2014 14th IEEE-RAS International Conference on
Humanoid Robots (Humanoids)
November 18-20, 2014. Madrid, Spain

A Robot Quizmaster that can Localize, Separate, and Recognize Simultaneous Utterances for a Fastest-Voice-First Quiz Game

Izaya Nishimuta[1], Naoki Hirayama[1], Kazuyoshi Yoshii[1], Katsutoshi Itoyama[1], and Hiroshi G. Okuno[2]

***Abstract*— This paper presents an interactive humanoid robot that can moderate a multi-player fastest-voice-first-type quiz game by leveraging state-of-the-art robot audition techniques such as sound source localization and separation and speech recognition. In this game, a player who says "Yes" first gets a right to answer a question, and players are allowed to barge in a questionary utterance of the quizmaster. The robot needs to identify which player says "Yes" first, even if multiple players respond at almost exactly the same time, and must judge the correctness of the answer given by the player. To enable natural human-robot interaction, we believe that the robot should use its own microphones (i.e., ears) embedded in the head, rather than having pin microphones attached to individual players. In this paper we use a robot audition system called HARK for separating the mixture of audio signals recorded by the ears into multiple source signals (i.e., almost the simultaneous utterances of "Yes" and the questionary utterance) and estimating the direction of each source. To judge the correctness of an answer, we use a speech recognizer called Julius. Experimental results showed that our robot can correctly identify which player spoke first when the players' utterances differed by 60 msec.**

I. Introduction

Robots that can interact with multiple people via speech media have actively been developed for performing various tasks. Asoh *et al.* [1], 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 [2], [3] or "edutainment" (education + entertainment) [4]. To make children familiar with and a robot, Tielman *et al.* [5] proposed a robot that adaptively expresses various emotions by using its voice and gestures. Schmitz *et al.* [6] developed a humanoid robot called ROMAN that is able to track and communicate with a human interaction partner using verbal and non-verbal features. Nakano *et al.* [7] developed a two-layer model for the behavior and dialogue planning module of conversational service robots that can engage in multi-domain conversation.

A major limitation of conventional spoken dialogue systems is that, although we want to speak directly to a facing robot, we are required to speak to microphones unnaturally close to our mouths [8]. This requirement, however, is not satisfied in real environments in which multiple people tend to make utterances simultaneously and the utterances of a robot are often overlapped by the utterances of users (called *barge-in*). It is therefore natural to assume that the robot hears a mixture of sounds that may include human and self-generating utterances with their reflections and environmental sounds through its own microphones (*i.e.*, ears).

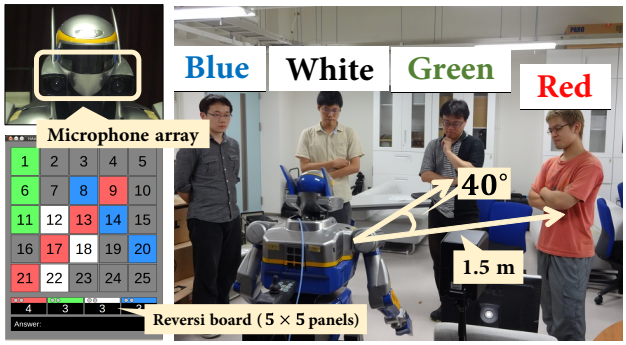

Fig. 1. A robot quizmaster (HRP-2) and four players in a fastest-voice-first quiz game called HATTACK25.

The quiz game is one of the most interesting forms of multi-party interaction and the robot quizmaster is an excellent research topic for developing speech-based interaction techniques [3]–[5], [9]–[14]. 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.* [11] showed that a robot could join quiz interaction with groups of Japanese and English people. Matsuyama *et al.* [9], [10] tackled task 2) and showed that a robot could promote the communication in a quiz game. We focused on task 1) and developed a robot quizmaster that can control the progress of a quiz game as humans do. To achieve this, a quizmaster should interact with multiple players through speech media. For example, the quizmaster reads a question aloud while waiting for responses from the players. In the answering phase, the player who reacts (*e.g.*, pushes a button or says "Yes") first is prompted by the quizmaster to answer the question. In the judgment phase, the quizmaster judges the correctness of the answer. Such speech-based interaction plays an important role in entertainment applications, including quiz games and contains the key elements of conversation in our daily lives.

