A Spoken Dialogue System for Noisy Environment

Kazuki Kogure, Masahiro Yoshinaga, Hikaru Suzuki, and Tetsuro Kitahara

Nihon University
3-25-40, Sakurajosui, Setagaya-ku, Tokyo, Japan
{kogure,yoshinaga,hikaru,kitahara}@kthrlab.jp

Abstract. One of the important challenges for achieving a spoken dialogue system in noisy environments is to make the system's speech audible for the user. Although there have been many studies on speech recognition in noisy environments, very few attempts to improve the audibility of the system's speech. In this paper, we develop a spoken dialogue system that has three functions: real-time volume adjustment, utterance delay, and re-utterance. Experimental results have shown that these three functions improve the audibility of the system's utterances.

1 Introduction

Spoken dialogue systems are expected to play a role in reducing digital divides because speech is one of the most natural methods of human communication. In particular, this will be very helpful to blind people. In order to develop spoken dialogue systems that work in the real world, a robustness to environmental noise is required. To achieve this kind of robustness, we have to resolve two issues: one is to recognize the user's speech in noisy environments and the second is to make the system's speech audible for the user. Although many studies have attempted to tackle the first issue[1,2], only a small number have attempted to overcome the second issue.

In this paper, we develop a spoken dialogue system that has the following three functions: (1) real-time volume adjustment according to noise levels, (2) utterance delay for a particularly loud noise, and (3) re-utterance when asked by the user. Through these three functions, we reduce cases for which the user cannot follow the system's utterances due to loud noises.

2 Proposed System

We developed a spoken-dialogue-based railway route search system. This system is intended for installation on the platform of a rail station. To allow blind people to use this system, the system uses only auditory modality (speech inputs and outputs), and never uses a graphical display or a touchscreen. The system overview is shown in Figure 1 . Once detecting a user's face in the camera, the system asks the user to what station he/she is travelling. Given the response

C. Stephanidis (Ed.): HCII 2014 Posters, Part II, CCIS 435, pp. 577–582, 2014.

of the user, the system searches for the shortest route from the current station to the destination station using a railway route search API, called Expert Web API. The system then utters the search result. An example of the dialogues is shown below:

System: Hi, this is a railway route guidance system. What station are you going to?
User: Shinjuku Station.
System: Shinjuku Station?
User: Yes.
System: First travel to Sasazuka Station via the Keio New Line.
System: Then change to the Keio Line.
System: It will take 15 minutes. The fare will be 150 yen.
System: This guidance is now over. Thank you.

In this example, the departure is assumed to be the Sakurajosui Station. We use Julius [3] for speech recognition and OpenJTalk [4] for speech synthesis. To improve the system's robustness to environmetal noise, as described in the Introduction, we introduced the following three functions:

1. **Real-time volume adjustment:** to measure the noise level via a microphone array and control the volume of the system's voice according to the noise level;
2. **Utterance delay:** to delay the system's utterance when a particularly loud noise occurs;
3. **Re-utterance:** to utter again when the user utters a word requesting a re-utterance such as "eh?"

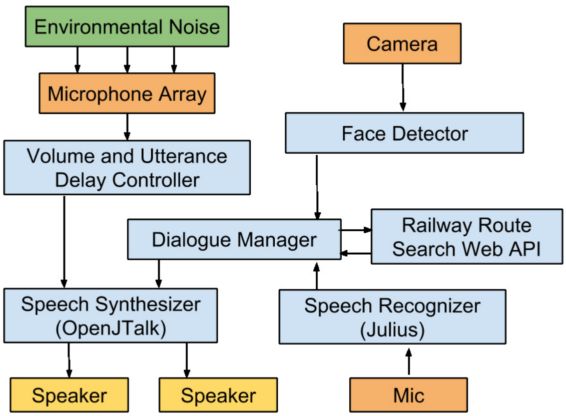

Fig. 1. System Overview

2.1 Real-Time Volume Adjustment

The environmental noise is observed by a 7-channel microphone array. For every second, the root mean square (RMS) of the audio signal of the environmental noise is calculated after the 7-channel signal is converted to monaural. The noise level is then estimated based on the RMS. If the sound pressure level of the system's utterances are lower than that of the noise, the system increases the volume setting of the system's utterances. If the sound pressure level of the system's utterances are higher than that of the noise, the system decreases the volume setting of the system's utterances.

