# Context-dependent modelling of English vowels in Sepedi code-switched speech

Thipe I. Modipa* †, Marelie H. Davel†, Febe de Wet*,
*Multilingual Speech Technologies, North-West University, Vanderbijlpark 1900, South Africa
†Human Language Technologies Research Group, CSIR Meraka Institute, Pretoria, South Africa
Email: {tmodipa,fdwet}@csir.co.za, marelie.davel@gmail.com

*Abstract*—**When modelling code-switched speech (utterances that contain a mixture of languages), the embedded language often contains phones not found in the matrix language. These are typically dealt with by either extending the phone set or mapping each phone to a matrix language counterpart. We use acoustic log likelihoods to assist us in identifying the optimal mapping strategy at a context-dependent level (that is, at triphone, rather than monophone level) and obtain new insights in the way English/Sepedi code-switched vowels are produced.**

## I. Introduction and Background

Code switching – using words and phrases from more than one language within a single utterance – is a common phenomenon among multilingual speakers. There are a number of reasons why multilingual speakers engage in code switching. In the case of Sepedi, speakers often use a foreign language (English) for numbers, dates and time, a phenomenon that has been observed in other South African languages as well [1].

For automatic speech recognition (ASR) systems, code-switched (CS) speech provides an interesting challenge. This can be dealt with by building fully multilingual systems (combining dictionaries, language and/or acoustic models from multiple languages) or by running more than one monolingual system in parallel, switching from the one to the other [2], [3]. We are interested in the first approach, and specifically where acoustic models are combined at the phone or sub-phone level.

Various techniques have been used when deciding how and when to combine the acoustic model of a phone from the embedded language (English in this case) with a phone from the matrix language (Sepedi in this case). One such technique consists of mapping the embedded phones to the matrix phones prior to system training. This can be achieved in different ways, specifically:

- Using IPA features directly: mapping phones based on existing linguistic knowledge. (IPA features classify sounds based on the phonetic characterisation of those speech sounds [2]).
- Using a confusion matrix from an existing ASR system: calculating the rate of confusion between two phones using a phoneme recogniser in the matrix language and acoustic data from the embedded language [3].
- Using log likelihood differences directly as a distance measure that tests how well two different models fit the same data [4], [2].
- Using acoustic distance measures such as Kullback-Liebler measure, Battacharyya distance metric, Mahalanobis measure or a simple Euclidean measure [5].
- Using a probabilistic phone mapping [6], that is, a model for mapping phones between source sequence X, and target sequence Y, where the model parameters are given by
$$PM(x \mid y) : x \in X, y \in Y \tag{1}$$
and this model is estimated from the results of a phoneme recogniser and the modelled pronunciations. Note that this model (like the current work) is context-sensitive.

In an earlier analysis of English/Sepedi CS speech [7], it was found that applying grapheme-to-phoneme (g2p) rules of the matrix language (Sepedi) to the code-switched words directly, outperformed more sophisticated mapping approaches, and specifically one whereby the g2p rules of the embedded language (English) is used to predict possible pronunciations and these then mapped on a per-phone basis to the closest matching Sepedi phone. This was an unexpected result: it could either mean that the mapping used (obtained from a confusion matrix, as described in [7] ) was not optimal, or that Sepedi speakers do interpret some English words according to Sepedi pronunciation rules, for example, pronouncing the word 'chocolate' as / S O k O l a t / rather than as / t S Q k l @ t / (using X-SAMPA notation).

In this work we investigate the process of obtaining a phone mapping from the embedded language to the matrix language. The main goal is to determine whether a better mapping can be obtained, given the specific corpus we are modelling, and to explore tools to analyse this task. We focus on English vowels (English consonant mappings are more predictable), and investigate the use of model likelihoods to guide the mapping choice at a context-dependent level. When unlimited training data is available, using all matrix language and embedded language models combined is expected to perform best; with constrained corpora, extending the phone set indiscriminately is expected to hurt performance due to data scarcity. The optimal mapping is therefore dependent on the specific speech corpus being modelled: our goal is to investigate tools that can guide this mapping process.

