A DATA COLLECTION SYSTEM USING A SPEECH RECOGNITION SYSTEM UNDER MULTIPLE ENVIRONMENTS

[1]*Sunao Hara,* [1]*Chiyomi Miyajima,* [2]*Katsunobu Itou and* [1]*Kazuya Takeda*

[1]Graduate School of Information Science, Nagoya University, Nagoya 464-8903, JAPAN
[2]Graduate School of Computer and Information Sciences, Hosei University, Tokyo 184-8584, JAPAN
[1]{hara,miyajima,takeda}@sp.m.is.nagoya-u.ac.jp, [2]itou@k.hosei.ac.jp

ABSTRACT

We implement a spoken dialogue system and a speech collection system for conducting field experiments in many and unspecified environment. The speech data are collected on each user's PC by using the spoken dialogue system, and transmitted to the speech collection system via the Internet. Moreover, users can easily customize their spoken dialogue systems by themselves. We conducted a two-month field experiment that was open to the public on the Internet and collected data that included segments of voice activity totaling 5 hours and 41 minutes, which comprised 6338 utterances that spoken to the system.

1. INTRODUCTION

We are developing a music retrieval system with a speech interface that can play music in users' PCs or from a commercial music store on the Internet. However, there is a point worth noting in relation to using automatic speech recognition (ASR) as an input interface: We cannot know which microphone users employ to input speech in a real environment, what acoustical features the environment has, or whether users can operate their microphone correctly. These features lead to a reduction in speech recognition accuracy. To improve speech recognition accuracy, it is crucial to collect speech data in the environment in which the system is used[1]. However, in most conventional studies on this topic, all subjects used a microphone of the same quality in the same environment. Our system, on the other hand, can be easily installed on users' PCs, it is equipped with functions that users can customize, and it can transmit speech data collected on users' PCs via the Internet. We have collected a large amount of speech data in various environments through an online demonstration experiment of our system.

2. ONLINE DATA COLLECTION SYSTEM

Our system features an Internet music retrieval system with a spoken dialogue interface called "MusicNavi2" and an online maintenance server called "Moscowmule." Fig. 1 shows the relationship between MusicNavi2 and Moscwomule.

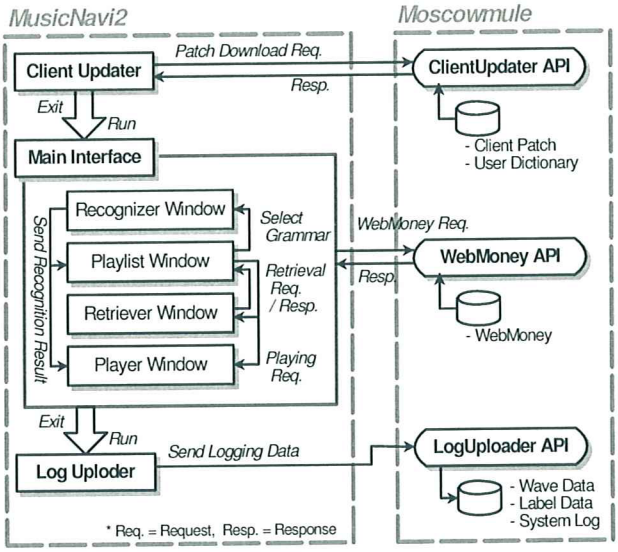

Fig. 1. Relationship between MusicNavi2 and Moscowmule.

2.1. MusicNavi2: an Internet Music Retrieval System with a Spoken Dialogue Interface

Our music retrieval system enables users to retrieve, choose, and play music through a conversation with a spoken dialogue interface by specifying music titles, album names, artist names, and commands. The system recognizes the user's request and obtains a music list from the music database on the user's PC or the "Mora" an Internet music retrieval service[1]. The contents of the music list are spoken by a speech synthesis module, and users can choose their favorite music from the list without actually seeing the list and without pushing any buttons. Therefore, users have the option of using our system as both a hands-free and eyes-free music retrieval system.

