Pronunciation Modelling of Foreign Words for Sepedi ASR

Thipe Modipa
Department of Electrical, Electronic,
and Computer Engineering,
University of Pretoria
Email: tmodipa@csir.co.za

Marelie H. Davel
Human Language Technology Competency Area
CSIR Meraka Institute,
Pretoria, South Africa
Email: mdavel@csir.co.za

***Abstract*—This study focuses on the effective pronunciation modelling of words from different languages encountered during the development of a Sepedi automatic speech recognition (ASR) system. While the speech corpus used for training the ASR system consists mostly of Sepedi utterances, many words from English (and other South African languages) are embedded within the Sepedi sentences. In order to model these words effectively, different approaches to pronunciation dictionary development are investigated, specifically: (1) using language-specific letter-to-sound rules to predict the pronunciation of each word (based on the language of the word) and mapping foreign phonemes to Sepedi phonemes using linguistically motivated mappings, (2) experimenting with data-driven foreign-to-Sepedi phoneme mappings, and (3) using Sepedi letter-to-sound rules to predict the pronunciation of all words irrespective of language. We find that the data-driven phoneme mappings are more accurate than the initial linguistically motivated mappings evaluated, and (with a slight margin) obtain our best result using Sepedi letter-to-sound rules across all words in the speech corpus.**

I. Introduction

Spoken dialog systems (SDSs) are automated systems that use voice as input and output when interacting with a user. These systems rely on speech technologies such as automatic speech recognition (ASR) and speech synthesis. SDSs are important tools for information service provision over the telephone, and are increasingly being developed for under-resourced languages in developing countries such as South Africa.

Amongst other things, the development of ASR systems relies on the accurate modelling of word pronunciations, typically using pronunciation dictionaries to map a word to its standard (or canonical) pronunciation [1]. Context-dependent phonetic effects are usually not modelled explicitly in the pronunciation dictionaries of speech recognition systems, as the statistical acoustic models are trained to take context-dependent effects into account.

One of the challenges encountered when developing a pronunciation dictionary in multilingual environments relates to the extent in which code-switching occurs: speakers naturally embed words or phrases from other languages. For example, even when constrained to a spoken dialogue, many speakers of South African languages would use English numbers, dates and times. In addition, many place names have pronunciations that are clearly linked to other languages spoken in the vicinity.

In this paper we focus on the pronunciation modelling of foreign (non-Sepedi) words encountered during the development of a Sepedi ASR system. Words are categorised according to their language and we experiment with different approaches that can be used to model the out-of-language words. We measure the effectiveness of our modelling approaches by measuring phoneme recognition accuracy.

The paper is structured as follows: In Section II we first discuss related research. We then describe our approach and experimental design in Section III, and present the results obtained in Section IV. The overall outcome of the experiments and future work are discussed in Section V.

II. Background

In this section we provide some background with regard to speech recognition for Sepedi and related languages, and discuss general approaches to modelling out-of-language words.

A. Sotho-Tswana speech recognition

Sepedi is one of the official South African languages and is spoken by approximately 4.2 million people. It is mostly spoken in the Limpopo province [2] and has more than 20 dialects [3]. Sepedi belongs to the Sotho-Tswana languages, with Setswana and Sesotho two other languages from this language family. These three languages share most of their phoneme inventories. Sesotho is spoken by approximately 3.5 million people and this language is dominant in the Free State province. On the other hand, Setswana is mostly spoken in the North West province, by approximately 3.6 million people [2].

A number of Sotho-Tswana ASR systems have already been developed: an initial Sepedi ASR system [4]; Sesotho, Sepedi and Setwana ASR systems as part of the Lwazi project [5]; and an improved Sepedi ASR system [6]. The latter work specifically investigated whether complex consonant clusters could be represented as sequences of simpler sounds. This process reduced the phoneme inventory of Sepedi from 45 to 32, resulting in simpler dictionary development and slightly more accurate acoustic modelling.

B. Recognising out-of-language words

Individuals from multilingual environments tend to use more than one language in their conversations and these utterances pose a challenge to monolingual automatic speech recognisers. (Monolingual recognisers are trained to recognise speech in one language only.) For these recognisers, foreign words are often ignored and regarded as out-of-vocabulary (OOV) words. Other options for recognising foreign words include:

- Recognising the occurrence of foreign words on-the-fly using confidence measures and language identification systems, and then switching to different monolingual recognisers for identified sections of utterances. This is typically used for longer phrases and sentences embedded within the primary language.
- Modeling foreign words explicitly by combining language models and dictionaries from multiple languages. This is the more typical approach.

