ROBOT AUDITION: ITS RISE AND PERSPECTIVES

Hiroshi G. Okuno† *and Kazuhiro Nakadai*‡

† Graduate Program for Embodiment Informatics, Waseda University,
‡ Honda Research Institute Japan, Co. Ltd./Tokyo Institute of Technology

ABSTRACT

The ability of robots to listen to several things at once with their own "ears", that is, *robot audition*, is an important factor 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 first overviews activities and issues related to robot audition. Then, it presents the "HARK" robot audition software, which provides three primary functions for robot audition, sound source localization, sound source separation, and separated sound recognition, and then reports their performance. Finally, it discusses future directions in new promising areas as well as robotics.

Index Terms— Robot Audition, Microphone array, Active audition, Ego-noise cancellation

1. WHAT IS ROBOT AUDITION AND WHY IS IT NEEDED?

Speech recognition plays an important role in communication and interaction, and people with normal hearing capabilities can listen to many kinds of sounds at once under various acoustic conditions. Robots should have hearing capability equivalent to that of humans to achieve effective and smooth human-robot communication, particularly robots expected to help us in our daily environments. In daily environments, there are many noise sources, including the robot's own motor noises, besides the target speech source. Many robot systems for social interaction avoid this problem by forcing interaction parcitipants to wear a headset microphone [9]. For smoother and more natural interactions, a robot should instead listen to sounds with its own "ears".

The robot audition concept proposed by Nakadai and Okuno [22] aims to realize recognition of noisy speech such as simultaneous speech by using microphones embedded on the robot. Robot audition is critical in bidirectional human robot interactions (HRI). Without robot audition, HRI is mostly one-directional. While robots can speak fluently and emotionally thanks to recent advances in speech synthesis technology, they have trouble hearing what people say simultaneously, under noisy conditions, and at non-personal distances.

Robot audition is expected to facilitate capabilities similar to human ones. For example, people can focus their listening attention on one speaker even in a noisy environment. This capability is known as the cocktail party effect. For a robot to have this capability it must be able to separate out a speech stream from a mixture of sounds. It may even achieve the hearing capability of "*Prince Shotoku*", who according to Japanese legend, could listen to the petitions of ten people at once.

Thanks to JSPS Kakenhi (S) No.24220006 and ImPACT "Tough Robotics Challenge" for funding.

1.1. Computational Auditory Scene Analysis

A robot should be able to "understand" many kinds of sounds as well as produce various sounds. Auditory scene analysis is the process of simulating useful intelligent behavior; it is required even when objects are not visible. While auditory research has traditionally focused on human speech understanding, understanding auditory scenes in general is receiving increasing attention. Researchers in the field of computational auditory scene analysis (CASA) have been studying a general framework of sound processing and understanding [33, 41]. Their goal is to enable an arbitrary sound mixture including speech, non-speech signals, and music to be understood in various acoustic environments.

Three primitive functions are needed for CASA: sound source localization (SSL), sound stream separation (SSS), and automatic speech recognition (ASR). Robot audition also makes use of these functions, but the critical requirements in robot audition are real-time processing and robustness against diversity of acoustic environments. CASA researchers are working on ways to implement robots with these capabilities and on their deployment in various acoustical environments.

1.2. Robot Audition in 20th Century

Let us imagine situations in which autonomous robots are used in various environments, such as a pet robot in one's living room, a service robot in the office, or a robot serving drinks at a party. The robot should be able to identify people in the room, distinguish their voices, look at them to identify them visually, and associate voices with visual images so that highly robust auditory scene analysis can be accomplished. These are the minimum requirements for social interaction [10].

Some robots are equipped with an advanced robot-human interface. *Jijo-2* [1], for example, can recognize a phrase command by using its speech-recognition system; *AMELLA* [40] can recognize pose and motion gestures. *Kismet* of MIT AI Lab [8] can recognize speech by using its speech-recognition system and express various kinds of sensation. *Hadaly* of Waseda University [19] can localize the speaker in multi-party interactions as well as recognize speech by using its speech-recognition system. *Takemaru* of the Nara Advanced Institute of Science and Technology served as a concierge robot for several years in public spaces [28]. For these robots, except for *Jijo-2*, it is assumed that the speaker is near to the robot's microphone. That is, they do not exploit robot audition in human-robot interactions.

