Proceedings of the 2014 IEEE/SICE International Symposium on System Integration, Chuo University, Tokyo, Japan, December 13-15, 2014

Development of a Robot Quizmaster with Auditory Functions for Speech-based Multiparty Interaction

Izaya Nishimuta, Kazuyoshi Yoshii, Katsutoshi Itoyama, and Hiroshi G. Okuno

***Abstract*— This paper presents a robot quizmaster that has auditory functions (i.e., ears) for moderating a multiplayer quiz game. The most basic form of oral interaction in a quiz game is that a quizmaster reads aloud a question, and each player is allowed to answer it whenever the answer comes to his or her mind. A critical problem in such oral interaction is that if multiple players speak almost simultaneously for answering, it is difficult for a "human" quizmaster to recognize overlapping answers and judge the correctness of each answer. To avoid this problem, players have conventionally been required to push a button, raise a hand, or say "Yes" to just get a right to answer a question before doing it. This requirement, however, inhibits natural oral interaction. In this paper we propose a "robot" quizmaster that can identify a player who correctly answers a question first, even when multiple players utter answers almost at the same time. Since our robot uses its own microphones (ears) embedded in the head, individual players are not required to wear small pin microphones close to their mouths. To localize, separate, and recognize overlapping utterances captured by the ears, we use a robot audition software called HARK and an automatic speech recognizer called Julius. Experimental results showed the effectiveness of our approach.**

I. Introduction

Partner robots that live together and interact with humans in a real daily environment should have not only vision (*i.e.*, eyes) but also audition (*i.e.*, ears) for flexibly and effectively gathering environmental information. Since humans are considered to obtain 90% of environmental information from eyes, real-time image processing techniques have intensively been studied in a sub-area of computer vision called *robot vision* [1]. Inspired by the concept of computational auditory scene analysis (CASA) [2], on the other hand, the field of *robot audition* was established in 2000 [3]. Environmental information obtained from ears is vital in many daily situations in which eyes cannot be used, *e.g.*, when robots are in a dark room or when a target to follow is hidden by other objects (called occlusion). In this paper we focus on speech-based interaction between robots and humans.

Robots that use their voices for interacting with humans have been developed for various purposes. Asoh *et al.* [4], for example, proposed a mobile robot that can gather environmental information through dialogue with humans in an office environment. Several robots were intended to interact with children for the purpose of education [5], [6] or edutainment (education + entertainment) [7]. Tielman *et al.* [8] proposed a robot that adaptively expresses various emotions by using its voice and gestures. Schmitz *et al.* [9] developed a humanoid robot called ROMAN that is able to track and communicate with a human partner using verbal and non-verbal features. Nakano *et al.* [10] proposed a two-layer model of behavior and dialogue planning for conversational service robots engaging in multi-domain guidance.

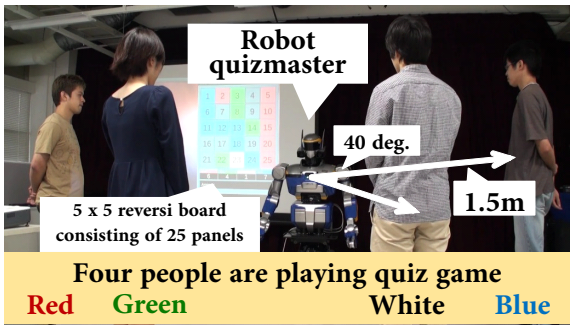

Fig. 1. A snapshot of our speech-based multiplayer quiz game moderated by a robot quizmaster having auditory functions. Four players compete to get as many panels as possible on a reversi board by correctly answering questions. The robot is capable of identifying a player who utters a correct answer first. The players are allowed to directly utter answers (*barge-in* utterances) without any sings even when the robot is reading questions.