We are interested in the latter approach, which again has two important variations, specifically with regard to the rules that are used to generate the pronunciations, and whether these originate from the primary language or the foreign language [7]. In the first case, the letter-to-sound rules of the primary language are applied to the word list and pronunciations are predicted. In the second case, the use of letter-to-sound rules from the foreign language is required to predict the pronunciation of the words, and the language-dependent phonemes mapped from the foreign to the primary language. (The primary language and the foreign language consist of some phonemes that are common to both languages and other phonemes that are language dependent. Common phonemes are simply retained.)

Where language-dependent phonemes are encountered, a mapping is required. Such a mapping is obtained according to one of the following main approaches [7]:

1) creating a manual mapping by hand,
2) using a linguistic feature-based automatic mapping, and
3) generating a data-driven mapping.

Manual mapping requires a phonetic expert to analyse the data; linguistic feature-based mappings rely on the accuracy and consistency with which international phonetic alphabets are applied across languages; and data driven mappings include the use of distance measures and the analysis of confusion matrices [8].

III. Experimental Design

In this section we describe the different approaches we use to model foreign words occurring in the Sepedi corpus, the Sepedi speech corpus itself and the various experiments conducted.

A. Approaches investigated

In the subsequent experiments, we compare the following three approaches for modelling foreign words:

1) Using language-specific letter-to-sound rules to predict the pronunciation of each word (based on the language of the word) and mapping foreign phonemes to Sepedi phonemes using linguistically motivated mappings,
2) Experimenting with data-driven foreign-to-Sepedi phoneme mappings based on the confusion matrices obtained in (1), and
3) Using Sepedi letter-to-sound rules to predict the pronunciation of all words, irrespective of language.

B. Data

Our experiments are based on the Lwazi ASR corpus [9]. The corpus contains speech data from each of the eleven official languages of South Africa. Approximately 200 speakers per language (2,200 speakers in total), contributed read and elicited speech, recorded over a telephone channel. Each speaker produced approximately 30 utterances; 16 of these were randomly selected from a phonetically balanced corpus and the remainder consist of short words and phrases.

We use the Sepedi subset of the Lwazi ASR corpus and develop our own pronunciation dictionary, by extending the Lwazi Sepedi pronunciation dictionary [10]. Each of the Lwazi dictionaries is accompanied by a set of letter-to-sound prediction rules, which can be used to predict words not contained in the original dictionary.

C. Baseline system

Our baseline ASR system follows a standard Hidden Markov Model (HMM) design. Acoustic models consist of cross-word tied-state triphones modelled using a 3-state continuous density HMM. Each HMM state distribution is modelled by a 6-mixture multivariate Gaussian with a diagonal covariance matrix. The 39-dimensional feature vector consists of 13 static Mel-Frequency Cepstral Coefficients (MFCCs) with 13 delta and 13 delta-delta coefficients appended. The final preprocessing step applies Cepstral Mean Normalization (CMN) which calculates a per utterance bias and removes it. The different HMM state distributions are estimated by running multiple iterations of the Baum-Welch re-estimation algorithm. Once the triphone acoustic models are trained, a 40-class semi-tied transform is estimated to further improve acoustic model robustness.

Phoneme recognition is performed using a flat language model (all phonemes are considered equally likely at all times) and phoneme accuracy is measured. This is a conservative measure: an accuracy of approximately 60% when performing flat phoneme recognition can translate into an accuracy of 90% when performing word recognition for a small ($<$ 100 word) vocabulary. Phoneme recognition provides a more robust measure than word recognition, which is heavily influenced by the recognition vocabulary.

Accuracy is measured using 10-fold cross validation. The set of 190 speakers is divided into 10 folds. Each training set consists of 9 folds (171 speakers) and the test set consists of the remaining 19 speakers (per cross validation run).

D. Word categorisation

A list of words is generated from the Sepedi Lwazi corpus. The full word list consists of various types of words, including partial words (the full word was not produced), standard

Sepedi words, words from different languages and proper names (such as names of people or places). The majority of non-Sepedi words were found to be English.

As a first pre-processing step, all partial words are removed from the word list and the remainder of the word list is categorised as Sepedi, English and Other words (where 'Other' refers to any language that is not Sepedi or English). The initial language categorization is performed automatically, using existing English and Sepedi word lists. This results in three word lists, which are then reviewed manually to ensure correct categorisation.

