Creative Commons Speech Repository
UNSW-CSE-TR-0441

Mohammed Waleed Kadous
School of Computer Science & Engineering
University of New South Wales
Sydney, NSW 2052, Australia
E-mail: waleed@cse.unsw.edu.au

December 10, 2004

Abstract

The availability of high quality open source solutions has often led to rapid growth of a particular area; most famously, the LAMP (Linux-Apache-MySQL) platform for Web development that has made it easy for small entities to set up rich web sites easily. Recent developments in open source speech recognition software, notably the Sphinx 4 system, mean that potentially the same could occur for speech. However, one outstanding issue is the availability of high-quality acoustic models required for such systems to function effectively.

This report outlines the possibilities for a new model for speech corpus repository, based on the principles of Open Source and the Creative Commons licenses.

Rather than the usual approach for gathering corpora that are done in a centralised location, using fixed high-quality equipment, we propose that donors record audio passages using their own equipment and upload them to a central repository for processing. The data in the speech repository could be used for training acoustic models for speech recognition by anyone.

In addition, such a speech repository, especially when combined with co-located computational processing facilities, such as a Beowulf cluster, create an opportunity for new applications, for example, the possibility of personalised acoustic models. Such a system would customise an acoustic model for a particular user, not just using that person's data, but by using people whose voices are similar.

Keywords: speech recognition, speech corpus, open source, creative commons.

1 Introduction

Many of today's speech recognition systems, particularly those involving large vocabulary continuous speech recognition, require large amounts of recorded speech data to train them to sufficient accuracy to classify speech correctly.

For this reason, speech corpora (collections of recording of speech) are very important for speech recognition. Companies go to great length to record high-quality audio of people speaking from a variety of different speech backgrounds and accents to develop their speech products. Organisation such as the Linguistic Data Consortium (LDC) have also come in to existence to distribute and act as repositories for speech corpora.

However, sometimes such speech corpora are held under extremely strict or limiting license, if at all. For example, many of the LDC corpora are licensed for only one researcher to analyse the results, and licenses are costly, on the order of several hundred dollars per CD of audio data.

The main reason for this is the cost of recording the data. Most of the recordings are done under extremely good sound conditions and high-quality equipment.

However, there are a number of significant technological developments that could help to solve this problem by devolving the recording of speech samples to individuals:

- **High-quality home audio equipment**: Many users now have high quality equipment installed as part of their computers, that can quite adequately record at 22kHz in 16-bit resolution with a signal to noise ratio in excess of 80dB. Many also have high-quality noise-cancelling microphones for either gaming, speech recognition or Internet telephony.
- **Popularity of broadband**: Recordings of speech tend to be large. One hour of compressed lossless audio takes up approximately 75 megabytes of space[1], and consequently over a modem would take approximately 7.5 hours to upload to a repository. However, with broadband's increasing popularity, this can be singificantly, reduced; for instance, using a typical ADSL connection (128K/s upstream), it would take approximately 1.3 hours.
- **Cheap storage**: Hard disk storage is now at the point where it is below A$1 per gigabyte. Even allowing for backups via RAID arrays, etc., this allows for many people's recordings to be stored in a cost-effective manner. Assuming 1 hour of audio per user, a simple file server offering 1 terabyte of data and costing less than A$4,000 could store data for more than 10,000 users.

We therefore propose the following idea: that average users could be provided with a sequence of sentences to read; that they could provide information about

[1] 22kHz x 2 bytes/sample x 60 seconds x 60 minutes with a compression ratio of 0.5 gives 75 megabytes

their speech background (native language, geographical location etc) and could then record audio of themselves speaking. This could then be uploaded to a central repository, where their data could be used by speech researchers to build better voice models.

These audio recordings would be made available under a Creative Commons [8, 1] license with the following conditions:

- The data could be used for commercial or non-commercial purposes freely without licensing fee.
- Using the data would require attribution.
- Any derived works (namely speech models developed using the data), i.e. hidden Markov model parameters, phoneme models etc. would also have to be released under a Creative Commons license.

The development of a creative commons speech repository could lead to the creation of a viable open source large vocabulary speech recognition solution.

