The RedDots Platform for Mobile Crowd-Sourcing of Speech Data

Kong Aik Lee, Guangsen Wang, Kam Pheng Ng, Hanwu Sun, Trung Hieu Nguyen, Ngoc Thuy Huong Thai, Bin Ma, Haizhou Li

Institute for Infocomm Research, A*STAR, Singapore

{kalee,wang-g,ngkp,hwsun,thnguyun,nthhthai,mabin,hli}@i2r.a-star.edu.sg

Abstract

With ever increasing computational power and availability of broadband connectivity to the Internet, mobile devices have become pervasive and gaining more popularity. Capitalizing on the mobile-Internet infrastructure, a speech data collection platform was developed as part of the RedDots project which is dedicated to the study of speaker recognition over mobile devices. The RedDots platform consists of a centralized Web server receiving inputs from the *crowd* through mobile devices. As of the time of this writing, we have recruited speakers from 17 countries in three-month period, showing the potential of the RedDots platform to collect speech data from the worldwide population.

Index Terms: speech corpus collection, mobile crowd sourcing platform, speaker recognition

1. Introduction

The collection of speech corpora is an expensive undertaking. Data collection is typically carried out by either requesting the speakers to be onsite [1], [2] or remotely through telephone calls [3], [4], [5], [6]. The former has the benefit of a controlled environment, where the channel and acoustic environment conditions can be kept consistent across speakers. The latter allows speakers to record speech from wherever they happen to be located, which has the benefit of a potentially wider coverage with greater diversity. Coupled with an automated data collection procedure, phone-based collection method has enable large scale speech corpus collection since the 90s [3].

Recently, we have seen a rising trend of collecting speech data remotely using the Internet as it becomes more widely accessible [7], [8], [9]. In [7], [8] speech data were collected via a Web-based interface. A slightly different methodology was reported in [9], where a dedicated mobile application (app) was developed for data collection. The presence of the Internet facilitates not only data collection but also other related tasks like data labeling and transcription, assessment of speech technology, speech perception study, where non-expert intelligence or inputs are consolidated from the crowd [10].

Another benefit of Internet-based data collection is its capability to extend beyond specific regions at a much lower cost compared to telephone-based approach. For the RedDots project[1], we focus on English speakers, both native and non-native, recruited worldwide. This is made possible through the use of a recording front-end consisting of an application running on mobile devices communicating with a centralized Web server at the back-end. Speech recordings are collected by having speakers read text prompts displayed on the screen of the mobile devices. Speakers record their voices offline and later on upload the recordings to an Web server when Internet connection is available. This has the benefit of enabling data to be collected in the field in the absence of a persistent Internet connection.

2. Collection procedure

The flow chart in Fig. 1 illustrates the process flow from speaker registration to data collection. Among others, the online registration form, mobile application and Web server form the three main functional blocks of the data collection infrastructure.

Speakers are recruited through personal contacts and various mailing lists. To begin with, prospective participants fill in an online registration form[2]. The information collected includes e-mail address, gender, age group, country of residence, native language, and user consent to allow their voice recordings to be made available for research purpose. The information received via the online form is verified by a moderator. Upon a successful registration, a user ID paired with a log-in passcode, a user guide and the download link for the mobile app are made available to the new recruit via e-mail.

On the first log-in, a designated list of sentences will be retrieved from the server. Once the list of sentences has been downloaded to the mobile app, recording can be carried out off-line. Speech recordings are collected by having speakers reading text prompts displayed on the screen of the mobile devices. The recordings are uploaded when network connectivity is available. A weekly e-mail is sent to all participants to give an update on the progress of the project. This has the benefit of sustaining the interest of the participants if the data collection is to be conducted over a long time span, for example, a few months to a year.

3. Client-server architecture

Also shown in Fig. 1 is the technology stack consisting of a set of software libraries and standards that provides the infrastructure for mobile crowd-sourcing platform. The mobile clients act as front-end recording devices communicating with a Web server at the backend through the Internet. The communication between the mobile devices and Web server is accomplished using the Hypertext Transfer Protocol (HTTP). In addition, the communication channel is encrypted through a Secure Sockets Layer (SSL). Once a secure communication channel is set up, text message and speech data could be sent from the client to the server. Example text messages sent from the client are the user ID and passcode. On the server side, the user ID and passcode received from the client are verified against the user credentials stored in the SQL database and a success or fail status is re-

