Implications of Sepedi/English code switching for ASR systems

Thipe I. Modipa
Human Language Technologies,
CSIR Meraka,
Pretoria, South Africa
Email: tmodipa@csir.co.za

Marelie H. Davel
Multilingual Speech Technologies,
North-West University,
Vanderbijlpark, South Africa
Email: marelie.davel@gmail.com

Febe de Wet
Human Language Technologies,
CSIR Meraka,
Pretoria, South Africa
Email: fdwet@csir.co.za

Abstract—**Code switching (the process of switching from one language to another during a conversation) is a common phenomenon in multilingual environments. Where a minority and dominant language coincide, code switching from the minority language to the dominant language can become particularly frequent. We analyse one such scenario: Sepedi spoken in South Africa, where English is the dominant language; and determine the frequency and mechanisms of code switching through the analysis of radio broadcasts. We also perform an initial acoustic analysis to determine the impact of such code switching on speech recognition performance. We find that the frequency of code switching is unexpectedly high, and that the continuum of code switching (from unmodified embedded words to loan words absorbed in the matrix language) makes this a particularly challenging task for speech recognition systems.**

Index Terms: code switching, speech recognition, multilingual speech recognition

I. Introduction

Code switching is a phenomenon observed around the world where speakers are exposed to more than one language. These, often multilingual, speakers spontaneously use words, phrases or sentences from one language (the *embedded* language) interspersed among words or sentences in the primary language (the *matrix* language) [1]. Code switching has significant implications for Automatic Speech Recognition (ASR) systems, since the acoustic models, pronunciation models and language models all need to be designed to accommodate words from different languages. Many ASR systems actually do not model code switching explicitly. Rather, general techniques used to deal with out of vocabulary (OOV) words are also applied to code-switched words [2]. This can result in information-rich words being ignored during speech processing, as code-switched words often do not have alternatives in the matrix language and can be key terms in an utterance [2], [3].

While code switching is a well-studied phenomenon for various language pairs (see, for example [4], [5]), much less work has been done analysing the implications of code switching for minority languages, and much fewer resources – such as large code-switched corpora – are available in these languages. The aim of this paper is to use corpus analysis to obtain an understanding of the prevalence of code switching in Sepedi, and to analyse the factors to consider when developing ASR systems capable of dealing with Sepedi/English code-switched speech. The paper is structured as follows: Section II provides background relevant to the modelling and analysis of code switching for ASR and introduces the Sepedi-English language task. Section III describes the approach we followed to design, develop and analyse the code-switched corpus. Analysis results are presented in Section IV, with the most pertinent findings summarised in Section V.

II. Background

Code switching is regarded as the process of switching from one language to the other during conversations [6]. These language alternations/switching can take place from one sentence to the other (*inter-sentential* code switching), or can occur within sentences where the secondary language is embedded within the primary language (*intra-sentential* code switching). In our work, the matrix language is Sepedi: this is the speaker's native language, and is dominant in the utterances. The embedded language is the speaker‘s non-native language, in this case mostly English. Both inter- and intra-sentential code switching are considered.

Since most languages also contain loan words – words from a different language incorporated into a recipient language as part of its accepted vocabulary – the distinction between loan words and code-switched words is not always easy to make. In addition, during code switching, speakers tend to either insert or delete vowels or consonants in order to reproduce a phonotactic structure comparable to their native language [7]. This process affects the pronunciation of the embedded words, and blurs the distinction between code-switched words, and words of foreign origin that have been incorporated into the matrix language.

ASR systems deal with code-switched speech in three main ways: (1) ignoring foreign words completely and dealing with them as out-of-vocabulary words, (2) switching amongst monolingual recognisers when encountering out-of-language words, and (3) modelling foreign words explicitly within a multilingual system; the latter being the most typical approach, and the one considered in this paper. The explicit modelling of code-switched speech can be performed at the pronunciation dictionary, acoustic and/or language model level.

The pronunciation dictionary provides a convenient level at which to add pronunciations for out-of-language words. These pronunciations can be generated manually or automatically, using the letter-to-sound rules of either the matrix and/or embedded language [8]. In addition, native speech can be used to generate non-native variants automatically by using a phoneme recogniser to derive variants from a training corpus [7]; or a direct mapping can be performed from the embedded language to the matrix language [5].

