2017 International Conference on Computational Science and Computational Intelligence

Multi-party Interactions by Quizmaster Robot in Speech-based Jeopardy! like Games

Izaya Nishimuta*†, Katsutoshi Itoyama*, Kazuyoshi Yoshii*
*Graduate School of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo, Kyoto 606-8501, Japan
Email: {nisimuta, itoyama, yoshii}@i.kyoto-u.ac.jp
†currently with Mitsubishi Electric.

Hiroshi G. Okuno
Graduate Program for Embodiment Informatics
Waseda University
2-4-12 Okubo, Shinjuku, Tokyo 169-0072, Japan
okuno@nue.org

***Abstract*—Robot Audition enables a robot to listen to simultaneous utterances via its own ears by localizing and separating sound sources and recognizing separated sounds. Although several applications are developed as a proof of concept, such capabilities of robot audition have not been well shaped in the context of applications, particularly, in multi-parson interactions. This paper focuses on the question "*what is the next step when a robot and/or system can listen to several utterances at once?*" As an example of multi-party interactions, the paper describes a quizmaster robot for speech-based Jeopardy! like games with two interaction models, school-class-type and auction-type. The player of the quiz answers a question by getting the right to answer in the former, while he/she can say an answer directly. Empirical evaluation of the system and lessons are discussed.**

***Keywords*-Robot audition, quizemaster robot, multi-party interaction, auction interaction, school-class interaction**

I. Introduction

Robot Audition, listening capabilities with ears (microphones) of the robot [1], enables a robot to listen to simultaneous utterances via its own ears by localizing and separating sound sources and recognizing separated sounds [2]. Several applications are developed as a proof of concept: Jijo2 robot for office-conversant mobile robot [3], Sparcus robot attending a conference [4], HRP-2 robot for playing an ensemble with human players [5]–[8] and so on.

A *cock-tail party robot* does not exploit the full capability of robot audition, because robot audition provides a capability of listening to simultaneous utterances, "*Three simultaneous utterances*" is used as a benchmark [9] to measure and improve the performance of sound source localization (SSL) [10], sound source separation (SSS) [11], and automatic speech recognition (ASR) of separated speeches [12], [47], [48]. Through these evaluations, robot audition open source software, HARK[1], has been developed [13], [14]. Since HARK provides various kinds of signal processing algorithms under a uniform interface, more than 90 K copies has been downloaded as of Aug. 31, 2017. HARK has been applied to search and rescue activities; for example, a unmanned aerial vehicle (UAV) equipped with a microphone array can localize a sound source from the air [15] and a hose-shaped robot with a microphone array can enhance voiced utterances [16]. HARKBird based on HARK can localize and separate bird songs and estimate the location of bird spots by using three microphone arrays [17]. However, such capabilities of robot audition have not been well shaped in the context of applications, particularly, in multi-parson interactions. This paper focuses on the question "*what is the next step when a robot and/or system can listen to several utterances at once?*"

This paper focuses on a quizmaster robot for speech-based Jeopardy! like games, HATTACK25, [20] with two interaction models: *school-class type* [18] and *auction-type* [19]. In the former, players compete the right to answer by saying "yes" as soon as possible and the fastest one obtains the right to say his/her answer. The quizmaster robot first localizes and separates each utterance, recognizes separated utterances and determines the fastest responding player. In the latter, they compete to say an answer as soon as possible. The robot should do additional task to judge whether an answer is correct or not. Because players say when the robot presenting a question or background music is replayed, that is, bargin utterances occur, the robot cancels its own utterance and background music to enhance players' utterance. The rest of the paper presents the design and implementation of HATTACK25, empirical evaluation and lessons.

II. Related Work and Issues

Human robot interactions have been studied well by means of *alternate initiative*; i.e., all players hear what a robot says and then one of players speaks [27]. Considerable studies have been reported recently in interactions between a single person and a robot [21]–[25]. Some studies recently reported multi-party interactions [26]–[31]. These systems assume that only one player utters at the same time, although some systems identify a speaker by SSL. This type of alternate initiative, or *hear-and-then-speak interaction* has been a convenient way in human robot interactions till robot audition emerged.