In order to realize such multi-player speech-based interaction in a real quiz-game environment, robot audition functions such as sound source localization and separation [15] are indispensable. Robots should be able to estimate the directions of multiple sound sources and separating a mixture of sounds into those sources. Those two functions have been demonstrated as useful for human-to-human interaction in the context of telepresence communication [16] and have also been applied to interactive robot dancing [17].

[1]Graduate School of Informatics, Kyoto University, Yoshida-honmachi, Sakyo-ku, Kyoto 606-8501, Japan {nisimuta, hirayama, yoshii, itoyama}@kuis.kyoto-u.ac.jp
[2]Graduate Program for Embodiment Informatics, Waseda University, Shinjuku, Tokyo 169-0072, Japan okuno@aoni.waseda.jp

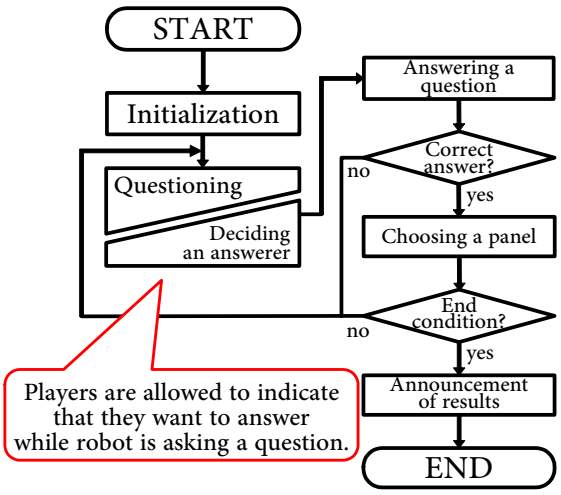

Fig. 2. The flow chart of HATTACK25.

In this paper we present an interactive robot quizmaster that can manage a speech-based fastest-voice-first quiz game called HATTACK25[1]. The use of a versatile robot audition software called HARK[2] [18] is a key to developing the robot quizmaster working in a real noisy environment. The player who is currently interacting with the robot is determined by using the localization result of players' utterances. 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 utterances of "Yes" by players and questionary utterances of the robot). A main contribution of our study is to integrate human-robot interaction techniques based on automatic speech recognition into the framework of robot audition.

II. The Robot Quizmaster

This section describes a multi-player fastest-voice-first quiz game called HATTACK25 and a proposed robot quizmaster. We then discuss the requirements for the robot quizmaster in terms of robot audition functions.

The robot quizmaster must be able to differentiate the players to determine who it will give the right to speak (*i.e.*, the right to answer the quiz, to hear only the utterance of the desired person). This identification is crucial in multi-player quiz games. If the robot fails to identify the correct speaker, the quiz game would quickly fall apart.

A. Specification of the Quiz Game "HATTACK25"

HATTACK25 is a speech-based quiz game played by four players competing for 25 panels of the reversi board (Fig. 1) by answering questions. The player who gets the most panels win 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 have been selected.

Due to the fact that HATTACK25 is a sound-based quiz game, it has the following specifications.

1) The questions are readable for the quizmaster. Visual and musical questions are not used.
2) The players say "Hai" ("Yes" in Japanese) to indicate that they want to answer. Devices such as push buttons are not used.
3) When more than one person says "Yes" and the fastest player answers incorrectly, the right to answer moves on to the second-fastest player.
4) The players are allowed to say "Yes" whenever they want to answer, even if the robot is still reading the question. This type of interruptive utterance is referred to as a barge-in.

The robot is given the information of the players' directions for identification purpose at the beginning of the game. We assume that the players do not change their directions until the game has finished.

