Using question-specific vocabularies with Salaam to support speech data collection

Kayokwa N Chibuye
University of Cape Town
Rondebosch, Cape Town, 7700
South Africa
chbnic001@myuct.ac.za

Brian DeRenzi
University of Cape Town
Rondebosch, Cape Town, 7700
South Africa
bderenzi@cs.uct.ac.za

Todd Rosenstock
World Agroforestry Centre
(ICRAF)
Kenya
t.rosenstock@cgiar.org

Abstract

Research suggests that voice-based systems may seem to be the only practical option for self-completion surveys in developing regions where a substantial proportion of respondents are illiterate. Voice user interfaces cater for the information needs of speakers of low-resource languages, whether or not these languages have a formal writing script. However, the language resources needed to train speech recognition engines are either limited or completely non-existent for languages in Africa. Historically, the process to obtain or construct new language resources is also long and costly. In spite of this, techniques have been developed that enable a speech recognition engine from a high resource and well-trained language to be used for speech recognition in a new low-resource language by leveraging the similarity of sounds between the two languages. Presented here are our on going efforts to establish the limitations and suitability of using this technique to support speech data collection in rural Zambia with a focus on Bantu languages.

Author Keywords

Automatic Speech Recognition, SALAAM, resource-scarce languages, small-vocabulary speech recognition.

ACM Classification Keywords

H.5.m [Information interfaces and presentation (e.g., HCI)]:

Miscellaneous; See [http://acm.org/about/class/1998/]: for full list of ACM classifiers. This section is required.

Introduction

Data collection via voice has become popular amongst research projects in poor resource countries over the past decade [7, 4, 9].

For example, the World Bank's Listening to Africa project uses this mode of data collection as a complement of paper-based household surveys in Madagascar, Malawi, Mali, Senegal, Tanzania, and Togo[1].

Voice data collection is especially popular because its flexibility [1]. In the case of a call center setup, an interviewer can use different languages during a session. Voice based interviews also accommodate illiterate respondents. An interviewer can ask complex questions and clarify matters that the respondent might find confusing. Instead of revisiting respondents, a supervisors can simply re-call respondents for quality control purposes. These reasons as well as the ability to accommodate respondents owning low-end phones has made voice data collection the technology of choice for most of sub-Saharan Africa [3, 1].

Automatic speech recognition has been demonstrated to be very useful in the design of speech driven interfaces for applications, especially in the developing world where literacy levels are low [4, 9, 1].

The main benefit of their use is their reach to a large illiterate or semi-literate population of the developing world where text-based interfaces may not be very useful [1]. Unfortunately, the languages spoken in these regions typically lack adequate resources needed to train speech recognition engines [8]. The process to train a speech recognition engine has been known to be difficult. It is expensive and demands expert knowledge in speech technology and linguistic expertise in the local language of interest, all of which are lacking in developing regions [5, 8, 10].

However, recent advances suggest that one can use an existing speech recognizer for a high resource language (HRL), such as French, to achieve small-vocabulary recognition tasks in a low-resource language (target language). This can be achieved by leveraging the similarity of sounds (phonemes) between the two languages though a technique called cross-language phoneme mapping. A phoneme is the simplest unit of sound in a language. By generating a pronunciation lexicon, representing the pronunciation of target language words based on the phonemes of the source language, speech recognition of the target language vocabulary can be achieved. Mapped pronunciations can be handwritten but they demand the use of an expert linguist who is fluent in both the source and target languages [10]. Considering this is a resource lacking in developing regions, the process of creating cross-language pronunciations have been developed [6]. Eliminating the need of linguist experts.

Speech-based Accent Learning And Articulation Mapping (SALAAM) is a technique that can automatically generate the aforementioned mapped pronunciations from a handful of training data and achieve high quality recognition accuracy [6]. SALAAM automatically generates a pronunciation lexicon that represents how a word (target language word) can be phonetically represented using the phonetic dictionary of another language (high resource language) that an underlying speech recognition engine is trained in. For example, "SAAN", might be the cross-language mapping result of the Mandarin (target language) word for "three"

[1] https://blogs.worldbank.org/africacan/measuring-the-pulse-of-africa-one-phone-call-at-a-time

using Microsoft Speech Platform U.S English (source language) recognizer [2]. The pronunciation lexicon produced by SALAAM can then be used with a speech recognizer to support speech recognition of resource-scarce language words they contain. This makes it easier for developers with no speech technology expertise to quickly develop small-vocabulary speech driven applications for low resource languages.