In the current work we first verify and extend the earlier English/Sepedi code-switched ASR results (as discussed above) to determine whether these were corpus-specific or whether trends are retained across corpora; we then use log likelihood ratios to analyse the possible context-dependent phone mappings from the embedded language phones to the matrix language phones.

The paper is organised as follows: In Section II we describe the approach we use to analyse context-dependent mappings. In Section III we describe the speech corpora used in a fair amount of detail, as this provides the context for the various experiments undertaken. Experiments and results are discussed in Section IV. Section V summarises the findings from this analysis and provides some suggestions for future work.

## II. Approach

The approach we use to determine an optimal phone mapping is fairly straightforward: we score the English vowels against context-dependent acoustic models of vowels from both the embedded and matrix language and compare the likelihood ratios. These ratios give us an indication of 'model closeness' and suggest mapping candidate(s) at a triphone level. We analyse these mapping candidates to determine whether a triphone should be mapped, and if so, to which matrix language triphone.

The specific process we use to determine mappings is as follows:

1) Context-dependent acoustic models are trained with pure Sepedi data (not containing any code-switched speech).
2) Context-dependent acoustic models are trained from the available Sepedi code-switched data by extending the Sepedi phone set with all English phones.
3) For each English phone, possible mapping candidates are selected using a confusion matrix (as described in more detail later in Section IV-A). Note that these mapping candidates are selected at the monophone level.
4) Analysis is performed at triphone level:
    a) The English data is force aligned using the English triphone model.
    b) The same data is similarly aligned using each of the Sepedi candidate triphone models. These models are constructed from the actual left and right contexts observed, with only the centre phone replaced.
    c) The likelihood ratio between (a) and (b) is evaluated per candidate triphone, per code-switched sample, in practice by calculating the difference in log likelihood, for each English triphone $e$, matching Sepedi candidate $s_e$ and data sample $d_e$, referred to from here onwards as $ll\_diff(e, s_e, d_e)$.
    d) The average of the values in (c) is obtained per candidate triphone $s_e$ by averaging over all data samples $d_e$, giving a single value of $ll\_mean(e, s_e)$ per English and Sepedi candidate triphone pair.
5) The relative scores are used to determine mappings:
    a) If there is a clear Sepedi triphone winner, only that candidate triphone is selected for mapping, that is, if the difference in $ll\_mean(e, s_e)$ between two candidate triphones exceeds a threshold $\alpha$.
    b) If there is not a clear winner, all triphone candidates that have a value of $ll\_mean(e, s_e)$ less than a second threshold value $\beta$ are selected as possible mappings (introducing variants for that specific context).
    c) For triphones that have no suitable counterpart (no candidate mappings that obtain a value of $ll\_mean(e, s_e)$ smaller than $\beta$), phone set extension is considered.

## III. Data

In this section we describe the data used during experiments: the audio corpora, phone sets and dictionaries.

### A. Audio corpora

We use two different audio corpora for the experiments: a general Sepedi corpus (NCHLT [8]) and a custom-designed code-switched corpus (SPCS [9]).

The NCHLT corpus was collected using a locally developed smart-phone based speech data collection tool, Woefzela [8]. The corpus consists of prompted speech, mostly in Sepedi but also including some English speech (generated from English text) as produced by Sepedi first language speakers. The corpus consists of 12 560 unique word tokens produced by 113 speakers. We use both the full corpus (referred to as $nchlt\_all$ from here onwards) consisting of all Sepedi and English data and create a subset ($nchtl\_sep$) consisting only of pure Sepedi utterances. This corpus contained no code-switched sentences. Table I shows the distribution of male and female speakers, and the duration of the train and test sets in the different corpora.

TABLE I
*Distribution of the number of male and female speakers.*

| | | Speakers | Duration (min) |
|---|---|---|---|
| nchlt_sep | Train | 92 (38 female, 54 male) | 1 417.62 |
| | Test | 20 (10 female, 10 male) | 247.28 |
| nchlt_all | Train | 82 (33 female, 49 male) | 2 782.48 |
| | Test | 30 (15 female, 15 male) | 1 055.68 |

The SPCS corpus was collected using prompts that were derived from code-switched transcriptions generated from actual radio broadcasts [9]. It was also collected using Woefzela. Twenty speakers (12 females, 8 males) each read approximately 450 utterances, resulting in 10 hours of prompted speech.