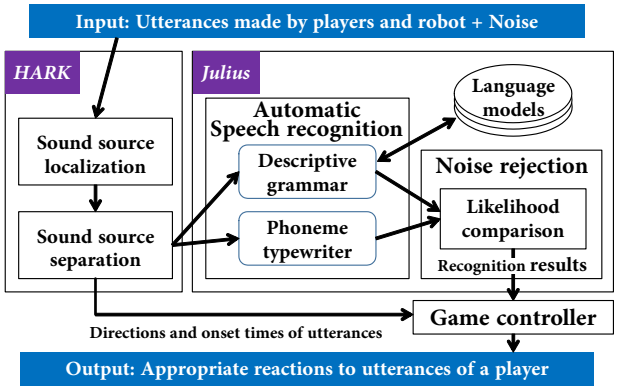

Fig. 3. Internal architecture of HATTACK25 robot quizmaster.

B. Main Functions Required for the Quizmaster

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

1) Identification of the speaker of each utterance
2) Recognition of the players' utterances

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' mouth, the input to the robot is affected by reflections and surrounding noise (such as sneezes, coughs and fan noise of air conditioner). Therefore, it is necessary that the automatic speech recognition (ASR) used be robust against such noise.

C. Challenges in Multi-Party Quiz Game

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 multi-party human-robot interaction. In

[1]HATTACK25 is a purely voice-based version of a popular Japanese TV program called Panel Quiz ATTACK25 (similar to a popular TV program in US called *Jeopardy!*). http://asahi.co.jp/attack25/index.html

[2]HARK: http://www.hark.jp

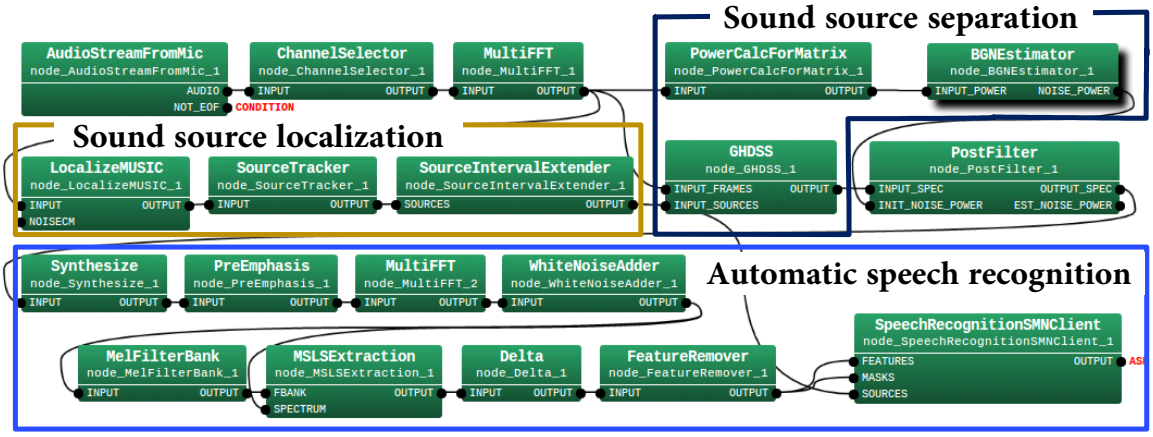

Fig. 4. Visual programming interface of HARK: Processing modules are roughly categorized into three parts.

the questioning phase, for example, the robot should accept a player's signal 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 does not have a right to answer if that player speaks before a player who does have the right. 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, this response utterance is wrongly accepted in turn. 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 players' utterances from its own questionary utterance and self-generating motor noise and determine which player has the right to answer on the basis of simultaneous signs.

III. System Architecture

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 [19] 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 [20] and using a noise rejection method [21].

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 indicating that they want to answer, answering a question, and choosing a panel, the mixture of audio signals that might include players' and the robot's own utterances are captured by the microphone array and then localized and separated using HARK. The network representation of input-output relationships between various modules in HARK is shown in Fig. 4. 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[3] [22] 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.