In addition to improving speech recognition, there are many other avenues of research open; for example, for existing speech recognition products, suppliers are forced to provide a "one-size-fits-all" (or at the very least, a "one-size-fits-many") acoustic model. By using a speech corpus consisting of potentially thousands of people, there is the potential to generate "customised" acoustic models to suit particular groups, e.g. female Australian accent with a Chinese inflection.

2 Components Required

A project such as this requires assistance across a wide breadth of speech, systems and network aspects.

2.1 Text corpus and database

To simplify comparison between different users for research purposes, it is necessary that a particular corpus of sentences be selected and used. Initially, there might be only one corpus, but there should be the facility to support multiple corpora.

Also very important is the amount of information about the person doing the recording, their audio setup, etc that needs to be stored with each user. Data of this kind really can be classified into two groups: those associated with the user and those associated with the recording process.

One initial possibility is to use as a model an existing database, such as ANDOSL [10].

2.2 Recording software

The recording software would:

- Assist the user in setting up appropriate volume levels for their equipment and let them know things such as the signal-to-noise ratio.
- Prompt the user for something to say.
- Record the audio, and label it appropriately.
- Manage the compression and transmission of audio back to the server.

2.3 Audio compression algorithms

Obviously, for quality reasons, lossy compression algorithms should not be used on the recorded speech data. However, there is nothing to preclude the use of lossless speech compression algorithms. So far, research on lossless compression for algorithms for speech applications is still developing. However, there are several general purpose lossless audio compression tools. According to many reports Monkey's Audio [2] is one of the best performing in terms of compression, but only slightly behind is FLAC (Free Lossless Audio Codec) [3], which has the advantage that it's cross-platform and open source.

In either case, typical reported compression rates are on the order of 0.5.

2.4 Server-side training tools

There are several reasons why the server need not just be a repository. It may also be a processing cluster with high-speed access to the data. Therefore it might be possible to create a system where users could, over the Web, select a subset of datasets for training, based on certain HMM parameters, and then generate an acoustic model based on the selected subset of instances. This could then be downloaded and/or evaluated by the user.

3 Background

3.1 Speech Corpora

There are many speech corpora available over the Web. However, they are often available only under strict licenses. Furthermore, they are frequently either small in (a) the number of people using the system or (b) the duration of recording.

One of the largest repositories of speech corpora is the Linguistic Data Consortium [9]. In the Australian context, there is the Australian National Database of Spoken Language (ANDOSL) [10].

3.2 Open Source/Creative Commons Licenses

One of the most interesting developments of the last decade has been the development of Free and Open Source Licenses; and more recently the development of the Creative Commons License.

Open Source Licenses, of which the most popular is the GNU Public License or GPL [4] , allows the release of source code under a number of freedoms. These freedoms are:

- The freedom to run the program, for any purpose.
- The freedom to study how the program works, and adapt it to your needs (freedom 1). Access to the source code is a precondition for this.
- The freedom to redistribute copies so you can help your neighbor (freedom 2).
- The freedom to improve the program, and release your improvements to the public, so that the whole community benefits (freedom 3). Access to the source code is a precondition for this.

While these conditions are appropriate for software, it is less clear how they apply in the context of data, or creative content. As a reaction to this Lessig developed the Creative Commons license [1]. It is a legal framework that provides an alternative to copyright. There are several variant licenses, differing in the following dimensions:

- **Attribution:** Does attribution need to be made to the original work?
- **Commercial use:** Is commercial use of the work allowed?
- **Derivative works:** Are derivative works allowed, and if so, does the derivative work also have to be released under a Creative Commons license (often referred to a ShareAlike clause).

The most appropriate license for this project would seem to be the Attribution-ShareAlike license [5]. This would require that any users of the data attribute it, commercial use is permitted, and that any derivative works also be released under a similar license.

3.3 Existing Software

Some of the component software parts are already available, and mostly under open source licenses. EMU [7] for example, is a piece of software designed for speech annotation, but also includes capabilities for recording and capturing data.

There are a number of audio codecs, as previously mentioned, such as FLAC and Monkey's Audio, designed for lossless encoding of audio.

As far as speech recognition is concerned, the Sphinx 4 toolkit, which is being worked on by Sun, HP, and CMU [6] is likely to be used as the platform for which derived acoustic models will be distributed. The Sphinx Train tools will be used for training the data.