1.3. Issues in Robot Audition

Several groups have studied robot audition, in particular, SSL and SSS [13, 15, 20, 23, 37, 39, 42, 43]. Since they focused on their

own robot platform, their systems are neither available nor transferable for other research groups. Thus, researchers who want to incorporate robot audition into their robot must make their own robot audition system from scratch. Valin released the "ManyEars" SSL and SSS software for robots as GPL open-source software. This is the first software that can provide generally applicable and customizable robot audition systems. ManyEars supports only SSL and SSS – ASR is not supported.

Robot audition software should support ASR by integrating SSL and SSS because ASR has many parameters that greatly affect the performance of a total robot audition system. Therefore, the technical issues in robot audition should focus on system-integration technology as well as individual technologies.

- *ManyEars* (Univ. of Sherbrooke, Canada): GSS is implemented on FlowDesigner and on DSP. ROS is supported. `http://ManyEars.sourceforge.net/`
- *The EAR project* (CNRS-LAAS, France): SSL is implemented using software and hardware [7].
- *OpenRTM* (CNRS-LAAS, France): SSL and SSS by beamformer, echo cancellation, and ASR are supported.
- *Kinect SDK* (Microsoft, USA): SSL, echo cancellation, noise cancellation, and ASR are supported.
- *BINAAHR* (CNRS-LAAS, France): SSL with 2 microphones, i.e., binaural SSL, is supported.
- *HARK*[1] (HRI-JP and Kyoto University): Most popular robot audition software with over 30,000 downloads in 2014 (as of 31 Aug. 2014).

The requirements for robot audition are summarized below [24]:

1. It should localize, separate, and recognize sound sources robustly even when there are multiple speech and noise sources.
2. It should provide a set of modules for signal processing including SSL, SSS, ASR, sound input devices, and other miscellaneous functions.
3. It should provide an easy way to choose and combine various kinds of modules.
4. It should support real-time processing or at least minimize the time delay by using a mechanism to share acoustic data between modules.
5. It should have high usability so that various kinds of researchers and developers can use it.

HARK has been maintained by the HARK support team and is used widely. Its various functions for robot audition are explained in Section 3.

Binaural audition for robots is commonly studied for three reasons. Since humans and most animals have two ears, robot audition with two microphones has been well studied [15]. Since stereo input devices are ubiquitously avaliable, binaural audition is extensively used when it can cope with a mixture of multiple sound sources. Since binaural audition can be considered as a computational model of binaural hearing, it is expected to contribute to an understanding of human and animal hearing capabilities.

Multiple sound sources usually degrades the perfomance of binaural audition while other media or multimodal information fusion may improve its performance. For example, active audition [22] is the auditory equivalent of active vision. Similar to active vision, a robot may move its microphone or body to improve auditory perception. With binaural hearing, like that of a person with two ears, it is usually difficult to determine whether the sound source is in front of or behind the listener. This ambiguity is refered to as "front-rear" confusion. Suppose that a listener moves his or her head to the right. If the sound source moves in the same direction, it is behind the listener. Otherwise, it is in front of the listener. The problem with active audition is the presence of motor noises caused by the robot's own movements.

[1] "Hark" is an old English word meaning "listen."

2. ACTIVITIES RELATED TO ROBOT AUDITION

In this section we summarize the meetings and sessions related to robot audition and the major robot audition projects of which we are aware of. Hiroshi G. Okuno, one of the authors, was elevated to IEEE Fellow for his contributions to robot audition technologies in 2012 upon the recommendation of the Robot and Automation Society. He is the first IEEE Fellow in the field of robot audition.

