‡Tokyo Insititute of Technology * Kyoto University

1. Robot Audition Software

Robot audition is a research field to realize the capability of a robot to listen to a general sound, i.e., a mixture of sound sources, by using its own microphones. This simultaneous listening function obviously makes human-robot interaction richer and more natural. Robot audition has only a ten-year history [1], and thus we have seldom found an *Open Source Software (OSS)* for auditory functions so that any person in the world can use them, while we can find OSS like OpenCV in computer vision, which has a longer history. We released open source robot audition software HARK[1], which aims at "OpenCV in robot audition" in 2008 as a compilation of our achievements including a function of listening to simultaneous utterances [2]. One of our dreams with HARK is to make and deploy "Prince Shotoku" robots in the world. Prince Shotoku is a Japanese legendary person who was able to understand 10 simultaneous petitions and deal with them appropriately.

So far, HARK has been downloaded from numerous universities, institutes and companies all over the world, and international collaboration through HARK has been established as mentioned in Sec. 5.. We also have continuous activities to propagate HARK via many presentations at international conferences and Japanese research meetings, and several one-day tutorials ranging from theoretical to practical topics. Such activities and international collaboration growth have given us a lot of know-how on applying HARK to robots, and have helped us to build the HARK community. Therefore, we feel that we enjoy the benefit from our OSS project for robot audition.

The rest of this paper is organized as follows: first, we discuss technical issues in robot audition. Secondly, we describe technical solutions for the issues such as sound source localization, sound source separation, speech enhancement, and automatic speech recognition of separated speech. After that, we show applications using HARK modules. Finally, we give a summary and future directions.

[1]HARK stands for HRI-JP Audition for Robots with Kyoto University, and it means "listen" in medieval English. It follows HARK license, i.e., it is currently open-sourced for research purposes, and can be licensed for commercial purposes (see also http://winnie.kuis.kyoto-u.ac.jp/HARK/).

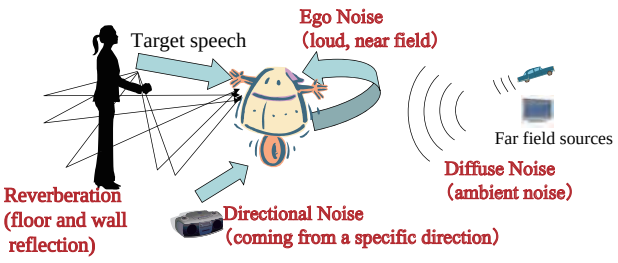


Figure 1: Noise types surrounding a robot

2. Issues in Robot Audition

Automatic Speech Recognition (ASR) systems recently have performed well under noisy environments as long as a microphone is located close to the mouth of a speaker. On the other hand, robot audition should deal with a distant speaker in the same way as hands-free speech recognition and distant-talking speech recognition. In this case, a signal-to-noise ratio (SNR) becomes considerably low because the power of input signals attenuates in inverse proportion to the square of the distance and the other noise sources are mixed with the input signals. It is difficult to deal with this situation only by improving a decoding algorithm, an acoustic model and a language model in ASR. Figure 1 shows four kinds of noises that a robot has to deal with in a general environment as follows:

1. Directional noise: noise coming from a specific direction such as TV sounds and human voice.
2. Diffuse noise: ambient noise, and background noise.
3. Reverberation: noise with time delays due to reflection in the ceiling, the floor and the walls.
4. Ego-noise: noise generated by a robot's fans and actuators, which has both aspects of directional and diffuse noise.

3. Approaches to Solve the Issues in HARK

HARK takes three approaches to solve the above noise issues as follows:

1. Introduction of a microphone array.
2. Use of microphone-array-based techniques as preprocessing of ASR such as sound source local-

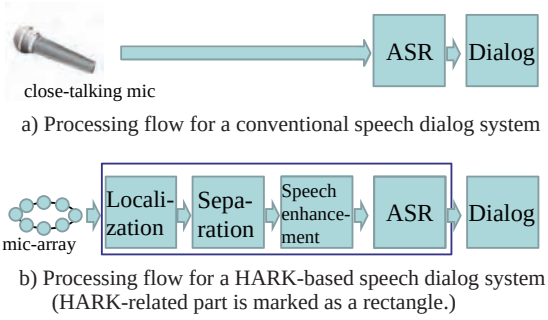


Figure 2: The Difference between Conventional Speech Dialog and HARK-based Systems