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

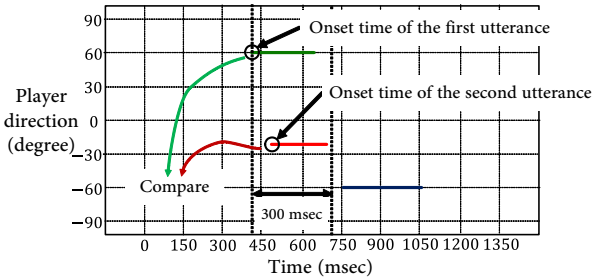

Fig. 5. Direction estimation and onset time comparison of two simultaneous utterances using HARK.

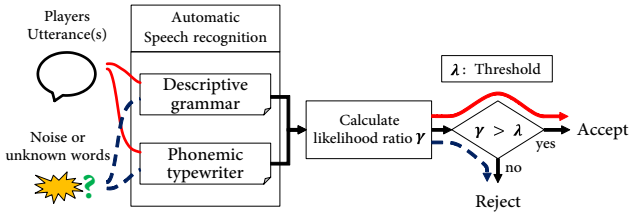

Fig. 6. Likelihood-comparison-based noise rejection.

To find the fastest-voice player who has a right to answer, the robot performs sound source localization. As shown in Fig. 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. 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. Since the user-input part of HATTACK25 consists of *deciding an answerer*, *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 that gets the highest likelihood (no word-level constraints used).

As shown in Fig. 6, 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: "Which sport is played by the most smallest number of players: soccer, baseball, or basketball?"
System: Switch to "deciding an answerer" model.
Red, Green: "Yes!", "Yes!"
System: Select the player who said "Yes" first.
Robot: "Red!"
System: Switch to "answering a question " model.
Red: "Baseball!"
Robot: "Wrong."
Robot: "Red cannot answer the next two questions." (penalty)
System: Select the player who said "yes" second.
Robot: "The second-fastest is Green!"
Green: "Basketball!"
Robot: "That's right! Basketball is correct."
System: Switch to "choosing a panel" model.
Robot: "Which panel do you want to select?"
Green: "16."
Robot: "16 and 12 turn green."
System: Change panels 16 and 12 to green.

Fig. 7. Quiz-game interaction between the robot and players.

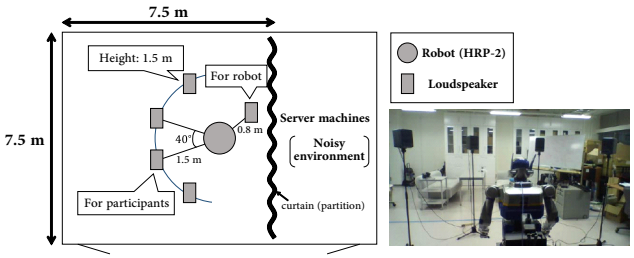

Fig. 8. Experimental setting.

C. An Example of Interaction in Quiz Game

Figure 7 shows an actual example of interaction between the robot and four players. In the figure, **Robot** shows the robot quizmaster, **Red**, **Green** indicate the two players, and *System* refer to the processing of the system. The robot asked a question and two players said "Yes" almost simultaneously, indicating that they both wanted to answer the question. The player who said "Yes" first answered the question but gave an incorrect response. The second-fastest player then answered with a correct response, chose a panel, and the robot announced which panels had changed color. A demo video is uploaded in our website.[4]

IV. Evaluation

To realize interaction shown in Section III-C, it is important to accurately detect the fastest speaker from simultaneous utterances with a slight time lag. We therefore evaluated the accuracy of the speaker identification method.

A. Experimental Conditions

As shown in Fig. 8, we constructed an experimental environment using loudspeakers instead of people for repeated evaluations under various conditions. The loudspeakers were

[4]http://winnie.kuis.kyoto-u.ac.jp/members/nishimuta/humanoids2014/

TABLE I

UTTERANCE CONDITIONS.