A main problem of conventional robots based on standard speech recognition and spoken dialogue systems is that input audio signals captured by microphones are assumed to be always clean isolated speech signals. In a real environment, however, multiple people often make utterances simultaneously and the utterances of a robot are often overlapped by the utterances of users (called *barge-in*). To avoid these situations, we are generally controlled to speak in turn to small microphones unnaturally close to our mouths [11], although we want to speak directly to a facing robot whenever we want to do so. This inhibits natural interaction with arbitrary multiple people who do not wear microphones. In addition, the input audio signals are still far from clean speech signals. The key feature of robot audition research, on the other hand, is that the robot is assumed to *always* hear mixed sounds that may contain multiple utterances made by humans and the robot, their reflections, background music, and environmental noise through its own microphones (*i.e.*, ears).

In this paper we present an interactive robot quizmaster that can manage a speech-based multiplayer quiz game using its own auditory functions (Fig. 1). This is an important first step to develop an ultimate partner robot having human-like

This work was supported by JSPS KAKENHI 24220006

I. Nishimuta, K. Yoshii and K. Itoyama is with Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto, Japan {nisimuta, yoshii, itoyama}@kuis.kyoto-u.ac.jp

H. G. Okuno is with Graduate Program for Embodiment Informatics, Waseda University, Shinjuku, Tokyo, Japan okuno@aoni.waseda.jp

978-1-4799-6943-2/14/$31.00 ©2014 IEEE

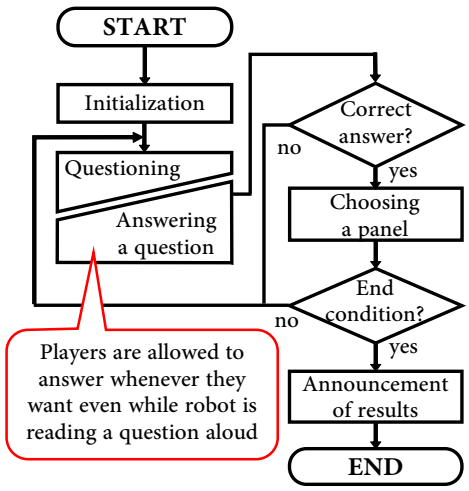

Fig. 2. The flow chart of our speech-based multiplayer quiz game.

intelligence because we humans sometimes enjoy playing quiz games or riddles using only our own voices in our daily lives as a casual way of multiparty interaction. Note that the quiz game we discuss here is different from TV-program-type quiz games that need special devices (*e.g.*, buttons) for identifying a player having the right to answer. Our speech-based quiz game allows players to directly answer a question by speaking whenever the answer comes up to their minds. To localize, separate, and recognize overlapping answers captured by the ears, we jointly use a robot audition software called HARK [12] and an automatic speech recognizer called Julius [13]. A main contribution of our study is to integrate human-robot interaction techniques into the framework of robot audition.

II. Multiparty Interaction in Quiz Game

The quiz game is one of the most interesting forms of multiparty interaction and the robot quizmaster is a good application of speech-based interaction techniques [6], [7], [8], [14], [15], [16], [17], [18], [19]. Required tasks of a quizmaster are 1) managing the progress of a quiz game and 2) livening up the players and spectators. As to task 1), for example, Fukushima *et al.* [16], showed that a robot could join quiz interaction with Japanese and English groups. Matsuyama *et al.* [14], [15] tackled task 2) and showed that a robot could promote the communication in a quiz game. In this paper we focus on task 1) and propose a robot quizmaster that can control the progress of a quiz game as humans do. To achieve this, the robot should be able to interact with multiple players through speech media. For example, the robot should be able to read a question aloud while waiting for answers uttered by players. In addition, the robot should be able to judge the correctness of each answer and identify a player who uttered a correct answer first. Such speech-based interaction plays an important role in various daily situations including quiz games.

In this section we specify a "fastest-voice-first-type" multiplayer quiz game. We then discuss the requirements for the robot quizmaster in terms of robot audition and present a brief overview of our approach.

A. Specification of the Speech-based Quiz Game

Our speech-based quiz game is typically played by four players competing for 25 panels of the reversi board (Fig. 1) by answering questions. The player who gets the most panels wins the game. As shown in Fig. 2, the basic flow of the game is 1) questioning by the quizmaster, 2) answering by a player, 3) judgment of the answer by the quizmaster, and 4) panel selection by the player. This speech-based interaction is repeated until all panels are taken by players.