While monolingual recognisers are affected by the performance of language identification systems, multilingual systems try to solve this problem through techniques such as retaining the pronunciation of the secondary language and using multilingual acoustic models [2], or mapping, adaptation or merging at the phone, state or model level [5], [9], [10].

In this work, we are focusing specifically on Sepedi/English code switching. 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 [11]. In South Africa, English is spoken as a first language by only about 3.6 million people [11], but it is widely spoken as an additional language. Code switching is an everyday phenomenon observed among bilingual Sepedi speakers [12], but limited results are available with regard to the analysis of such code-switched speech. (Specific studies to mention include [8], [13] and [14].)

III. Approach

Since no corpora of naturally-occurring Sepedi speech were available prior to this analysis, we first develop such a corpus. There are many factors that influence the frequency, mechanisms and reasons why code switching occurs. One of these is the setting in which the language is used, such as formal, informal, academic or social. Code switching is also highly speaker-dependent. It would therefore be impossible to compile a corpus of code-switched speech that represents all possible uses of code switching for all the speakers of a language. For the purposes of the current study, it was decided to focus on radio broadcasts because many different communication scenarios and styles are used on the radio.

We use a two-step process: We first record and review a set of radio broadcasts, counting the number of code switching events that occur, and transcribing examples of code-switched speech. We then use the specific examples of code switching observed as prompts for recording additional samples from multiple speakers in order to study speaker-specific pronunciation differences. As the quality of the recorded prompts is influenced by speaker error, we validate the quality of the recordings through a combination of automated and manual means.

We use the first of these corpora, referred to as the Sepedi Radio (SR) corpus, to analyse the frequency of code switching, mechanisms of code switching and the reasons why code switching occurs in these broadcasts. We use the second of the corpora, the Sepedi Prompted Code Switching (SPCS) corpus to perform a first acoustic analysis of the effect of Sepedi/English code switching on ASR performance.

A. Data collection - radio broadcasts

We first compiled the SR corpus containing examples of code-switched Sepedi by recording and transcribing radio broadcasts. A number of programmes that are broadcast between 7 am and 4 pm were selected to be recorded. These included a general breakfast show, youth and current affairs programmes as well as an afternoon show. The level of literacy of the radio broadcast speakers was not determined.

The recorded audio files were reviewed and orthographic transcriptions created manually. The corpus was divided into three portions, namely, code-switched, Sepedi, and 'other' data. For the code-switched portion, the starting and end times of the utterances or phrases that contained code-switched words were captured. In many instances it was not easy to determine sentence boundaries (as is typical in conversational speech). In such cases sentence boundaries were estimated based on naturally occurring phrases within the range of sentence lengths as observed in the set of more clearly delineated sentences. The starting and end times of Sepedi portions that did not contain code-switched words were also marked, but the corresponding transcriptions were not created. Music and advertisements were marked as 'other' and were not considered for analysis in this experiment. The transcriptions corresponding to the code-switched speech sections were used to compile a list of phrases. These phrases were subsequently used as prompts to collect an acoustic database of code-switched speech (see Section III-C).

B. Transcription analysis

First language speakers of Sepedi validated the transcriptions as well as the word lists that were extracted from the transcriptions. The word lists were classified as English and 'semi-transformed'. This term is defined to refer to words that are clearly of English origin, not part of existing Sepedi vocabulary and transformed from the original English so that they are no longer the exact English word, for example, 'diwheelchair'. The duration of the sections of speech that are tagged as instances of code switching were calculated to quantify the frequency of code switching.

The first step in analysing the mechanisms of code switching was to create a word list from the transcriptions of the code-switched data. A number of labels were assigned to each word in the word list: (1) English words with and without a Sepedi alternative; (2) words that are semi-transformed (as defined above); (3) part of speech per code-switched word; and (4) whether the word forms part of a phrase that is a multi-word example of code switching, or whether it is a single word example. The frequency of occurrence of each category was subsequently derived from these labelled transcriptions.

To determine some of the reasons why code switching occurs, events were identified where English words were used in conjunction with Sepedi words. Frequency counts were then compiled for events where English was used for emphasis (where speakers use the matrix language for a word or phrase and then repeat the concept using the embedded language) and events where English was used because a Sepedi alternative does not exist.

