Interspeech 2018
2-6 September 2018, Hyderabad

Cross-language Phoneme Mapping for Low-resource Languages: An Exploration of Benefits and Trade-offs

Nick K Chibuye[1]*,Todd S Rosenstock*[2]*,Brian DeRenzi*[1]

[1]University of Cape Town
[2]World Agroforestry Centre (ICRAF)

chbnic001@myuct.ac.za, t.rosenstock@cgiar.org, bderenzi@cs.uct.ac.za

Abstract

Voice-based systems are an essential approach for engaging directly with low-literate and underrepresented populations. Previous work has taken advantage of high-resource speech recognition technology for low-resource language speech recognition through cross-language phoneme mapping. Unfortunately, there is little guidance in how to deploy these systems across a range of languages. We present a systematic exploration of four source languages and five target languages to understand the trade-offs and performance of different source languages and training techniques. We find that one can improve recognition accuracy by selecting a source language that has similar linguistic properties to that of the target language. We also find that the number of alternative pronunciations per word and gender of participants also impact recognition accuracy. Our work will allow other researchers and practitioners to quickly develop high-quality small-vocabulary speech-based applications for under-resourced languages.

Index Terms: speech recognition, low-resource languages, human-computer interaction, cross-language phoneme mapping, spoken dialog systems, SALAAM, spoken language processing, nutrition

1. Introduction

Across the world, low-literate users have been categorically excluded from many of the benefits of digital and computing technology [1]. Globally, 750 million people are considered low-literate, with 27% of the world's low-literate adults coming from sub-Saharan Africa. Traditionally, there have been three main approaches to improving the use of digital technology amongst low-literate users: mediated input [2], where a literate user assists a low-literate user; graphical user interfaces [3], which rely heavily on iconography and graphical representations as opposed to text; and speech-based systems [4, 5, 6], which allow users to interact using audio and voice. The latter has the benefit of reaching people directly without the need of third party (important for sensitive information) or a smartphone (critical as smartphone penetration in Africa is still limited outside urban centres) for a graphical interface. However, voice-based systems are largely constrained to major languages, which limit their utility for much of Africa and other multilingual societies. There is a need to develop ways to adapt voice-based systems to be able to reach low-literate populations in their mother tongue.

Automatic speech recognition (ASR) plays a key role in the design and development of speech-driven interfaces for spoken dialogue systems. Oftentimes, spoken dialog systems and ASR technologies can be used as tools to bridge the gap between the low-literate populations of developing regions and information technology [7]. With the widespread adoption of the mobile phone in low-income countries across the world, the use of speech technology represent an increasingly feasible approach to directly reach the large low-literate populations in low-income countries [8].

However, the majority of languages spoken in these settings lack adequate resources needed to train speech recognition engines [9]. Training a speech recognition engine is expensive and demands a deep understanding of speech technology and linguistic expertise in the local language of interest, all of which are more difficult to find in regions with low-resource languages [7, 9, 10]. These constraints makes it difficult to develop applications suitable for the populations who need them most.

However, recent work has demonstrated that a speech recognizer trained in a high-resource language (HRL)–such as English or French–can be repurposed to achieve small-vocabulary automatic speech recognition in a low-resource language (LRL) by using similarity of sounds (phonemes) between the two languages [11]. This process is known as cross-language phoneme mapping. Using this technique, a pronunciation lexicon representing the pronunciation of target language word types based on the phonetic alphabet of the HRL is generated and used to achieve speech recognition over the LRL vocabulary [10]. These pronunciation maps could be handwritten, but they demand the use of an expert linguist who is fluent in both the source and target languages but often do not yield high recognition accuracy [10]. Therefore, processes of automatically creating cross-language phoneme mappings between languages were developed [12, 13, 14], eliminating the requirement of an expert linguist. This technique has been used in a number of information, communication, and technology for development (ICT4D) projects in the health sector [4], agriculture sector [8, 6] and for research purposes [14, 10]. Despite the development and use of the automated cross-language phoneme mapping, there remains little guidance of how the approach behaves in different conditions, i.e., different source-target language pairings, and training techniques.