Table II lists the number of unique English and Sepedi words found in the corpus. As discussed in [9], we also list *semi-modified* words (giving a total of 787 unique words): English words that are transformed when embedded in Sepedi speech, for example the word *graduate* that can be pronounced as *graduata* when used within general Sepedi speech.

TABLE II
*Number of unique words and total number of utterances in the SPCS corpus.*

| # Semi-modified | # Eng words | # Sepedi words | # Utterances |
|---|---|---|---|
| 58 | 345 | 384 | 12 386 |

### B. Phone sets and dictionaries

The pronunciation rules were obtained from two sources:

1) Standard Sepedi g2p rules (Default&Refine [10] trained on the 5 000-word Lwazi dictionary [11]). In addition, affricates were split according to [12] resulting in 32 Sepedi phones being used in practice.
2) English g2p rules (Default&Refine trained on a South African English (SAE) dictionary created using manually created British-to-SAE phone-to-phone (p2p) mappings [13])

All pronunciations of words occurring in the SPCS corpus were manually verified and corrected, where necessary. The final dictionary contained 29 phones that occur in English but are not found in the Sepedi phone set, as shown in Table III.

TABLE III
*Number of phones of different categories found in the various phone sets used.*

| | Sepedi standard | Sepedi split affricates | English | English phones not occurring in Sepedi |
|---|---|---|---|---|
| Affricates | 9 | - | 2 | 1 |
| Fricatives | 11 | 11 | 10 | 5 |
| Stops | 8 | 8 | 6 | 6 |
| Nasals | 5 | 5 | 3 | - |
| Vowels | 7 | 7 | 12 | 8 |
| Trill | 1 | 1 | - | - |
| Approximants | 4 | 4 | 4 | 1 |
| Diphthongs | - | - | 8 | 8 |
| Total | 45 | 36 | 45 | 29 |

## IV. Experiments and Results

First, we repeat the experiments as performed in [7] on the NCHLT corpus, for two reasons: to determine whether trends are consistent across corpora, and to obtain a comparable baseline for the phone mapping analysis. Once the baseline has been established we analyse the context-dependent likelihood ratios for the English vowels to obtain a possible mapping.

### A. Baseline ASR systems

As a baseline implementation we create four systems on the same training data ($nchlt\_all$) using four standard approaches, the first three of which were used in [7]:

1) Sepedi-only phone set: all words (English and Sepedi) are predicted using Sepedi g2p.
2) Extended phone set: English words are predicted using English g2p, Sepedi words are predicted using Sepedi g2p and all phones retained.
3) Mapped phone set: All English phones (from (2)) are mapped to the single best candidate based on a confusion matrix; no English phones are retained. The confusion matrix was obtained as follows:
   - Freely decoded phone-level labels are obtained from the Sepedi system (using $nchlt\_all$, but only Sepedi phones).
   - The SCPS data is aligned using a dictionary containing the extended phone set (English and Sepedi phones).
   - Iterative dynamic programming (using tools from [14]) is used to obtain an accurate confusion matrix at phone-level.
   - For every English phone, the Sepedi phone with the highest confusability is selected.
4) Code switched variants: Sepedi pronunciations from (1) and English mapped pronunciations from (3) are added as variants both during training and testing.

All four systems are created in a similar way: a fairly standard Hidden Markov Model (HMM) based ASR system is implemented using the HTK toolkit [15]. Acoustic models consist of cross-word tied-state triphones modelled using a 3-state continuous density HMM. Each HMM state distribution is modelled by an 8-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 acceleration coefficients appended. The Cepstral Mean and Variance Normalisation (CMVN) preprocessing is used and Semi-tied transforms applied.

These four systems are then tested on three different test sets, obtained from the Sepedi-only NCHLT data ($nchlt\_sep$), all NCHLT data ($nchlt\_all$) and all SPCS data ($spcs$), respectively. Note that the SPCS data is always used as a test set: it is never included in data used either for training or system tuning.

