Number pronunciation in a multilingual environment and implications for an ASR system

Raymond Molapo
Human Language Technologies
Research Group Meraka Institute
CSIR, South Africa
Multilingual Speech Technologies Group
North-West University
Vanderbijlpark
South Africa
Email: rmolapo@csir.co.za

Etienne Barnard
Multilingual Speech Technologies Group
North-West University
Vanderbijlpark
South Africa
Email: etienne.barnard@nwu.ac.za

***Abstract*—The purpose of this paper is to address the challenges and describe step-by-step solutions faced when developing an automatic speech recognition system in multilingual societies. We give a brief statistical analysis of the data that have been harvested from the internet. The harvesting process operates in a multilingual environment where code-switching is the norm. We specifically focus our attention on the challenge of number normalization, pronunciation and the variations associated with it. We then develop various systems to illustrate the effects of different approaches to modelling the pronunciation of numbers.**

I. Introduction

AUTOMATIC speech recognition implementation in a multilingual society has proven to be a challenging task with added factors that need to be taken into account. In a country such as South Africa, there are eleven official languages each with its own variation of accents and dialects. Although speakers may be classified under a certain language group, their geographic location within the country could heavily influence how they pronounce certain words. In addition, the movement of migrant workers to big cities has slowly decreased the number of homogeneous societies around the country. As different societies mingled with one another, certain words from one society found their way into the other. Therefore, it is not unusual for a standard South African conversation to consist of words from several different languages. It should however be noted that different factors such as age group, levels of education and geographical location may heavily influence the homogeneity of a conversation. For instance, people who live in rural areas of South Africa, which is mostly homogeneous, may converse using very little use of words from other languages.

The occurrence of mixing words and sentences from different languages within a conversation, described above, is termed code switching [1]. It is mostly prominent in multilingual societies especially on the African continent where the dominant language or lingua-franca is embedded within normal day to day conversations. A significant amount of literature exists on the development of ASR systems in multilingual environments. Challenges such as the development of a pronunciation dictionary are often encountered due to the number of out-of-vocabulary (OOV) words present in the corpus [2]. These OOV words may include numbers that are found in the corpus. Numbers are normally left in the transcriptions for the sole purpose of hearing how people call them out in case of uncertainty. As a consequence, the accuracy of a monolingual ASR system may be affected negatively since what is contained in the audio is quiet different to what is in the transcription.

In this paper we focus on the implications of having numbers that are not normalized in the corpus. We first give a background on the use of code switching and its applications on day to day conversations in Section II. In Section III, we give a statistical analysis of the collected text data and the audio acquired. Section IV conducts several experiments which include passing all the text data through an English number normalizer to generate word representation of the numbers. Thereafter we train various grapheme-based ASR systems to compare the results from a system with no special provision for numbers, a system with normalized English numbers and a system with transliteration.

II. Background

A study conducted by [3] has highlighted the use of lingua-franca languages when Interactive Voice Response (IVR) users were given an option to choose between their mother tongue and English. The study showed that 84% of the users chose to interact with the system in English, although they were not English mother tongue speakers. Such preferences are likely to result from the fact that the African or native languages have not evolved to accommodate new Western influences such as technological gadgets, literature etc. To compensate for this inadequacy, most native speakers have devised ways to include these words into their vocabulary by adding prefixes and/or suffixes to make the words sound more like their dialect. This process takes a significant amount of time and the words may go through a number of transformations to reach a level of acceptance.

Words that do not go through this morphological process are used without being changed. These words may even extend to complete sentences within a conversation. However, most government institutions and broadcast corporations have put

strict regulations on news bulletins and certain shows in an attempt to preserve African languages. To avoid the complexities and uncertainties resulting from this state of affairs, users may simply opt for English versions of an IVR system.

According to [3], users exclusively opted to use English when stating numbers, telephone numbers, temperature etc. This phenomenon called code-switching makes the process of ASR development more challenging. To develop an ASR system without foreknowledge of the target language, requires that some aspects of prompt generation especially language dependent aspects, to be overlooked. These reasons could be wanting to hear how the respondents pronounce certain words that are not in their vocabulary. This in turn introduces other difficulties that requires another level of post-processing.

In general, the most widely and accepted way to spot embedded languages within a code-switched text, is to perform language identification (LID) on the collected text [4] [5]. This identification permits the system to know beforehand which speech engine to execute during recognition. The approach only works when there are large amounts of data to train a language model, which is not the case for our under resourced conditions.

III. Methods and Data

Our text data collection process requires for a starting point that the language under investigation should at least have written orthography. The main target languages are those that have little or no collected text or speech corpora and hence classified as under resourced. The language also needs to have some form of internet text that can be harvested by web crawlers.

