Tools for Collecting Speech Corpora via Mechanical-Turk

Ian Lane[1,2], Alex Waibel[1,2]
[1]Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213, USA
{ianlane,ahw}@cs.cmu.edu

Matthias Eck[2], Kay Rottmann[2]
[2]Mobile Technologies LLC
Pittsburgh, PA, USA
matthias.eck@jibbigo.com
kay.rottmann@jibbigo.com

Abstract

To rapidly port speech applications to new languages one of the most difficult tasks is the initial collection of sufficient speech corpora. State-of-the-art automatic speech recognition systems are typical trained on hundreds of hours of speech data. While pre-existing corpora do exist for major languages, a sufficient amount of quality speech data is not available for most world languages. While previous works have focused on the collection of translations and the transcription of audio via Mechanical-Turk mechanisms, in this paper we introduce two tools which enable the collection of speech data remotely. We then compare the quality of audio collected from paid part-time staff and unsupervised volunteers, and determine that basic user training is critical to obtain usable data.

1 Introduction

In order to port a spoken language application to a new language, first an automatic speech recognition (ASR) system must be developed. For many languages pre-existing corpora do not exist and thus speech data must be collected before development can begin. The collection of speech corpora is an expensive undertaking and obtaining this data rapidly, for example in response to a disaster, cannot be done using the typical methodology in which corpora are collected in controlled environments.

To build an ASR system for a new language, two sets of data are required; first, a text corpus consisting of written transcriptions of utterances users are likely to speak to the system, this is used to train the language model (LM) applied during ASR; and second, a corpora of recordings of speech, which are used to train an acoustic model (AM). Text corpora for a new language can be created by manually translating a pre-existing corpus (or a sub-set of that corpus) into the new language and crowd-sourcing methodologies can be used to rapidly perform this task. Rapidly creating corpora of speech data, however, is not trivial. Generally speech corpora are collected in controlled environments where speakers are supervised by experts to ensure the equipment is setup correctly and recordings are performed adequately. However, for most languages performing this task on-site, where developers are located, is impractical as there may not be a local community of speakers of the required language. An alternative is to perform the data collection remotely, allowing speakers to record speech on their own PCs or mobile devices in their home country or wherever they are located. While previous works have focused on the generation of translations (Razavian, 2009) and transcribing of audio (Marge, 2010) via Mechanical-Turk, in this paper we focus on the collection of speech corpora using a Mechanical-Turk type framework.

Previous works (Voxforge), (Gruenstein, 2009), (Schultz, 2007) have developed solutions for collecting speech data remotely via web-based interfaces. A web-based system for the collection of open-source speech corpora has been developed by the group at www.voxforge.org. Speech recordings are collected for ten major European languages and speakers can either record audio directly on the website or they can call in on a dedicated phone line. In (Gruenstein, 2009) spontaneous speech (US English) was collected via a web-based memory game. In this system speech prompts were not provided, but rather a voice-based memory game was used to gather and partially annotate

Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, pages 184–187,
Los Angeles, California, June 2010.