[1] http://www.hark.jp/

978-1-5386-2652-8/17 $31.00 © 2017 IEEE
DOI 10.1109/CSCI.2017.313

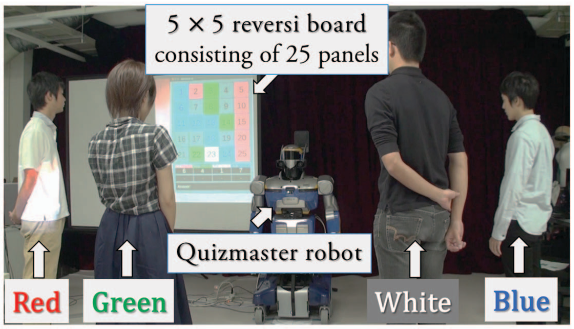

Figure 1. Four people are playing a quiz game, HATTACK25, with a quizmaster robot. Several number-crunching and file servers are running behind the dark curtain.

When a player is allowed to utter at any time in multi-party interactions, several critical issues emerge in signal processing: (1) *overlapped and barge-in utterances*, (2) *rejection of non-voice*, (3) *identification of speaker*, and (4) *recognition of each utterance*. These issues are implicitly avoided in hear-and-then-speak interactions. They are also important when a robot always accepts players' utterances at any time. Without them, a robot may respond to its own uttearance in spoken-dialogue systems, which may cause an infinite loop of interactions, like a *howling* at the spoken dialogue. Matsusaka et al. [27], for example, used a pair of microphone to localize a speaker and a microphone near to each player in two-person robot interactions. Therefore, the issue (3) was achieved whereas (4) is avoided by placing a microphone near the speaker.

The HARK team developed an attendant robot that accepts three simultaneous meal orders on various platforms; SIG-2, Robovie-II [32], [48], HRP-2 [33], ASIMO, and HEARBOT. Nakadai et al. extended the system on HEARBOT that could take eleven-people's simultaneous meal orders. HEARBOT exceeds the listening capabilities of the Price Shotoku (574-622) who, as the Japanese legend says, could listen to and judge ten-people's simultaneous petitions. Another application of HARK is a robot referee for rock-paper-scissors sound games [34]. In this game, three players say one of three words, "rock", "paper", and "scissors" and ASIMO judges who wins the game under the rule that "rock" wins "scissors", "scissors" wins "paper", and "paper" wins "rock". When three players say three different words, the game is drawn. These systems demonstrate *uni-directional* interactions, not *bi-directional*.

For bi-directional interactions, we focus on a multi-party quiz game using only voices. Looije et al. [23] developed a robot that helped children of chronic disease learn about the disease through a quiz game interaction. Oh et al. [21] reported an 'edutainment', i.e., integration of education and entertainment, robot, with which people could learn while they were enjoying playing a quiz game. Fukushima et al. [24] developed a robot that manages a quiz game interaction with a Japanese group and an English group. Matsuyama et al. [28], [29] developed a robot that persuaded group quiz game communication. However, these studies have not coped with the four issues by means of robot audition.

III. HATTACK25 Quizmaster Robot

In this paper, we developed a quizemaster robot that manages a speech-based multiparty fastest-**voice**-first-type quiz game called "HATTACK25" (Fig. 1). HATTACK25 was inspired by a popular long-running Japanese TV program of the fastest-**hand**-first-type quiz show type called 'Panel Quiz Attack 25' (similar to the popular US program 'Jeopardy!').

For investigating the variety of interaction, we model two types of fastest-voice-first-type interaction:

(1) **School-class-type interaction** [18]: Players compete the right to answer by saying "yes" as soon as possible and the fastest responding player obtains the right to say his/her answer. The way of requesting the right of answering is often seen in school classes.

(2) **Auction-type interaction** [19]: Players compete to say an answer as soon as possible even when the quizmaster robot is reading a question. This is similar to an auction.

Original Attack25 and Jeopardy! adopt school-class-type interaction. Players compete the right to answer by saying "yes" as soon as possible and the fastest one obtains the right to say his/her answer. The quizmaster robot first localizes and separates each utterance and recognizes it and determines the fastest responding player. In auction-type interaction, they compete to say an answer as soon as possible. The robot should do additional task to judge whether an answer is correct or not. Because players say when the robot presenting a question or background music is replayed, that is, barge-in utterances occur, the robot cancels its own utterance and background music to enhance players' utterance.