The aim of this paper is to report on an investigation into the performance and tradeoffs of different source languages and training techniques for speech-based applications that use cross-language phoneme mapping.

2. Related Work

Our work builds directly on previous work, namely the Speech-based Automated Learning of Accent and Articulation Mapping (SALAAM) [12, 11] algorithm, as implemented in the open source tool Lex4All [14, 13].

The SALAAM technique was developed to facilitate small-vocabulary speech recognition for under-resourced languages by using cross-language phoneme mapping and a high-resource

10.21437/Interspeech.2018-2454

language speech recognizer [10].

The primary idea behind the SALAAM technique is to find the best pronunciation sequence for a given word in a target language from one or more audio samples by using a source language speech recognizer to perform phone decoding (decoding by phoneme) [10]. Since most commercial speech recognizers do not directly support phone decoding, the SALAAM technique uses a specially-designed grammar to mimic phone-decoding [10, 13]. This is achieved by creating a recognition grammar representing a phoneme *super wildcard* to guide pronunciation discovery. The grammar enables the speech recognizer to break down a word in the target language into a series of one to ten 'sounds'. Each of these 'sounds' are then matched a sequence of one to three source language phonemes [10, 12]. The SALAAM heuristic accepts, as input, a set of audio samples of the same word or short phrase and a requested number of k pronunciations. Using an iterative process, the heuristic builds a set of phoneme strings, returning the top-k performing pronunciations based on the phonetic inventory of the underlying speech recognizer [10, 12]. This results in the pronunciation(s) of each word or phrase being represented as a set of phoneme sequences. For example, using SALAAM with the English (US) source language to generate the top three pronunciations for *Mkate*, the Kiswahili word for bread, would result in the following phoneme sequences: *M K AA T I*, *M K AA CH I*, or *M K AH CH E*, which are then written to a lexicon file that is used later during the speech recognition process.

Previous studies have only used English (US) and French (France) as the source languages, with a maximum of 10 alternative pronunciation per target language word type [10, 12, 11, 13]. We investigate the performance and tradeoffs of different source languages and training techniques using four source languages: English (US), French (France), German (Germany) and Mandarin (China), and five target languages: English (South African), Sotho, Afrikaans, chiShona and Kiswahili. Three of these languages–English, Sotho and Afrikaans–are drawn from the 11 official languages of South Africa [15], while chiShona is spoken widely in neighboring Zimbabwe and Kiswahili is spoken widely across East Africa. Afrikaans and English have European roots and English(South Africa) is used as a control language. Sotho, Chishona and Kiswahili are indigenous to Africa and are representative of the Bantu language family. Bantu languages are a group of languages indigenous to Africa, from the south of Nigeria, covering most of central, east, and southern Africa [16]. There are Bantu language speaking communities in 27 of the continent's 54 countries, representing about 240 million speakers. The number of languages ranges from 300 to 680, depending on the criteria used to differentiate between dialects and languages [16]. They are agglutinating, have concordial agreement systems, and all nouns are assigned to a noun class [16].

3. Methodology

The study had two main phases: a data collection phase, where we recorded over 50,000 words from native language speakers; and an analysis phase where we ran a series of experiments across the dataset.

3.1. Data collection

One hundred four (53 female and 46 male) native language speakers were recruited for the five target languages. Participants were undergraduate students at the University of Cape Town (UCT) and were recruited via e-mail and through word of mouth. Participation in the study was voluntary, and the study was granted ethics approval through the IRB at UCT.