Due to the purely speech-based nature of the quiz game, we pose the following rules.

1) All questions are readable for the quizmaster. Unreadable questions using images and audio signals are not given to players.
2) The players are allowed to directly utter answers without any advance notice (*e.g.*, pushing buttons, raising hands, or saying "Yes") whenever they want to answer. Special devices such as buttons are not used.
3) When multiple players utter correct or wrong answers almost at the same time, a player who utters a correct answer first gets a right to select a panel.
4) The players are allowed to answer even if the robot is still reading questions aloud. This type of interruptive utterances is referred to as barge-in.

The robot needs to register the direction of each player at the beginning of the game. We assume that the players do not change their directions until the game has finished. In addition, background music is played back during thinking time until some players utter answers.

B. Auditory Functions of the Quizmaster

There are two main functions that are required for enabling the robot to manage the quiz game through spoken dialogue:

1) Speaker identification for each utterance
2) Speech recognition for each utterance

To target a player who is speaking and avoid mistaking the utterances of irrelevant players and those of the robot for the target player's utterance, the robot needs to always distinguish players and itself. Since the microphones are always active and away from players' mouths, the input to the robot is affected by reflections and surrounding noise. Therefore, it is necessary that the automatic speech recognition (ASR) used be robust against such noise.

C. Technical Challenges in a Real Environment

While typical spoken dialogue systems are based on "hear-and-then-speak" communication, a key feature of our robot quizmaster is that microphones are always active and can accept input at any time. Such an all-time-input situation poses interesting issues in multiparty human-robot interaction. In the questioning phase, for example, the robot should accept a player's response exactly even if the robot is still reading a question, and in the answering phase, the robot should reject the utterance of a player who made a wrong answer

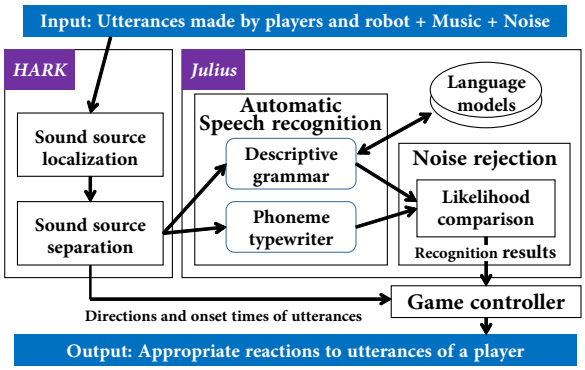

Fig. 3. The internal architecture of our robot quizmaster.

even if that player spoke before a player who made a correct answer. In the judgment phase, we need to tackle the issue of *self-utterance howling*. If the robot wrongly accepts its own utterance as a player's utterance, the response utterance of the robot is wrongly accepted again. To prevent such howling effect, the robot should reject its own utterance.

The discussion above leads to two technical requirements for the auditory functions of a robot that can interact with multiple people through speech media:

- Sound source localization: The robot should be able to identify which player has made an utterance so as to determine which player to interact with.
- Sound source separation: The robot should be able to distinguish the utterances of individual players from its own questionary utterance, self-generating motor noise, and background music for speech recognition.

D. Our Approach based on Robot Audition Techniques

In the speech-based quiz game, robot audition functions such as sound source localization and separation form the basis of multiplayer interaction. Robots should be able to estimate the directions of multiple sound sources and separating a mixture of sounds into those sources [3]. Those two functions have also been demonstrated as useful for human-to-human interaction in the context of telepresence communication [20] and have also been applied to interactive robot dancing [21].

The use of a versatile open-source robot audition software called HARK (http://www.hark.jp) [12] is a key to developing the robot quizmaster working in a real noisy environment. A player to interact with is determined by localizing players who are speaking. In the questioning phase, the player who has spoken first can be identified by separating the recorded mixture signals into multiple source signals (*i.e.*, almost simultaneous answers made by players and questionary utterances of the robot).