IV. Implementation of HATTACK25

The system structure of HATTACK25 is depicted in Fig. 2. The humanoid robot, HRP-2 [36], has an 8-channel microphone array embedded in its head. An input sound captured by the microphone array is usually a mixture of a robot's utterance, players' utterances, background music replayed during the response timeframe, and background noise. Please note that the curtain separates the robot and several cluster machines. HATTACK25 consists of HARK for SSL and SSS, Julius [35] for ASR, Noise rejection and Game controller.

A. HARK for SSL and SSS

The HARK subsystem localizes and separates sound sources. Fig. 3 shows the HARK network for HATTACK25 quizmaster robot, because HARK provides network-based

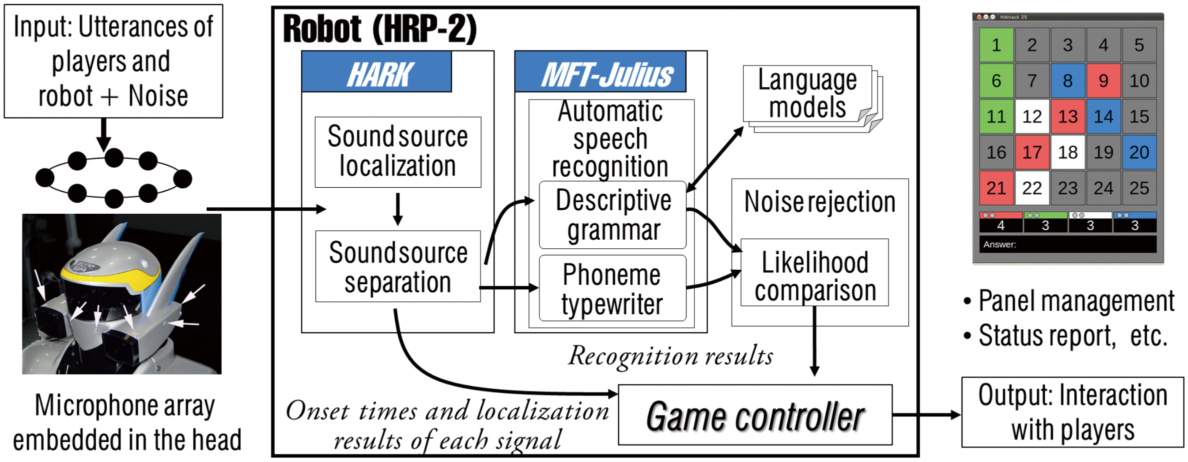

Figure 2. Overview of HATTACK25 quizmaster robot: HRP-2 Humanoid, HARK for signal processing, MFT-Julius for automatic speech recognition, Noise rejection, and Game controller.

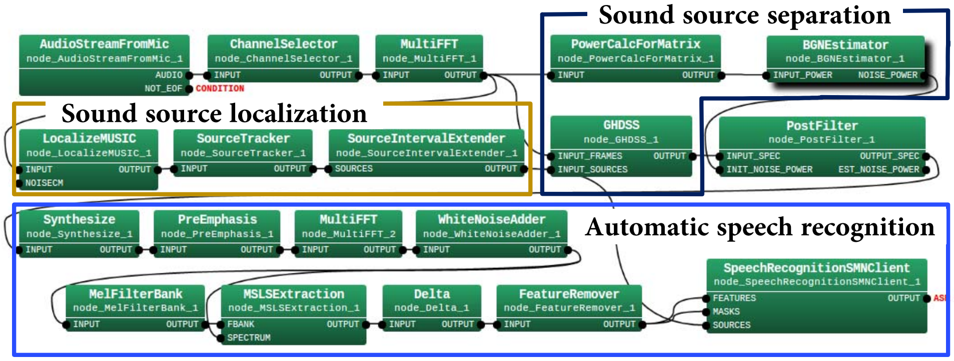

Figure 3. HARK network for HATTACK25: SSL, SSS and Interface to ASR.

programming. It consists of three components: SSL, SSS and interface to ASR.