TABLE IV
*Phone error rates of different baseline systems on each of three test sets.*

| Test set | Sepedi-only | Extended phone set | Mapped phone set | CS variants |
|---|---|---|---|---|
| nchlt_sep | 30.72 | 45.09 | 33.65 | 31.92 |
| nchlt_all | 33.37 | 42.54 | 34.32 | 35.88 |
| spcs | 39.63 | 56.46 | 44.16 | 42.27 |

The phone error rates (PER) of NCHLT and SPCS test data using different approaches to modelling code switched words are obtained as shown in Table IV. Utterances that cannot be decoded by any of of the systems are removed from the corpus to ensure a fair comparison across systems.

In this careful analysis across different test sets, we see that the previously observed trends remain consistent: Sepedi-only g2p provides the most effective approach to dealing with code-switched speech. Simply extending the phone set results in a large increase in error rate. When the English phones are mapped to their Sepedi counterparts, error rate decreases (compared to the extended phone set); error rate again decreases when two variants (the English remapped version and the Sepedi g2p version) are added per code-switched word. Even though error rates decrease during this process, the best results are still obtained when using a straightforward Sepedi g2p prediction.

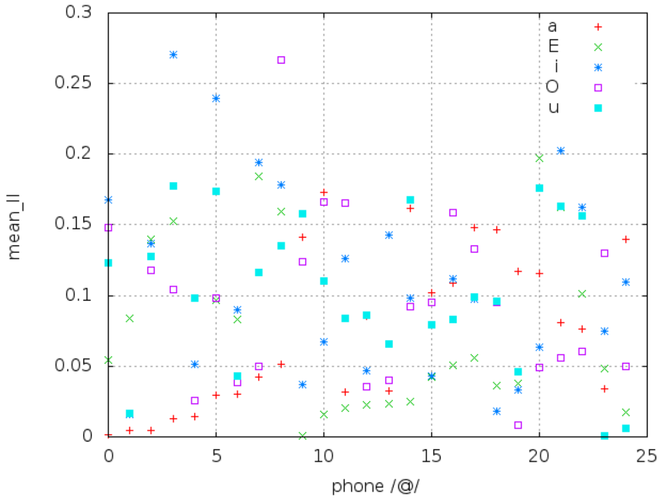

Fig. 2. Mean log likelihood differences (*ll_mean*) for one phone /@/ in different context and mapping candidates /a/, /E/, /i/, /O/ and /u/.

### *B. Selecting candidate mappings*

We obtain mapping candidates from the same confusion matrix described in section IV-A (previously used to identify a single best match). This time, we flag all phones that are confused with the target phone more than 20% of the target phone occurrences.

Table V lists the frequency of occurrence of the English vowels in the NCHLT training set, and the SPCS corpus, respectively. For each vowel, the mapping candidates are identified and per candidate, the number of times a target phone to mapping candidate pair was observed in the confusion matrix is provided in brackets. We also show the number of unique phone contexts observed in the SPCS corpus.

TABLE V

*Phone mapping candidates obtained from confusion matrix. For each English vowel, the number of times it was observed in each corpus is provided. For each phone-candidate pair, the number of times that the confusion was observed in the testing data is provided in brackets.*

| phone | train counts<br>(nchlt_all) | test counts<br>(spcs) | candidates | unique phone<br>contexts |
|---|---|---|---|---|
| @ | 59 652 | 10 445 | a (4448), E (2534)<br>i (1165), O (1156)<br>u(78) | 121 |
| i: | 21 789 | 711 | i (389), E (205) | 15 |
| A: | 2 731 | 749 | a (635), E (51) | 11 |
| { | 2 265 | 2 479 | a (1775), E (536) | 39 |
| u: | 1 220 | 1065 | u (434), O (216 ) | 23 |
| Q | 1 214 | 1811 | O (1208), a (429) | 32 |
| O: | 1 174 | 1 333 | O (1009), a (283) | 19 |
| E: | 972 | 991 | E (663), a (196) | 18 |

### *C. Context-dependent analysis*

Once the mapping candidates have been identified, the triphone analysis as described in Section II (4) can be performed. The English models are obtained from the $nchlt\_all$ corpus and the Sepedi models from the $nchlt\_sep$ corpus. The $ll\_mean(e, s_e)$ values are calculated for all the vowels $e$ and mapping candidates $s_e$ as listed in Table V. In this work, we only consider contexts where the left and right contextual phones occur in both the English and Sepedi phone sets. (This means, for example, that we do not include a triphone such as /T-Q+@/ in the current analysis.)

