Robot Auditon: Missing Feature Theory Approach and Active Audition

Hiroshi G. Okuno, Kazuhiro Nakadai, and Hyun-Don Kim

Hark, hark, I hear! The Tempest, Willian Shakespeare

Abstract Robot capability of listening to several things at once by its own ears, that is, *robot audition*, is important in improving interaction and symbiosis between humans and robots. The critical issue in robot audition is real-time processing and robustness against noisy environments with high flexibility to support various kinds of robots and hardware configurations. This paper presents two important aspects of robot audition; Missing-Feature-Theory (MFT) approach and active audition. HARK open-source robot audition incorporates MFT approach to recognize speech signals that are localized and separated from a mixture of sound captured by 8-channel microphone array. HARK is ported to four robots, Honda ASIMO, SIG2, Robovie-R2 and HRP-2, with different microphone configurations and recognizes three simultaneous utterances with 1.9 sec latency. In binaural hearing, the most famous problem is a front-back confusion of sound sources. Active binaural robot audition implemented on SIG2 disambiguates the problem well by rotating its head with pitting. This active audition improves the localization for the periphery.

1 Robot Audition – Why, What and How?

Speech recognition plays an important role in communication and interaction, and people with normal hearing capabilities can listen to many kinds of sounds under various acoustic conditions. Robots should have hearing capability equivalent to humans to realize human-robot communication, when they are expected to help us in a daily environment. In daily environments, there exist a lot of noise sources including robot's own motor noises besides a target speech source. Many robot systems avoided this problem by forcing interaction parcitipants to wear a headset mi-

Horoshi G. Okuno and Hyun-Don Kim are with the Graduate School of Informatics, Kyoto University, Yoshida-Honmachi, Sakyo, Kyoto 606-8501 JAPAN, e-mail: {okuno, hyundon}@kuis.kyoto-u.ac.jp. Kazuhiro Nakadai is with Honda Research Institute Japan Co. Ltd. 8-1 Honmachi, Wako, Saitama 351-0188 Japana, e-mail: nakadai@jp.honda-ri.com and Tokyo Institute of Technology.

crophone [7]. For smoother and more natural interactions, a robot should listen to sounds by its own ears instead of using parcitipants' headset microphones.

"*Robot Audition*" research we have proposed [18] aims to realize recognition of noisy speech such as simultaneous speech by using robot-embedded microphones. As the robot audition research community is gradually growing, we have organized sessions on robot audition at IEEE/RSJ IROS 2004-2009 to promote robot audition research worldwide and special session on robot audition at IEEE-Signal Processing Society ICASSP-2009 to trigger cooperation between IEEE-RAS and -SPS.

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 [30]. Its goal is to understand an arbitrary sound mixture including speech, non-speech signals, and music in various acoustic environments.

Two key ideas for CASA are 1) *the Missing Feature Theory (MFT) approach* [29] and 2) *active audition*. MFT approach treats each feature as either reliable or unreliable. Since noise or distortion still carries information, unreliable features may have some information. In Figure 1a), People can not see a letter "A." On the contrary, other information such as occlusion and noise helps the organization of fragments as is shown in Figure 1b). It is known that in the human auditory system noises that pad temporal gaps between sound fragments help auditory perception organization [8]. This is called "perceptual closure"in Gestalt psychology.

Active audition [18] is the auditory equivalent of acive vision. Similar to active vision, a robot may move its microphone or body to improve auditory perception. For binaural hearing, like human with two microphones, it is usually difficult to determine whether the sound source is ahead of or at the rear. This ambiguity is

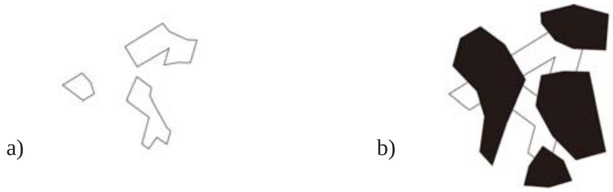

Fig. 1 Perceptual closure in Gestalt psychology. *Noises provide information on preception.* For the left part a), without noises, people cannot recognize the letter easily. Three fragments do not organize. For the right part b), with noises, people can do it easily. Noises help organization.

referred to as “front-rear” confusion. Suppose that the head moves to right. If the sound source moves to the same direction, it is behind. Otherwise, it is ahead of. The problem with active audition is motor noises caused by robot’s own movements.

Three primitive functions for CASA are *sound source localization (SSL)*, *sound stream separation (SSS)*, and its recognition including *automatic speech recognition (ASR)*. Although robot audition share CASA’s primitive functions as listed above, the critical requirements in robot audition are *real-time processing* and *robustness against diversity of acoustic environments*. These requirements are pursued in implementation on robots and their deployment to various acoustical environments.

Several groups have studied robot audition, in particular, SSL and SSS [11, 13, 16, 20, 34, 36, 38, 40]. Since they focused on *their own robot platform*, their systems are neither available nor portable for other research groups. Thus, researchers who want to incorporate robot audition in their robot had to make their own robot audition system from scratch. Valin released SSL and SSS software for robots called ”ManyEars”[1] as GPL open-sourced software. This is the first software which can provide generally-applicable and customizable robot audition systems. ManyEars convers only SSL and SSS, but ASR is not included. Robot audition software should support ASR by integrating SSL and SSS, because ASR has a lot of parameters that affect the performance of a total robot audition system severely.

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”[2] (HRI-JP Audition for Robots with Kyoto University).

