Analysis of Long-Term and Large-Scale Experiments on Robot Dialogues Using a Cloud Robotics Platform

Komei Sugiura and Koji Zettsu
National Institute of Information and Communications Technology
3-5 Hikaridai, Seika, Soraku, Kyoto 619-0289, Japan
Email: {komei.sugiura, zettsu}@nict.go.jp

***Abstract*—To build conversational robots, roboticists are required to have deep knowledge of both robotics and spoken dialogue systems. Unlike using stand-alone speech recognition/synthesis toolkits, a cloud robotics platform for human-robot communication enables high-quality speech recognition and synthesis that is optimized to human-robot interactions. This is challenging because we need to build a wide variety of functionalities ranging from a stable cloud platform to high-quality multilingual speech recognition and synthesis engines. From this background, we constructed rospeex [1], which is a cloud robotics platform for multilingual spoken dialogues with robots. Over 20,000 unique users have used rospeex in the two years since it was launched. In this paper, we propose a method to reduce the response time in rospeex; and analyze its effectiveness. We also analyze the server logs of rospeex that we have collected.**

I. Introduction

Building conversational robots requires deep knowledge of both robotics and spoken dialogue systems. This prevents roboticists from spending more time on their own research topics for real applications.

In this study, our target use case is spoken dialogues with service robots that mainly work in domestic environments (Fig.1). An example use case is where a user asks a robot to fetch a plastic bottle from a table in RoboCup@Home [2].

Most roboticists have been using speech recognition and speech synthesis toolkits running on stand-alone computers. However, high-quality speech recognition is very difficult since stand-alone computers have CPUs and memory limitations. Although they can use existing cloud-based APIs that were built for other services, e.g., voice search, it is difficult for service providers to continuously improve the performance specific for robotics tasks because robotics-related logs will get buried in non-robotics-related logs.

Based on the above background, we constructed a cloud robotics platform called rospeex [1] that we have been operating for over two years. Any roboticist can use it without payment or authentication.

It is crucial for such a cloud robotics platform to reduce the response time between the robot and the cloud server because longer waiting times will diminish the naturalness in dialogues. In this paper, we analyze response time by a method which divides the utterances into fragments and sends them to the rospeex server. We also analyze the logs that we collected by operating rospeex.

The following are our key contributions:

- A speech fragmentation method that reduces the response time in rospeex.
- Analysis of robotics-specific logs collected as rospeex's server logs.

II. Cloud Robotics Platform: rospeex

The details of rospeex's structure were described in a previous paper [1]. We provide the browser user interface, the rospeex modules (noise reduction, voice activity detection, and speech synthesis), and the rospeex cloud services. We assume that the developer writes code on dialogue-related functions including language understanding, dialogue management, and response generation.

A. Cloud-based Speech Recognition

The performance of the speech recognition engine was previously described in [3]. The Word Error Rate (WER) was reported to be 8.2% for the IWSLT ASR 2012 test data. The rospeex cloud servers for speech recognition and speech synthesis were originally developed for "VoiceTra," a speech-to-speech translation system [4]. Our cloud services support multilingual speech communication in English, Chinese, Japanese, and Korean.

For usability, we also provide a way to select other cloud services. Developers can switch cloud engines with very little effort and integrate rospeex with their software assets written for ROS.

B. Speech Fragmentation

It is crucial to reduce the response time of speech recognition since longer response times diminish the naturalness in

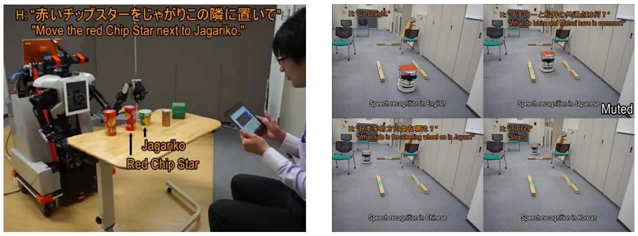
H: "赤いチップスターをじゃがりこの隣に置いて"
"Move the red Chip Star next to Jagariko."
Jagariko
Red Chip Star
H: "Go straight."
Speech recognition in English
H: "イチローと松井の共通点は何？"
"What do Ichiro and Matsui have in common?"
Speech recognition in Japanese
Muted
H: "日本车的方向盘在哪边？"
"What side is the steering wheel on in Japan?"
Speech recognition in Chinese
H: "기다려"
"Stop."
Speech recognition in Korean

Fig. 1. Sample target use cases of rospeex, which is available at http://rospeex.org/

978-1-4673-8370-7/16/$31.00 © 2016 IEEE

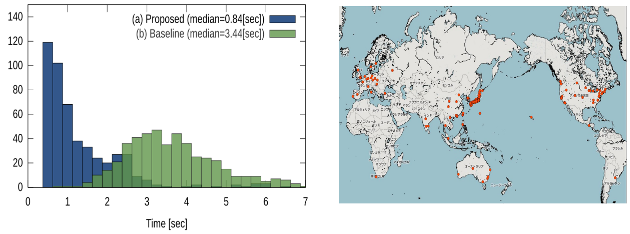

Fig. 2. Left: Histogram of response time of (a) proposed and (b) baseline methods. Vertical axis represents the frequency. Right: IP distribution of rospeex users shown as orange dots.

human-robot dialogues. Here we propose a speech fragmentation method in which speakers' utterances are fragmented into smaller segments and sent to the cloud server while they are speaking. Such speech fragmentation is widely used in non-robotics fields.