SALAAM has been used to support VUIs in a number of agriculture, health and entertainment related projects in India and Pakistan [9, 4, 7]. In [1] a mobile video search application that was developed for farmers in India to ease the distribution of agriculture best-practice demonstration videos. SALAAM was used to support voice user interaction of the application in Hindi, achieving up to 90% recognition accuracy. This demonstrates that with careful application design, cross-language phoneme techniques can be used to support speech data collection. The techniques require minimal training data and support quick VUI development.The process is cheap and relatively easy [6, 9].

Speech recognition for low-resource languages using SALAAM

Our work aims to establish the limitations and suitability of using cross-language phoneme mapping techniques to support speech data collection in low-resource languages, with a focus on Bantu languages. Our interest in Voice User Interfaces (VUIs) as human-computer interfaces stems from their ability to enhance data collection in developing regions were text-based interfaces are ineffective because most of the population is illiterate. [1]. To support our research, we chose an open source implementation of SALAAM called Lex4All [11].

In order to use Lex4all for lexicon generation, a user needs to submit a name of a word and the audio file(s) containing a pronunciation of that word. Once submitted, the program passes the audio file(s) and name to the SALAAM algorithm through the grammar control. The pronunciation generation is based on the language model the underlying speech engine is trained in, such as U.S English. A pronunciation lexicon is then given to the user as output. The user can then use it to support speech recognition of the target language words it contains. Figure 1 shows a simplified illustration of lexicon generation with SALAAM.

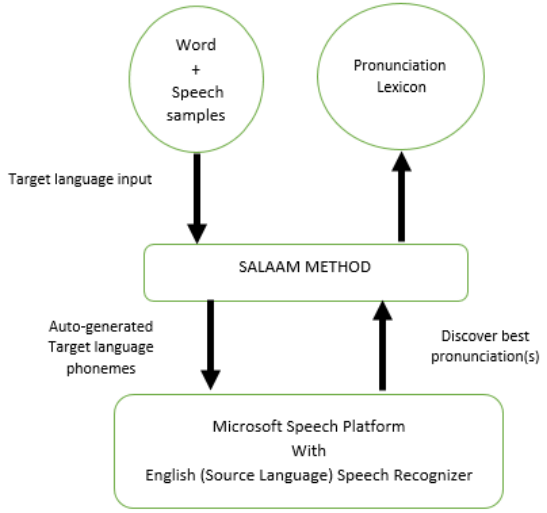

Figure 1: Lexicon generation with SALAAM.

A user can generate a lexicon containing multiple words and their respective pronunciations at the same time [11, 6].

An example of a pronunciation lexicon generated is shown in figure 2 below:

```
<?xml version="1.0" encoding="utf-8" standalone="no"?>
<lexicon version="1.0" xml:lang="en-US" alphabet="x-microsoft-ups"
xmlns="http://www.w3.org/2005/01/pronunciation-lexicon">
    <lexeme>
        <grapheme>amenshi</grapheme>
        <phoneme>SH EI S S</phoneme>
        <phoneme>SH E S S</phoneme>
        <phoneme>SH EI S</phoneme>
        <phoneme>SH E S</phoneme>
    </lexeme>

    <lexeme>
        <grapheme>umuntu</grapheme>
        <phoneme>U</phoneme>
        <phoneme>U U</phoneme>
        <phoneme>JH U</phoneme>
        <phoneme>Z U</phoneme>
    </lexeme>
</lexicon>
```

Figure 2: Lexicon generated by SALAAM using Lex4all.

The example above shows a pronunciation lexicon of two words, *amenshi* (a Bemba word for water) and *umuntu* (a Bemba word for person). Each word is represented as a lexeme, a combination of a grapheme and phonemes. The names of the words submitted during the lexicon generation stage are indicated in the grapheme element. The generated pronunciations are represented by the phonemes that each grapheme contains.

The pronunciation lexicon is then used in combination with application grammar to support speech recognition as shown in figure 3 below:

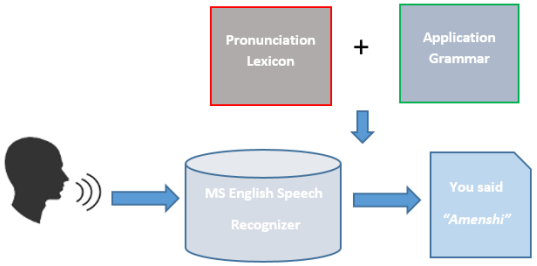

Figure 3: Use of pronunciation lexicon.

Lexicon evaluation is similar to lexicon generation. However, The system then gives the number of incorrect, correct or unrecognized words. In our study, when a recognizer is tested for recognition accuracy of a word, there are three potential answers; correct, incorrect and unrecognized.

A word is said to be correct if the name (grapheme) of the word and its pronunciation (phoneme) have been found in the recognizer's phoneme inventory. A word that is said to be incorrect is one with a matching sound (pronunciation) but wrong name (grapheme) or vice versa. A word is said to be unrecognized if neither the name nor the sound match the recognizer's phoneme inventory.