HARK provides noise-robust SSL algorithms based on MUltiple SIgnal Classification (MUSIC) [41], GEVD-MUSIC and GSVD-MUSIC [42]. Both methods use precalculated noise correlation matrix and noise is reduced by applying either generalized eigenvalue decomposition (GEVD) or generalized singular value decomposition (GSVD). These two algorithms are provided to the user to adopt an appropriate one according to the tradeoff between the performance and speed. 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 of SNR offline, and GSVD-MUSIC can cope with -10 dB in real-time. Thus, HATTACK25 uses GSVD-MUSIC for real-time processing.

Among 11 SSS algorithms provided by HARK [2], [14], we use Geometrically-constrained higher-order decorrelation-based source separation with adaptive step-size control (GHDSS-AS) [43]. GHDSS is a hybrid algorithm between blind separation and beamforming developed by extending geometric source separation (GSS) [44] to improve separation performance and to deal with dynamically changing sound sources. Blind separation in GSS relies merely on cross-power correlation whereas GHDSS-AS uses higher-order correlation similar to independent component analysis (ICA). In addition, GHDSS-AS adopts an adaptive step-size control method to speed up the processing time. Because GHDSS-AS shows good performance and response time with a robot, most of our demonstrations with HARK used GHDSS-AS.

B. MFT-Julius for ASR

Missing feature theory (MFT) [45], [46] 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 [32], [47], [48]. HARK provides the interface to the ASR module, called

"MFT-Julius", an extension of Julius[2]. An acoustic feature vector for MFT-Julius consists of 13-dimensional MSLS and ΔMSLS and Δ power, which are calculated in the HARK module.

Language model switching is adopted to improve the accuracy of speech recognition [37], [38]. Because the basic cycle of the interaction is a repetition of *asking a question*, *choosing an answerer*, *answering a question*, and *choosing a panel* phases, we prepared three types of language model corresponding to the input part of players, that is, *choosing an answerer*, *answering a question*, and *choosing a panel* phases, and then execute speech recognition with the appropriate language model. Thus, for example, an inappropriate number, for example, is rejected in the phase of *choosing a panel* by using the language model specific to the phase.

C. Noise Rejection using Phoneme Typewriter

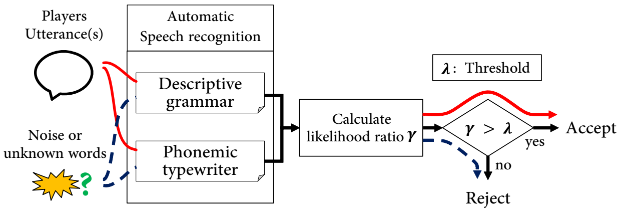

Figure 4. Strategy for Noise Rejection

To determine whether a separated audio signal is an actual utterance or environmental noise, we utilize a noise rejection function by using a phoneme typewriter [39], [40]. In speech recognition, both a standard speech recognition using a descriptive grammar and a speech recognition using phoneme typewriter are performed in parallel.

The input is accepted if the ratio of the likelihood obtained by speech recognition using the descriptive grammar to that obtained by speech recognition using the phoneme typewriter-based grammar is higher than a certain threshold (see Fig. 4). Note that speech recognition using the descriptive grammar for an unknown word or noise obtains a low likelihood and that for an actual utterance included in the grammar obtains a high likelihood. In contrast, speech recognition using the phoneme typewriter obtains the upper limit of likelihood that can obtained by the recognizer for any input.

This function distinguishes actual utterances from unknown words and environmental noise that are not included in the grammar. Owing to this, it is possible to avoid the acceptance of noise and unknown words as actual utterances.

Although HARK provides semi-blind ego-noise suppression [49] by exploiting the knowledge of signals that is replayed at a loudspeaker, HATTACK25 does not use it. Because self-utterance or background music replayed at the loudspeaker behind the robot, HATTACK25 simply filters separated sounds by the directions of players.

D. Game controller

To moderate the game, the robot uses the onset times, the localization results, and the recognition results of each utterance obtained by HARK and Julius. The onset times and directions are used to detect which player spoke first and the recognition results are used to judge the correctness of an answer and to accept the number of the panel chosen by the player.