2.2 Utterance Delay

If the noise level is higher than 64dB, the system delays the utterance as opposed to changing the volume because the noise is too loud. Specifically, the system stops the utterance and then restarts the utterance after the noise ends.

2.3 Re-utterance

When the user utters a word asking for a re-utterance, such as "eh?" and "*kikoenai*"(I can't hear it), the system utters the same sentence again. When the user utters "*kikoenai*", the system increases the volume setting as well as re-utters the reply.

3 Experiment 1

3.1 Experimental Method

The first experiment aimed to confirm improvements for the audibility of the system's utterances by two functions, volume adjustment and utterance delay. We installed a 6-channel speaker array around our system and virtually reproduced the noisy environment using a recording of noise from a Tokyo railway station platform. We asked 14 participants with normal hearing ability (Age: 21–24; 7 males and 7 females) to listen to the system's utterances under the noisy environment and rate the audibility, and choose from four options of what they believed the system uttered. As the aim of the experiment was to confirm the audibility of the system's utterances, face detection with a camera, users uttering a destination station, and producing an actual route were skipped. The system simply uttered pre-designed sentences. Examples of the sentences uttered include the following:

Type I
"*Aoto kara Aoi made no ryokin wa ni hyaku san ju en desu.*"
(The fare from Aoto Station to Aoi Station is 230 yen.)
"*Aoi kara Aoto made no ryokin wa ni hyaku san ju en desu.*"
(The fare from Aoi Station to Aoto Station is 230 yen.)

"*Aoto kara Aoi made no ryokin wa go hyaku san ju en desu.*"
(The fare from Aoto Station to Aoi Station is 530 yen.)
"*Aoi kara Aoto made no ryokin wa go hyaku san ju en desu.*"
(The fare from Aoi Station to Aono Station is 530 yen.)

Type II
"*Kameari kara Seibu Tamagawa-sen ni nori, Kameido e mukaimasu.*"
(Get on the Seibu Tamagawa Line at Kameari Station, then get off at the Kameido Station.)
"*Kameido kara Seibu Tamagawa-sen ni nori, Kameari e mukaimasu.*"
(Get on the Seibu Tamagawa Line at Kameido Station, then get off at the Kameari Station.)
"*Kameari kara Seibu Tamako-sen ni nori, Kameido e mukaimasu.*"
(Get on the Seibu Tamako Line at Kameari Station, then get off at the Kameido Station.)

As above, phonologically similar confusing stations/lines were included. To avoid issues resulting from the participants already knowing the station and/or lines, untrue fares/routes were uttered. The subjective rating of audibility was based on the following criteria:

1. Impossible to hear
2. Possible to hear but impossible to recognize the content
3. Possible to hear but difficult to recognize the content
4. Partly difficult to hear
5. Appropriate volume
6. Too loud
7. Much too loud

The experiment was conducted both with normal utterances (without volume adjustments or utterance delays; our *baseline*) and with utterances of volume adjustments and utterance delays (our *proposed*). For each condition, the experiment was repeated 21 times.

3.2 Experimental Results

The experimental results are shown in Figures 2 and 3. Figure 2 shows the ratings of audibility for utterances using both methods. The system rarely failed in picking up the utterances. Such cases have been removed from the results. For the proposed method, the ratings of 1 and 2 were decreased by 6% and the rating of 5 was increased by 14% when compared to the baseline method. Figure 3 shows the accuracy of the multiple-choice tests for listening to the system's utterances. The accuracy for the proposed method was 8% higher than that of the baseline method.

