Special edition paper

Talking Humanoid Robot Verification in Tokyo Station (Technical Verification of a Multilingual Talking Humanoid Robot)

Manabu Sugasawa* Sei Sakairi**

* Researcher, Frontier Service Development Laboratory, Research and Development Center of JR East Group
** Chief Researcher, Frontier Service Development Laboratory, Research and Development Center of JR East Group

Abstract

We conducted technical verification of a Hitachi-made humanoid robot that recognizes English, Chinese and Japanese at the JR East Travel Service Center of Tokyo station. We verified the performance of switching language according to speech by the other party and performance of noise reduction. Also, we confirmed some technical problems in realizing smooth conversation, especially processing time. As a result, the success ratio of language switching was about 60%, dialog failure due to noise reduction failure was about 1%, and interval time required between the user speaking and the robot replying was about 3 seconds.

●***Keywords*:** Voice interaction technology, Artificial intelligence, Humanoid robot, Multilingual dialog, Question answering

1. Objective

Research is underway on customer guidance by robots and artificial intelligence (AI) as a measure to handle changes in the business environment with the declining birthrate and aging population and to handle the sudden increase in visitors from abroad. A decision has already been made at JR East to introduce a call center business support system that uses AI, but stations differ from call centers in that there is much noise. We thus focused on voice recognition technology and conducted tests using a robot under development by Hitachi, Ltd.

One objective of the tests was to confirm performance at current technology levels for language switching with a view to multilingual guidance and for noise reduction important in station environments with much noise. And another was to identify issues that need to be overcome in order to achieve smooth conversation between human and robot.

1.1 Confirmation of Performance of Language Switching Based on Speech

Usability differs greatly according to the method used for switching the language used for the dialog. This time, we confirmed the practicality of a mechanism to switch language between English, Chinese, and Japanese according to the user's speech instead of selecting the language by touch panel, thereby reducing an action needed.

1.2 Confirmation of Noise Reduction Performance

There is much noise in stations that can interfere with dialogs, such as train noise, announcements, and crowd noise. Technology to extract only the user's voice and reduce other noise from sound picked up by a microphone is imperative, so we confirmed the performance of that.

1.3 Confirmation of Issues in Achieving Smooth Dialog

Processing time from the user asking a question until the robot answers is very important when considering practicality. We thus analyzed dialog logs to identify to what degree smooth interaction is possible in a station environment.

2. Test Overview

2.1 Equipment Used and Test Location

In the tests, we used the EMIEW3 humanoid robot under development by Hitachi for customer service and guidance in public spaces and the like. With that, it was relatively easy to prepare for confirmation of performance of language switching based on speech, and the robot has high noise reduction performance. We also placed a human assistant near the robot to help out if comprehensible dialog was not possible. We selected the JR East Travel Service Center in Tokyo Station where guidance is given to visitors from abroad as the location for tests.

2.2 Question and Response List

To identify questions frequently asked by visitors from abroad, we held interviews with the Travel Service Center staff and GranSta commercial facility guidance staff. From that, we produced a list of 216 types of questions and answers to those in the three languages of English, Chinese, and Japanese and programed them to the EMIEW3. The majority were one-to-one question and answers regarding directions. However, “directions to the airport”, “directions to the Shinkansen platform”, and the like required further question to be answered, so the robot would actually ask the user whether they meant Narita or Haneda airport or where they intended to get off the Shinkansen.

2.3 Screen Display

In order to further facilitate conveyance of information, screens to display images and videos were installed on the front and back of the robot. Initially, detailed maps of Tokyo Station and the surrounding area were provided for answers to questions regarding directions, and the route to the destination was shown, but a simple design (Fig. 2) was subsequently adopted due to concerns that those maps could actually be confusing. Additionally, the robot displayed an introductory screen (Fig. 3) to express the scope in which it could answer questions.

2.4 Improvements to Language Switching and Flow

For language switching in the tests, we designed so that the language was switched according to certain phrases when the users spoke them. In order to make the dialog go as smooth as possible, phrases that would be used when speaking to human staff were used, even though the other party of the dialog was a robot. For example, if the users said, “Can I ask you something” in English to indicate desire to use English, the language would switch to English. The same would apply for a similar phrase in Chinese (“Wǒ kěyǐ wèn nǐ jǐgè wèntí ma?”) and Japanese (“Shitsumon ga arimasu”) respectively. Guidance on the language switch flow was provided on the screen to facilitate understanding by users, but two weeks into the tests we found that many users looked at the robot without reading the screen and started asking questions without understanding the language switch phase. Also, we made English the default language for dialogs as we found that about 80% of foreign users selected English. We changed the dialog flow to switch language only when where was a request to change to Chinese or Japanese. As with the language switch flow, specific examples of questions the robot handled were shown on an introductory screen, but we found that many users would ask questions exactly as shown on the screen. We were interested in what sort of questions users would ask to the robot, so we changed to show just the types of questions that could be answered and deleted examples of questions.