Direction-based Speaker Identification: To distinguish a player from other players and from the robot, the localization result of each utterance and the stored directions of the players are compared. At the beginning of the game, the players line up from approximately 1.5 m from the robot at at intervals of approximately 40°. Then, each localization result for the reply of a player to the confirmation of the robot is registered as the direction of the corresponding player θ_i $(1 \leq i \leq 4)$.

During playing a game, the player i is identified as the speaker when the difference between the localization result of an utterance and the registered direction θ_i is less than ε. We set $\varepsilon = 15°$ so as not to overlap the allowable range for each player. When the source tracker of HARK detects a sound source using the results of SSL and SSS, HATACK25 uses the start time as the onset time of separated signal. Then, its direction is used to identify the speaker.

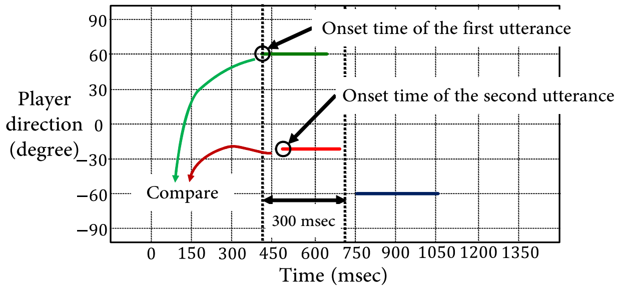

Figure 5. Direction estimation and onset time comparison of two simultaneous utterances using HARK.

To find the fastest responding player who has a right to answer, HATTACK25 quizmaster robot performs sound source localization. As shown in Fig. 5, the onset time of a separated audio stream is defined as its first frame (circled in the figure). HARK can detect the fastest utterance saying "Yes" even if multiple utterances are made almost simultaneously, i.e., the onset difference is 100 ms. The onset times of multiple utterances within 300 msec are compared and the robot gives a priority to each speaker. That is, if a player makes a wrong answer, the right to answer is moved to the next player.

Answer Manipulation: In school-class type interaction of HATTACK25, the robot first determines the fastest player

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

and gives the right to answer to him/her. Then, the robot judges whether an answer that the player with the right to answer says is correct or not. Needless to say, the robot also checks whether the right player says a right answer. If the answer is correct, proceed to the *choosing a panel* phase. Otherwise, the right to answer is given to the second fastest player.

In auction-class type interaction, the robot's task is simpler. The robot incrementally judges whether an answer is correct or not. If correct, it says "The fastest answer, RED said, is correct." Otherwise, it says "The fastest answer, RED said, is wrong. The next fastest answer, Green said, is correct."

Panel management: In choosing a panel, a valid number, i.e., between 1 and 25, is filtered by ASR with a particular language model. Then, the robot checks whether the number fulfils the constraints of the board. Finally, the board is updated and a new question starts.

V. Evaluation and Observation

This section summarizes the results of evaluation obtained by the experiment that one to four louds speakers replay answers at random [20]. Success rate indicates a simple ratio of correct identification over total utterances.

The fastest speaker identification: In school-class type interaction, the success rate is approximately 90.0% or more when the time difference between the first and second fastest players is 60.0 ms. With a time difference of more than 60.0 ms, the success rate under background music is almost the same as without it.

In auction type interaction, the success rate exceeds 90.0% when the time difference is more than 100 ms. The success rate remains the same under background music. Note that each player says a different word, whereas the same in school-class type.

Speech recognition of the fastest speaker: In school-class type interaction, the success rate is almost 100.0%, whereas in auction type interaction, the success rate is approximately 70.0% and degrades to 60.0% under background music. Note that the success rate depends on questions and simultaneous utterances.

Observation: A similar experiments were conducted by five male and one female graduate students as a human quizmaster. The results show that the average success rate of the fastest responding player identification is similar to that of the robot. Under background music or noisy situation, the robot outperforms humans. On the contrary, human subjects outperform the robot in *speech recognition* under any condition.

