TELEPHONE DATA COLLECTION USING THE WORLD WIDE WEB[1]

Edward Hurley, Joseph Polifroni, and James Glass

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139 USA
http://www.sls.lcs.mit.edu

ABSTRACT

Over the past year our group has begun development of telephone-based speech understanding capability for our GALAXY conversational system. An important part of this process has been the collection of telephone speech which was used for training and evaluation. In the first phase of data collection our goal was to collect read speech from a wide variety of talkers, telephone handsets, and noise/channel conditions. In the second phase of data collection our additional goal was to collect spontaneous telephone speech from subjects actually using the system. In order to maximize variation in telephone conditions, as well as ease of use for subjects, the data collection software was designed to telephone subjects at their specified phone numbers around North America. Subjects initiate the data collection session by submitting an electronic form accessible by a WWW browser. For read speech collection, a set of prompts is automatically generated for the subject. This paper describes the design of the data collection system we are using for these purposes. To date we have collected over 9,000 utterances from over 270 subjects.

1. INTRODUCTION

One of the desiderata for our research in conversational spoken language technology is to enable mobile and affordable access to information using voice input [1]. To this end, we believe it is natural and desirable to allow telephone-based access to spoken language systems. Telephones are prevalent in our society, they are convenient and, very importantly, the average person knows how to use them properly. The telephone-based input model fits well with the client/server architecture used in our GALAXY system [2]. The user can be working with a lightweight client, while their voice proceeds directly without delay to a speech recognizer compute server. In addition, telephone-based input is a first step towards displayless conversational systems, another area of research we are interested in.

In order to achieve telephone-based capability, we needed to obtain a speech corpus we could use for training and evaluating the GALAXY system. Our group has been involved in the creation of many types of read and spontaneous speech corpora [3, 4, 5, 6]. In the case of our GALAXY system, we have been using a wizard-based data collection framework [5]. Subjects come to our laboratory, and interact with the system after being given GALAXY scenarios to solve. A wizard in another room listens to the queries and types them to the system, which then responds to the user.

Using the wizard-based framework, we collect spontaneous speech needed to properly train and evaluate the speech recognition and natural language components of our system. It was relatively straightforward to augment this framework to simultaneously collect speech in a laboratory setting from both a noise-cancelling microphone and a telephone. However, since we wanted to be able to use the GALAXY system anywhere, we felt this approach suffered from a lack of variation in telephone handsets and line conditions. We therefore desired an alternative method which, although not on the same scale as major telephone data collection efforts [7, 8, 9, 10], could be used to augment our telephone speech corpora.

The approach that we have taken is to collect data from remote users, using the internet to display either prompts or the GALAXY client for read or spontaneous data collection respectively. This method provides us with the variation in handset and line conditions that we desired. It is convenient to the users, since they can provide data at any time via an electronic form accessible on any World Wide Web (WWW) browser. Finally, it is low-cost and low-maintenance, since we do not need to supervise the data collection process itself.

In the following sections we describe the hardware and software we used for data collection. We then outline the automatic prompt generation mechanism used for read data collection. We then describe the utterance recording and verification stages and report on the current status of our data collection efforts. Finally we discuss our future plans in this area.

2. ARCHITECTURE

In order to maximize the range of telephone handset and line conditions, as well as the ease of use for subjects, our data collection software actually telephones subjects at their various locations around the continent. A set of prompts are automatically generated, and the user is prompted to read them. The following subsections describe the design and setup in more detail.

[1] This research was supported by DARPA under contract N66001-94-C-6040, monitored though Naval Command, Control and Ocean Surveillance Center.

2.1. Hardware

The current phone data collection system consists of a commercially available hardware component, called Computerfone, that provides an interface between a workstation and an analog phone line. Line-in and line-out of the Computerfone are connected to the workstation. This allows programs running on the workstation to receive speech from the telephone line and also to play back prompts and responses via the phone line. The logic for Computerfone is controlled via a serial connection to one of the serial ports on the workstation. A simple set of commands is used to answer the phone, establish a connection, and hang up the phone when the call is completed.

2.2. Software

Access to the data collection facility is provided by a dedicated HTTP server. As can be seen in Figure 1, the read data collection page provides both instructions and a form interface for users to enter their phone numbers and email addresses. The phone number is used by the program to call the subject and email addresses are saved to a log-file. The email address is used to enter subjects in a drawing for prizes that are used as recruiting incentives. The read data collection homepage also contains a consent form, which subjects are asked to read and which explains how the data will be used. By clicking on an icon reading "I accept," subjects agree to participate in data collection. The same button submits the form to the HTTP server, which in turn launches a CGI script.