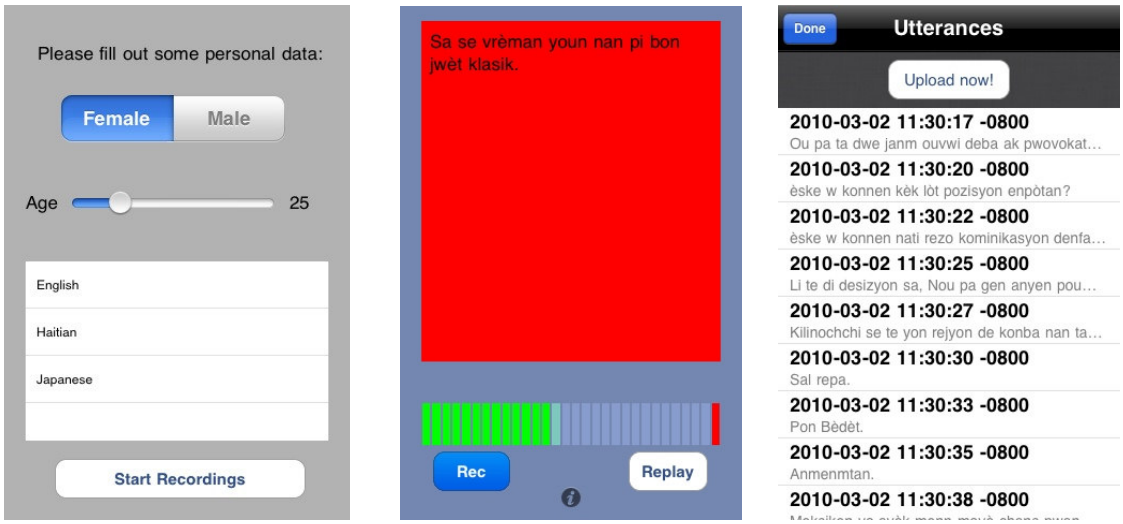

Figure 1: Screenshots from Speech Collection iPhone App

spontaneous speech. In comparison to the above works which focus on the collection of data for major languages, the SPICE project (Schultz, 2007) provides a set of web-based tools to enable developers to create voice-based applications for less-common languages. In addition to tools for defining the phonetic units of a language and creating pronunciation dictionaries, this system also includes tools to create prompts and collect speech data from volunteers over the web.

In this paper, we describe two tools we have developed to collect speech corpora remotely. The first, a Mobile smart-phone based system which allows speakers to record prompted speech directly on their phones and second, a web-based system which allows recordings to be collected remotely on PCs. We compare the quality of audio collected from paid part-time staff and unsupervised volunteers and determine that basic user training and automatic feedback mechanisms are required to obtain usable data.

2 Collection of Speech on Mobile Devices

Today's smart-phones are able to record quality audio onboard and generally have the ability to connect to the internet via a fast wifi-connection. This makes them an ideal platform for collecting speech data in the field. Speech data can be collected by a user at any time in any location, and the data can be uploaded at a later time when a wireless connection is available. At Mobile Technologies we have developed an iPhone application to perform this task.

The collection procedure consists of three steps. First, on start-up a small amount of personal information, namely, gender and age, are requested from the user. They then select the language for which they intend to provide speech data. The mobile-device ID, personal information and language selected is used as an identifier for individual speakers. Next, collection of speech data is performed. Collection is performed offline, enabling data to be collected in the field where there may not be a persistent internet connection. A prompt is randomly selected from an onboard database of sentences and is presented to the user, who reads the sentence aloud holding down a push-to-talk button while speaking. During the speech collection stage, the system automatically proceeds to the following prompt when the current recording is complete. The user however has the ability to go back to previous recordings, listen to it and re-speak the sentence if any issues are found. Finally, the speech data is uploaded using a wireless collection. Data is uploaded one utterance at a time to an FTP server. Uploading each utterance individually allows the user to halt the upload and continue it at a later time if required.

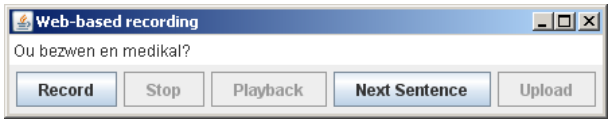

Figure 2: Java applet for Web-based recording

3 Collection via Web-based Recording

One of the most popular websites for crowd-sourcing is Amazon Mechanical Turk (AMT). "Requesters" post Human Intelligence Tasks (HITs) to this website and "Workers" browse the HITs, perform tasks and get paid a predefined amount after submitting their work. It has been reported that over 100,000 workers from 100 countries are using AMT (Pontin, 2007).

AMT allows two general types of HITs. A Question Form HIT is based on a provided XML template and only allows certain elements in the HIT. However, it is possible to integrate an external JAVA applet within a Question Form HIT which allows for some flexibility. Questions can also be hosted on an external website which increases flexibility for the HIT developer while remaining tightly integrated in the AMT environment.

For collection of audio data Amazon does not offer any integrated tools. We thus designed and implemented a Java applet for web based speech collection. The Java applet can easily be incorporated in the AMT Question-Form mechanism and could also be used as part of an External-Question HIT. Currently the Java applet provides the same basic functionality as outlined for the iPhone application. The applet sequentially shows a number of prompts to record. The user can skip a sentence, playback a recording to check the quality and also redo the recording for the current sentence (see screenshot in Figure 2).

After the user is finished, the recorded sentences are uploaded to a web-server using an HTTP Post request. An important difference is the necessity to be online during the speech recordings.

4 Evaluation of Recorded Audio

One issue when collecting speech data remotely is the quality of the resulting audio. When collection is performed in a controlled environment, the developer can ensure that the recording equipment is setup correctly, background noise is kept to a minimum and the speaker is adequately trained to use the recording equipment. However, the same is not guaranteed when collecting speech remotely via mechanical-turk frameworks.