The processing time of speech recognition with/without speech fragmentation is almost constant. The difference is whether the speech is sent while the speaker is talking to the robot or after the VAD detects the end-of-speech.

The size of a fragment, s, should be carefully designed in the method. On the one hand, the fragments needs to be small enough to reduce the response time. On the other hand, a decrease in the response time converges at a certain point, and the server's load increases instead if the fragments are too small. In this paper, we set s to be 3.52kB to balance the trade-off. When using the ATR-503 corpus, an utterance will be fragmented into approximately 50 segments on average.

C. Non-Monologue HMM Speech Synthesis

Although natural conversations are desirable in human-robot interactions, most speech synthesis engines are not optimized for them. Therefore, the robot voices do not sound friendly and natural. When intonation is inappropriate, the speakers might not realize that they are being asked a question.

For expressive speech synthesis, we built the Non-Monologue HMM-based Speech Synthesis method [5]. Since rospeex is compatible with this method, developers can use it for free. Unlike conventional methods, the robots using it can synthesize friendly and natural voices. In previous research, we evaluated it with the standard MOS metric and showed that its performance almost approached the theoretical upper limit. For example, in robot dialogue tasks, the MOS values of the theoretical upper limit, the baseline, and the non-monologue HMM speech synthesis were 3.86, 2.03, and 3.65, respectively. The details of the results were previously explained [5].

III. Experiments

Analyses were performed to investigate the feasibility of our proposed platform. The performance of the speech recognition/synthesis was described in detail in other publications, and is beyond this paper's focus. The WER of the speech recognition engine was 8.2% for the IWSLT ASR 2012 test data [3]. The quantitative results of our speech synthesis method were also shown in [5].

A. Analysis of Response Times

Next we investigate the feasibility of our speech fragmentation method. We used the ATR-503 corpus read by a voice talent and calculated the response times for 495 sentences successfully detected by the VAD module. We defined the response time as the difference between the response obtained from rospeex's cloud service and the end time of the detected utterance.

We simulated a model use case where utterances are sent from a WiFi-connected PC to rospeex's server through an Internet connection. We did not measure the response times at all the possible bandwidth conditions since this is not our focus.

The left-hand figure of Fig.2 shows a histogram of the response times for the speech recognition. We compared our proposed speech fragmentation method with a baseline method that did not use the speech fragmentation method. The (median) response times of the baseline and the proposed methods were 3.44 [sec] and 0.84 [sec].

B. Analysis of Human-Robot Dialogues

The rospeex cloud platform was launched on September 1, 2013. As of November 30, 2015, it has obtained over 20,000 unique users. Here we define a unique user as an access to our platform from an IP address; multiple accesses from the same IP address on a single day are counted as one unique user. The distribution of the addresses is shown on the right-hand in Fig.2.

Finally we explain the utterance logs collected by the server. The total number of utterances was 44960; we filtered out the speech files that did not contain any sound. Among them, 31.7% were simple conversations (e.g., "hello") and 19.3% were simple QAs (e.g., "give me today's weather forecasts.") The categorization details are shown in [1].

IV. Conclusion

We analyzed the server logs we have collected by operating rospeex. Unlike merely using cloud services, our approach aims to construct our own cloud services that roboticists can freely use through robot middleware so that we can store a huge amount of robotics-related logs in the cloud server. Future directions include the improvement of language models for robotic tasks, and sharing the log corpus with other roboticists.

V. Acknowledgments

This work was partially supported by JSPS KAKENHI Grant Number 15K16074 and JST CREST.

References

[1] K. Sugiura and K. Zettsu, "Rospeex: A cloud robotics platform for human-robot spoken dialogues," in *Proc. IEEE/RSJ IROS*, 2015, pp. 6155–6160.

[2] L. Iocchi *et al.*, "RoboCup@Home: Analysis and results of evolving competitions for domestic and service robots," *Artificial Intelligence*, vol. 229, pp. 258–281, 2015.

[3] C.-L. Huang *et al.*, "The NICT ASR System for IWSLT 2013," in *Proc. of IWSLT*, 2013.

[4] S. Matsuda *et al.*, "Multilingual Speech-to-Speech Translation System "VoiceTra"," in *Proc. Workshop on Field Speech and Mobile Data*, 2013, pp. 229–233.

[5] K. Sugiura *et al.*, "A Cloud Robotics Approach towards Dialogue-Oriented Robot Speech," *Advanced Robotics*, vol. 29, no. 7, pp. 449–456, 2015.