Evaluation of recognition accuracy

Our work will take a quantitative approach to recognition accuracy evaluation by looking at three aspects of it: (1) number of correctly recognized words, (2) number of incorrectly recognized words and (3) number of unrecognized words. This information will be recorded for each experiment as we vary the following test conditions: (1) gender of the participants during training and evaluation (2) vocabulary size of

each target language during evaluation (3) sample rates of the audio data (4) source language choices (5) and type of training technique.

Our study shall use iciBemba, ciTonga, ciNyanja/Chichewa and siLozi, four of six major local languages spoken in rural Zambia. Some of these languages are also spoken in Malawi (Chichewa), Democratic Republic of Congo (iciBemba) and Zimbabwe (Chichewa and ciTonga). The Bantu language family was chosen to aid our research for two reasons: (1) the SALAAM technique has never been used to support speech recognition with the language family [6, 9, 7] and (2) Bantu languages are widely spoken languages in rural Africa. Apart from US English and French, our study shall employ German and Mandarin, source languages that have never been used with SALAAM [5, 9, 1]. We plan to collect the audio data from Lusaka, Zambia which provides a centralized region where many of the native speakers of the aforementioned Bantu languages live.

Conclusion

Speech recognition technology can further enhance data collection methods and survey designs especially for developing regions where text-base interfaces are ineffective. It can be used to support self-completion surveys and used in the development of spoken dialog systems. Using cross-language phoneme mapping techniques, researchers and practitioners can achieve high quality speech recognition in low-resource languages in a short amount of time, cheaply and without a lot of effort. This consequently helps to foster development through information dissemination in areas such as health and agriculture. It also enhances the collection of actionable data such as the nutritional status of a population sample of a country or region.

References

[1] Kalika Bali, Sunayana Sitaram, Sebastien Cuendet, and Indrani Medhi. 2013. A Hindi speech recognizer for an agricultural video search application. In *Proceedings of the 3rd ACM Symposium on Computing for Development*. ACM, 5.

[2] Hao Yee Chan and Roni Rosenfeld. 2012. Discriminative pronunciation learning for speech recognition for resource scarce languages. In *Proceedings of the 2nd ACM Symposium on Computing for Development*. ACM, 12.

[3] Kevin Croke, Andrew Dabalen, Gabriel Demombynes, Marcelo M Giugale, and JGM Hoogeveen. 2012. Collecting high frequency panel data in Africa using mobile phone interviews. *World Bank Policy Research Working Paper* 6097 (2012).

[4] Somani Patnaik, Emma Brunskill, and William Thies. 2009. Evaluating the accuracy of data collection on mobile phones: A study of forms, SMS, and voice. In *Information and Communication Technologies and Development (ICTD), 2009 International Conference on*. IEEE, 74–84.

[5] Fang Qiao, Roni Rosenfeld, and Jahanzeb Sherwani. 2010a. Layperson-Trained Speech Recognition for Resource Scarce Languages. (2010).

[6] Fang Qiao, Jahanzeb Sherwani, and Roni Rosenfeld. 2010b. Small-vocabulary speech recognition for resource-scarce languages. In *Proceedings of the First ACM Symposium on Computing for Development*. ACM, 3.

[7] Agha Ali Raza, Rajat Kulshreshtha, Spandana Gella, Sean Blagsvedt, Maya Chandrasekaran, Bhiksha Raj, and Roni Rosenfeld. 2016. Viral spread via entertainment and voice-messaging among telephone users in india. In *Proceedings of the Eighth International Con-*

ference on Information and Communication Technologies and Development. ACM, 1.

[8] Frederick Webera Kalika Balib Roni Rosenfeldc and Kentaro Toyamab. UNEXPLORED DIRECTIONS IN SPOKEN LANGUAGE TECHNOLOGY FOR DEVELOPMENT. (????).

[9] Jahanzeb Sherwani, Sooraj Palijo, Sarwat Mirza, Tanveer Ahmed, Nosheen Ali, and Roni Rosenfeld. 2009. Speech vs. touch-tone: Telephony interfaces for information access by low literate users. *ICTD* 9 (2009), 447–457.

[10] Anjana Vakil and Alexis Palmer. 2014. Crosslanguage mapping for small-vocabulary ASR in under-resourced languages: Investigating the impact of source language choice. In *Spoken Language Technologies for Under-Resourced Languages*.

[11] Anjana Vakil, Max Paulus, Alexis Palmer, and Michaela Regneri. 2014. lex4all: A language-independent tool for building and evaluating pronunciation lexicons for small-vocabulary speech recognition.. In *ACL (System Demonstrations)*. 109–114.