III. The Robot Quizmaster

This section describes implementation of our robot quizmaster with a focus on the main functions listed in Section II-B. Our robot is a humanoid called HRP-2 [22] with an 8-channel microphone array embedded in the head, a loudspeaker to generate synthesized speech of the robot, and a large screen to show the reversi board consisting of 5×5 panels. Multiple players who are speaking simultaneously can be identified in real time by using techniques of sound source localization and separation. Robust automatic speech recognition is achieved by switching language models [23] and using a noise rejection method [24].

First, we present the configuration of the robot from both the hardware and software point of view and then we discuss how we implement the intelligent functions.

A. Overview

The internal architecture of the robot is shown in Fig. 3. When one or multiple players speak for answering a question or choosing a panel, the mixture of audio signals that might include players' and the robot's own utterances are captured by the microphone array. Individual sound sources are then localized and separated using HARK. This network consists of sound source localization and separation and automatic speech recognition (bridge to Julius).

Instead of just using an automatic speech recognizer called Julius (http://julius.sourceforge.jp/) [13] with a single general language model, we prepare multiple language models and switch those models. We also use a noise rejection method based on a phoneme typewriter to improve the recognition performance.

The direction and onset time of each utterance obtained by HARK and the recognition result obtained by Julius are used for managing the game, *i.e.*, determining the priority order of the players to answer a question, to judge the correctness of an answer, and to accept a panel chosen by the player. The robot then changes panels on the reversi board according to the player's request and outputs synthetic speech from the loudspeaker to explain the current game status.

B. Requirements and Solutions

We implement the two main functions of the robot quizmaster (*i.e.*, speaker identification and speech recognition) described in Section II-B by using three techniques.

1) Direction-based Speaker Identification: The players and the robot can be identified by comparing their registered directions with the estimated directions of the utterances.

- **Initialization**: At the beginning of the game, the players line up in an arc at intervals of 40° (Fig. 1). Then, each player is asked to reply to the confirmation of the robot. The localization results for the replies are registered as the directions of the players θ_i $(1 \leq i \leq 4)$.
- **Identification**: If the difference between a registered direction θ_i and the estimated direction of an utterance is less than ε, the i'th player is identified as the speaker. We set $\varepsilon = 15°$ so as not to overlap the allowable range for each players.

Standing at the back of the robot is not recommended because the type of interaction is a quiz game. However, players are allowed to stand at the back of the robot except the in front of the loudspeaker for the robot if it is suitable for the situation of the interaction.

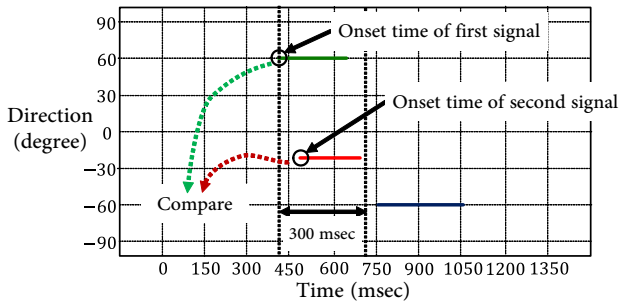

Fig. 4. Direction estimation of two simultaneous utterances using HARK.

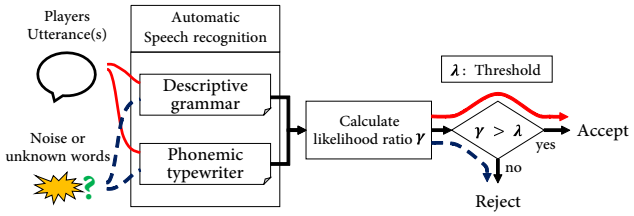

Fig. 5. Likelihood-comparison-based noise rejection.

To find the fastest-voice player who has a right to answer, the robot performs sound source localization. As shown in Fig. 4, the onset time of a separated audio stream is defined as its first frame (circled in the figure). HARK can detect the fastest utterance even if multiple utterances are made almost simultaneously. The onset times of multiple utterances within 300 msec are compared and the robot gives a priority to each speaker (if a player makes a wrong answer, the right to answer is moved to the next player).