The CGI script extracts the user's phone number and email, and passes these arguments to the data collection application. The CGI script then selects one of 100 files of 50 prompts, dynamically creates an HTML page with the selected prompts, and returns it to the user's web browser (see Figure 2). The data collection application sends commands to the Computerfone, which calls the user back at the specified phone number. When the user answers the call, the program enters a loop prompting the user to speak each utterance in turn, and saving the recorded waveform in a session-specific directory. When all 50 utterances are finished, the application thanks the user and exits. The entire process takes an average of 11 minutes.

3. PROMPT GENERATION

Although we would like to collect spontaneous telephone speech in order to obtain both acoustic and linguistic data, we began our data collection effort with read speech. There were several reasons which made read data collection preferable, not the least of which was the fact that GALAXY could not handle telephone speech input. In addition, running a GALAXY client remotely was more complicated (we had similar problems running a predecessor collection scheme which required an XWindow display). Finally, we wanted coverage of all words in the GALAXY vocabulary. For these reasons, we decided to collect read speech, and therefore needed text prompts.

In designing the prompts for read telephone data collection, we tried to maximize for coverage of new words in the GALAXY lexicon and minimize the length of each utterance, so that subjects could speak each in as natural a way as possible. The algorithm for generating these prompts is straightforward. It starts with utterance templates

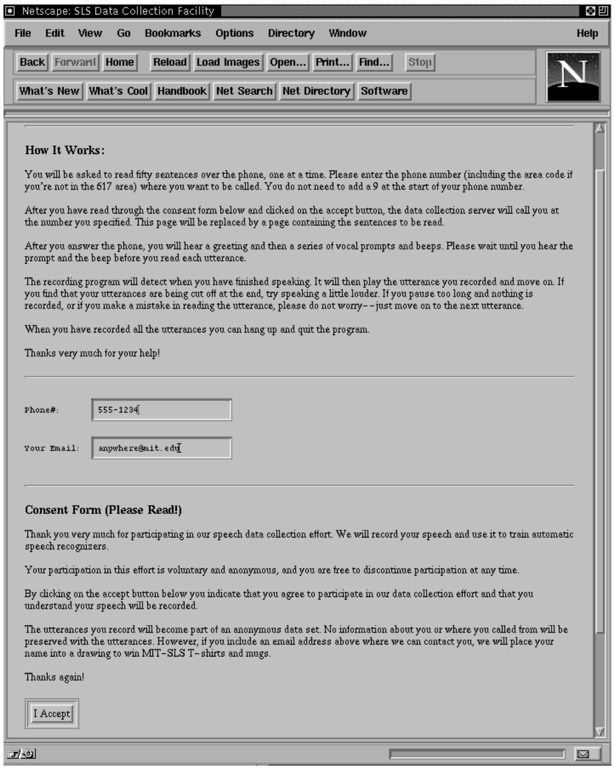

Figure 1: Web page form for data collection.
The top portion of this page provides instructions for the subject on how to use the data collection facility. The middle portion contains forms for the subject to enter their telephone number and email address (optional). The lower portion of the page consists of a consent form and an accept button. Data collection is initiated when the subject acknowledges understanding the purpose and terms of the data collection by clicking on the accept button.

containing a mix of words and variable names. These utterance templates are selected at random and each variable name is instantiated, by going in order through a corresponding list of the appropriate values. Example utterance templates are shown in Figure 3.

The variables in the utterance templates in Figure 3 are indicated by capital letters with a preceding colon. In the example, each variable is instantiated from a list containing words appropriate for the category. An example of how these lists are represented is shown in Figure 4. From the first utterance template in Figure 3 and the first variable in each category in Figure 4, we generate the prompt, "How do I walk to Joyce Chen in the Symphony Hall area." In addition to the utterance templates, there are a smaller number of prompts that do not need variables (e.g., "Zoom in"). These prompts can be inserted into the larger set at a rate determined by the developer.

The process that creates these utterances cycles through each list in order and retains a record of how many times each variable and specific word is used. In this way, the developer has the ability to modify the prompts to increase coverage in any given categories. Once a set of prompts has been generated, they are broken up into individ-

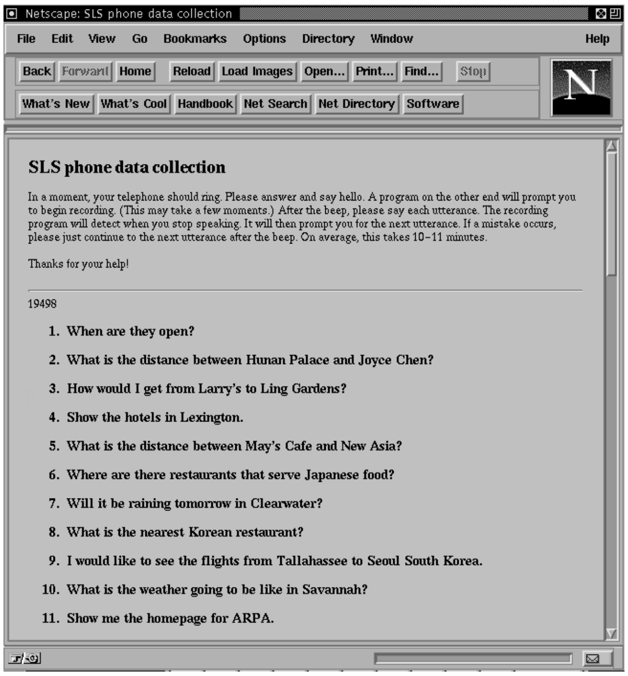

Figure 2: Example prompts page for data collection.