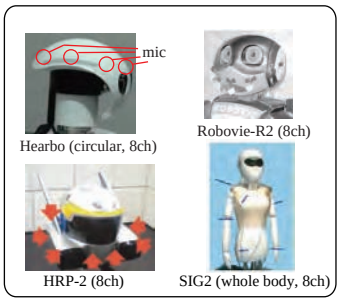


Figure 3: Microphone layouts

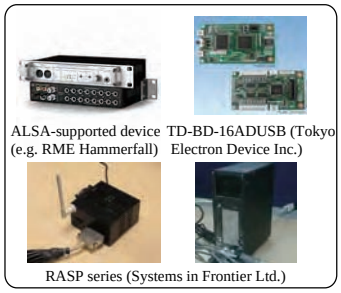


Figure 4: HARK Devices

ization, sound source separation and speech enhancement.

3. Improvement of ASR systems by introducing Missing Feature Theory (MFT).

Figure 2 shows the difference between a conventional speech dialog system and HARK-based speech dialog system. Because the SNR of the target speech drastically improves with HARK, the HARK-based speech dialog system is able to work properly under a highly-noisy environment. Note that since a conventional ASR system sometimes utilizes single channel noise reduction and speech enhancement, it is noise-robust to some extent when the noise level is low.

Directional noise can be separated using directions of arrival for sound sources as long as sound sources are sparse. For example, when a direction interval between two sound sources is more than 20 degrees, and the number of sound sources is less than that of the microphones, every sound source can be separated theoretically [2]. Since diffuse noise does not include direction information explicitly, another technique like speech enhancement is necessary. For reverberation, we divide it into two factors; early reverberation and late reverberation. The former is called inter-frame reverberation because its effect is limited within a frame which is a temporal unit of ASR (usually 25 ms). Thus, improvement in an acoustic model of ASR can relax this effect. On the other hand, the latter is difficult to suppress because we cannot assume the fixed model for reverberation. For ego-noise suppression, we propose a template-based method using joint status information. Although this method is still under development, we will introduce it to HARK in the near future.

3·1 Algorithm selection policy

Actually, there are a lot of algorithms for noise reduction based on microphone arrays. The selection of algorithms is a key issue. Some algorithms like beamforming are robust for environmental changes, but their peak performance is relatively low. Adaptive algorithms have high peak performance, but they easily deteriorate when they face unexpected situations. Recently, Microsoft released Kinect SDK which includes microphone array processing such as beamforming using four microphones embedded in Kinect. It seems that they selected the former algorithms because such algorithms are free from optimization and parameter tuning, and thus it is easy to use them. We took a greedy strategy, that is, we basically selected the latter adaptive algorithms, because their performance is crucial to ASR, and at the same time, we improved the algorithms so that they had less parameters, in addition, some parameters were adapted to be optimal for practical use.

4. Overview of HARK

HARK 0.1.7 was released in 2008 and major version-up to 1.0.0 was made in Nov. 2010. HARK is featured by the following points:

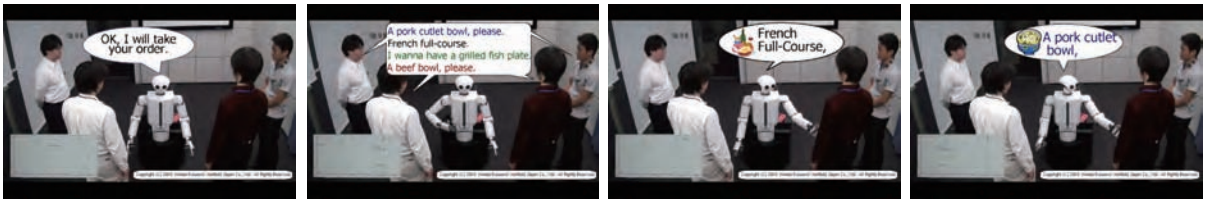
1. GUI-based flexible customization and integration based on dataflow-oriented programming environment Flowdesigner [3].
2. A wide range of functional modules for robot audition.
3. Support of multi-channel A/D converters.
4. Online and real-time processing.
5. Full documentation (manual and cookbook) in Japanese and English.

Currently, HARK supports Linux OS (e.g. Ubuntu 10.04), and we are going to support Windows OS and new sound devices like Kinect.