2.1. Meetings and Sessions

1. Organized Session on "Robot Audition" at the IEEE/RSJ Intern'l Conferences on Intelligent Robots and Systems (IROS) from 2004 to 2013. Organizers: Hiroshi G. Okuno, Kazuhiro Nakadai and others.
2. Session on "Audition" at IROS-2014. IROS-2014 did not have any organized sessions. Note that "robot audition" was added to the keyword list for IROS-2014. This is the first time robot audition has been added to the list for IROS/ICRAs.
3. Special Session on "Robot and Signal Processing" at ICASSP-2009. Organizers: Akihiro Sugiyama and H. G. Okuno.
4. Special Session on "Audio for Robots – Robots for Audio" at ICASSP-2015. Organizers: Emmanuel Vincent and Jonathan Le Roux.
5. Organized Meetings on "Robot Audition" in the AI Challenges SIG of the Japanese Society of Artificial Intelligence every year from 2000 to 2014. Organizer: H. G. Okuno and K. Nakadai.
6. Organized Sessions on "Robot Audition" at the Annual Meeting of the Robotic Society of Japan every year from 2005 to 2014 in Japan. Organizers: H. G. Okuno, K. Nakadai and others.

2.2. Projects

1. "Robot Audition from Computational Auditory Scene Analysis" (PI: Hiroshi G. Okuno) Apr. 2007 – Mar. 2012. JSPS Grant-in-Aid for Scientific Research (KAKENHI) (S) (Total amount: 119M yen[2]) `http://winnie.kuis.kyoto-u.ac.jp/HARK/`
2. "Multiple Development of Robot Audition" (PI: Hiroshi G. Okuno) Apr. 2012 – Mar. 2017 (expected). JSPS KAKEHI (S) (Total amount: 218M yen) `http://www.hark.jp/`
3. "Binaural Active Audition for Humanoid Robots (BINAAHR)" (PI-JP: Hiroshi G. Okuno, PI-FR: Patrick Danès). Sep. 2009 – Mar. 2013 (Total amount: 26M yen) (FR: May 2010 – Oct. 20132) JST-ANS Japan-France Intern'l Collaborative Research. `http://projects.laas.fr/BINAAHR/`

[2] 1 USD was equal to around 90 yen and 109 yen in 2010 and 2014, respectively.

4. "Embodied Audition for Robots (EARS)" (PI: Radu Patrice Horoud) Jan. 2014 – Dec. 2016. as a European project. http://robot-ears.eu/
5. FP7 FET-Open Two!Ears project, "Reading the World with Two!Ears" (PI: Alexander Raake) Dec. 2013 – Nov. 2016. 3M Euro. http://www.twoears.eu/
6. "Robot Audition for Extreme Environments" (GL: Hiroshi G. Okuno) Oct. 2014 – Dec. 2018. Total amount: tentative 145M yen. As the Intelligent Component Team of ImPACT (IMpulsing PAradigm Change through disruptive Technology, Cabiet Office, Government of Japan) "Tough Robotics Challenge" (PM: Satoshi Tadokoro, tentative 3,500M yen).

3. HARK: OPEN-SOURCE ROBOT AUDITION

The structure of HARK is shown in Figure 1. The GUI is based on HARK Designer and most modules are interfaced with BatchFlow middleware. ASR is independent from HARK. Some tools are provided as independent programs.

3.1. Sound source localization

HARK provides several noise-robust SSL algorithms based on MUltiple SIgnal Classification (MUSIC) [35]. MUSIC is based on eigenvalue decomposition. Since MUSIC has sharper peaks for sound source directions than conventional beamformers such as a delay-and-sum beamformer, it is a noise-robust algorithm. However, when the level of the noise signals is higher than that of the target signals, the performance of SSL deteriorates. This is because MUSIC assumes that the eigenvalues of target signals are larger than those of the noise signals; that is, the target signal should be theoretically loud enough in MUSIC.

To solve this problem, HARK extended conventional MUSIC by applying generalized eigenvalue decomposition (GEVD) and generalized singular value decomposition (GSVD), i.e., GEVD-MUSIC and GSVD-MUSIC, respectively [27]. These methods use knowledge about noise sources stored in a noise correlation matrix calculated from noise signals captured in advance. Both GEVD-MUSIC and GSVD-MUSIC can deal with extremely noisy environments in which the signal-to-noise ratio is less than 0 dB. In fact, GEVD-MUSIC can cope with -20 dB offline, and GSVD-MUSIC can cope with -10 dB in realtime.