With the drastic increase of foreign nationals coming into South Africa from neighbouring countries in recent years, and to avoid the relatively well-resourced South African languages for our investigation, we decided to explore a common non-South African language. With millions of these migrant workers coming from Southern Africa, mostly Zimbabwe, the choice was between ChiShona and the Zimbabwean dialect of IsiNdebele. The IsiNdebele language is closely related to South African IsiNdebele. So for the purposes of this investigation we opted to explore the Shona language. As described in [6], Shona is an umbrella term used to represent a family of dialects. It is the dominant language and one of the official languages of Zimbabwe. Shona has about 10.8 million first language speakers across Southern Africa. Through prior investigation it was established there were several Shona internet sites that were still in operation, although the number was significantly reduced due to the political climate in Zimbabwe.

The web crawlers were directed to harvest Shona text data from various internet sources [7] [8]. We managed to collect about 19 MB of raw text data from the internet, which contained 267 000 sentences. The data was then post-processed and reduced to 8581 sentences with approximately 52 250 word tokens. The reduction was due to the amount of English content that was present and needed to be filtered out. It should be noted that only sentences which contained only English were discarded and those that were mixed or contained Shona only were considered to reflect the code-switching nature of normal conversations. This was not surprising since most internet sites in a multilingual society have a mixture of text from different languages.

From the resulting corpus, over seven thousand 3-word prompts were generated and prepared for recording. The text post-processing stage only focused on removing punctuation marks and the embedded English text as mentioned above. However we opted not to normalize any numbers that were found within the text. This decision was made with two considerations: we intended to learn about how people call out numbers, and under resourced languages do not have the luxury of having number normalizers or readily available linguists to do the translation. Consequently, our prompts contained a substantial amount of numbers which were present throughout the recording process. The recording process took place over a period of two months and resulted in over 7 and a half hours of speech data from 22 speakers.

IV. Results and Observations

In order to evaluate how useful our data is for the purposes of ASR (and to create a basic Shona recognizer for further development), we have conducted several experiments. Our results were conducted using threefold cross validation with no speaker overlap.

A. Pronunciation Dictionary

In general, the development of an efficient ASR system requires a proper pronunciation dictionary. However, for under resourced languages that do not have many of the relevant resources, a proper pronunciation dictionary may not be easily compiled.[9] describes using mappings from one language to generate pronunciations for the other language. Though this approach may give acceptable results, it proved to be less effective than generating pronunciations for all the words regardless of the language. For this reason, a grapheme-based method of generating pronunciations by representing a word with its corresponding space-separated grapheme sequence was used, as first proposed by [10]. This method is mostly effective when used for languages that are regular or have a close phone to grapheme mapping.

B. Data Preparation

After all the other resources were in place, we used HTK [11] tools to perform feature extraction from the collected audio. The recognizer used a system based on standard Hidden Markov Models (HMM). For feature extraction, a standard 39 dimensional feature vector composed of 12 Mel Frequency Cepstral Coefficients (MFCCs) as well as their delta and double delta values coefficients extracted for each frame and also one energy feature along with its delta and double delta. The MFCCs were extracted using a 25 milliseconds window size and a 10 milliseconds shift. Due to the lack of text data resources, we opted to use a flat language model for grapheme recognition.

C. Experiment 1: Baseline Results

The baseline system was built using all the data from the corpus. The training and test set contained sentences with English in them. At this stage, all the numbers were present in

the corpus. There were 8581 files in total and 266 of them contained numbers. On close inspection it was found that the numbers mostly represented years, days, bible verses and temperatures.

As mentioned in III, there were 22 speakers who were recruited for the recording process. The respondents consisted of university students, security guards and a professional with a post graduate qualification. Despite this wide pool of individuals, all of them switched to English to call out numbers regardless of their qualification or social status, and also regardless of the semantic role of the numbers.

It is common for different speakers to call out a string of numbers in a different manner depending on how the string is structured. We found that the pronunciation of years and phone numbers turns out to be most varying among respondents/speakers. A number such as 2010, could be pronounced as twenty ten or two thousand and ten. Variations such as zero and oh when calling out phone numbers were also observed in the corpus.

TABLE I. *Baseline results before number normalization.*

Corpus	Accuracy	Correct	Duration
Shona + English	59%	68.74%	7.67 hours

Table I shows the accuracy results for the system trained on all the data. The grapheme results are comparable to the ones achieved for phone recognition on the 11 official South African languages during the Lwazi project [12], despite the presence of English in the training and test corpus. We find a slight increase in accuracy by transliterating the English content in the corpus, as reported in [6].

D. Experiment 2: Number Normalization

During the recording process, it was observed that all the numbers in the prompts were read or called out in English regardless of the context. Consequently, all the numbers in the corpus were normalized using an English number normalizer. The number normalizer uses the context in which the number is represented to normalize it. For instance, a number such as the year 2012 was found to have been called out differently by different respondents during the recording process. For the same number 2012, variations such as *twenty twelve*, *two thousand and twelve* or *two zero one two* were encountered. This variation makes it difficult for the ASR system to correctly recognize what is contained in the audio files. Subsequent to performing number normalization, we conducted another experiment to determine the importance of this effect.

TABLE II. *Baseline results after normalization.*

Corpus	Accuracy	Correct	Duration
Shona + English	55.49%	66.81%	7.67 hours

The results displayed in Table II show a decrease in the system's grapheme accuracy; this outcome could be expected since the normalization of numbers resulted in an increase of English content in the corpus. This is because the grapheme recognition results of an irregular language such as English are poor [13].