We developed a vocabulary of 100 word types in English, based primarily on words commonly used with agriculture and nutrition, in addition to standard words such as numbers, months and days of the week. The vocabulary included a mixture of words and short phrases (collectively called word types). The selected word types reflected the words and phrases likely to be used in our intended use case. A vocabulary size of 100 word types per target language provided a basis for ease of statistical significance assessment [13]. The vocabulary was then translated to the other target languages with the aid of Google Translate, and validated by native speakers. For each word type, we collected five audio samples recorded by each of the speakers [12]. The recordings were done in a quiet room using a mobile phone recording at 44.1kHz. A mobile phone was used for data collection because they are prevalent in developing regions and the audio quality is similar to what one expects when users are interacting with a spoken dialog system in this context [12]. The mobile phone also allowed us to maintain the same audio quality throughout the entire data collection process and ensure a uniform dataset.

3.2. Data analysis

We conducted three experiments to evaluate performance related to the sensitivity of: (1) source language, (2) training technique and (3) number of pronunciations. Each of these factors may affect the training of the phoneme, and ultimately the ability for the voice system to accurately recognize the words.

1. **Source language impact on recognition accuracy**: the generation of pronunciation lexicons that map each term from a target language to one or more sequences of phonemes in the source language depends on the phonemes the high resource language speech recognizer can model [10, 12, 7]. Therefore, we hypothesized that if the target and source languages were of similar phonemic properties then the overlap between the source and target language phoneme inventory would be maximized. This would in turn reduce the difficulty of phoneme mapping by finding better pronunciations and yielding better recognition accuracy.
2. **Effect of training technique on recognition accuracy with respect to gender**: we hypothesized that, for applications developed using cross-language phoneme mapping, gender may have a confounding effect on recognition accuracy, as it has in previous studies [17]. Gender-sensitivity may be due to the different acoustic properties of pronunciation between men and women, which may affect the audio signal interpretation by the underlying speech recognizer [17]. We evaluated recognition accuracy across three experimental setups: same-gender pairs (training and testing datasets comprised of a single gender), multi-gender pairs (mixed-gender training and testing datasets) and cross-gender pairs (training with a single gender and testing with the other gender).
3. **Impact of number of alternative pronunciations on recognition accuracy**: we hypothesized that increasing the number of alternative pronunciations would improve recognition accuracy, as demonstrated in previous work [12], up to an inflection point, after which recognition accuracy would decrease. We expected that the re-

duction in accuracy improvements at the margin would occur due to the inevitable overlap of alternative pronunciations for words with similar phonetic structure [10].

3.3. Experimental Setup

We used the SALAAM method as implemented in the open source tool Lex4All [14]. To test hypothesis (1) source language, we used four source languages recognizers: English (American), French (France), Mandarin (Mainland China) and German (Germany), selected because of availability of phonetic alphabets. We accessed these recognizers through Microsoft Speech Platform SDK 11 [18], a technology developed by Microsoft for server-side recognition of telephone-quality audio. We used this system because of its robustness and to reproduce the experimental environment of previous studies [14, 12, 10]. No additional modifications to the underlying models of this system were made –our goal was to test a system that was feasible for user groups to implement without technical modifications.

With respect to hypothesis (2) training technique, for each target language, we created three training datasets: male-only, female-only and mixed-gender. We created the single gender datasets by randomly selecting four participants per gender. The mixed-gender datasets were made up of 2 male speakers and two female speakers from the male-only and female-only datasets, all of which were randomly selected. We created a testing dataset by randomly selecting two female and two male speakers whose data were not used to form any of the training datasets. We used these datasets to evaluate the effect of training technique on recognition accuracy with respect to gender. The total number of speakers per dataset was capped to four for uniform testing conditions across all target languages. The randomized selection of speakers for each of these datasets is achieved using the 'sample' function from the R software environment [19].

To investigate hypothesis (3) number of pronunciations, during the training phase, we generated pronunciations for each target language word type using audio data from the participants whose audio samples formed the single-gender and multi-gender datasets. We achieved this by using the SALAAM method [14] and each of the four source language recognizers.