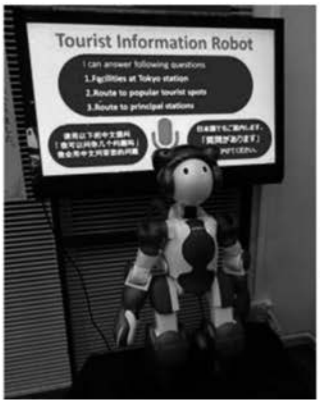

Tourist Information Robot
I can answer following questions
1.Facilities at Tokyo station
2.Route to popular tourist spots
3.Route to principal stations

Fig. 1 EMIEW3

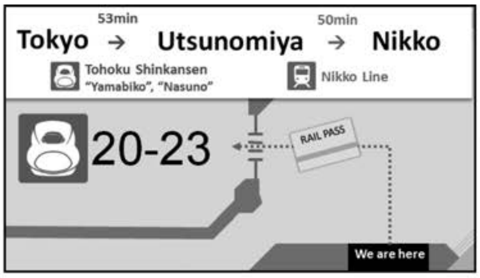
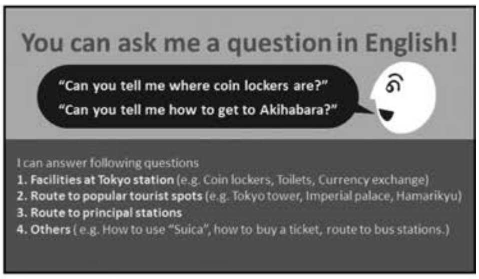


Fig. 2 Route Guidance



You can ask me a question in English!

“Can you tell me where coin lockers are?”
“Can you tell me how to get to Akihabara?”

I can answer following questions
1. **Facilities at Tokyo station** (e.g. Coin lockers, Toilets, Currency exchange)
2. **Route to popular tourist spots** (e.g. Tokyo tower, Imperial palace, Hamarikyu)
3. **Route to principal stations**
4. **Others** (e.g. How to use “Suica”, how to buy a ticket, route to bus stations.)

Fig. 3 Introductory Screen

3. Test Results

As a result of tabulating dialog logs, we found that 634 people participated in the tests with 43% selecting English, 8% selecting Chinese, and 49% selecting Japanese.

3.1 Confirmation of Performance of Language Switching Based on Speech

Table 1 shows the success rate of language switching and of answering questions. Number of attempts in the table indicates the number of times the participant spoke to the robot, and number of successes indicates the number of times language switching was successful or that answers to questions were successfully made. Success and failure were judged from dialog logs. As a result of changing the program in the second half to ignore utterances with no meaning such as "eh" and "ah" typically seen in Japanese speech, switching to Japanese was improved by 23 points.

Table 1 Dialog Success Rate

		Language switching			Question answering		
		Number of attempts	Number of successes	Success rate	Number of attempts	Number of successes	Success rate
First half	English	282	176	62 %	636	174	27 %
	Chinese	50	40	80 %	68	22	32 %
	Japanese	407	205	50 %	689	223	32 %
	Subtotal	739	421	57 %	1393	419	30 %
Second half	English	-	-	-	290	88	30 %
	Chinese	12	11	92 %	27	6	22 %
	Japanese	153	111	73 %	336	164	49 %
	Subtotal	165	122	74 %	653	258	40 %

3.2 Confirmation of Noise Reduction Performance

Table 2 shows the causes identified for instances not counted as successes in Table 1. Dialogs that were judged to be failures due to unsuccessful reduction of noise were classified as "V". Accurate detection of the vocalization part (accurate processing to extract just the user's voice from sound data) greatly depends on noise reduction precision. Values of the V column are extremely small compared to other classifications, demonstrating that noise reduction performance was high.