[1] http://www1.i2r.a-star.edu.sg/~kalee/RedDots/

[2] http://goo.gl/forms/2KmktzgVV9

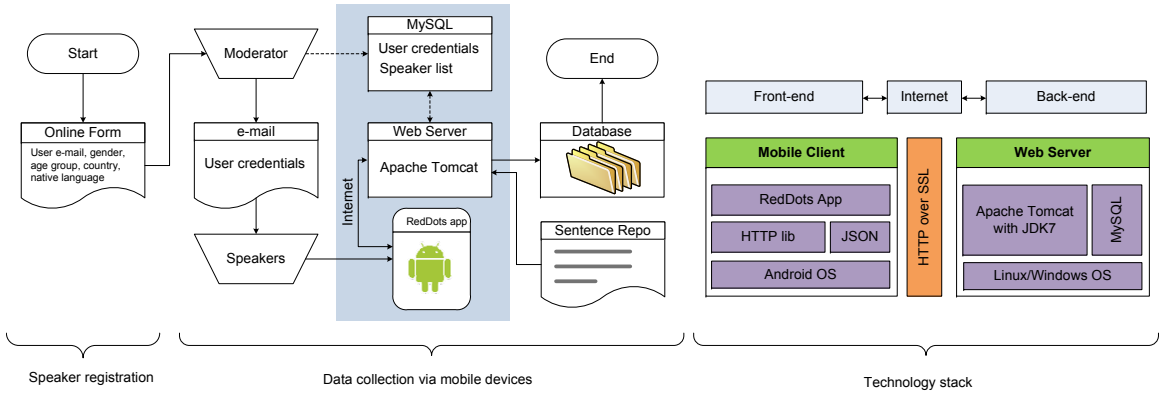

Figure 1: *The acquisition protocol, process flow, and platform for data collection over mobile devices.*

turned to the client. In our implementation, the text messages are formated in the JSON (JavaScript Object Notation) format.

Apart from the user credentials, speech recordings together with meta-data (time stamp, mobile device information, etc.) are also sent to the Web server in the same HTTP session. These two streams of data are stored in separate files – a raw PCM file for the speech data and an ASCII text file for the meta-data. In the database on the server side, speech recordings are tagged to the user ID.

4. Sentence repository

The RedDots project aims to collect speech data preliminary for text-dependent speaker recognition research. One typical application is access control where users are required to pronounce a given passphrase. Considering this use case, the platform was designed to collect read speech data where speakers pronounce sequence of sentence prompts.

In our current implementation, the sentence repository consists of 116,010 unique sentences. These sentences are stored in a single text file referred to as the *master* list. Each speaker is assigned a subset of sentences referred to as the *speaker* list, which are tagged to the user ID in the SQL database. To save memory, the speaker list contains only the indices corresponding to the entries in the master list. Upon the first login, sentences are retrieved from the master list and pushed to the client according to the sentence indices in the speaker list. The sentences are downloaded once and stored locally on the mobile device. This mechanism allows recordings to be done offline anytime everywhere. The recordings are uploaded when Internet connection is available. Using the indices in the speaker list, we could then map each recording to the master list to get the corresponding transcription of the utterance.

5. Conclusions

We have reported the core elements of a client-server platform for speech data collection over mobile devices. As of the time of this writing, we have recruited 67 speakers from 17 countries in 3 months period (see Fig. 2). This shows the potential of the RedDots platform for speech data collection covering a worldwide scope. Though we focus on speech data collection, one possible extension to the current platform is the collection of face images and videos since camera is readily available on mobile devices.

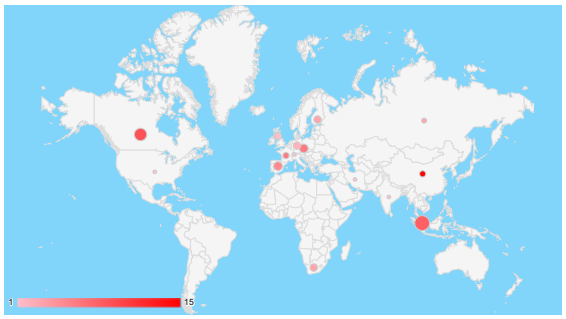

Figure 2: *Distribution of speakers per country as of April 13 2015.*

6. References

[1] V. Zue, S. Seneff, and J. R. Glass, "Speech database development at MIT: Timit and beyond," *Speech Communication*, vol. 9, no. 4, pp. 351–356, 1990.

[2] A. Larcher, K. A. Lee, B. Ma, and H. Li, "Text-dependent speaker verification: Classifiers, databases and RSR2015," *Speech Communucation*, vol. 60, pp. 56 – 77, 2014.

[3] B. Wheatley and J. Picone, "Voice across America: toward robust speaker-independent speech recognition for telecommunications application," *Digital Signal Processing*, vol. 1, no. 2, pp. 45–63, 1991.

[4] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *Proc. ICASSP*, 1992, pp. 517–520.

[5] J. Bernstein, K. Taussig, and J. Godfrey, "Macrophone: an American English telephone speech corpus for polyphone project," in *Proc. ICASSP*, 1994, pp. 81–84.

[6] C. Cieri, L. Corson, D. Graff, and K. Walker, "Resources for new research directions in speaker recognition: the Mixer 3, 4 and 5 corpora," in *Proc. INTERSPEECH*, 2007, pp. 950–953.

[7] "LibriVox - free public domain audiobooks," https://librivox.org/, accessed: 2015-02-27.

[8] "VoxForge," http://www.voxforge.org, accessed: 2015-02-27.

[9] I. Lane, A. Waibel, M. Eck, and K. Rottmann, "Tools for collecting corpora via Mechanical Turk," in *Proc. NAACL HLT*, Jun. 2010, pp. 184–187.

[10] G. Parent and M. Eskenazi, "Speaking to the crowd: looking at the past achievements in using crowdsourcing for speech and predicting future challenges," in *Proc. INTERSPEECH*, Aug. 2011, pp. 3037–3040.