For each target language and source language pair, we obtained lexicon pronunciation files for the female-only (single-gender), male-only (single-gender) and multi-gender training sets. To address our three hypotheses, we used the pronunciation lexicons generated during training, with respect to the source language, target language and training techniques used, to perform recognition accuracy evaluation. In each instance, we used lexicon files containing 5, 10, 20, 40, 60, 80 and 100 alternative pronunciations per word. When a word contains multiple alternative pronunciations, the underlying speech recognizer will match any of those pronunciations without making any distinction or preference among them [10]. We used the R software environment [19] for statistical analysis and Seaborn [20] for data visualization using the Python programming language [21].

4. Results & Discussion

We present results obtained from investigating the impact of source language, effect of training technique with respect to gender and impact of the number of alternative pronunciations on recognition accuracy.

4.1. Source language effect on recognition accuracy

Figure 1 shows the results obtained from this experiment. We see that using English (US) as the source language produced the best results for all target languages with an exception of Sotho, which had a higher recognition accuracy when the source language was Mandarin.

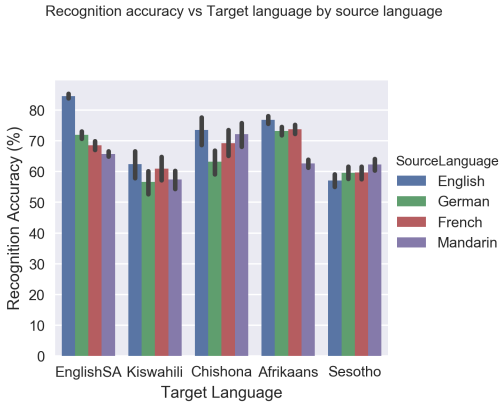

Figure 1: ***Recognition accuracy versus Target language by source language***

The results were further analyzed using several statistical methods. A Shapiro-Wilk test determined our data was not normally distributed. As a result, we used the Kruskal-wallis test, a nonparametric test, for statistical analyses. The first evaluation looked at the results irrespective of the target language, we achieved this by aggregating the results based on the source language and running statistical analyses on them. The tests revealed a significant overall effect of source language on recognition accuracy ($x^2(3) = 110.29, p < 2.2e-16$), supporting our hypothesis. Performing Post-hoc pairwise comparisons using the Wilcox sum rank test with Bonferroni correction showed a significant effect of source language on recognition accuracy among all source-target language pairs except when French and German were used as source languages. Thus our findings suggest that there are optimal pairings between source and target languages for phoneme matching. English(South Africa) recorded the highest recognition accuracy with English(US) as a source language. This is what we expected as the languages share the same phoneme inventory. Contrary to previous findings that used the SALAAM technique [10], our findings seem to indicate that choosing a source language whose phoneme inventory overlaps more with the target language would yield significantly higher recognition accuracy. Follow up work would do well to trace the linguistic roots of the languages to try to determine which properties are most important for matching and thus can create more general predictions for which source language is best suited for particular target languages, as this was beyond the scope of the current paper.

4.2. Impact of technique on recognition accuracy

We evaluated recognition accuracy for: same-gender pairs (single-gendered training and testing datasets comprised of a single gender), multi-gender pairs (a mixed-gender training dataset was used) and cross-gender pairs (single-gendered training and testing datasets comprised of different genders).

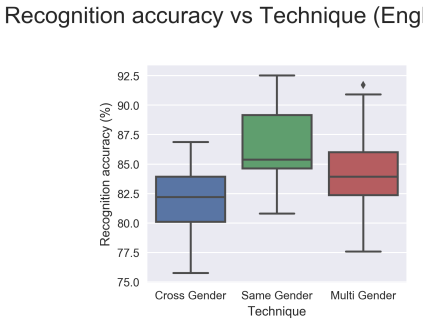

Figure 2: ***Recognition accuracy versus Technique***

Figure 2 shows a box plot representing our findings when English (US) is used as a source language. The same-gender technique recorded the best recognition accuracy followed by multi-gender and cross-gender techniques.