Table 2 Causes of Dialog Failure

		I	II	Ⅲ	Ⅳ	V	VI	Ⅶ
First half	English	9.9 %	41.2 %	6.2 %	14.1 %	0.2 %	16.3 %	12.1 %
	Chinese	16.0 %	42.0 %	10.0 %	6.0 %	2.0 %	2.0 %	22.0 %
	Japanese	10.8 %	30.7 %	12.3 %	14.1 %	0.6 %	16.8 %	14.7 %
	Subtotal	10.6 %	36.2 %	9.4 %	13.7 %	0.5 %	15.9 %	13.8 %
Second half	English	1.0 %	45.5 %	4.0 %	6.9 %	4.0 %	36.1 %	25.0 %
	Chinese	12.5 %	45.8 %	16.7 %	8.3 %	0.0 %	4.2 %	12.5 %
	Japanese	4.6 %	20.7 %	17.8 %	5.7 %	2.9 %	43.1 %	5.2 %
	Subtotal	5.6 %	23.7 %	17.7 %	6.1 %	2.5 %	38.4 %	6.1 %
Total		8.6 %	35.8 %	9.7 %	11.7 %	1.2 %	21.7 %	11.2 %

I. Poor language recognition at question
II. Poor recognition of question content
III. Inaccurate answer
IV. Poor facial recognition at question
V. Poor voice activity detection
VI. No answer information
VII. Other

3.3 Confirmation of Issues in Achieving Smooth Conversation

Table 3 shows the average times from the participant finishing speaking to the robot starting to speak. That includes noise reduction and other-language processing time. There are differences in processing for language switching and answering questions, so those are shown separately. As there was approx. 3 seconds between question and answer, many participants repeated their question thinking that the robot did not comprehend their question, and "slow reply speed" was the No. 2 reason for dissatisfaction by both Japanese and foreigners shown in questionnaire surveys taken after the tests.

Table 3 Processing Time (seconds)

		Language switching	Question answering
First half	English	3.0	2.6
	Chinese	3.5	3.0
	Japanese	3.2	2.5
	Subtotal	3.1	2.6
Second half	English	-	2.6
	Chinese	3.1	1.9
	Japanese	3.1	2.2
	Subtotal	3.1	2.4

4. Consideration

4.1 Language Switching Based on Speech

Ideally, language switching and identification of question content would be processed simultaneously, but such simultaneous processing is not actually possible. In tests, users did not look at the screen, so many participants were at a loss as to what to do. Even if proceeding in the prescribed order, success rate for language switching was 57% in the first half and 74% in the second half of the tests. So, it is difficult for dialog with the robot to proceed by voice alone for reasons of ease of understanding and level of maturity of speech recognition technology. Therefore, using touch panel input along with voice input is assumed to currently be an adequate approach.

4.2 Noise Reduction Performance

As shown in Table 3, the rate of dialog failure caused by noise reduction was extremely low. However, even if human voices could be separated from other sound, the technology is not at a level where voices of multiple people can be differentiated, so dialog can currently only be performed with one person standing in front of the robot.

4.3 Issues in Smooth Conversation

Processing time is affected by noise level in the conversation environment, so it is assumed that time required for noise reduction can be shortened in very quiet spaces. However, pre-test investigation showed that nose level in the Travel Service Center was a maximum of 65 dB and average of 55 dB, with noise at ticket gates and concourses assumed to be even greater and processing time thus even longer. From such perspectives as well, a posture of flexibility without being insistent on voice recognition is needed with measures taken such as adopting touch screen input in dialogs where selections such as language switching are made. Also, informing users by voice or illustration that voice recognition processing is being done is assumed to be important in reducing dissatisfaction.

5. Conclusion

Human guides can deal with questions they have not heard before by background knowledge and common sense, but robots do not have those abilities and are not able to accomplish dialogs when topics and phrasings not registered are included in questions. This time, we taught the robot 216 frequently asked questions and their answers in advance for the tests, but there were many cases where dialog could not be accomplished for questions the robot was not taught and cases where the intent of the question could not be recognized if phrasing was different even for topics that included the 216 questions (Table 3, II and IV). We who set up the robot cannot identify in advance all of the topics that users may ask about and the phrasing of those, so we assume that we need to make the robot learn more from actual dialogs in order to be able to handle a variety of topics and phrasings. However, if there is an attendant nearby the robot, questioners tend to gather around the attendant instead of the robot, so we need to come up with some modifications. We are currently working to develop elemental technologies to be able to deal with that issue. And we plan to continue to make efforts jointly with manufacturers and vendors in R&D necessary to achieve customer guidance by robots and AI.