To illustrate the concept, we first plot the results for a single context /S-@+n/ when found in different words. Results are averaged over all speakers. As can be seen in Fig. 1, the best matching context (/S-E+n/) is always the closest match, irrespective of the word in which it is used. The runner up is /S-i+n/: this context always provides a poorer match than /S-E+n/, with results most comparable in the word ‘national’, which interestingly, does have a different morphological construct than the others. The results displayed in Fig. 1 is better contextualised by considering the mean log likelihood difference between standard Sepedi /S-E+n/ contexts and the Sepedi /S-E+n/ model, which is 0.004 (indicated in Fig. 1) by a horizontal line.

In Fig. 2 we provide the same results, but now averaged over all words that contain a specific context. We plot the results for one phone /@/ when found in different contexts. Again, results are averaged over all speakers. From Fig. 2 it is clear that /E/ provides the best match in general, but that there are some contexts where other phones are better mapping candidates. The phones /a/, /O/ and /i/ also provide best matches in a limited number of contexts, whereas the phone /u/ only provides a best match in two instances.

This process was repeated for all the vowels. Two more examples are shown in Figures 3 and 4, illustrating the mean log likelihood differences for vowels /Q/ and /{/, respectively.

When this process is repeated for additional contexts, we are able to identify additional context-dependent Sepedi candidates that provide the best match to each of the context-dependent English vowels.

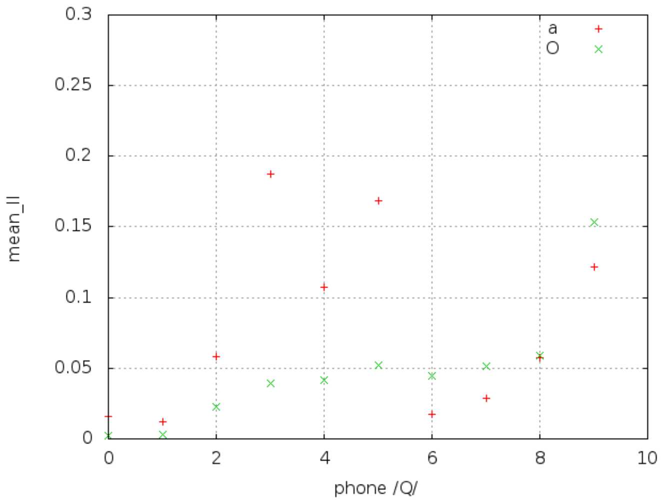

Fig. 3. Mean log likelihood differences (*ll_mean*) for phone /Q/ in different context and mapping candidates /a/ and /O/.