Since these methods require a pre-estimated noise correlation matrix, incremental estimation of the noise correlation matrix was introduced so that dynamic changes in noise can be coped with. The resulting iGEVD-MUSIC and iGSVD-MUSIC ("i" stands for "incremental") where shown to be effective by using a quadrotor with

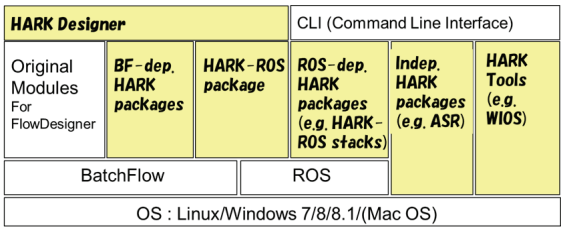

Fig. 1. Structure of HARK wrt OS, BatchFlow and ROS. HARK consists of hatched parts.

Table 1. SSS algorithms supported by HARK-SSS
[AS] means that adaptive step-size control is supported.

Fixed Beamforming
1. Delay-and-Sum Beamformer (DS)
2. Null Beamformer (NULL)
3. Weighted Delay-and-Sum Beamformer (WDS)
4. Indefinite Least Square Estimator Beamformer (ILSE)
Explicit Use of Noise Information
5. Maximum Likelihood Beamformer (ML) [6, 36]
6. Maximum SNR Beamformer (MSNR) [21]
Linearly Constrained Minimum Variance [AS]
7. Linear Constrained Minimum Variance Beamformer (LCMV) [11]
8. Griffith-Jim Beamformer (GJ) [12]
Linearly Constrained Blind Separation [AS]
9. Geometric Source Separation (GSS) [31]
10. Geometric Independent Component Analysis (GICA) [16]
11. Geometric High-order Decorrelation based Source Separation (GHDSS) [supported by main package of HARK] [26]

a microphone array. The results showed that a sound source at 20 m can be detected under dynamically changing noise conditions produced by propellers and wind noise [29].

3.2. Sound source separation

Geometrically-constrained higher-order decorrelation-based source separation with adaptive step-size control (GHDSS-AS) [26] is included in the main package of HARK. This is a hybrid algorithm between blind separation and beamforming developed by extending geometric source separation (GSS) [31] to improve separation performance and to deal with dynamically changing sound sources. Blind separation in GSS relies merely on cross-power correlation while GHDSS-AS uses higher-order correlation similar to independent component analysis (ICA). The original GSS supports only offline methods while GHDSS-AS extends online GSS [37] by using an adaptive step-size control method. It maintains a step-size parameter that incrementally updates a separation matrix to be optimal by using Newton's method. Because GHDSS-AS shows good performance with a robot, most of our demonstrations with HARK used GHDSS-AS. However, other algorithms may have better performance in different situations. Therefore, we released another package called HARK-SSS which includes ten standard SSS algorithms besides GHDSS-AS, as shown in Table 1.

3.3. Automatic speech recognition

Missing feature theory (MFT) [5, 32] was introduced as an ASR module for HARK. MFT is able to cope with distortion caused by microphone array processing and speech enhancement by masking out unreliable features on recognition. In combination with a spectral acoustic feature called Mel-frequency cepstrum coefficient (MSLS), MFT achieves simultaneous speech recognition [38]. The implementation of the ASR module is based on Julius[3], so the module is called "MFT-Julius". Currently, conventional GMM-HMM is supported while DNN-HMM based ASR should be available in the near future.