A Shapiro-Wilk test determined our data was not normally distributed, so a Kruskal-wallis test was again used to perform statistical analyses on the data. It revealed an overall significant difference in recognition accuracy among the different techniques ($x^2(6) = 36.69, p < 2.021e-06$), supporting our hypothesis. This suggests that when using this approach it is important to know the gender composition of the target speaker group. Despite recording the highest recognition accuracy, the same-gender technique results did not significantly differ from those obtained using the multi-gender technique. This implies that either technique could be used, however, the multi-gender technique would be better as it would be more robust against a gender bias. These findings underscore the need for researchers and practitioners to gender into consideration when developing datasets that are to be used with the SALAAM method.

4.3. Impact of pronunciations on recognition accuracy

Figure 3 shows a series of figures showing how recognition accuracy varies with alternative pronunciations with respect to source language when English (South Africa) is used as a target language .

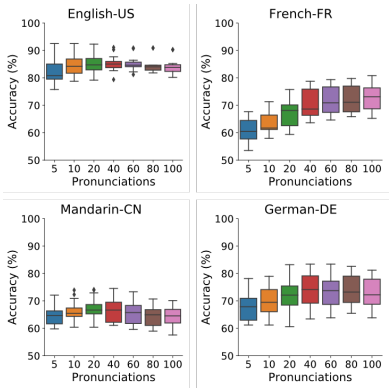

Figure 3: ***Recognition accuracy versus number of alternative pronunciation***

For all source languages, we observed an increase in recognition accuracy with an increase in alternative pronunciations up to 20 alternative pronunciations. However, from 40 pronunciations onwards, we observed a small decline in recognition accuracy for both English (US) and Mandarin. The results obtained using English (US) and Mandarin as source languages appear to support our hypothesis. However, we did not observe a decline in recognition accuracy for French and German as the number of alternative pronunciations increases. Running statistical analysis revealed that pronunciation only had a significant effect on recognition accuracy for the *English (South Africa) - French* source-target language pair ($x^2(6) = 36.69, p < 2.021e-06$). These results suggest a source-language dependent behavior. Our observations regarding the relationship between multiple pronunciations and recognition accuracy do not differ from previous studies [12]. Recognition accuracy is generally improved with an increase in the number of alternative pronunciations per word type. However, this improvement also comes at a cost in terms of system response time during use. The greater the numbers of alternative pronunciations per word type, the larger the search space and, consequently, the longer the response time. Presented in table 1 is a summary of our observations of mean evaluation time for 2000 English (South Africa) word type evaluations with varying alternative pronunciations. There appeared to be a roughly linear relationship between the number of pronunciations used and the amount of time required for the system to match words.

Table 1: *Evaluation time vs number of pronunciations*

Number of pronunciations	Mean evaluation time (Minutes)
5	2:04s
10	2:04s
20	4:04s
40	8.45s
60	13:15s
80	17:24s
100	21.04s

5. Conclusions and Future Work

This paper explored the performance of using cross-language phoneme mapping for the development of speech based applications. We used four source languages and five target languages. In the first experiment, we establish that source language choice has a significant impact on recognition accuracy. That recognition accuracy can be improved if a target and source language share similar phonemic properties. In the second experiment, we establish that recognition accuracy generally improves with the use of multiple pronunciation but at the expense of a longer response time. Lastly, we find that gender also has a confounding effect on recognition accuracy for speech-based applications developed using the SALAAM technique. The results show that using the single-gender training technique yields the best recognition accuracy, though not significantly different from those obtained using the multi-gender technique. This study represents a next step into using phoneme mapping for voice recognition of under-served African languages. Given the diversity of languages and code-mixing behavior–such as Sheng in Kenya, which is a combination of Kiswahili and English–to Bantu languages with click-sounds–such as isiZulu and isiXhosa–there is still substantial work to be done. However, a systematic approach based on a typology of linguistic properties and future use cases could move this technique quickly into use, allowing a much greater segment of the population to be reached and understood.