2) Language Model Switching: To improve the accuracy of speech recognition, we switch multiple language models according to the progress of the quiz game. Since the user-input part consists of *answering a question* and *choosing a panel* (Fig. 2), we prepare the corresponding specialized models. Since the utterances required for each situation are different, only a suitable language model is activated.

3) Phoneme-Typewriter-based Noise Rejection: To determine whether a segregated audio stream is an actual utterance or noise, we use both a phoneme typewriter and a standard speech recognizer with a descriptive grammar. The phoneme typewriter is a special kind of speech recognizers that directly converts an input audio signal into a phoneme sequence (no word-level constraints used).

As shown in Fig. 5, an input audio stream is rejected as irrelevant if the likelihood ratio of the descriptive-grammar-based speech recognizer to the phoneme typewriter is lower than a certain threshold. Note that the likelihood obtained by the the phoneme typewriter is unaffected by whether an uttered word is defined in the descriptive grammar. The likelihood obtained by the descriptive-grammar-based speech recognizer, on the other hand, is small if the uttered word is not defined in the grammar. This technique reduces the influence of surrounding noise and unknown words that are not included in the grammar, thus making it possible to improve the accuracy of speech recognition.

Robot: "Next question"
Robot: "What is the capital of Brazil?"
System: Switch to "answering a question" model.
Red: "Rio de Janeiro!"
Green: "Brasilia!"
Blue: "Brasilia!"
(*Three players answered almost simultaneously*)
System: Determine the order of the utterances (answerers) from their onset times.
Robot: "The answer of the fastest Red was wrong"
Robot: "The answer of the second-fastest Green was correct"
Robot: "Green, which panel do you want to select?"
System: Switch to "choosing a panel" model.
Green: "16."
System: Change the colors of panels 16 and 12 to green.
Robot: "16 and 12 turned green"

Fig. 6. An example of multiplayer interaction in the quiz game.

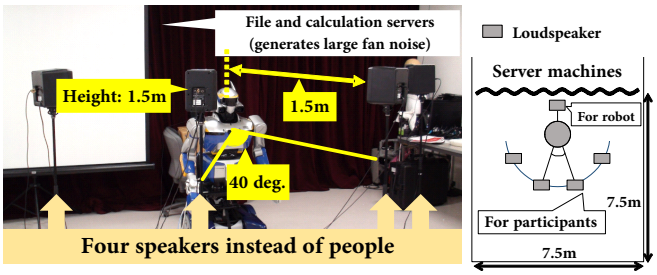

Fig. 7. Experimental conditions.

C. An Example of Interaction in Quiz Game

Figure 6 shows an example of interaction between the robot and players. **Robot** indicates an utterance of the robot quizmaster, **Red**, **Green**, and **Blue** indicate those of players, and *System* shows an internal process of the system. In this example, the robot asked a question and the three players answered almost simultaneously. The player who spoke first made an incorrect answer. The second fastest speaker who made a correct answer thus got a right to select a panel. A demo video will be uploaded in our website.[1]

IV. Evaluation

We conducted several experiments to evaluate the success rates of identifying the fastest speaker and recognizing his or her utterance in different conditions.

A. Experimental Conditions

We prepared 30 questions including multiple-choice questions and recorded the corresponding correct answers uttered by four players (three males and a female in their twenties). As shown in Fig. 7, four loudspeakers used for playing back