For some of the words, accurate pronunciations are already known (words occurring in the Lwazi dictionaries) but for 60% of the Sepedi words and most of the 'Other' words, this was not the case. Table I shows the number of words contained in the Lwazi dictionaries, for the three main word categories.

In addition to these categories, proper nouns can be identified directly from the transcriptions, based on capitalisation. For each of the above categories, proper names are also flagged for special attention, as discussed in more detail below.

TABLE I
Number of words with pronunciations contained in the Lwazi dictionaries and number of words with unknown pronunciations

	Lwazi dict	unknown
Sepedi	1 232	1 865
English	144	42
Other	12	116

The correct categorisation of words is not always a clear-cut task. Sepedi, like many other languages, has a number of words originally borrowed from another language and now used as primary Sepedi words, for example, *divositse* ('divorced'), a loan word from English. Such a word often mixes Sepedi and English spelling and pronunciation conventions, and are difficult to deal with generically. While *divositse* is clearly no longer the original English word, it also does not follow Sepedi writing conventions. (Consider for example the letter 'v' found in this word, even though this letter does not occur naturally in Sepedi words.)

While some loan words, such as *divositse* for *hladile*, have Sepedi indigenous version, other words do not. For example, the word *domain* is written in Sepedi as *domeine* and has no other Sepedi counterpart. Where both words do exist, a loan word sometimes has preference over its indigenous counterpart. For example, *Janaware* is mostly used instead of *Pherekgong* which refers to January in English.

Problematic words were categorised according to the spelling system used. Words such as *Janaware* were categorised as Sepedi, while an unchanged English word such as *eight* occurring within a Sepedi utterance would be categorised as English.

Partial words (words that are cut at the beginning or end of an utterance) are also problematic since it is difficult to determine to which language they belong. Short words are treated as Sepedi. This is done because the original speech corpus annotaters (all Sepedi first language speakers) used Sepedi writing conventions to transcribe word fragments.

Upon completion, the final categorisation was verified by a second reviewer. (The second reviewer evaluated 1 450 words and edited the categories of 47 words.)

E. Extending the phoneme set

In the first experiment, the extended version of the dictionary is developed as follows:

- All words in the ASR transcriptions are categorized according to language origin (Sepedi, English or other) and type of word (general word or proper name) as described above.
- Pronunciations for Sepedi words (both general words and proper names) are automatically generated based on the Lwazi Sepedi letter-to-sound rules.
- Pronunciations for English and other words are similarly generated using the Lwazi English letter-to-sound rules.
- The problematic word lists (all proper names and the general words that are neither from Sepedi or English origin) are reviewed manually, and errors found are corrected.
- The ASR system is trained with a phoneme set containing all phonemes from both English and Sepedi. Note that none of the Other words utilised phonemes not occurring in either English or Sepedi. (For the rest of the paper all foreign phonemes are therefore referred to as English phonemes.)

F. Linguistically motivated mappings

In this experiment we follow the same procedure as described in section III-E but this time we define a mapping that maps each English phoneme to its closest matching Sepedi phoneme, as described in [6] and listed in Table II. Initial mappings are based on SAMPA notation, and phonemes that are similar for Sepedi and English are not shown. (Phone inventories of the languages of the world contain both language-dependent and language-independent sounds. Phonetic experts documented these sounds in phonetic inventories, such as IPA or SAMPA [11].) Where no close match can be found and an English phoneme occurs sufficiently frequently in the corpus, the phoneme inventory is extended with an English phoneme.

Finally, the problematic word lists (all proper names and the general words that are neither from Sepedi or English origin) are reviewed manually, and pronunciation errors found are corrected, prior to ASR system training. While pronunciation errors that were found were corrected, it was not always clear how a word should be pronounced. In these cases the most probable pronunciations were selected.

G. Data-driven mappings

In this experiment we follow the same procedure as described in section III-F but this time the English-to-Sepedi phoneme mapping is developed based on the confusion matrix obtained when training a system by including all English and Sepedi phonemes. For each English phoneme, the most

TABLE II
Phoneme substitution choices for English words occurring in the Sepedi corpus [6].