E. Experiment 3: Manual verification

It was observed that the automatic normalization of numbers did not always reflect what was said in the audio files. This was due to the different variations during the prompt recording process. For that reason, we decided to manually verify the normalized transcriptions by listening to what was actually spoken in the audio. Because of the vast mismatch, most of the automatically normalized numbers had to be manually transcribed to match what was said in the audio files. We then conducted another experiment to see the effect of manual transcription.

Table III shows an increase in system's grapheme accuracy after the manual verification process.The transliterated results were obtained by mapping the English phones to Shona graphemes. A number such as 41, was first normalized to *fourty one* then each pronunciation was transliterated to *foti* and *wan* respectively.

TABLE III. *Manually verified results.*

Corpus	Accuracy	Correct	Duration
Shona+English	60%	70.64%	7.67 hours
Transliterated	61.42%	71.72%	7.67 hours

V. CONCLUSION

We have explored the effects of having numbers and content from other languages on an ASR system. The process required us to have no prior knowledge of the language of choice, implying that we could not perform any language specific text normalization before speech data was collected. This meant that we had to leave the numbers in the prompts in unnormalized form.

In addition, we found that most Bantu language speakers read out numbers in the local lingua franca during the recording process. In the case of Shona speakers, English was the language of choice when code switching. These findings will make it simpler to determine which language to use when normalizing numbers beforehand and it will prevent the tedious process of manual verification. This will also provide consistency on how users call numbers out during the recording.

Furthermore, we have developed three systems to illustrate the effect of having numbers in the corpus. The first system gave acceptable grapheme results compared to the second system. This is indicative of the poor grapheme accuracy introduced by English content. We have shown that a simple transliteration can greatly enhance overall system performance in a multilingual code-switching environment particularly when dealing with a regular and an embedded irregular language.

ACKNOWLEDGMENT

Tim Schlippe, Ngoc Thang Vu, Charl van Heerden, Willem D. Basson, Nic de Vries, Neil Kleynhans, and the HLT Research Group, Meraka Institute, CSIR contributed to this project in various ways. Financial support from a Google Research Award is gratefully acknowledged.

REFERENCES

[1] T. I. Modipa, M. H. Davel, and F. De Wet, "Implications of sepedi/english code switching for asr systems," in *24th Annual Symposium of the Pattern Recognition Association of South Africa*. Pretoria, South Africa: PRASA 2013 Proceedings, December 2013, pp. 64–69.

[2] C. M. White, S. Khudanpur, and J. K. Baker, "An investigation of acoustic models for multilingual code-switching." in *INTERSPEECH*, Brisbane, Australia, September 2008, pp. 2691–2694.

[3] T. Ndwe, E. Barnard, and M. De Villiers, "Admixture practises in south african languages: Impact on speech-enabled technology design," in *IST-Africa Conference Proceedings, 2011*. IEEE, 2011, pp. 1–8.

[4] N. T. Vu, H. Adel, and T. Schultz, "An investigation of code-switching attitude dependent language modeling," in *Statistical Language and Speech Processing*. Springer, 2013, pp. 297–308.

[5] H. Adel, K. Kirchhoff, D. Telaar, N. T. Vu, T. Schlippe, and T. Schultz, "Features for factored language models for code-switching speech," in *SLTU*, St. Petersburg, Russia, May 2014, pp. 32–38.

[6] R. Molapo, E. Barnard, and F. De Wet, "Speech data collection in an under-resourced language within a multilingual context," in *SLTU*. St. Petersburg, Russia: International Research Insitute, May 2014, pp. 238–242.

[7] A. Kivaisi and A. Mbogho, "Web-based corpus acquisition for Swahili language modelling," in *3rd workshop on Spoken Languages Technologies for Under-resourced languages*, 2012, pp. 42–47.

[8] T. Schlippe, C. Zhu, J. Gebhardt, and T. Schultz, "Text normalization based on statistical machine translation and internet user support." in *INTERSPEECH*. Makuhari, Japan: Citeseer, Sept 2010, pp. 1816–1819.

[9] T. Modipa and M. H. Davel, "Pronunciation modelling of foreign words for sepedi asr," in *21st Annual Symposium of the Pattern Recognition Association of South Africa*. Stellenbosch, South Africa: PRASA 2010, 2010, pp. 185–189.

[10] S. Kanthak and H. Ney, "Context-dependent acoustic modeling using graphemes for large vocabulary speech recognition," in *ICASSP*, vol. 2. Citeseer, 2002, pp. 845–848.

[11] S. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, "The HTK book," vol. 3, 2002, p. 175.

[12] Meraka-Institute. (2009) Lwazi ASR corpus. [Online]. Available: http://www.meraka.org.za/lwazi

[13] W. D. Basson and M. H. Davel, "Comparing grapheme-based and phoneme-based speech recognition for Afrikaans," in *23rd Annual Symposium of the Pattern Recognition Association of South Africa*. PRASA 2012, 2012, pp. 144–148.