How do I :MOVE to :SPECIFIC_PLACE in :LOCAL_REGION?
Give me a weather forecast for :CITY_STATE.
I want to book a flight from :CITY_STATE to :CITY_STATE.

Figure 3: Example utterance templates used to generate prompts for read speech data collection.

ual files, one of which is read in by the data collection software when each session begins. This allows for a quick turn-around in targeting new words/phrases for data collection and actually having those words appear in the read prompts.

4. UTTERANCE RECORDING

At the start of the phone call, a short two second recording is made to estimate the noise level for the automatic endpoint detector. This step was necessary to adapt to varying channel conditions. After an initial greeting, the user is told which utterance to speak (e.g., "utterance ten") and then prompted with a beep. Utterances are recorded and then played back to the user. Although there is no mechanism for the user to repeat, we felt it was important for the user to hear what was being recorded. Part of the instructions suggest that the user speak louder, for example, if utterances are being clipped.

MOVE
walk, drive, get, go

LOCAL_REGION
the Symphony Hall area, Chinatown, Back Bay, Beacon Hill, the Boston Harbor area, ...

SPECIFIC_PLACE
Joyce Chen, M.I.T., the Gardner Museum, the Royal East, ...

CITY_STATE
Fresno California, Los Angeles, Caracas Venezuela, Dakar, ...

Figure 4: Example lists of variables for utterance templates.

5. UTTERANCE VERIFICATION

Even though the data were read, each utterance still had to be listened to, verified, and modified if necessary. A transcription tool was developed for these purposes. Our goal was to produce software which was easy to use, intuitive, and allowed for processing a large number of utterances quickly and accurately. We used a Tcl/Tk interface to provide an editable window that the transcriber can use to examine and modify each transcription. Below the window are mouse-activated widgets for playing the utterances multiple times, for moving forward and backward within the set of utterances being transcribed, and for saving transcriptions that have been modified.

We also wanted to use the same tool to check transcription quality as the transcriber was using it, so that typographical errors, for example, could be detected and dealt with as the words were being entered or modified. When a transcription is entered, the program checks each word to make sure that it is in the lexicon. If the word does not appear in the transcription lexicon, the transcriber is warned and given the option of either adding the word to the lexicon or changing it in the orthography file. In this way, typographical errors and misspellings can be found and dealt with while a human is still in the loop and the file is active. In the case of specialized markings for truncated words or filled pauses, the transcriber can add these as needed to the lexicon and the program will not produce a warning when it next sees them.

6. CURRENT STATUS

6.1. Read Speech

Our data collection system has been operational since the fall of 1995. To date, we have collected a total of 6,287 prompted utterances from 179 separate calls. Although the calls tend to peak around the times we actively solicit new data, we are still receiving callers on a regular basis (URL: http://www.sls.lcs.mit.edu).

After the data collection system had been up for several months, we analyzed the transcription from 5,215 read utterances. We found that, for the great majority of utterances, the transcriber had to make no changes at all to the prompt transcription as presented to the subject. The transcriber made changes in about 4% of the utterances. A breakdown of the types of modifications that needed to be made can be found in Table 1. Fewer than 1% of the utterances contained filled pauses (e.g., [uh], [er]), which is not surprising in read data. The miscellaneous category contains utterances with word substitutions (e.g., "Sunday" for "Saturday") and insertions, (e.g., "What is the weather going to be like *no I mean going to be* in Boston"). The two largest categories of modifications in transcriptions were mispronunciations and truncated words.

The prompts often contained words that were either foreign (e.g., "Quito, Ecuador"), and possibly difficult to pronounce, or specific to the Boston area (e.g., "Faneuil Hall") and also possibly difficult

Category of: Error	Total occurrences	Percentage of total
Filled pauses	42	.8
Truncated words	86	1.6
Mispronunciations	62	1.2
Misc. errors	10	.2
Total	200	3.8

Table 1: Analysis of errors found by transcriber in read data.

to pronounce. Most of the truncated words were either a member of one of these two classes or occurred immediately preceding such a word. In a similar fashion most of the words marked as mispronunciations were either city names or restaurant names in the Boston area. We asked our transcribers to be fairly lenient in classifying a word as mispronounced. Pronunciations that deviated from standard American English as a result of dialectal variation or foreign accent were not marked as mispronunciations. For example, many native speakers of American English pronounce the Boston restaurant "Giacomo" with four syllables, accenting the penultimate. Someone more familiar with Italian would pronounce it with three syllables, accenting the first. We accepted either as options for this word.