C. Data collection - prompted code-switched speech

The prompts that were derived from the transcribed code-switched radio data were used to compile the SPCS corpus. Broadband speech data was collected using Woefzela, a locally-developed, smartphone-based speech data collection tool [15]. Twenty speakers (12 males, 8 females) each read approximately 450 utterances, resulting in 10 hours of prompted speech. The ages of the participants ranged between 17 and 27 years old.

D. SPCS corpus evaluation

The quality of the SPCS corpus was verified using Phone-based Dynamic Programming (PDP) scores, as described in [16]. The technique consists of developing a phone-based ASR system, and then comparing the phone labels obtained when (a) decoding an utterance using a phone-loop grammar and (b) aligning the same utterance at phone-level using the intended prompt. The two phone strings are aligned using dynamic programming and either a flat or a variable scoring matrix (obtained from the data being scored, as described in more detail in [16]). The alignment score is then used as a direct indication of both audio and transcription quality.

Once alignment scores were obtained, all utterances were ordered according to these scores, and the quality of the corpus at specific points (according to this ordering) was verified manually.

1) Data: We used the data described in Section III-C to perform four-fold cross-validation. (75% of the data was used for training and the remaining 25% for testing; this is repeated four times.)

2) Dictionary development: For verification, we use a straightforward approach to develop the pronunciation dictionary. We develop two versions: one in which all the words in the SPCS corpus were predicted using Sepedi grapheme-to-phoneme (g2p) rules (the *sep_g2p_1* dictionary) and another where the pronunciations of English words were manually corrected where gross errors occurred (the *sep_g2p_2* dictionary). In both dictionaries, the affricates were split, as described in [17], thereby resulting in a reduced Sepedi phone set.

3) ASR system development: A baseline ASR system was implemented using the HTK toolkit [18]. A standard 3-state left to right Hidden Markov Model architecture was used to develop context-dependent, tied-state, triphone models. A 39-dimensional feature vector was used (13 static Mel-Frequency Cepstral Coefficients, with delta and delta-delta coefficients appended). Speaker-specific cepstral mean and variance normalisation, as well as semi-tied transforms were applied.

These systems were only used to perform corpus evaluation; custom-designed systems are built for acoustic analysis.

E. Acoustic analysis of code-switched speech

In order to obtain a first indication of the acoustic impact of code switching on speech recognition performance, we develop a basic ASR system using a prior corpus of Sepedi data, and evaluate the difference in accuracy when recognising different test sets: (1) Sepedi-only data, (2) code-switched data and (3) a combination of the two data sets.

1) Data: For training, we use the NCHLT corpus [15] which consists of prompted speech in Sepedi, but which also includes some English speech. (The latter was generated from general English text and are not examples of actual code switching.) The corpus consists of 113 speakers and 12,560 unique word tokens. We train our system on both English and Sepedi speech, and model code switching at the pronunciation dictionary level. We use the *sep_g2p_1* dictionary described above as a benchmark, but obtain additional results using a more sophisticated dictionary (as described below).

Both the SPCS corpus and the Sepedi portion of the NCHLT test set are used during evaluation, and three separate results are produced: for only the code-switched (SPCS) corpus, for the Sepedi portion of the NCHLT test set, and for the former and the latter data combined.

2) Dictionary development: In order to obtain a credible result, we develop a more sophisticated dictionary for the acoustic analysis, following the process described in [8]. In order to map the English phonemes to Sepedi phonemes, we first train an ASR system containing Sepedi-only phonemes and then decode the code-switched speech using a phone-loop grammar. The resulting phone strings are then aligned against the language-specific pronunciations of all words – English pronunciations of embedded words and Sepedi pronunciations of matrix words – and the phoneme substitutions counted. The resulting matrix of alignment counts clearly shows which substitutions occur most frequently, and each English phoneme is then remapped to its closest Sepedi counterpart. The resulting mapping is used to generate additional Sepedi variants for all English words found in the data; these variants are added to a standard Sepedi version of all words, generated using g2p, as described in [8]. (In the final dictionary, each English word would therefore include at least 2 variants.)

3) ASR system development: The new ASR system is also implemented as described in Section III-D3.