The rest of the paper is organized as follows. Section 2 describes the HARK open-source Robot Audition Software. Section 3 evaluates the performance of HARK. Section 4 describes the active binaural robot audition system. Section 5 presents the evaluation of active audition to disambiguate the front-rear confusion . Section 6 concludes the paper.

2 Open-source Robot Audition Software HARK

HARK provides a complete module set for robot audition (see Fig 2). The modules are categorized into six categories as is shown in Table 1: multi-channel audio input, sound source localization and tracking, sound source separation, acoustic feature extraction for automatic speech recognition (ASR), automatic missing feature mask (MFM) generateion, and ASR interface. MFT based ASR (MFT-ASR) is also provided as a patched source code for a Japanese/English open source ASR, Julius/Julian[3]. All modules except MFT-ASR run on FlowDesigner[4], because these

[1] http://ManyEars.sourceforge.net/

[2] The word “hark” is an old English that stands for “listen.” HARK of the current verison 0.1.17 is available at the following URL. HARK 1.0.0 will be released this mid-autumn .
http://winnie.kuis.kyoto-u.ac.jp/HARK/

[3] http://julius.sourceforge.jp

[4] http://flowdesigner.sourceforge.net/

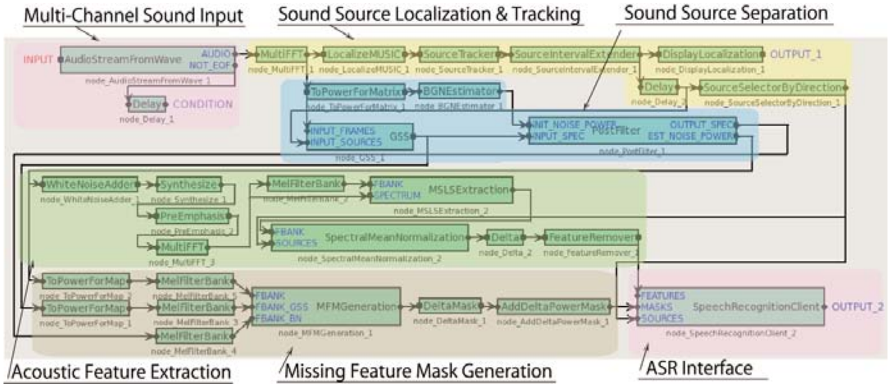

Fig. 2 An Instance of robot audition system using HARK 1.0.0 with FlowDesigner Interface. It comprises six main modules from multi-channel sound input to interface to Automatic Speech Recognition (ASR), which runs as a separate module.

modules and MFT-ASR do not share audio data. MFT-ASR runs separately through ASR interface over the network between PCs. HARK also provides a set of support modules under the category of data conversion and operation.

FlowDesigner is open-sourced and integrates modules by using shared objects, that is, function-call-based integration. It allows us to create reusable modules and to connect them together using a standardized mechanism to create a network of modules, as is shown in Figure 2. The modules are connected dynamically at run time. A network composed of multiple modules can be used in other networks. This mechanism makes the whole system easier to maintain since everything is grouped into smaller functional modules. If a program is created by connecting the modules, the user can execute the program from the GUI interface or from a shell terminal.

When two modules have matching interfaces, they are able to be connected regardless of their internal processes. One-to-many and many-to-many connections are also possible. A module is coded in programming language C++ and implemented as an inherited class of the fundamental module. Dynamic linking at run time is realized as a shared object in the case of Linux. Since data communication is done by using a pointer, it is much faster than socket communication. Therefore, FlowDesigner maintains a well-balanced trade-off between independence and processing speed. We have extended FlowDesigner to be more informable and robust against erros to use it as a programming environment for HARK.

In the remaining of this section, we explain main categories. Since a lot of audio signal processing methods and technologies have been developed for particular conditions under particular assumptions, HARK is designed to provide some of promising methods for robot audition. Please note that there is no panacea for robot audition and the tradeoff between generality and performance is critical.

Table 1 Modules provided by HARK 1.0.0
HARK 1.0.0 has six categories of FlowDesigner modules, one non-FlowDesigner module and one miscellaneous modules.

Category Name	Module Name
Multi-channel Audio I/O	AudioStreamFromMic AudioStreamFromWave SaveRawPCM
Sound Source Localization and Tracking	LocalizeMUSIC ConstantLocalization SourceTracker DisplayLocalization SaveSourceLocation LoadSourceLocation SourceIntervalExtender
Sound Source Separation	DSBeamformer GSS Postfilter BGNEstimator
Acoustic Feature Extraction	MelFilterBank MFCCExtraction MSLSExtraction SpectralMeanNormalization Delta FeatureRemover PreEmphasis SaveFeatures
Automatic Missing Feature Mask Generation	MFMGeneration DeltaMask DeltaPowerMask
ASR Interface	SpeechRecognitionClient SpeechRecognitionSMNClient
MFT-ASR	Multiband Julius/Julian (non-FlowDesigner module)
Data Conversion and Operation	MultiFFT Synthesize WhiteNoiseAdder ChannelSelector SourceSelectorByDirection SourceSelectorByID MatrixToMap PowerCalcForMap PowerCalcForMatrix

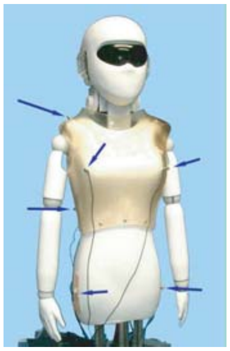

Fig. 3 SIG2 has 8 microphones on its body.