4·1 GUI-based flexible customization and integration

Required functions for robot audition depend on the target applications. This means that robot audition software should support the users such that they can flexibly select the required functions and easily integrate them. Since Flowdesigner [3] supports both GUI-based integration and batch-based execution, we decided to use it as middleware to fulfill this requirement, and implemented robot audition functions as modules for Flowdesigner.[2] To build a robot audition

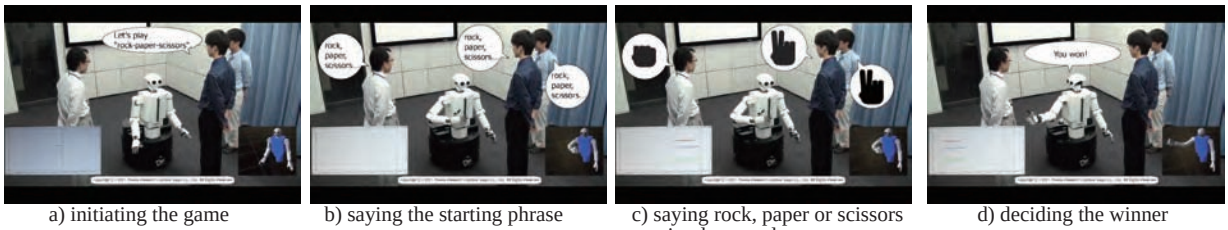
[2]We also released the revised version of Flowdesigner to fit HARK modules better.



a) taking order b) ordering simultaneously c) confirming the leftmost order d) confirming the second left one

e) confirming the third left one f) confirming the rightmost one g) showing the total price h) completing order

Figure 5: Meal order taking



a) initiating the game b) saying the starting phrase c) saying rock, paper or scissors simultaneously d) deciding the winner

Figure 6: A referee for rock-paper-scissors sound game

application with HARK, a user simply selects necessary modules from the module list and connects the selected modules. If necessary, parameters for each module such as a microphone layout can be set in a property window. For beginners, several sample networks which are typical in robot audition applications can be downloaded from our web site.

4·2 Functional modules for robot audition

According to the policy of algorithm selection, we provide MUltiple SIgnal Classification (MUSIC), Geometric High-order Decorrelation-based Source Separation with Adaptive Stepsize (GHDSS-AS)[4], Histogram-based Recursive Level Estimation (HRLE)[5] and MFT-ASR for sound source localization, sound source separation, speech enhancement and ASR, respectively. MUSIC is an adaptive beamformer, and it is known that it provides noise-robust sound source localization. GHDSS-AS is a hybrid algorithm of beamforming and blind separation, and thus, this method can benefit from the merits of these two methods. In addition, a step-size parameter to update a separation matrix, which is manually configured in most cases, is automatically optimized with GHDSS-AS. HRLE is a practical speech enhancement technique. It has only two parameters to be tuned, while "postfilter" provided with the previous version of HARK has 13 parameters. MFT-ASR can cope with distortions caused by sound source separation and speech enhancement by using missing feature masks. The details and the completed list of HARK modules are shown in [2] and the HARK manual.

4·3 Multi-channel audio devices

HARK supports three types of multi-channel A/D converters shown in Figure 4 as follows:

- Audio devices supporting Advanced Linux Sound Architecture (ALSA). e.g. RME Hammerfall series.
- TD-BD-16ADUSB developed by Tokyo Electron Device, Ltd. (16 ch A/D converter with USB connection for embedded use)
- RASP series developed by System in Frontier Inc. (8 ch and 16 ch A/D converter with wired/wireless LAN connection)

We usually use 8 ch microphone array shown in Figure 3. However, HARK does not specify the number of microphones, and users can select it according to their target application.

4·4 Online and realtime processing

Online and realtime processing is crucial for robot applications. We usually use 30 ms and 10 ms for frame length and frame shift, because these are standard values for ASR. We implemented each module so that the processing time can be less than 10 ms. Most of the integrated systems as we provided with sample networks are guaranteed to have a total processing time less than 10 ms.