IV. Results

In this section, we first present the results from the analysis of the initial SR corpus. The verification and acoustic analysis of the SPCS corpus follow in Sections IV-D and IV-E.

A. SR corpus code switching frequency

The transcriptions were generated from about 10 hours of audio, 3.6 of which contain speech content. The remainder is non-speech content such as music, silence and advertisements. From the content portion, the code-switched portion was almost 31%. The remaining part constituted Sepedi-only speech. The speakers used, on average, 3.4 embedded words per utterance (with the average length of the utterances being 15.8). Most of the observed code-switched words were numbers. Of those code-switched words, about 91% were pure English and the remaining 9% were semi-transformed words. In Figure 1, we show the number of English words per utterance, which ranged between 1 and 22. There were 245 utterances which contained a single English word. The utterances with close to 20 English words where actually telephone numbers (different telephone numbers mentioned in one utterance).

B. The mechanisms of code switching

In most instances (922 of 1 018 code-switched words observed), speakers used English words without any modification. There were also instances where English words were modified to conform to the Sepedi consonant-vowel (CV) syllable structure, mostly by adding vowels at the end of the words. Quite frequently, English words were appended with suffixes such as *-e*, *-a*, *-ing* and the prefix *di-* as shown in

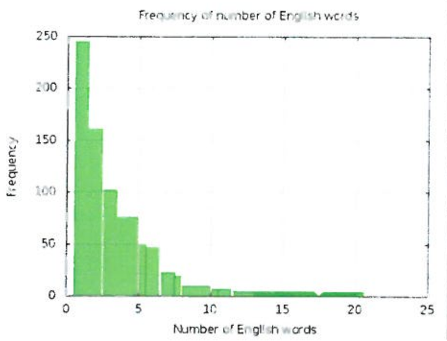

Fig. 1. The frequency of English words per sentence in the code-switched portion of the SR corpus.

Table I. The prefix *di-* had a dominant percentage distribution of 66.67%.

TABLE I. PREFIX/SUFFIX APPENDED TO MODIFIED ENGLISH WORDS AND THEIR PERCENTAGE DISTRIBUTION.

Prefix/suffix	% distribution
di-	66.67
-a	12.50
-e	9.38
-ing	2.08
other	9.38

The suffixes were observed when English verbs were embedded in Sepedi, for example *enjoy* became *enjoya.* The prefix *di-* was used for nouns in order to create a plural, for example: 'Ba bangwe ba bona ba feleletsa ba dula mo *diwheelchair* (Some sit on wheelchairs)'. Table II shows the occurrence of the different parts of speech of English words observed. It is evident that the most code-switched words are nouns.

TABLE II. PART OF SPEECH OF EMBEDDED ENGLISH WORDS.

Total	Noun	Adjective	Verb	Adverb	Other
1 018	789	89	72	33	35

C. The reasons for code switching

There are a number of reasons why multilingual speakers code switch in their conversations. We observed that code switching often occurs where the concept being discussed does not exist in the vocabulary of the matrix language. In Table III we show the number of English word tokens that do not have a Sepedi alternative. It must be noted that speakers still use code switching even for words that have Sepedi alternatives. Reference to time and age can be uttered in English by speakers for clarity or emphasis. There were 18 instances where speakers used both English and Sepedi for time and age (uttered first the Sepedi phrase, then repeating it in English).

TABLE III. NUMBER OF ENGLISH WORD TOKENS WITH AND WITHOUT SEPEDI ALTERNATIVES.

	# words (with numbers)	# words (without numbers)
Has Sepedi alternative	444	412
No Sepedi alternative	478	478

D. SPCS corpus evaluation

The ASR systems used to evaluate the SPCS corpus were implemented as described in Section III-D3. These systems were used to decode and align utterances in order to compare the phone strings of the utterances. For the two systems developed using the two different dictionaries, four-fold cross-validated phone accuracies obtained were 59.30% and 65.11%, using the *sep_g2p_1* and the improved *sep_g2p_2* dictionaries, respectively. (See the 'all utterances' row in Table IV.)

TABLE IV. PHONE ACCURACIES OBTAINED WHEN CROSS-VALIDATING DIFFERENT SUBSETS OF THE SPCS CORPUS. HERE '10K' INDICATES THAT ONLY THE BEST 10K UTTERANCES WERE RETAINED.