3.4. Performance of Sound Source Separation

The word accuracy results of ASR with GHDSS-AS and eigth other SSS algorithms selected from HARK-SSS except ILSE are summarized in Figure 2. The evaluation was performed under three types

[3]http://julius.sourceforge.jp/en_index.php?q=index-en.html

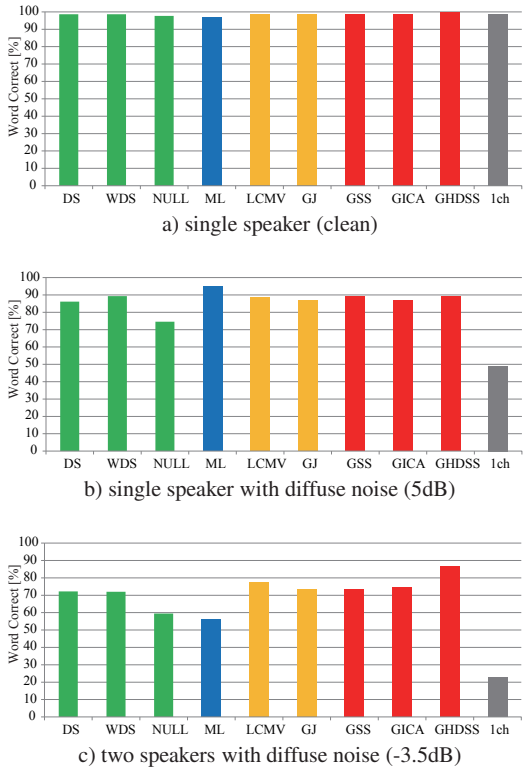

Fig. 2. Comparison of 9 sound source separation algorithms supported by HARK-SSS. An 8-element circular microphone array embedded on the head of Hearbo. The room is 4m by 7m with RT_{60} is 0.2 sec. Word correct rate is shown with isolated word recognition.

of noise conditions: a) single speaker recorded in a quiet room, b) single speaker recorded with robot's ego noise (dffusive noise), and c) two simultaneous speakers. The robot was a human-size humanoid robot with 32 degrees-of-freedom called HEARBO developed in HRI-JP. An eight channel circular microphone array was embedded around the top of the head. 1 channel data is obtained through the microhpone nearest to the target speaker. The acoustic model for ASR is trained with JNAS clean database. The test data is ATR 216 words. Both corpra are available from NII-SRC.

As shown in Figure 2a), the word correct rate for all algorithms was over 90%, which is almost equivalent to the single channel result without microphone array processing (not shown). This shows that microphone array processing does not degrade in the case of a high signal-to-noise ratio. In more noisy environments, microphone array processing dealt with directional sound sources well, as shown in Figure 2c). It did have difficulty with diffuse noise shown in Figure 2b). To deal with diffuse noise, HARK also provides speech enhancement algorithms [25, 14]. As shown in the figure, the linearly constrained blind separation algorithms had good performance. Null beamforming also showed good performance, but test data in a stationary environment was used. In a dynamic environment, algorithms featuring adaptive step-size control would be better.

4. FUTURE DIRECTIONS ON ROBOT AUDITION

Although a robot audition system has been designed and implemented on the basis of the requirements listed in Section 1.3, its potential application area is much wider than humanoid robots. Several potential application areas are summarized below:

1. Multimodal and transmodal treatments of emotion [17]
   Emotion should be treated uniformly among modalities, e.g., voice, music, gait, and gesture, by the SIRE (strength, intensity, regularity, and extent) model. Emotion represented by the SIRE model can be transfered to other media.
2. Auditory map generation: [18, 34]
   A mobile robot creates a geographical map by using SLAM (simultaneous localization and mapping), a laser range finder, and/or a LiDAR (Light Detection and Ranging) range finder with cameras. The Peacock mobile robot of AIST with a 64-channel microphone array created a map with LiDAR and also localized and traced moving talkers. This kind of sound source localization and separation is very difficult, partially because it should cope with an unknown time-varying number of sound sources and partially because the difference in sound volume is quite large. In the latter case, sounds with lower volume are difficult to localize and separate. Bayesian nonparametric microphone array processing [30] can simultaneously localize and separate such sounds [4].
3. Deployment in extreme environments:
   Robot audition has been applied to a hose-shaped rescue robot [2] with a set of alternately positioned microphones and loudspeakers. Sound signal processing is used to obtain the posture of the robot by estimating the positions of the microphones and then to localize sound sources by using the set of microphones.
   Robot audition has been also applied to unmanned aerial viehicles [29]. A quadrotor was given a microphone array, and iGSVD-MUSIC was used to suppress rotor noise by estimating a noise correlation matrix incrementally.
4. Deployment in natural environments:
   Frog chorusing was recorded on the Gold Coast of Australia and analysed using Bayesian nonparametric microphone array processing [30] to discover phenomena of alternate calling of one species of tree frog [3]. Bird calling in a community natural park was also analyzed using the same method.