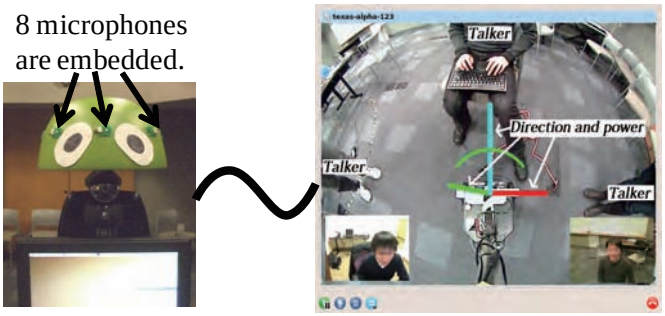


Figure 7: Texai's GUI for a remote user

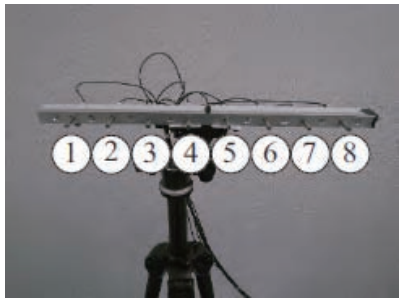


Figure 8: The Ear Sensor in LAAS-CNRS

5. Application of HARK

5·1 Speech dialog tasks

HARK is applied to several speech dialog tasks. Figure 5 shows a simultaneous meal order taking task. In this task, the robot called Hearbo first listens to simultaneous orders. Using sound source localization, Hearbo separates them into each utterance, and recognizes each separated utterance. Finally, it confirms the orders and calculates the total price. Figure 6 shows a robot referee for a rock-paper-scissors sound game. Hearbo hosts a rock-paper-scissors sound game. Each player says one of rock, paper, and scissors synchronously. Hearbo listens to their utterances and decides who is a winner. These tasks used almost the same HARK network, and two different dialog programs, that is, for meal order taking and rock-paper-scissors sound game. These tasks are already working on several robot platforms like Robovie and HRP-2 simply by changing some parameters like robot acoustic model, i.e., transfer functions. This shows general applicability and high reusability of HARK network. Performance evaluation of HARK is shown in [2, 4, 5, 6].

5·2 Our activities with HARK

In Mar. 2009, we implemented several functions using HARK on tele-presence robot Texai developed by Willow Garage, Inc. [7] This tele-presence robot is controlled from a remote user, and it had a problem that it was difficult to know who is speaking because the remote user had to listen to speech contaminated by all surrounding noise sources. Figure 7 shows the functions provided by HARK for the remote user, that is, sound source localization and separation. Sound source localization provides the direction of the speaking person shown as lines. With sound source separation, the remote user can control a sound for listening by changing the direction and the size of an arc. [3]. In Nov., 2010, two students in the HARK support team have visited LAAS-CNRS for a month, and they ported HARK to the ear sensor developed by LAAS-CNRS as shown in Figure 8. We will collaborate to evaluate the sensor in a real environment in terms of sound source localization. Another activity to distribute HARK, we are continuously planning to have HARK tutorials. So far, we had two international tutorials in Korea and in France (at Humanoids 2009), and three domestic tutorials in Japan. These tutorials are also useful for us to improve HARK based on feedback from participants.

[3] See http://www.willowgarage.com/blog/2010/03/25/hark-texai

6. Summary and Future directions

This paper introduced HARK, which is open-sourced robot audition software. HARK provides basic functions for robot audition like sound source localization, sound source separation, speech enhancement, and automatic speech recognition of separated speech. Users can make their own program using GUI thanks to our middleware Flowdesigner. We also showed some applications of HARK and our activities to deploy HARK. In the near future, we will support Windows OS for HARK and other sound devices like Kinect. We hope that HARK is widely used in robotics and also in any other field.

Acknowledgment

We thank Prof. H.Nakajima of Kogakuin Univ. for his contribution to HARK. We also thank Dr. J.-M. Valin, Dr. S. Yamamoto and every member in Okuno Lab., Kyoto Univ. and in HRI-JP.

References

[1] K. Nakadai *et al.*: Active Audition for Humanoid, AAAI-2000, pp. 832-839, AAAI.
[2] K. Nakadai *et al.*: Design and Implementation of Robot Audition System "HARK", Advanced Robotics, vol.24, pp.739-761 (2010).
[3] C. Côté *et al.*: Code Reusability Tools for Programming Mobile Robots, IEEE/RSJ IROS 2004, pp.1820-1825.
[4] H. Nakajima *et al.*: Blind Source Separation with Prameter-Free Adaptive Step-Size Method for Robot Audition, IEEE T-ASLP, 18(6), pp.1476-1484 (2010)
[5] H. Nakajima *et al.*: An Easily-Configurable Robot Audition System Using Histogram-Based Recursive Level Estimation, IEEE/RSJ IROS 2010, pp.958-963.
[6] S. Yamamoto *et al.*: Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory. IEEE-RAS ICRA 2005, pp.1427-1482.
[7] T. Mizumoto *et al.*: Design and Implementation of Selectable Sound Separation on a Texai Telepresence System Using HARK. IEEE-RAS ICRA 2011, pp.2130-2137.