MusicNavi2 also includes functions for buying songs, uploading recorded speech data via the Internet, and customiz-

[1]Mora: http://mora.jp/

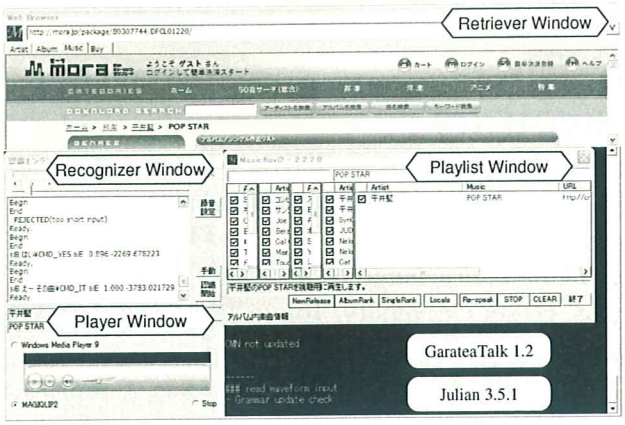

Fig. 2. Screenshot of MusicNavi2.

ing the recognition dictionary for every user. MusicNavi2 is distributed with an installation wizard and is automatically updated via the Internet. These features were effectively demonstrated in field experiments conducted on users' PCs.

Fig. 2 shows a screenshot of MusicNavi2. The *Retriever Window* manages the music retrieval from users' PCs or the Internet according to a request to MusicNavi2, and it sends the retrieval result to the *Playlist Window*. The *Playlist Window* manages retrieval results, manages dialogue state transitions, and controls a speech synthesis module. A retrieval result list is then created when the retrieval result is acquired. The *Recognizer Window* manages speech recording and recognition. It is always recording data from the microphone while the system is running. The recorded data are automatically compressed with no less by FLAC [2] and are written into a speech log file in the user's PC. For Voice Activity Detection(VAD), we adopt a threshold detection by the speech amplitude and the zero-cross count. This threshold can be freely altered by users; only the VAD segments are sent to the speech recognition module. The time of each VAD segment and the speech recognition results are written in a speech label file in the user's PC. The *Player Window* manages a music player for downloading and playing music from Mora.

The speech recognition module uses "Julian 3.5.2," which is a grammar-driven parser[3]. Our system employs a gender-independent acoustic model of phonetically tied mixtures with 3,000 states (129 codebooks) and 64 Gaussians. The grammar network accepts isolated words or several command sentences. The word dictionary is created by users on Moscowmule, furthermore, new word entries can be added from the retrieval result list if necessary. The speech synthesis module uses "GalateaTalk 1.2[4]."

2.2. Moscowmule: an Online Maintenance Server

Moscowmule is a World-Wide-Web application server that runs under a LAMP (Redhat Linux9, Apache2, MySQL4.1 and PHP5) environment. Users can access Moscowmule with a Web browser and register a user account, download the MusicNavi2 software package, answer relevant questionnaires, and customize a speech recognition dictionary for each user. To customize the dictionary, users choose their favorite artists from among the artist lists. Artist lists are grouped by the first characters of artists' names and sorted in Japanese pronunciation order. They can also narrow down their search area by entering a part of an artist's name. Chosen artist names, their album titles, and their music titles are registered in the users' dictionary. To date we have collected 4,533 artist names, 18,463 album titles and 135,775 music titles from Mora, and we have added pronunciations to the entries.

MusicNavi2 also accesses Moscowmule, automatically uploads the log files, downloads each user's speech recognition dictionary, and updates itself. For security reasons, all of the communications are encrypted by SSL.

3. ONLINE DEMONSTRATION EXPERIMENT

We have been distributing the system and collecting system logs including speech data since February 1, 2006, and have also been conducting some questionnaire-based surveys about users' environments. To March 31, 2006, about fifty-nine hours of system logs have been collected. These data will be analyzed to evaluate the achievement of users' tasks, acoustical environments, and speech recognition qualities, but it might be necessary to collect more system logs in many environments by many users to gain an accurate analysis.

4. CONCLUSION AND FUTURE WORK

In this paper, we presented a data collection system, that contains an Internet music retrieval system which includes a spoken dialogue interface and an online maintenance server for the spoken dialogue interface. We conducted a field experiment to collect speech logs in the users' environments, collecting approximately fifty-nine hours of system logs. Future work will include considering the continuation of data collection and data analysis, and upgrading of the system.

5. REFERENCES

[1] James Glass *et* al., "Data collection and performance evaluation of spoken dialogue systems: The MIT experience," in *Proc. of ICSLP-00*, Oct. 2000, vol. 4, pp. 1–4.

[2] Josh Coalson, "Free Lossless Audio Codec (FLAC)," http://flac.sourceforge.net/.

[3] T. Kawahara *et* al., "Recent progress of open-source LVCSR engine Julius and japanese model repository; software of continuous speech recognition consortium," in *Proc. of ICSLP-04*, Oct. 2004, vol. 2, pp. 3069–3072.

[4] S. Kawamoto *et* al., "Open-source software for developing anthropomorphic spoken dialog agent," in *Proc. of PRICAI-02*, Aug. 2002, pp. 64–69.