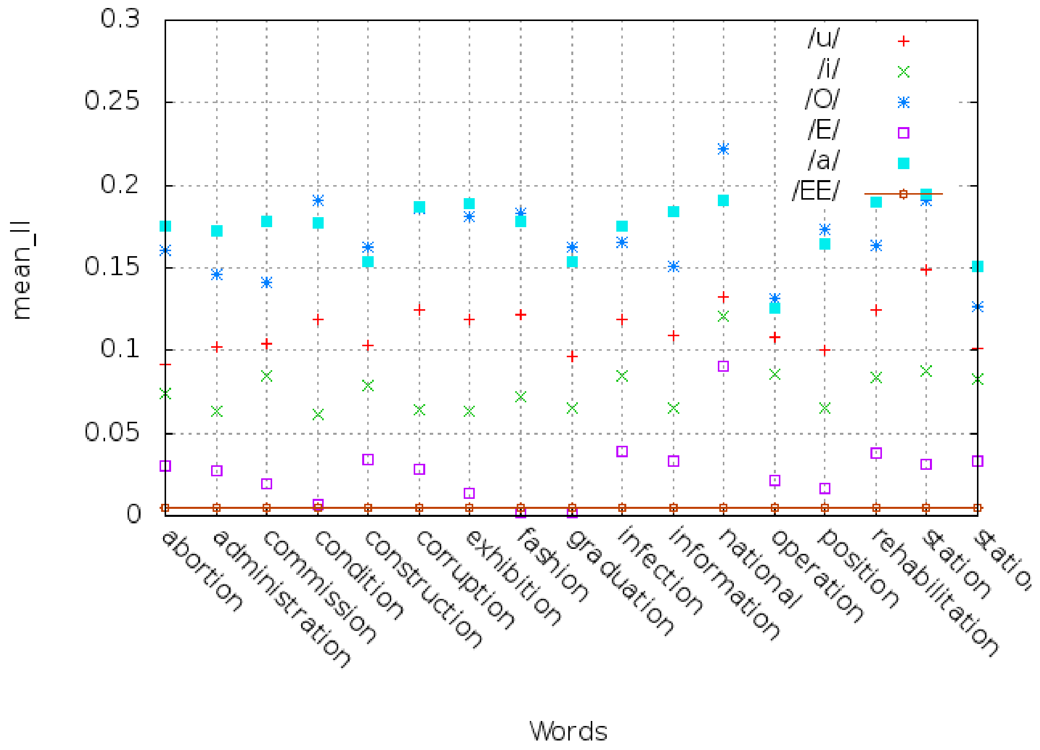

Fig. 1. Mean log likelihood differences ($ll\_mean$) for one context /S-@+n/. Each mapping candidate is displayed using a different colour. /EE/ is displayed as calibration: the $ll\_mean$ of standard Sepedi /E/ data measured against the standard Sepedi /E/ model.

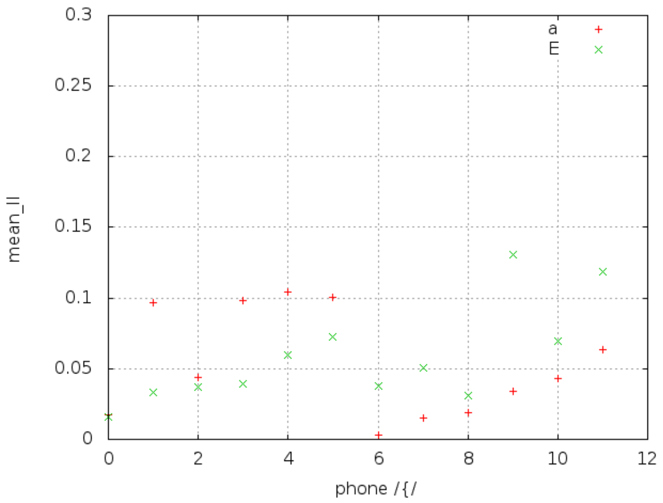

Fig. 4. Mean log likelihood differences ($ll\_mean$) for phone /{/ in different context and mapping candidates /a/ and /E/.

### D. Obtaining a mapping from likelihood results

The above likelihood results are used to determine possible actions to take with regard to the English vowels. As mentioned in Section II, the possible options per phone context are to:

1) Extend the matrix language phone set by adding the embedded language phone (if no candidate with an $ll\_mean$ value of less than $\beta$);
2) Map the embedded language phone to the single closest matrix language phone; or
3) Map the embedded phone to more than one candidate matrix phone (if candidates closer than $\alpha$).

Both $\alpha$ and $\beta$ can be tuned on a development set. The context-dependent mapping is obtained by finding the most appropriate candidate triphones using these thresholds. For every winning candidate triphone (see 2), we determine which other candidate triphone is within the defined threshold.

In order to illustrate the concept, we use the analysis in IV-C to select thresholds such that $\alpha$ is 0.02 and $\beta$ is 0.1 (implying that the phone set is not extended).This results in the mappings determined for /@/, as shown in Table VI.