5. CONCLUSION

This paper overviewed the research on robot audition. In spite of the long history of audition and signal processing, robot audition research is quite new and has been acknowledged as a keyword in robotics communities. Since the concept of robot audition is universal, it can be deployed in wide areas of research and development including rescue and surveillance robots, animal acoustics as well as human robot interactions. We hope that the signal processing academia embarks on the road to robot audition or at least real-world applications.

6. REFERENCES

[1] Asano F, et al. (1999) Sound source localization and signal separation for office robot "Jijo-2". in *IEEE Intern'l Conf. on Multisensor Fusion and Integration for Intelligent Systems*, 243–248.

[2] Bando Y, et al. (2013) Posture Estimation of Horse-Shaped Robot using Microphone Array Localization. in *IEEE/RSJ IROS-2013*, 3446–3451.

[3] Bando Y, et al. (2015) Recognition of In-field Frog Chorusing using Bayesian Nonparametric Microphone Array Processing. Tech. Report of AAAI-2015 Workshop on Computational Sustainability.

[4] Bando Y, et al. (2015) Baysian Nonparametic Simultneous Localization and Separation of Unknown Time-Varying Number of Sources with Big Volume Difference. in *IEEE ICASSP-2015*, *in print*.

[5] Barker J, et al. (2001) Robust ASR Based on Clean Speech Models: An Evaluation of Missing Data Techniques for Connected Digit Recognition in Noise. in *EuroSpeech-2001*, 213–216.

[6] Barroso V, and Moura JMF (1991) Maximum likelihood beamforming in the presence of outliers. in *ICASSP-1991*, Vol.2, 1409–1312.

[7] Bonnal J, et al. (2010) The EAR Project. *J. of RSJ*, special issue on "robot audition", 28(1):10-13.

[8] Breazeal C, and Scassellati B (1999) A context-dependent attention system for a social robot. in *IJCAI-1999*, 1146–1151.

[9] Breazeal C (2001) Emotive Qualities in Robot Speech. in *IEEE/RSJ IROS-2001*, 1389–1394.

[10] Brooks RA, et al. (1998) Alternative essences of intelligence. in *AAAI-1998*, 961–968.

[11] Frost OL (1972) An algorithm for linearly constrained adaptive array processing. *Proc. of IEEE*, 60(8):926–935.

[12] Griffth LJ and Jim CW (1982) An Alternative Approach to Linearly Constrained Adaptive Beamforming. *IEEE TAP*, 30(1):27–34.

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

[14] Ince G and Nakadai K (2011) Assessment of single-channel ego noise estimation methods. in *IEEE/RSJ IROS-2011*, 106–111.

[15] Kim H-D, et al. (2009) Human Tracking System Integrating Sound and Face Localization using EM Algorithm in Real Environments. *Advanced Robotics*, 23(6):629–653.

[16] Knaak M, et al. (2007) Geometrically Constrained Independent Component Analysis. *IEEE TSAP*, 15(2):715–726.

[17] Lim A and Okuno HG (2014) The MEI Robot: Towards Using Motherese to Develop Multimodal Emotional Intelligence. *IEEE TAMD*, 6(2):126–138.

[18] Martinson E and Brock D (2007) Auditory Perspective Taking, *IEEE Tr. Cybernetics*, 43(3):957–969.

[19] Matsusaka Y, et al. (1999) Multi-person conversation via multi-modal interface — a robot who communicates with multi-user. in *EUROSPEECH-99*, 1723–1726.