Number of simultaneous speakers (players)	Two	Three	Four
Number of loudspeakers to use	two of four (6 ways)	three of four (4 ways)	all of four (1 way)
Time difference between utterances	20-200 msec (10 ways)	20-200 msec (10 ways)	20-200 msec (10 ways)
Loudspeakers to be given a delay	either (2 ways)	two of three (3 ways)	three of four (4 ways)
Number of trials under each condition	5	5	15
Total	$(6 * 10 * 2) * 5 = 600$	$(4 * 10 * 3) * 5 = 600$	$(1 * 10 * 4) * 15 = 600$

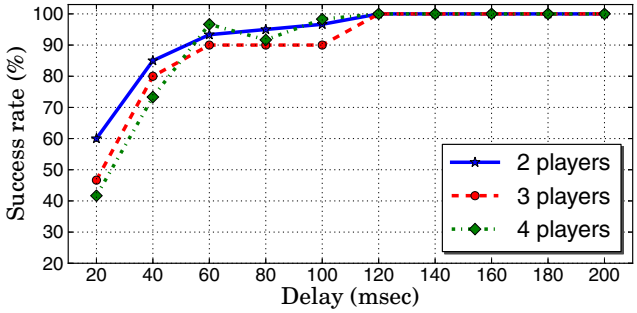

Fig. 9. Success rate for each delay (normal conditions).

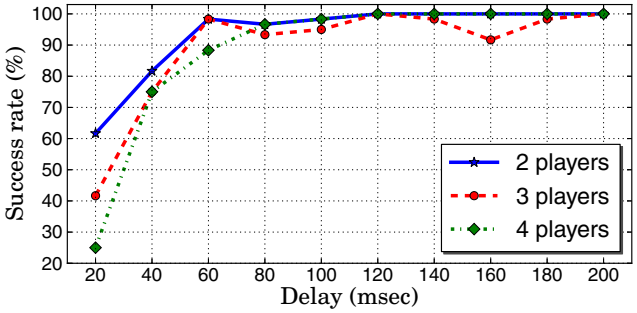

Fig. 10. Success rate for each delay (barge-in conditions).

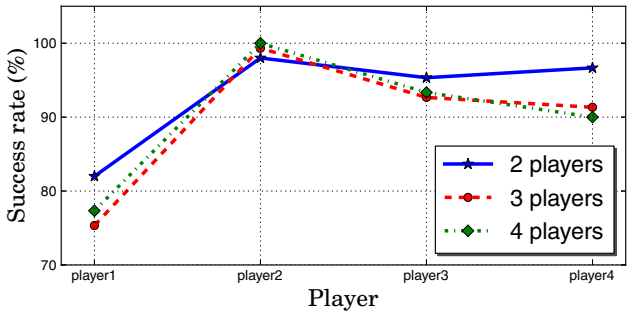

Fig. 11. Success rate for each player (normal conditions).

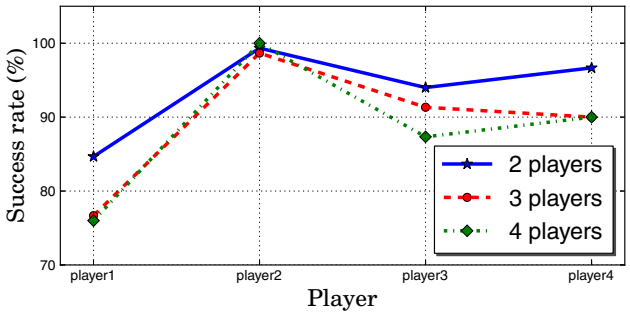

Fig. 12. Success rate for each player (barge-in conditions).

located along a 120° arc in front of the robot to mimic the 120° binocular field of view of the human eye. We placed loudspeakers at 40° intervals. Each loudspeaker was located 1.5 m away from the microphone array in the robot head according the theory on interpersonal distance [23] which states that the relation of the quizmaster and the players in multi-party interaction corresponds to the social distance. Each loudspeaker was set up at a height of 1.5 m from the ground, to mimic the height of a human mouth. The room was filled with large fan noise generated from calculation and file servers. The reverberation time (RT60) of this room is 470 msec. Prior to the experiment, we recorded "Yes" of each player (male, 20–29 years old).