6. References

[1] E. Brewer, M. Demmer, M. Ho, R. Honicky, J. Pal, M. Plauche, and S. Surana, "The challenges of technology research for developing regions," *IEEE Pervasive Computing*, vol. 5, no. 2, pp. 15–23, 2006.

[2] N. Sambasivan, E. Cutrell, K. Toyama, and B. Nardi, "Intermediated technology use in developing communities," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 2583–2592.

[3] I. Medhi, A. Sagar, and K. Toyama, "Text-free user interfaces for illiterate and semi-literate users," in *Information and Communication Technologies and Development, 2006. ICTD'06. International Conference on*. IEEE, 2006, pp. 72–82.

[4] J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, and R. Rosenfeld, "Healthline: Speech-based access to health information by low-literate users," in *Information and Communication Technologies and Development, 2007. ICTD 2007. International Conference on*. IEEE, 2007, pp. 1–9.

[5] S. Patnaik, E. Brunskill, and W. Thies, "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, 2009, pp. 74–84.

[6] N. Patel, D. Chittamuru, A. Jain, P. Dave, and T. S. Parikh, "Avaaj otalo: a field study of an interactive voice forum for small farmers in rural india," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2010, pp. 733–742.

[7] F. Qiao, R. Rosenfeld, and J. Sherwani, "Layperson-trained speech recognition for resource scarce languages," 2010.

[8] K. Bali, S. Sitaram, S. Cuendet, and I. Medhi, "A hindi speech recognizer for an agricultural video search application," in *Proceedings of the 3rd ACM Symposium on Computing for Development*. ACM, 2013, p. 5.

[9] F. W. K. B. R. Rosenfeldc and K. Toyamab, "Unexplored directions in spoken language technology for development."

[10] A. Vakil and A. Palmer, "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*, 2014.

[11] J. Sherwani, S. Palijo, S. Mirza, T. Ahmed, N. Ali, and R. Rosenfeld, "Speech vs. touch-tone: Telephony interfaces for information access by low literate users." *ICTD*, vol. 9, pp. 447–457, 2009.

[12] F. Qiao, J. Sherwani, and R. Rosenfeld, "Small-vocabulary speech recognition for resource-scarce languages," in *Proceedings of the First ACM Symposium on Computing for Development*. ACM, 2010, p. 3.

[13] H. Y. Chan and R. Rosenfeld, "Discriminative pronunciation learning for speech recognition for resource scarce languages," in *Proceedings of the 2nd ACM Symposium on Computing for Development*. ACM, 2012, p. 12.

[14] A. Vakil, M. Paulus, A. Palmer, and M. Regneri, "lex4all: A language-independent tool for building and evaluating pronunciation lexicons for small-vocabulary speech recognition." in *ACL (System Demonstrations)*, 2014, pp. 109–114.

[15] T. Niesler, P. Louw, and J. Roux, "Phonetic analysis of afrikaans, english, xhosa and zulu using south african speech databases," *Southern African Linguistics and Applied Language Studies*, vol. 23, no. 4, pp. 459–474, 2005.

[16] D. Nurse and G. Philippson, *The bantu languages*. Routledge, 2006.

[17] W. Abdulla, N. Kasabov, and D.-N. Zealand, "Improving speech recognition performance through gender separation," *changes*, vol. 9, p. 10, 2001.

[18] "Microsoft speech platform," https://msdn.microsoft.com/en-us/library/office/hh361572(v=office.14).aspx, (Accessed on 03/19/2018).

[19] "R: The r project for statistical computing," https://www.r-project.org/, (Accessed on 03/19/2018).

[20] "seaborn: statistical data visualization seaborn 0.8.1 documentation," https://seaborn.pydata.org/, (Accessed on 03/19/2018).

[21] "Welcome to python.org," https://www.python.org/, (Accessed on 03/19/2018).