4 Issues

There may be a number of issues to do with the implementation of the system.

4.1 Privacy

One of the greatest concerns is the privacy of people who donate their voice recordings to CCSR. It is important to provide some avenue of communication with the donors, while protecting their privacy. Some donors may be sensitive about the potential for their voice data to be used for concatenative speech synthesis.

4.2 Sample bias

The distribution of donors is unlikely to match that of the general populace, and would typically match what is often called the "Slashdot demographic", i.e., young, technically savvy but ethnically diverse males.

4.3 Longevity and stability

One concern is that any funding sources to create the repository might at one point "dry up". The solution to this is mirroring of the repository in multiple locations. While it may at first appear to be prohibitively expensive to mirror a terabyte of data, it is now possible to use hard disks as a viable transport medium. Including parity, 5 drives at a cost of some A$250 each are sufficient to copy a terabyte of data.

4.4 Quality of recordings

One issue is that compared to high-quality audio recordings taken in specialised studios, data recorded by people in their homes may not be of adequate quality.

This is definitely the case. However, for speech recognition applications, as distinct from phonological and speech annotation applications, this may be less of an issue. Many of the speech corpora available such as TIDIGITS were recorded over telephones.

Furthermore, some of the research has shown that *injecting* noise as part of the training process often improves accuracy. Therefore, it may be that the data is of sufficient quality for effective speech recognition.

4.5 Apathy

What would lead people to want to upload their audio data to a repository? Aside from altruism and enlightened self-interest, there are other possible factors that could motivate someone to upload data to a server. One possibility is that in exchange for uploading data onto the server, we provide the donor with a customised acoustic model that they can "plug in" to their open source speech recognition system locally that should function more effectively for their voice than the "standard" package.

5 Practical outcomes

The CCSR would have a number of practical outcomes:

5.1 High-quality open source speech recognition systems

One thing holding back wider deployment of speech recognition systems is that there is no high-quality out-of-the-box speech recognition systems appropriate for personal use (e.g. for dictation). With tools like Sphinx 4 being developed, this is not a problem of software; but rather one of corpus.

It is not that corpora that come with Sphinx 4 are poor, it is rather that they are too specific: they are for American English, for example. Its performance on, say, Australian English is not very good, let alone other languages.

By developing a Creative Commons corpus and deriving acoustic models that can be used together with these open source recognition systems, the scope for the use of such speech recognition tools is greatly expanded.

5.2 An extensive corpus for research

A corpus such as this would, despite the quality problems, still form an interesting database for the exploration of speech.

6 Research potential

In addition to the practical benefits mentioned above in terms of enabling future speech research, there are several other research opportunities present in the work.

6.1 Personalised acoustic model building

Given a sample of a person's speech, and a massive database of thousands of other users, and a knowledge of the demographic characteristics of that person, is it possible to select a subset of other voice models, combine them, and then return them to the user? How would such a system compare with a speech model built over the entire population?

One possibility particularly worthy of exploring is the potential to use data mining and machine learning techniques to detect which features are most important in determining whether a particular speaker has a "vocal analogue"?

7 Conclusion

This project has the potential to enable open source speech recognition to expand and grow rapidly in cooperation with existing projects. Interested parties are encouraged to contact the author with either positive or negative constructive criticism.

References

[1] Web: http://www.creativecommons.org/.

[2] Web: http://www.monkeysaudio.com/.

[3] Web: http://www.flac.org/.

[4] Web: http://www.gnu.org/.

[5] Web: http://creativecommons.org/licenses/by-sa/2.0/.

[6] Web: http://cmusphinx.sourceforge.net/sphinx4/.

[7] Steve Cassidy and J. Harrington. Multi-level annotation in the emu speech database management system. *Speech Communication*, 33:61–77, January 2001.

[8] Lawrence Lessig. The creative commons. *RBL (Tokyo)*, 2003.

[9] Mark Lieberman and Christopher Cieri. The creation, distribution and use of linguistic data. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, 1998.

[10] J. B. Millar, P. Dermody, J. M. Harrington, and J. Vonwiller. A national spoken language database: concept, design, and implementation. In *Proceedings of International Conference on Spoken Language Processing (ICSLP-90)*, pages 1281–1284, 1990.