Corpus	Accuracy (sep_g2p_1)	Accuracy (sep_g2p_2)
10K	56.77	63.40
11K	57.29	64.27
12K	58.07	64.86
all utterances	59.30	65.11

The evaluation of the corpus was performed with the two systems described above, and the PDP scores obtained are shown in Figure 2. Positive PDP scores show that the audio can be decoded and is closely matched in content to the given transcriptions. There were two scoring matrices used, a flat matrix, and a trained matrix. In Figure 2, 'flat' and 'trained' refer to the scoring matrices used, while 'sep_g2p_1' and 'sep_g2p_2' refer to the dictionaries used. For the evaluation of the corpus we only used scores obtained using the flat matrix and the *sep_g2p_2* dictionary (Sepedi g2p rules with manual verification of English words).

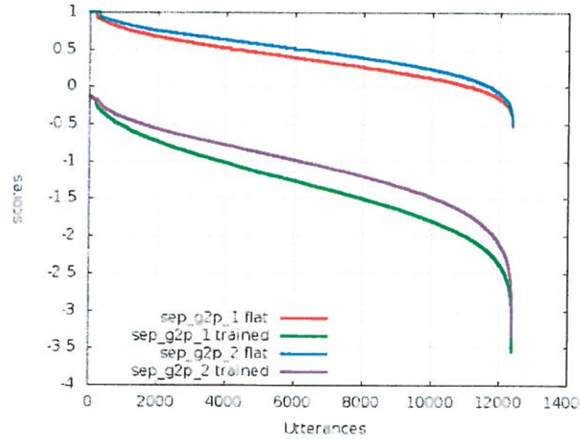

Fig. 2. PDP scores using using the *sep_g2p_1* and *sep_g2p_2* dictionaries with either a flat or trained scoring matrix.

While the graphs in Figure 2 provide an indication of the distribution of good utterances (to the left of the graph) as well as poor utterances (to the right of the graph), the threshold at which data becomes unusable for a specific application can only be determined through a systematic analysis of data at different scoring levels. Therefore, selected utterances were manually reviewed at different scoring levels: the utterance list was sorted according to the PDP scores, and at specific data points, 20 utterances were selected and listened to in order to rate the audio and transcription quality. Only one person listened to the selected utterances.

In Table V, we provide a summary of categories of problems observed at different data points. The first 20 utterances were listened to at the 0, 2K, 4K, 6K, 8K, 10K, 11K, 12K, and end-of-corpus data points. The first errors were only encountered at the 8K data point. Most of the utterances were good, with clipping and low volume affecting the PDP scoring. All the categories listed in Table V were considered as errors when evaluating the corpus, with the exception of the low volume category, which was regarded as acceptable audio.

TABLE V. THE PERCENTAGE OF GOOD UTTERANCES AT DIFFERENT DATA POINTS

Data points	No. of good utterances (%)
0 - 20	100
2,000 - 2,020	100
4,000 - 4,020	100
6,000 - 6,020	100
8,000 - 8,020	90
10,000 - 10,020	95
11,000 - 11,020	55
12,000 - 12,020	35
12,364 - 12,384	10

The quality of the utterances seems to deteriorate below the PDP score of 0.120. Even though there are good utterances below this score, many of these utterances have a sound artifact at the end, where the sound of a button being pressed is clearly audible. This effect would have affected the rating of the PDP scores (due to poorer alignment). Other utterances indeed contain errors.

The clean corpus size obtained from this evaluation process consists of utterances with PDP scores above the threshold of 0.120, thereby resulting in the corpus size of 11K utterances.

The following categories of errors were identified when manually listening to each utterance at different data points. Most of the low volume errors encountered came from one speaker:

- Correct but low volume
- Correct but clipping at the end
- Blank utterance
- Cut audio
- Background noise
- Speech repetitions
- Mispronunciations
- Hissing sound (channel effects)

The counts per each error category are shown in Table VI.