Paid Employees	
Language	English
Number of Speakers	10
Utterances Evaluated	445
Volunteers	
Language	Haitian Creole
Number of Speakers	3
Utterances Evaluated	167

Table 1: Details of Evaluated Corpora

1	Recorded utterance is empty
2	Utterance is not segmented correctly
3	Recording is clipped
4	Recording contains audible echo
5	Recording contains audible noise

Table 2: Annotations used to label poor quality recordings

When recording prompted speech there are three types of issues that result in unsuitable data:

- **Garbage Audio:** recordings that are empty, clipped, have insufficient power, or are incorrectly segmented.
- **Low quality recordings:** low Signal-to-Noise recordings due to poor equipment or large background noise
- **Speaker errors:** Misspeaking of prompts, both accidental and malicious

To verify the quality of audio recorded in unsupervised environments we compared two sets of speech data. First, in an earlier data collection task we collected 445 prompted utterances from 10 US-English speakers. This data collection was performed in a quiet office environment with technical supervision. Speakers were paid a fee for their time. As a comparison a similar collection of Haitian Creole was performed. In this case data was collected on a volunteer basis and supervision was limited. Details of the collected data are shown in Table 1.

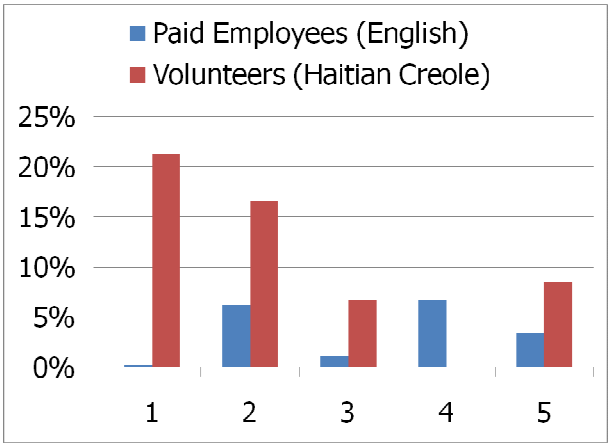

Figure 3: Percentage of recorded utterances determined to be inadequate for acoustic model training. Annotations limited to five issues listed in Table 1.

To determine the frequency of the quality issues listed above, we manually verified the two sets of collected speech. The recording of each utterance was listened to and if the audio file was determined to be of low quality it was annotated with one of the tags listed in Table 2. The percentage of utterances labeled with each annotation is shown for the English and volunteer Haitian Creole cases in Figure 3.

Around 10% of the English recordings were found to have issues. Clipping occurred in approximately 5% and a distinct echo was present in the recordings for one speaker. For the Haitian Creole case the yield of useable audio was significantly lower than that obtained for English. For all three speakers clipping was more prevalent and the level of background noise was higher. We discovered that due to lack of training, one of the volunteers had significant issues with the push-to-talk interface in our system. This led to many empty or incorrectly segmented recordings. In both cases, prompts were generally spoken accurately and technical problems caused poor quality recordings.

We believe the large difference in the yield of high quality recordings, 90% for English compared to 65% for Haitian Creole case, is directly due to the lack of training speakers received and the volunteer nature of the Haitian Creole task. By incorporating a basic tutorial when users first start our tools and an explicit feedback mechanism which automatically detects quality issues and prompts users to correct them we expect the yield of high quality recordings to increase significantly. In the near future we plan to use the tools to collect data from large communities of remote users.

5 Conclusions and Future Work

In this work, we have described two applications that allow speech corpora to be collected remotely, either directly on Mobile smart-phones or on a PC via a web-based interface. We also investigated the quality of recordings made by unsupervised volunteers and found that although prompts were generally read accurately, lack of training led to a significantly lower yield of high quality recordings.

In the near future we plan to use the tools to collect data from large communities of remote users. We will also investigate the user of tutorials and feedback to improve the yield of high quality data.

Acknowledgements

We would like to thank the Haitian volunteers who gave their time to help with this data collection.

References

N. S. Razavian, S Vogel, "The Web as a Platform to Build Machine Translation Resources", IWIC2009

M. Marge, S. Banerjee and A. Rudnicky, "Using the Amazon Mechanical Turk for Transcription of Spoken Language", IEEE-ICASSP, 2010

Voxforge, www.voxforge.org

A. Gruenstein, I. McGraw, and A. Sutherland, "A self-transcribing speech corpus: collecting continuous speech with an online educational game," Submitted to the Speech and Language Technology in Education (SLaTE) Workshop, 2009.

T. Schultz, et. al, "SPICE: Web-based Tools for Rapid Language Adaptation in Speech Processing Systems", In the Proceedings of INTERSPEECH, Antwerp, Belgium, 2007.

J. Pontin, "Artificial Intelligence, With Help From the Humans", The New York Times, 25 March 2007