We tested our method of fastest-voice speaker identification described in Section III-B.1 by changing a condition on the number of simultaneous utterances, as shown in Table I. The utterances were almost simultaneously played back from at least two loudspeakers under an assumption that only one utterance slightly preceded the other utterances. The onset difference (delay) ranged from 20–200 msec in 20-msec increments. To evaluate the robustness of our method to overlapped utterances made by the robot, we tested two conditions on players' utterances:

- *Normal condition*: Some players made utterances (SNR 10.0 [dB]) while the robot was silent.
- *Barge-in condition*: The utterances were made (SNR 0.0 [dB]) when the robot generated sounds continuously. An audio signal assumed as a question spoken by the robot was played back from another loudspeaker.

The speaker identification was performed for a total of 600 trials under all possible conditions.

The success rate (accuracy) of identifying the fastest-voice player, R, was calculated as follows:

$$R = \frac{N_{success}}{N_{all}}, \tag{1}$$

where $N_{success}$ is the number of successful identifications and N_{all} the total number of utterances.

B. Experimental Results

Figures 9 and 10 show the experimental results under the normal and barge-in conditions. As shown in Fig. 9, the robot could identify the direction of the fastest-voice player with a success rate of about 90.0% when the delay was 60 msec. The robot achieved perfect identification when the delay was larger than 120 msec under the normal condition. Figure 10 shows that the success rate at a delay of 20 msec under the barge-in condition was lower than that under the

normal condition, while with a delay of more than 60 msec, the success rate under the barge-in condition was almost the same as that under the normal condition. These results showed the robustness of our method to the robot's own utterances and surrounding noise.

Figures 11 and 12 show the success rate for each direction (player position) under the normal and barge-in conditions, respectively. The success rate of player 1 was lower than that of player 2 under both conditions. This may be attributed to the different acoustic characteristics of the players' voices; While the voice of player 1 was calm and drawling, the voice of player 2 was clear and sharp. The sharp one was often identified as the fastest by mistake when sharp and drawling voices were uttered almost simultaneously.

These results show that the robot could identify a player who has a right to answer a question with sufficient accuracy under a realistic assumption that multiple utterances made by players differ by more than 60 msec, and demonstrate that the robot has sufficient auditory ability to act as a quizmaster. We will perform the same kind of evaluation with a human quizmaster and compare the results to clarify any difference in ability of detecting when and where the person speaks. It is also important to investigate behavioral difference between the robot quizmaster a human quizmaster for developing a robot having a good ability of interaction.

V. Conclusion

This paper presented an interactive robot quizmaster based on auditory functions for a fastest-voice-first-type quiz game called HATTACK25. The results of sound source localization and sound source separation obtained by the robot audition system called HARK are used to identify the directions of utterances made by players. The robot can determine a player who speaks first and has a right to answer from audio signals including simultaneous utterances by estimating the onset times of those utterances. To accurately recognize a player's answer using an automatic speech recognizer called Julius, 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 which player says "Yes" first with a success rate of more than 90.0% in a noisy environment even under a barge-in condition when the delay was 60 msec.

Future work includes conducting a psycho-acoustic experiment to acquire new knowledge about multi-party 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.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI 24220006.

References

[1] H. Asoh, S. Hayamizu, I. Hara, Y. Motomura, S. Akaho, and T. Matsui, "Socially embedded learning of the office-conversant mobile robot Jijo-2," in *Proceedings of IJCAI-97*, vol. 1. AAAI, 1997, pp. 880–885.

[2] E. Hsiao-Kuang Wu, H. Chi-Yu Wu, Y.-K. Chiang, Y.-C. Hsieh, J.-C. Chiu, and K.-R. Peng, "A context aware interactive robot educational platform," in *Proceedings of IEEE-DIGITEL 2008*, 2008, pp. 205–206.

[3] R. Looije, A. van der Zalm, M. A. Neerincx, and R.-J. Beun, "Help, I need some body the effect of embodiment on playful learning," in *Proceedings of IEEE-ROMAN-12*, 2012, pp. 718–724.

[4] H.-J. Oh, C.-H. Lee, Y.-G. Hwang, M.-G. Jang, J. G. Park, and Y. K. Lee, "A case study of edutainment robot: Applying voice question answering to intelligent robot," in *Proceedings of IEEE-ROMAN-07*, 2007, pp. 410–415.