[1]http://winnie.kuis.kyoto-u.ac.jp/members/nishimuta/sii2014/

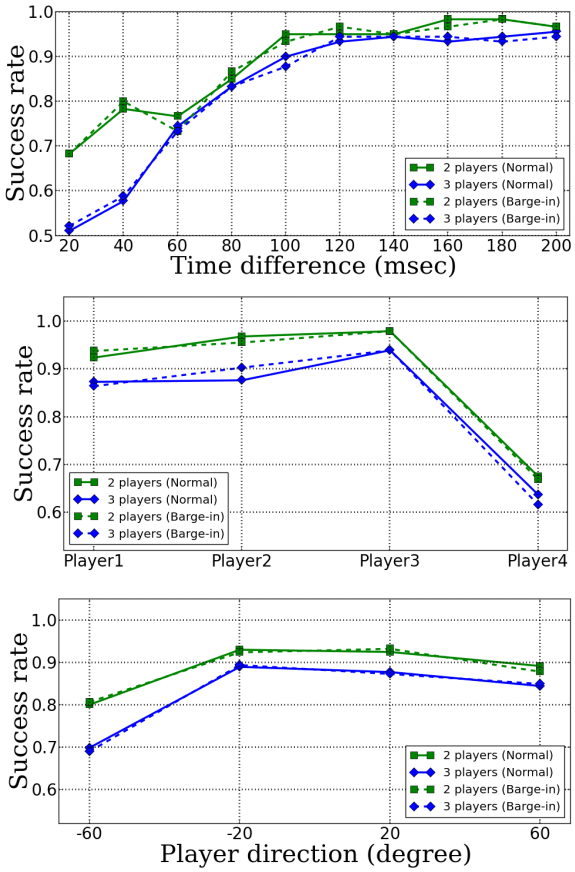

Fig. 8. The average success rates of fastest-speaker identification.

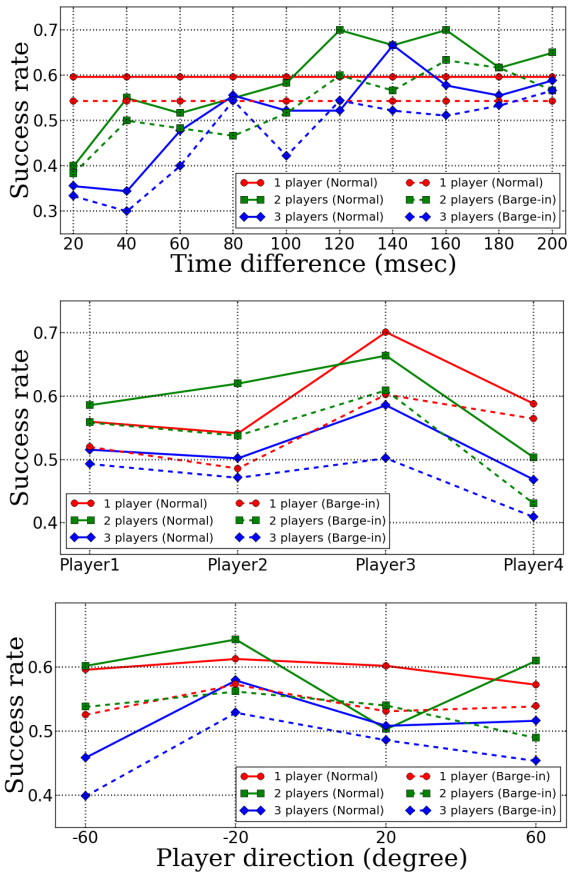

Fig. 9. The average success rates of speech recognition.

the recorded answers (assumed as utterances of players) were located along a 120° arc in front of the robot at 40° intervals and 1.5m away (social distance[25]) from the microphone array in the robot head. Another loudspeaker was used for playing back synthesized speech of the robot and background music during thinking time. Each loudspeaker was set up at a height of 1.5m that was almost same as the height of human mouth. The experimental room was 7.5m square and filled with large fan noise generated from server machines.

We evaluated the success rate of fastest-speaker identification and that of speech recognition for the fastest speaker in various conditions. The number of players who uttered answers almost at the same time was set to from one to three. When multiple (two or three) players uttered answers, only one player preceded the other player(s) by a small time difference that was set to from 20 to 200 msec in 20 msec increments. The position of players and a loudspeaker (direction) playing back the fastest answer was chosen at random.

The success rate of fastest-speaker identification, R_{fp}, and that of speech recognition for the fastest speaker, R_{sr}, were calculated as follows:

$$R_{fp} = \frac{M_{fs}}{N_{all}}, \quad R_{sr} = \frac{M_{sr}}{N_{all}}, \tag{1}$$

where N_{all} is the total number of utterances, M_{fs} is the number of utterances that were correctly identified as the fastest ones, and M_{sr} is the number of the fastest utterances that were correctly recognized.