The experience of HATTACK25 demonstrates that SSL and SSS are effective to localize and separate player's utterances and MTF-Julius works well to recognize separated utterances. HARK is proved as a good toolbox for speech-related applications. Functions not utilized in HATTACK25 are semi-blind ego-noise suppression and interface to DNN-HMM. Recently, HARK support the interface to Kaldi, DNN-HMM, as well as HARK-SaaS [14]. The next step is to develop cloud-based applications of HARK without installing HARK and ASR locally.

VI. Conclusion

This paper overviewed applications of HARK to multi-party interactions, and presented two possible interaction modes, school-class type and auction type, in HATTACK25. The HATTACK25 quizmaster robot provides a many-to-one interaction. Our previous applications of the function of "*listening to several things at once*" provided by HARK demostrates passive interactions. The critical issue in applying the function resides in the lack of methodology for many-to-many interactions. If the robot has several heads like "*eight-headed serpent* in Japanese mythology," each head can interact with each correspondent. This scheme is considered as a *collective behavior of one-to-one interactions*. Along this scheme, a robotic discussion facilitator may be a next candidate of HARK applications. Since the concept of robot audition is universal, we hope that HARK will be deployed to real-world applications.

> — *Harvest success and gain experience in special application areas first, pursuing research to make* robot audition *more general purpose.* —

Acknowledgment

The authors would like to thank to Mr. Naoki Hirayama for his first design and implementation of HATTACK25 and Dr. Kazuhiro Nakadai, Honda Research Institute Japan, for his comments on the usage of HARK. This research is partially supported by JSPS Kakenhi Kiban (S) No.2400006, ImPACT Tough Robotics Challenge, and Graduate Program for Embodiment Informatics, Waseda University.

References

[1] K. Nakadai, et al., "Active audition for humanoid," *AAAI-2000*, 832–839.

[2] H.G. Okuno, and K.Nakadai, "Robot Audition: Its Rise and Perspectives," *IEEE ICASSP-2015*, 5610–5614.

[3] H. Asoh, et al., "Socially embedded learning of the office-conversant mobile robot Jijo-2," *IJCAI 1997*, 880–885.

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

[5] A. Lim, et al., "Robot Musical Accompaniment: Integrating Audio and Visual Cues for Real-time Synchronization with a Human Flutist," *IEEE/RSJ IROS-2010*, 1964–1969.

[6] T. Itohara, et al., Improvement of Audio-Visual Score Following in Robot Ensemble with Human Guitarist, *IEEE Humanoids-2012*, 574–579.

[7] T. Mizumoto, et al., "Who is the leader in a multiperson ensemble? —Multiperson human-robot ensemble model with leaderness—," *IEEE/RSJ IROS-2012*, 1413–1419.

[8] J.L. Oliveira, et al., "Beat Tracking for Interactive Dancing Robots," *Int. J. Humanoid Robotics*, **12** (2015) 24p.

[9] H.G. Okuno, et al., "Understanding Three Simultaneous Speeches," *IJCAI-1997*, 30–35.

[10] C. Rascon, et al., Localization of Sound Sources in Robotics: A Review," *Robotics and Autonomous Systems*, *in print*, 2017.

[11] K. Nakadai, K. Nakamura, "SOUND SOURCE LOCALIZATION AND SEPARATION," Wiley Encyclopedia of Electrical and Electronics Engineering, June 2015.

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

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

[14] K. Nakadai, et al., "Development, Deployment and Applications of Robot Audition Open Source Software HARK," *Journal of Robotics and Mechatronics*, **27**(1):16–25, 2017.

[15] K. Hoshiba, et al., "Design of UAV-embedded Microphone Array System for Sound Source Localization in Outdoor Environments," *Sensors*, **17**(11):2535, 2017.

[16] Y. Bando, et al., "Low-Latency and High-Quality Two-Stage Human-Voice-Enhancement System for a Hose-Shaped Rescue Robot," *Journal of Robotics and Mechatronics*, **27**(1):198–212, 2017.

[17] R. Suzuki, et al., "A spatiotemporal analysis of acoustic interactions between great reed warblers (*Acrocephalus arundinaceus*) using microphone arrays and robot audition software HARK," *Ecology and Evolution*, *accepted*, 2017.