[5] M. Tielman, M. Neerincx, J.-J. Meyer, and R. Looije, "Adaptive emotional expression in robot-child interaction," in *Proceedings of IEEE-HRI-14*, 2014, pp. 407–414.

[6] N. Schmitz, J. Hirth, and K. Berns, "Realization of natural interaction dialogs in public environments using the humanoid robot roman," in *Proceedings of IEEE-HUMANOIDS-08*, 2008, pp. 579–584.

[7] M. Nakano, Y. Hasegawa, K. Nakadai, T. Nakamura, J. Takeuchi, T. Torii, and H. Tsujino, "A two-layer model for behavior and dialogue planning in conversational service robots," in *Proceedings of IEEE-IROS-05*, 2005, pp. 1542–1547.

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

[9] Y. Matsuyama, H. Taniyama, S. Fujie, and T. Kobayashi, "Designing communication activation system in group communication," in *Proceedings of IEEE-HUMANOIDS-08*, 2008, pp. 629–634.

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

[11] M. Fukushima, R. Fujita, M. Kurihara, T. Suzuki, K. Yamazaki, A. Yamazaki, K. Ikeda, Y. Kuno, Y. Kobayashi, T. Ohyama, and E. Yoshida, "Question strategy and interculturality in human-robot interaction," in *Proceedings of IEEE-HRI-13*, 2013, pp. 125–126.

[12] D. B. Jayagopi, S. Sheiki, D. Klotz, J. Wienke, J.-M. Odobez, S. Wrede, V. Khalidov, L. Nyugen, B. Wrede, and D. Gatica-Perez, "The vernissage corpus: A conversational human-robot-interaction dataset," in *Proceedings of IEEE-HRI-13*, 2013, pp. 149–150.

[13] D. B. Jayapogi and J.-M. Odobez, "Given that, should I respond? contextual addressee estimation in multi-party human-robot interactions," in *Proceedings of IEEE-HRI-13*, 2013, pp. 147–148.

[14] D. Klotz, J. Wienke, J. Peltason, B. Wrede, S. Wrede, V. Khalidov, and J.-M. Odobez, "Engagement-based multi-party dialog with a humanoid robot," in *Proceedings of the SIGDIAL-11: the 12th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2011, pp. 341–343.

[15] K. Nakadai, T. Lourens, H. G. Okuno, and H. Kitano, "Active audition for humanoid," in *Proceedings of AAAI-00*, 2000, pp. 832–839.

[16] T. Mizumoto, K. Nakadai, T. Yoshida, R. Takeda, T. Otsuka, T. Takahashi, and H. G. Okuno, "Design and implementation of selectable sound separation on the texai telepresence system using HARK," in *Proceedings of IEEE-ICRA-11*, 2011, pp. 2130–2137.

[17] J. L. Oliveira, G. Ince, K. Nakamura, K. Nakadai, H. G. Okuno, L. P. Reis, and F. Gouyon, "An active audition framework for auditory-driven HRI: Application to interactive robot dancing," in *Proceedings of IEEE-ROMAN-12*, 2012, pp. 1078–1085.

[18] K. Nakadai and T. Takahashi, "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.

[19] K. Kaneko, F. Kanehiro, S. Kajita, H. Hirukawa, T. Kawasaki, M. Hirata, K. Akachi, and T. Isozumi, "Humanoid robot HRP-2," in *Proceedings of IEEE-ICRA-04*, vol. 2, 2004, pp. 1083–1090.

[20] M. Santos-Pérez, E. González-Parada, and J. Cano-García, "Topic-dependent language model switching for embedded automatic speech recognition," in *Ambient Intelligence - Software and Applications*, 2012, vol. 153, pp. 235–242.

[21] T. Jitsuhiro, S. Takahashi, and K. Aikawa, "Rejection of out-of-vocabulary words using phoneme confidence likelihood," in *Proceedings of IEEE-ICASSP-98*, vol. 1, 1998, pp. 217–220.

[22] A. Lee and T. Kawahara, "Recent development of open-source speech recognition engine Julius," in *Proceedings of APSIPA-ASC-09*, 2009, pp. 131–137.

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