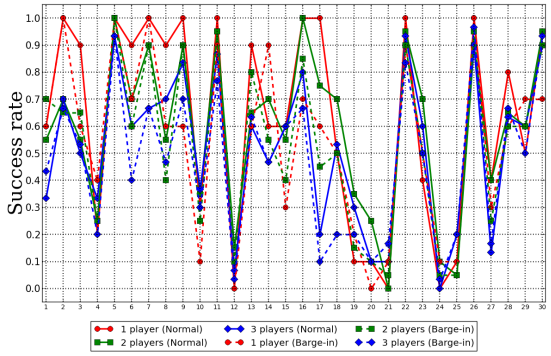

Fig. 10. The success rates of speech recognition for individual questions.

In this experiment, we used descriptive language models each of which was specialized for recognizing the answer of each question and an acoustic model trained by using separated speech signals. To evaluate the robustness of the robot to irrelevant sounds other than player utterances, we tested two conditions as follows:

1) *Normal condition*
   The recorded answers were played back while the robot was silent (SNR 10.0 dB).
2) *Barge-in condition*
   The recorded answers were played back while back-

gound music was continuously played back from the loudspeaker (SNR 0.0 dB).

B. Experimental Results

Fig. 8 shows the experimental results of fastest-speaker identification. The top, middle, and bottom figures indicate the success rates with respect to time differences, players, and player directions, respectively. The success rates under the barge-in condition remained almost the same as those under the normal condition. The robot achieved the success rate of 90% when the time difference was more than 100 msec, and the success rates were scarcely affected by player directions. An interesting fact was that the robot often failed to identify player 4 (female). This was considered to be attributed to the male-to-female ratio. In order to confirm our conjecture, we will conduct additional detailed experiments by changing the male-to-female ratio.

Fig. 9 shows the experimental results of speech recognition. The top, middle, and bottom figures indicate the success rates with respect to time differences, players, and player directions, respectively. The success rates under the barge-in condition were degraded from those under the normal condition. Nonetheless, the utterances of the fastest speakers were recognized with almost the same success rates regardless of the number of simultaneous answers and the existence of background music when the time difference was more than 120 msec. In contrast to fastest-speaker identification, the success rates were significantly degraded when more than two players spoke simultaneously. As shown in Fig. 10, the recognition difficulty varies according to how to answer questions. For example, questions 17-21 asked the players to choose one of the twelve months (*e.g.*, Q: "When the new term begins?" and A: "April"). Since the names of months are acoustically similar to each other in Japanese (*e.g.*, April: Shigatsu, February: Nigatsu), it was difficult to distinguish those names in a real noisy environment. This problem should be tackled in the future.

V. Conclusion

This paper presented a robot quizmaster having auditory functions for multiplayer interaction in a speech-based quiz game. A robot audition software called HARK was used to identify the directions of utterances made by players (sound source localization) in a noise-robust manner. The robot can determine the order of almost simultaneous utterances by estimating the onset times of those utterances. The utterance of each player is then extracted from noise-contaminated mixture signals captured by the robot's own microphones (sound source separation). To improve the accuracy of speech recognition in a real noisy environment, we used two techniques of language model switching and phoneme-typewriter-based noise rejection. Experimental results showed that our robot quizmaster is capable of identifying a player who speaks first with a success rate of more than 90.0% in a noisy environment even under a barge-in condition.

Future work includes conducting a psycho-acoustic experiment to acquire new knowledge about multiparty human-robot interaction from the perceptual and cognitive point of view. In addition, we plan to implement further interactions using sound source localization and separation and speech recognition for livening up the players and spectators of the quiz game as a human quizmaster does.

References

[1] N. Kyriakoulis *et al.*, "Color-Based Monocular Visuoinertial 3-D Pose Estimation of a Volant Robot," *Instrumentation and Measurement, IEEE Transactions on*, vol. 59, no. 10, pp. 2706–2715, 2010.

[2] A. S. Bregman, *Auditory Scene Analysis: The perceptual organization of sound*. MIT press, 1994.

[3] K. Nakadai, *et al.*, "Active audition for humanoid," in *Proc. of AAAI*, 2000, pp. 832–839.