After the evaluation of the corpus was completed, a second sanity check was performed by evaluating phone recognition accuracies on different subsets of the corpus, as shown in Table IV. The same test set (from 'all utterances') were retained but the training sets used were reduced by removing all utterances that fell below a given threshold. In Table IV we show the phone recognition accuracies of ASR systems developed using data with 10K, 11K, and 12K utterances after the removal of flagged problematic utterances using the *sep_g2p_1* and *sep_g2p_2* dictionaries. Accuracies deteriorate slightly, as the data set used for training becomes smaller and smaller. This indicates that the lower scored utterances (even though they contain some form of error) still contribute meaningful portions of audio, and are useful to retain for ASR purposes, even though they may not be ideal for detailed acoustic analysis of individual words. The full corpus is therefore used to obtain the results presented in Section IV-E.

E. Acoustic analysis

The effect of the addition of the code-switched speech on the ASR system is analysed by measuring the phone recognition accuracies when the test set contains (1) only the matrix language, (2) only the code-switched corpus, and (3) a combination of the two data sets (matrix language speech and code-switched corpus).

We perform a basic acoustic analysis of the new (SPCS) corpus using a flat phone-loop grammar. (While recognition accuracy is relatively low, this helps to show a better comparison of acoustic difficulty, without results being influenced by the choice of a language model, a topic that requires further study.) Table VII shows the results obtained for three different pronunciation modelling approaches: when matrix language g2p rules are used for all words including embedded words (as used in Section IV-D), when embedded words are mapped to the matrix language, and when two variants are added per embedded word: one using the matrix language g2p rules, another the (mapped) embedded language g2p rules.

Table VII shows the effect of code-switched speech: as expected, overall phone recognition accuracy decreases when code-switched speech is added. Adding more variants improves the recognition accuracy on the combined data from 59.26% to 65.52%, but even when adding variants, the large gap between matrix language results (68.47%) and embedded language results (64.07%) remains. Note that better modelling of the code-switched portion also improves the recognition results of the matrix language initially (the 'Sepedi g2p' experiment), but that the additional variants (the 'Variants added' experiment) introduce some additional confusability seen in the final result, thereby resulting in a recognition accuracy decrease for matrix language words.

TABLE VII. PHONE RECOGNITION ACCURACY FOR CODE-SWITCHED (SPCS), NON-CODE-SWITCHED (NCHLT SEPEDI) AND COMBINED DATA.

	SPCS	NCHLT Sepedi	combination
Mapping only	55.84	66.35	59.26
Sepedi g2p	60.37	69.28	63.27
Variants added	64.07	68.47	65.52

V. CONCLUSION

In this paper, we presented a new corpus that was developed to better understand the implications of English/Sepedi code switching for ASR systems. The corpus development process consisted of first recording and transcribing radio broadcasts. This data was then used to analyse the frequency, mechanisms, and reasons for code switching.

In addition, samples of the transcriptions (containing true code-switched events) were then re-recorded by multiple speakers, in order to obtain data that is useful for studying pronunciation variation in code-switched speech. In order to verify the quality of the recorded corpus, PDP scoring was

TABLE VI. THE COUNTS PER EACH ERROR CATEGORY AT DIFFERENT DATA POINTS

Data points	Low Volume	Clipping	Blank	Cut Audio	Background Noise	Speech Repetitions	Mispron	Hissing
0 - 20	0	0	0	0	0	0	0	0
2,000 - 2,020	0	0	0	0	0	0	0	0
4,000 - 4,020	0	0	0	0	0	0	0	0
6,000 - 6,020	0	0	0	0	0	0	0	0
8,000 - 8,020	2	2	0	0	0	0	0	0
10,000 - 10,020	1	1	0	0	0	0	0	0
11,000 - 11,020	0	5	0	2	1	1	0	0
12,000 - 12,020	4	2	0	3	0	2	1	1
12,364 - 12,384	12	1	4	0	0	0	0	0

used. While a cleaner corpus is not required for ASR purposes, such a corpus can be useful for detailed phonetic analysis of code-switched events – an area of ongoing interest [13].

As could be expected, it was found that nouns, numbers and dates were the most important categories of words where code switching occurred. More surprisingly, we found that there were no Sepedi alternatives for over 50% of the English words observed, which predicts that many of these words will be incorporated into Sepedi over time. In addition, about 10% of English words observed were still recognisable as English, but 'semi-transformed' into Sepedi words through the addition or transformation of syllables.