Substitutions			
from	to	from	to
{	E	Oi	O i
3:	E	p	p_h
A:	a	Q	O
ai	a i	r\	r
au	a u	t	t_h
d	l`	T	f
D	l`	tS	tS_h
e@	E @	@u	O
g	k_>	u:	u
@i	@ i	u@	u
i:	i	U	u
i@	i @	v	B
k	k_h	z	s
O:	O	Z	d_0Z
Additions			
@		b	

TABLE III
Phoneme substitution choices for English words occurring in the Sepedi corpus using confusion matrix.

Substitutions			
from	to	from	to
{	**a**	**Oi**	**E**
3:	E	**p**	**p_>**
A:	a	Q	O
ai	**i**	r\	r
au	**u**	t	t_h
d	l`	T	f
D	l`	tS	tS_h
e@	**E**	@u	O
g	**G**	u:	u
@i	**E**	**u@**	**O**
i:	**E**	U	u
i@	**a**	v	B
k	k_>	z	s
O:	O	Z	d_0Z
@	**E/a**	**b**	**B**

confusable Sepedi phoneme is selected for the mapping. Table III lists the final mapping selected, with changes from the previous mapping indicated in bold. The most significant changes to the mapping relate to the modelling of the diphthongs and the schwa. Note that the schwa (@) now maps to one of two possible phonemes: a or E. Pronunciation variants are included for all words containing schwas, and the best variant is automatically selected by the ASR system during training and use. A further cycle of confusion matrix analysis resulted in no further candidates for possible re-mapping.

H. Applying Sepedi letter-to-sound rules

In the final experiment, the categorisation is not used and all words are simply dealt with as if they were Sepedi words: the Sepedi letter-to-sound rules are applied, irrespective of the language the word is from. This is the simplest of all the strategies and is based on the assumption that the way an English word is spelled may influence its target pronunciation by a Sepedi speaker.

IV. Results

We first analyse the number of times a specific word occurs in the audio corpus: if a word occurs very frequently, an accurate pronunciation will have significantly more effect than if a word occurs only once. In Figure 1, the number of time each single word is observed in the audio corpus is shown. Close to 1 000 words have a frequency (appearance in the corpus) of over 10. Among those, the English words that appear to be more prevalent are the numbers and dates such as *one, two, three* or *September.*

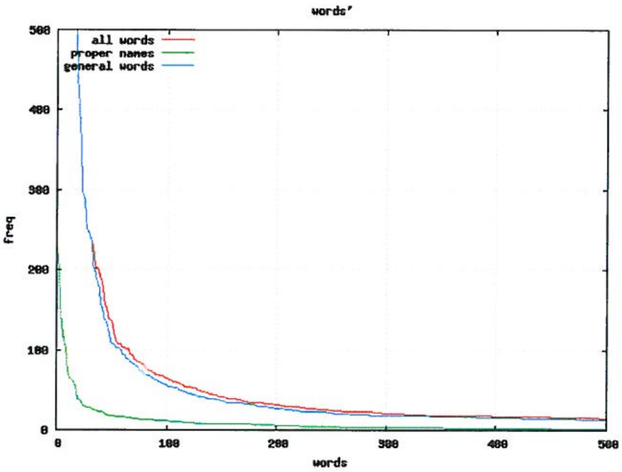

Fig. 1. Frequency with which different categories of words occur in the corpus

The phoneme recognition results obtained from the different experiments are shown in Table IV.[1] It is clear that different pronunciation modelling approaches do have a direct effect on ASR accuracy. The first experiment (combining all phonemes from both English and Sepedi) results in a large set of phonemes that occur rarely and that are not well estimated. Better accuracies are obtained as phoneme mappings are introduced, with a higher accuracy obtained when the phoneme mapping is guided by the confusion matrix.

Surprisingly, the best result is obtained with the simplest approach: using the letter-to-sound rules of the target language to predict all words. The difference between the best result (using Sepedi letter-to-sound rules on all words) and the second-best result (using data-driven phoneme mappings) is slight, if the standard deviation of the mean accuracy across the 10 cross-validation runs (σ_{10}) is taken into account[2]. However,

[1] The result obtained for the linguistically motivated mapping is not directly comparable with the others, as a change in the number of phonemes associated with a specific sound segment has an immediate effect on *measured* phoneme accuracy, even if the same recognition is performed. An adjusted baseline was calculated for this experiment (which takes the change in number of phonemes into account), but at 50.8%, the difference from the unadjusted baseline of 51.3% is not significant.

[2] The average of x independent measurements is expected to be distributed with a standard deviation of $\sigma/\sqrt{x}$ where σ is the measured standard deviation of the x measurements themselves

as the amount of effort involved in the latter approach can be prohibitive, the former approach becomes even more attractive.