## V. Conclusion

In this investigation, we have shown that acoustic log likelihoods provide a useful tool when analysing the optimal mapping of embedded language phones to matrix language phones, and that context is important when applying such mappings. We also introduced a new corpus of Sepedi/English codes-switched speech, and confirmed that (for this corpus, as found earlier in [7]), Sepedi g2p predictions of the pronunciations of English words provide a viable alternative to more sophisticated modelling approaches, and that, in fact, it is difficult to obtain a better alternative with context-insensitive mappings.

TABLE VI

*The context-dependent mapping for phone /@/.*

| Phone | Mapping |
|---|---|
| n-@+S | a |
| s-@+m | a,i,O,u |
| m-@+f | a |
| S-@+l | a |
| m-@+sil | a,O |
| n-@+l | a |
| d_0Z-@+l | a,O,u |
| d_0Z-@+h_b | a,O |
| s-@+d_0Z | a |
| f-@+f | E |
| S-@+n | E |
| s-@+l | a,E |
| n-@+m | E,O |
| s-@+n | a,E,O |
| h_b-@+l | E |
| n-@+s | E,i |
| d_0Z-@+n | E |
| m-@+n | E |
| s-@+s | E,i |
| l-@+s | O |
| s-@+w | i,O |
| i-@+w | O |
| l-@+n | a,O |
| i-@+f | u |
| l-@+d_0Z | E,u |

The next step in our research will be to determine the impact of the identified mappings on ASR system performance. This will also require a thorough investigation of the thresholds $\alpha$ and $\beta$, balancing the need for accurate mappings with the additional confusability introduced by extra pronunciation variants.

Future work will include extending the phone mapping analysis to contexts where the left and right phones themselves are only in one of the two phone sets. This will also allow us to extend the analysis to the full phone set by iteratively mapping phones, in the process increasing the matched phone sets. In addition, we would like to analyse whether some of the observed mappings are speaker-specific, or robust across speakers (the current assumption); and whether the graphemic context of the triphone also plays a role in producing an optimal mapping.

While the above would provide a practical (and more nuanced) tool when producing phone mappings for code-switched speech, the current analysis already provides some interesting insights with regard to the acoustic properties of English/Sepedi code-switched speech.

## REFERENCES

[1] T. Niesler and D. Willett, “Language identification and multilingual speech recognition using discriminatively trained acoustic models,” in *Multilingual Speech and Language Processing*, 2006.

[2] D. Yu, L. Deng, P. Liu, J. Wu, Y. Gong, and A. Acero, “Cross-lingual speech recognition under runtime resource constraints,” 2009.

[3] V. B. Le, L. Besacier, and T. Schultz, “Acoustic-phonetic unit similarities for context dependent acoustic model portability,” in *Proc. ICASSP*, 2006.

[4] 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.

[5] J. J. Sooful and E. C. Botha, “Comparison of acoustic distance measures for automatic cross-language phoneme mapping,” in *Proc. ICSLP*, 2002.

[6] K. C. Sim and H. Li, “Context-sensitive probabilistic phone mapping model for cross-lingual speech recognition,” in *Proc. Interspeech*, 2008, pp. 2715–2718.

[7] T. Modipa and M. H. Davel, “Pronunciation modelling of foreign words for Sepedi ASR,” in *Proc. 21st Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, 2010, pp. 185–189.

[8] 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 *Proceedings of Interspeech*, 2011, pp. 3177–3180.

[9] T. Modipa, F. de Wet, and M. H. Davel, “An acoustic corpus of english/sepedi code-switched speech,” *South African Journal of African Languages*, in preparation.

[10] M. Davel and E. Barnard, “Pronunciation prediction with Default&Refine,” *Computer Speech and Language*, vol. 22, pp. 374–393, 2008.

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

[12] T. Modipa, M. H. Davel, and F. de Wet, “Acoustic modelling of Sepedi affricates for ASR,” in *Proc. Annual Research Conference of the South African Institute of Computer Scientist and Information Technologists (SAICSIT 2010)*, 2010, pp. 394–398.

[13] L. Loots, M. H. Davel, E. Barnard, and T. Niesler, “Comparing manually-developed and data-driven rules for p2p learning,” in *Proc. 20th Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, 2009, pp. 35–40.

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

[15] HTK, “The Hidden Markov Model Toolkit (HTK),” 2009.