6.2. Spontaneous Speech

Due to the success of our initial read speech data collection efforts, we have been able to train up a telephone-based recognition system for our GALAXY system. From the time the system was deployed in late 1995, we have been collecting spontaneous speech from subjects using the GALAXY system remotely via the telephone. As mentioned earlier, this system is more complicated to run, but we have had some success in collecting data, both locally and from remote users. We have recently combined our read speech data collection with spontaneous speech collection by allowing users to speak directly to the GALAXY system. An icon at the bottom of the prompts screen allows users to connect directly to GALAXY and speak to it. These utterances are also saved and can be used for training. To date, we have collected approximately 2,000 spontaneous utterances in this fashion (in addition to approximately 1,000 spontaneous utterances collected in wizard mode).

More recently, we have begun to modify our read speech collection effort by asking users to paraphrase the prompt text. We believe this will be extremely useful for collecting multilingual speech corpora. In a preliminary study, we asked bilingual speakers to speak a *translation* (into their native language) of the read data collection prompts. In addition to providing us with acoustic data, these translations give us a variety of ways to ask for GALAXY-type information in a second language. We can use these data to help train both the speech recognition and natural language components of a spoken language system.

7. DISCUSSION AND FUTURE PLANS

The ability to collect data ourselves without bringing people into a laboratory environment has not only made it easier for us to collect data but has expanded our potential audience. Also, the collection is hands-off, meaning that our software handles a session without our intervention. We merely need to set the system up and advertise it. Not having to involve staff members with the collection process itself is a major advantage.

Our experience with collecting spontaneous speech remotely has shown that we need to make it easier for subjects to launch the GALAXY system and successfully use the system in an unsupervised manner. One approach we have been pursuing is to add more detailed instructions and examples dependent on the dialogue state. In the future, we are considering implementing the GALAXY client in a form that can run entirely within a WWW browser to simplify launching as much as possible. We also plan to increase the number of telephone connections to allow for multiple simultaneous data collection sessions.

8. REFERENCES

1. V. Zue, "Navigating the Information Superhighway Using Spoken Language Interfaces," *IEEE Expert*, 39–43, October, 1995.
2. D. Goddeau, E. Brill, J. Glass, C. Pao, M. Phillips, J. Polifroni, S. Seneff, and V. Zue, "GALAXY: A Human-language Interface to On-line Travel Information," *Proc. ICSLP*, 707–710, Yokohama, Japan, September, 1994.
3. L. Lamel, R. Kassel, and S. Seneff, "Speech Database Development: Design and Analysis of the Acoustic Phonetic Corpus," *Proc. DARPA Speech Recognition Workshop*, 100–109, February, 1986.
4. V. Zue, N. Daly, J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni, S. Seneff, and M. Soclof, "The Collection and Preliminary Analysis of a Spontaneous Speech Database," *Proc. DARPA Speech and Natural Language Workshop*, 126-134, October, 1989.
5. J. Polifroni, S. Seneff, and V. Zue, "Collection of Spontaneous Speech for the ATIS Domain and Comparative Analyses of Data Collected at MIT and TI," *Proc. DARPA Speech and Natural Language Workshop*, 360–365, February, 1991.
6. M. Phillips, J. Glass, J. Polifroni, and V. Zue, "Collection and Analyses of WSJ-CSR Data at MIT," *Proc. Fifth DARPA Workshop on Speech and Natural Language*, 367–372, Harriman, NY, February, 1992.
7. J. Bernstein, K. Taussig, J. Godfrey, "Macrophone: An American English Telephone Speech Corpus for the Polyphone Project," *Proc. ICASSP*, 81–84, Adelaide, Australia, April, 1994.
8. J. Godfrey, E. Holliman, J. McDaniel, "SWITCHBOARD: Telephone Speech Corpus for Research and Development," *Proc. ICASSP*, 517–520, San Francisco, CA, March, 1992.
9. T. Staples, J. Picone, N. Arai, "The Voice Across Japan Database–The Japanese Language Contribution to Polyphone," *Proc. ICASSP*, 89–92, Adelaide, Australia, April, 1994.
10. Y. Muthusamy, E. Holliman, B. Wheatley, J. Picone, and J. Godfrey, "Voice Across Hispanic America: A Telephone Speech Corpus of American Spanish," *Proc. ICASSP*, 85–88, Detroit, MI, April, 1995.