[20] Michaud F, et al. (2007) Spartacus attending the 2005 AAAI Conference, *Autonomous Robots*, 22(4):369–383, 2007.

[21] Monzingo RA and Miller TW (1980) *Introduction to Adaptive Arrays*. SciTech Pub., 543p.

[22] Nakadai K, Okuno HG (2000) Active audition for humanoid. in *AAAI-2000*, 832–839.

[23] Nakadai K, et al. (2004) Improvement of recognition of simultaneous speech signals using AV integration and scattering theory for humanoid robots. *Speech Comm.*, 44(4):97–112.

[24] Nakadai K, et al. (2010) Design and Implementation of Robot Audition SYstem "HARK" – Open Source Software for Listening to Three Simultaneous Speakers. *Advanced Robotics*, 24(5-6):739–761.

[25] Nakajima H, et al. (2010) Sound Source Separation and Automatic Speech Recognition. in *IEEE/RSJ IROS-2010*, 976–981.

[26] Nakajima H, et al. (2010) Blind Source Separation With Parameter-Free Adaptive Step-Size Method for Robot Audition. *IEEE TASLP*, 18(6): 1476–1485.

[27] Nakamura K, et al. (2009) Intelligent sound source localization for dynamic environments. in *IEEE/RSJ IROS-2009*, 664–669.

[28] Nishimura R, et al. (2004) Public Speech-Oriented Guidance System with Adult and Child Discrimination Capability. in *IEEE ICASSP-2004*, Vol.I, 433–436.

[29] Ohata T, et al. (2014) Improvement in Outdoor Sound Source Detection Using a Quadrotor-Embedded Microphone Array. in *IEEE/RSJ IROS-2014*, 1902–1907.

[30] Otsuka T, et al. (2014) Bayesian Nonparametrics for Microphone Array Processing. *IEEE/ACM TASLP*, 22(2):493–504.

[31] Parra LC and Alvino CV (2002) Geometric source separation: Margin convolutive source separation with geometric beamforming. *IEEE TSAP*, 10(6):352–362.

[32] Raj H and Sterm RM (2005) Missing-feature approaches in speech recognition. *IEEE Signal Proc. Mag.*, 22(5):101–116.

[33] Rosenthal D and Okuno HG (1998) *Computational Auditory Scene Analysis*, CRC Press, Harshey, NJ.

[34] Sasaki Y, et al. (2013) Nested iGMM recognition and multiple hypothesis tracking of moving sound sources for mobile robot audition. in *IEEE/RSJ IROS-2013*, 3930–3936.

[35] Schmidt RO (1986) Multiple Emitter Location and Signals Parameter Estimation. *IEEE TAP*, AP-34:276–280.

[36] Selzer ML, et al. (2004) A Bayesian Framework for Spectrographic Mask Estimation for Missing Feature Speech Recognition, *Speech Comm.*, 43(4):379-393.

[37] Valin J-M, et al. (2004) Enhanced Robot Audition Based on Microphone Array Source Separation with Post-Filter. in *IEEE/RSJ IROS-2004*, 2123–2128.

[38] Valin J-M, et al.(2007) Robust Recognition of Simultaneous Speech by a Mobile Robot. *IEEE Tr. Robotics*, 23(4):742–752.

[39] Valin J-M, et al. (2007) Robust localization and tracking of simultaneous moving sound sources using beamforming and particle filtering. *Robotics and Autonomous Systems J.*, 55(3):216–228.

[40] Waldherr S, et al. (1998) Template-Based Recoginition of Pose and Motion Gestures On a Mobile Robot. in *AAAI-1998*, 977–982.

[41] Wang D and Brown GJ (2006) *Computational Auditory Scene Analysis: Principles, Algorithms, and Applications*, Wiley-IEEE Press.

[42] Yamamoto S, et al. (2005) Enhanced Robot Speech Recognition Based on Microphone Array Source Separation and Missing Feature Theory. in *IEEE ICRA-2005*, 1477–1482.

[43] Yamamoto S, et al. (2006) Real-time robot audition system that recognizes simultaneous speech in the real world. in *IEEE/RSJ IROS-2006*, 5333–5338.