[18] I. Nishimuta, "A robot quizmaster that can localize, separate, and recognize simultaneous utterances for a fastest-voice-first quiz game," *IEEE Humanoids 2014*, 967–972.

[19] I. Nishimuta, et al., "Development of a robot quizmaster with auditory functions for speech-based multiparty interaction," *IEEE/SICE SII 2014*, 328–333.

[20] I. Nishimuta, et al., "Toward a quizmaster robot for speech-based multiparty interaction," *Advanced Robotics*, **29**(18):1205–1219, 2015.

[21] H. Oh, et al., "A case study of edutainment robot: Applying voice question answering to intelligent robot," *IEEE ROMAN 2007*, 410–415.

[22] N. Schmitz, et al. "Realization of natural interaction dialogs in public environments using the humanoid robot roman," *IEEE Humanoids 2008*, 579–584.

[23] R. Looije, et al., "Help, I need some body the effect of embodiment on playful learning," *IEEE ROMAN 2012*, 718–724.

[24] K. Fukushima, et al., "Question strategy and interculturality in human-robot interaction," *HRI 2013*, 125–126.

[25] M. Tielman, et al., "Adaptive emotional expression in robot-child interaction," *HRI 2014*, 407–414.

[26] Y. MatsusakY, et al., "Multi-person conversation via multimodal interface — a robot who communicates with multi-user," *EUROSPEECH-1999*, 1723–1726.

[27] Y. Matsusaka, et al., "Conversation robot participating in group conversation (*in Japanese*)," *IEICE Trans. Inf. and Sys.*, **E86-D**(1):26–36, 2003.

[28] Y. Matsuyama, et al., "Designing communication activation system in group communication," *Humanoids 2008*, 629–634.

[29] Y. Matsuyama, et al., "Framework of communication activation robot participating in multiparty conversation," *AAAI Fall Sym. 2010*, 68–73.

[30] D. Klotz, et al.,n "Engagement-based multi-party dialog with a humanoid robot," *SIGDIAL 2011*.

[31] D B. Jayagopi, and J-M. Odobez J-M, "Given that, should I respond? contextual addressee estimation in multi-party human-robot interactions," *ACM/IEEE HRI 2013*, 147–148.

[32] J-M. Valin, et al., "Robust Recognition of Simultaneous Speech By a Mobile Robot," *IEEE Tran. on Robotics*, **23**(4):742–752, 2007.

[33] H.G. Okuno, et al., "Robot Audition: Missing Feature Theory Approach and Active Audition," *Robotic Research*, STAR 70, 227–244, 2011.

[34] K. Nakadai, et al., "A robot referee for rock-paper-scissors sound games," *IEEE ICRA-2008*, 3469–3474.

[35] A. Lee, and T. Kawahara, "Recent development of open-source speech recognition engine Julius," *APSIPA-ASC 2009*, 131–137.

[36] K. Kaneko, et al., "Humanoid robot HRP-2," *IEEE ICRA-2004*, 1083–1090.

[37] L.R. Lane, et al., "Language model switching based on topic detection for dialog speech recognition," *ICASSP 2003*, 616–619.

[38] M. Santos-Pérez, et al., "Topic-dependent language model switching for embedded automatic speech recognition," *Ambient Intelligence – Software and Applications*, 235–242, 2012.

[39] K. Kita, et al., "Processing unknown words in continuous speech recognition," *IEICE Trans. Fund.* **E74-A**(7):1811–16.

[40] T. Jitsuhiro, et al., "Rejection of out-of-vocabulary words using phoneme confidence likelihood," *ICASSP 1998*, 217–220.

[41] R.O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE TAP*, **34**(3):276–280, 1986.

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

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

[44] L.C. Parra, and C.V. Alvino, "Geometric source separation: Margin convolutive source separation with geometric beamforming," *IEEE TSAP*, 10(6):352–362, 2002.

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

[46] H. Raj, and R.M. Sterm, "Missing-feature approaches in speech recognition," *Sig. Proc. Mag.*, **22**(5):101–116, 2005.

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

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

[49] R. Takeda, et al., "Efficient Blind Dereverberation and Echo Cancellation based on Independent Component Analysis for Actual Acoustic Signals," *Neural Computation*, **24**:1, 234–272, 2012.