[4] H. Asoh, *et al.*, "Socially embedded learning of the office-conversant mobile robot Jijo-2," in *Proc. of IJCAI*, vol. 1, 1997, pp. 880–885.

[5] E. Hsiao-Kuang Wu, *et al.*, "A context aware interactive robot educational platform," in *Proc. of IEEE-DIGITEL*, 2008, pp. 205–206.

[6] R. Looije, *et al.*, "Help, I need some body the effect of embodiment on playful learning," in *Proc. of IEEE-RO-MAN*, 2012, pp. 718–724.

[7] H.-J. Oh, *et al.*, "A case study of edutainment robot: Applying voice question answering to intelligent robot," in *Proc. of IEEE-RO-MAN*, 2007, pp. 410–415.

[8] M. Tielman, *et al.*, "Adaptive emotional expression in robot-child interaction," in *Proc. of IEEE-HRI*, 2014, pp. 407–414.

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

[10] M. Nakano, *et al.*, "A two-layer model for behavior and dialogue planning in conversational service robots," in *Proc. of IEEE-IROS*, 2005, pp. 1542–1547.

[11] Y. Matsusaka, *et al.*, "Conversation robot participating in group conversation," *IEICE TRANSACTIONS on Information and Systems*, vol. E86-D, no. 1, pp. 26–36, 2003.

[12] K. Nakadai *et al.*, "Design and implementation of robot audition system 'HARK' –open source software for listening to three simultaneous speakers," *Advanced Robotics*, vol. 24, no. 5-6, pp. 739–761, 2010.

[13] A. Lee *et al.*, "Recent development of open-source speech recognition engine Julius," in *Proc. of APSIPA-ASC*, 2009, pp. 131–137.

[14] Y. Matsuyama, *et al.*, "Designing communication activation system in group communication," in *Proc. of IEEE-HUMANOIDS*, 2008, pp. 629–634.

[15] Matsuyama, Yoichi and Taniyama, Hikaru and Fujie, Shinya and Kobayashi, Tetsunori, "Framework of communication activation robot participating in multiparty conversation," in *AAAI Fall Symposia*, 2010, pp. 68–73.

[16] M. Fukushima, *et al.*, "Question strategy and interculturality in human-robot interaction," in *Proc. of IEEE-HRI*, 2013, pp. 125–126.

[17] D. B. Jayagopi, *et al.*, "The vernissage corpus: A conversational human-robot-interaction dataset," in *Proc. of IEEE-HRI*, 2013, pp. 149–150.

[18] D. B. Jayapogi *et al.*, "Given that, should I respond? contextual addressee estimation in multi-party human-robot interactions," in *Proc. of IEEE-HRI*, 2013, pp. 147–148.

[19] D. Klotz, *et al.*, "Engagement-based multi-party dialog with a humanoid robot," in *Proc. of the SIGDIAL 2011: the 12th Annual Meeting of the SIGDIAL*, 2011, pp. 341–343.

[20] T. Mizumoto, *et al.*, "Design and implementation of selectable sound separation on the texai telepresence system using HARK," in *Proc. of IEEE-ICRA*, 2011, pp. 2130–2137.

[21] J. L. Oliveira, *et al.*, "An active audition framework for auditory-driven HRI: Application to interactive robot dancing," in *Proc. of IEEE-RO-MAN*, 2012, pp. 1078–1085.

[22] K. Kaneko, *et al.*, "Humanoid robot HRP-2," in *Proc. of IEEE-ICRA*, vol. 2, 2004, pp. 1083–1090.

[23] M. Santos-Pérez, *et al.*, "Topic-dependent language model switching for embedded automatic speech recognition," in *Ambient Intelligence - Software and Applications*, 2012, vol. 153, pp. 235–242.

[24] T. Jitsuhiro, *et al.*, "Rejection of out-of-vocabulary words using phoneme confidence likelihood," in *Proc. of IEEE-ICASSP*, vol. 1, 1998, pp. 217–220.

[25] E. T. Hall, *The hidden dimension*. Doubleday, 1966.