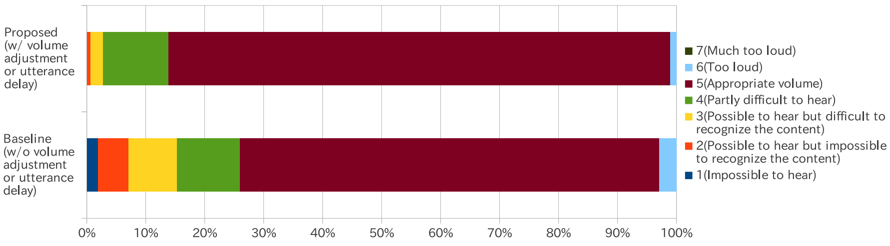

Fig. 2. Results of Experiment 1 (Audibility Ratings)

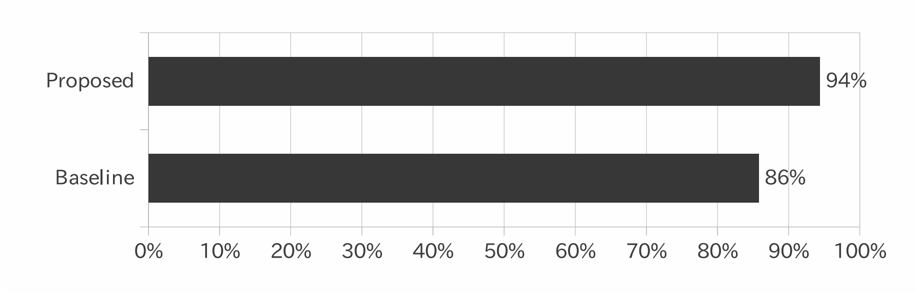

Fig. 3. Results of Experiment 2 (Accuracy of Listening)

4 Experiment 2

4.1 Experimental Method

The second experiment aimed to confirm the effectiveness of our system in actual dialogues. In the same environment as Experiment 1, we asked participants to search and write down a route (the name(s) of line(s) and transfer station(s), total fare) to a station. This was repeated 10 times for each of the following three systems:

System A: without any of the three functions
System B: with volume adjustment and utterance delay, without re-utterance
System C: with all three functions.

The experiment for the three systems was continuously conducted and the systems were chosen at random for each trial. The participants were not told which system was running. The participants were three people with normal hearing ability (Age: 21–24; two males and one female).

4.2 Experimental Results

The experimental results are shown in Table 1. From this table, it can be seen that the accuracy of the transcriptions of the system's utterances by the

Table 1. Results of Experiment 2

		System A	System B	System C
System's features	Vol. adj.	—	√	√
	Utt. delay	—	√	√
	Re-utterance	—	—	√
Accuracy of transcription of the system's utterances	Participant 1	65%	86%	96%
	Participant 2	80%	89%	96%
	Participant 3	64%	86%	95%

participant improved by the three functions. The three participants asked for re-utterance 17 times in total. Of the 17 re-utterances, 82.3% were successfully recognized by the participants. There are three reasons why the participants did not recognize 17.7% of the re-utterances. The first reason is the inappropriate timings of the participant uttering "eh?". At the exact time the participant uttered "eh?", the next utterance of the system had sometimes already started. In this case, the system re-uttered a different sentence from what the participant wanted to listen to again. The second reason is failure of speech recognition. The system recognized "eh?" or "*kikoenai*"(I can't hear it) as a different word, and as such the re-utterance was not executed. The third reason is inaudibility of the synthesized speech. Although the re-utterances were successfully executed, the synthesized speech was not of sufficient quality, so the participant did not understand it.

5 Conclusion

We illustrated the importance of making the system's utterances audible when developing a spoken dialogue system for a noisy environment, and thereafter proposed a system that has three functions: real-time volume adjustment, utterance delay, and re-utterance. Future work will include further experiments in various types of environments and when faced with different kinds of noise.

Acknowledgments. This research was supported by SCAT, Japan. We appreciate the VAL Laboratory Corporation for giving us permission to use Expert Web API.

References

1. Hinamoto, Y., Mino, K., Saruwatari, H., Shikano, K.: Interface for Barge-in Free Spoken Dialogue System Based on Sound Field Control and Microphone Array. IEEE Troc. ICASSP 2, 505–508 (2003)
2. Ishi, T., Matsuda, S., Kanda, T., Jitsuhiro, T., Ishiguro, H., Nakamura, S., Hagita, N.: A Robust Speech Recognition System for Communication Robots in Noisy Environments. IEEE Transactions on Robotics 3, 759–763 (2008)
3. Lee, A., Kawahara, T.: Recent Development of Open-Source Speech Recognition Engine Julius. In: APSIPA ASC 2009, pp. 131–137 (2009)
4. `http://open-jtalk.sourceforge.net/`