TABLE IV
Sepedi phoneme accuracy using different pronunciation modelling approaches

	% Accuracy	$\pm\sigma_{10}$
Extended phone set	51.3	1.0
Linguistically motivated mappings	57.5	0.8
Data-driven mappings	59.9	0.7
Sepedi letter-to-sound	60.9	0.8

In order to verify the result obtained, we repeat the first and last experiments (as described in III-E and III-H) for two related Sotho-Tswana languages. This time we obtain results using the Setswana and Sesotho letter-to-sound rules respectively, on all words in each speech corpus. Results are shown in Table V and show a similar tendency but a somewhat less pronounced increase in accuracye, with observed phoneme recognition accuracies for Sesotho and Setswana improving from 55.0% to 58.3% and 60.3% to 63.2%, respectively. (Note that the full experiment III-G was not conducted in this case, as manual word categorisation and pronunciation checking were not performed.)

TABLE V
Phoneme recognition accuracies for Sotho-Tswana languages.

	% Accuracy	$\pm\sigma_{10}$
Sepedi		
Extended phone set	51.3	1.0
Sepedi letter-to-sound	60.9	0.8
Sesotho		
Extended phone set	55.0	0.5
Sesotho letter-to-sound	58.3	0.5
Setswana		
Extended phone set	60.3	0.0
Setswana letter-to-sound	63.2	0.9

V. Conclusion

In this study we investigated the effect of different approaches to the pronunciation modelling of foreign words in a Sepedi ASR system. Interestingly, the simplest approach – the prediction of the pronunciation of foreign words using Sepedi letter-to-sound rules directly – provided the best results. Within a small margin, these results were comparable to those obtained when first predicting the pronunciation of foreign words, and then using a data-driven mapping to map foreign phonemes to Sepedi phonemes: a process that is significantly more labour intensive.

We realise that the pronunciation of the words currently categorised as 'Other' may still have an effect on the accuracy of the recogniser. Most of the words in this category are proper names that emanate from different languages (neither English nor Sepedi), and determining the accurate pronunciation of proper names remains a challenging task.

Future work will repeat some of the experiments described in this paper in more detail for the other Sotho-Tswana languages, in order to understand whether the results obtained are Sepedi- (or corpus-) specific, or whether these results indeed generalise across languages. A more detailed audio-based analysis of the frequently occurring English words (for example, using acoustic confidence measures or goodness of pronunciation scores) may shed additional light on the pronunciation phenomena being observed.

References

[1] S. Goronzy, *Robust Adaptation to non-native accents in automatic speech recognition.* Lecture Notes on Artificial Intelligence. Springer Verlag, 2002.

[2] P. Lehohla, *Census 2001: Census in brief.* Statistics South Africa, 2003.

[3] H. J. Oosthuizen, M. A. Mapeka, and M. J. Manamela, "Investigation into automatic continuous speech recognition of different dialects of Northern Sotho," *Proc. Seventeenth Annual Symposium of the Pattern Recognition Association of South Africa*, pp. 57–60, 2006.

[4] T. M. Modiba, "Aspects of automatic speech recognition with respect to Northern Sotho," Master's thesis, University of the North, South Africa, 2004.

[5] C. V. Heerden, E. Barnard, and M. Davel, "Basic speech recognition for spoken dialogues," in *Interspeech*, Brighton, UK, September 2009, pp. 3003–3006.

[6] T. Modipa, M. Davel, and F. de Wet, "Acoustic modelling of Sepedi affricates for ASR," *submitted for SAICSIT*, 2010.

[7] C. White, S. Khudanpur, and J. Baker, "An investigation of acoustic models for multilingual code switching," in *Proc. Interspeech*, 2008.

[8] T. Schultz and A. Waibel, "Experiments on cross-language acoustic modeling," in *Proc. Eurospeech*, 2001, pp. 2721–2724.

[9] E. Barnard, M. Davel, and C. van Heerden, "ASR corpus design for resource-scarce languages," in *Proc. Interspeech*, Brighton, UK, Sept. 2009, pp. 2847–2850.

[10] M. Davel and O. Martirosian, "Pronunciation dictionary development in resource-scarce environments," in *Proc. Interspeech*, 2009, pp. 2851–2854.

[11] J. Kohler, "Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds," in *Proc. International conference on spoken language processing (ICSLP 96)*, 1996, pp. 2195–2198.