The most unexpected result from this work was the high frequency of code switching that was observed. (See Figure 1.) Most of the embedded words were single English words, and mostly these were not transformed from standard English: while this makes them fairly easy to model, these words still had a significant influence on ASR performance.

An initial acoustic analysis of the effect of code-switched data on ASR performance was conducted. As expected, the addition of the code-switched corpus decreases the recognition accuracy of the ASR system. However, better lexical modelling of the code-switched portion of the corpus can (to an extent) compensate for the decrease in performance, as shown in Table VII.

In future work, we would like to use the newly created reference corpus to investigate more sophisticated approaches to the acoustic modelling of Sepedi code-switched speech. We are particularly interested in the various categories of code-switched speech (standard English, semi-transformed English, Sepedi loan words) and ways in which these can be modelled separately. In this work, we had complete control over the transcription process, but in a general text the transformation in spelling of code-switched words also becomes an issue, for example 'block' transforming to 'blocka' and then to 'bloka', a process we would like to model more precisely.

REFERENCES

[1] P. Li, "Spoken word recognition of code-switched words by Chinese-English bilinguals," *Journal of memory and language*, vol. 35, pp. 757–774, 1996.

[2] C. White, S. Khudanpur, and J. Baker, "An investigation of acoustic models for multilingual code switching," in *Proc. Interspeech*, 2008, pp. 2691–2694.

[3] C. Yeh, C. Huang, and L. Lee, "Bilingual acoustic model adaptation by unit merging on different levels and cross-level integration," in *Proc. Interspeech*, 2011, pp. 2317–2320.

[4] J. Chan, H. Cao, P. Ching, and T. Lee, "Automatic recognition of Cantonese-English code-mixing speech," *Computational Linguistics and Chinese Language Processing*, vol. 14, pp. 281–304, 2009.

[5] G. Stemmer, E. Noth, and H. Niemann, "Acoustic modeling of foreign words in a German speech recognition system," in *Proc. Eurospeech*, 2001, pp. 2745–2748.

[6] C. Myers-Scotton, *Social motivations for Codeswitching: Evidence from Africa*. Clarendon Press, Oxford, 1993.

[7] S. Goronzy, *Robust adaptation to non-native accents in automatic speech recognition*. Lecture Notes in Artificial Intelligence 2560, 2002.

[8] T. Modipa and M. H. Davel, "Pronunciation modelling of foreign words for Sepedi ASR," in *Proc. (PRASA)*, 2010, pp. 185–189.

[9] Y. Li, P. Fung, P. Xu, and Y. Liu, "Asymmetric acoustic modeling of mixed language speech," in *Proc. ICASSP*, 2011, pp. 5004–5007.

[10] C. Yeh, L. Sun, C. Huang, and L. Lee, "Bilingual acoustic modeling with state mapping and three-stage adaptation for transcribing unbalanced code-mixed lectures," in *Proc. ICASSP*, 2011, pp. 5020–5023.

[11] P. Lehohla, *Census 2011: Census in brief*. Statistics South Africa, 2012.

[12] M. L. Mokwana, "The melting pot in Ga-Matlala Maserumule with special reference to the Bapedi culture, language and dialects," Master's thesis, University of South Africa, South Africa, 2009.

[13] T. I. Modipa, M. H. Davel, and F. de Wet, "Context-dependent modelling of English vowels in Sepedi code-switched speech," in *Proc. PRASA*, 2012, pp. 173–178.

[14] E. Lebese, J. Manamela, and N. Gasela, "Towards a multilingual recognition system based on phone-clustering scheme for decoding local languages," in *Proc. SATNAC*, 2012, pp. 237–242.

[15] N. J. de Vries, J. Badenhorst, M. H. Davel, E. Barnard, and A. de Waal, "Woefzela: An open-source platform for ASR data collection in the developing world," in *Proc. Interspeech*, 2011, pp. 3177–3180.

[16] M. H. Davel, C. J. van Heerden, and E. Barnard, "Validating smartphone-collected speech corpora," in *Proc. SLTU*, 2012, pp. 68–75.

[17] T. Modipa, M. H. Davel, and F. de Wet, "Acoustic modelling of Sepedi affricates for ASR," in *Proc. SAICSIT*, 2010, pp. 394–398.

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