Predicting vowel substitution in code-switched speech

Thipe I. Modipa[1,2] and Marelie H. Davel[2,3]
[1]HLT Research Group, CSIR Meraka, South Africa.
[2]Multilingual Speech Technologies, North-West University, Vanderbijlpark, South Africa.
[3]CAIR, CSIR Meraka, South Africa.
tmodipa@csir.co.za, marelie.davel@gmail.com

***Abstract*— The accuracy of automatic speech recognition (ASR) systems typically degrades when encountering code-switched speech. Some of this degradation is due to the unexpected pronunciation effects introduced when languages are mixed. Embedded (foreign) phonemes typically show more variation than phonemes from the matrix language: either approximating the embedded language pronunciation fairly closely, or realised as any of a set of phonemic counterparts from the matrix language. In this paper we describe a technique for predicting the phoneme substitutions that are expected to occur during code-switching, using non-acoustic features only. As case study we consider Sepedi/English code switching and analyse the different realisations of the English schwa. A code-switched speech corpus is used as input and vowel substitutions identified by auto-tagging this corpus based on *acoustic* characteristics. We first evaluate the accuracy of our auto-tagging process, before determining the predictability of our auto-tagged corpus, using *non-acoustic* features.**

I. Introduction

Code switching tends to occur wherever speakers are exposed to more than one language. That is, multilingual speakers tend to mix languages naturally: embedding words or phrases from one language into speech primarily produced in a different language. As code switching introduces additional variability with regard to all aspects of speech – vocabulary, word usage and pronunciation – the presence of code-switched speech poses a challenge to automatic speech recognition (ASR) systems.

In an earlier study of Sepedi/English code-switched speech speech [1], an analysis of Sepedi radio broadcasts indicated that an unexpectedly high percentage (approximately 30%) of Sepedi utterances contained English words or phrases. It was also found that these words/phrases degrade ASR recognition significantly, resulting in an approximately 10% absolute degradation in performance. In the study, no special provision was made for these code-switched content, apart from including both English and Sepedi speech in the training data. Interestingly, by adding a 'Sepedi' pronunciation of English words by blindly performing grapheme-to-phoneme (G2P) prediction using Sepedi G2P models, some recognition improvement was observed by the authors.

This last observation was the primary motivation for the current study. In order to determine whether a more sophisticated approach to pronunciation modeling of the English words could improve results further, we first ask how predictable phoneme substitutions really are. Specifically we would like to determine whether vowel substitutions can be predicted based on non-acoustic features such as vowel context, word orthography or even speaker characteristics. We focus on vowels, as they exhibit significantly more variability in pronunciation than consonants, and specifically on the schwa, as its pronunciation tends to be the most unpredictable of all English vowels found in code-switched speech [2]. We approach this task by (a) first labeling a speech corpus with vowel tags based on an acoustic analysis, and then (b) determining how predictable these tags are using non-acoustic features.

The paper is structured as follows: Section II provides related background, specifically with regard to recognising code-switched speech, Sepedi/English code-switching and the 'Goodness of Pronunciation' score: a technique utilised in this study. Section III describes the approach we followed for the analysis and prediction of vowel substitution in code-switched speech. Findings are presented in Section IV, before we conclude with a summary of our findings in Section V.

II. Background

Code switching can be defined as the use of words or phrases from more than one language [3]. Such *code switching* can occur within a sentence, which is normally referred to as *intra-sentential* code switching. When the switching of language occurs between sentences the process is referred to as *inter-sentential* code switching [4]. During code switching, speakers may use words, phrases or sentences from one language (the *embedded language*) in combination with words or sentences in their primary or *matrix language* [3].

Code switching is a common phenomenon and brings with it challenges for ASR systems. These challenges can be overcome by building multilingual speech recognition systems consisting of multilingual pronunciation dictionaries, multilingual acoustic models, and multilingual language models. As an alternative, monolingual speech recognition systems can be run in parallel thereby switching from one system to the other [5], [6].

We study the development of combined, multilingual systems, and specifically the process to construct a multilingual phone set. The most popular approaches to the development of the phone set and phone-to-phone mappings, are as follows:

- Combining the phone set from multiple languages [7].

- Mapping the embedded phone set to the matrix phone set using IPA features directly. (IPA features classify sounds based on the phonetic characterisation of those speech sounds [5])
- Mapping highly confusable phones from the embedded to the matrix language based on a confusion matrix obtained from an existing ASR system.
- Merging language-dependent phone sets using hierarchical phone clustering algorithms and acoustic distance measures [7].
- Mapping phones between source and target sequences using probabilistic phone mapping [8]

In earlier work [1], [2], initial results with regard to the implications of Sepedi-English code switching for ASR systems were obtained on two custom-developed corpora. Specifically, in [2] a subset of an existing Sepedi corpus was selected and processed in order to isolate English events; and in [1] a new corpus was developed containing instances of code-switching as observed in radio broadcasts. This corpus, Sepedi Prompted Code-Switched (SPCS), was aimed to capture a speaker-specific pronunciation variability introduced during code switching.

The Goodness-of-Pronunciation (GOP) score was initially developed by Witt and Young in the context of phone-level pronunciation assessment [9]. Defined as the duration-normalised log of the posterior probability (that the speaker uttered a specific phone, given the acoustic data), it is approximated by the difference in log likelihood of the target and best matching phone, divided by the number of frames in the segment, that is:

$$GOP(q) = \left| log \frac{(p(x|q)}{(p(x|q')} \right| /NF(x) \qquad (1)$$

where q is the target phone, x the observed data, $NF(x)$ the number of frames obseved and q' the model that matches the observed data best. In practice, the log likelihood scores are obtained directly from the ASR system, a hidden Markov model (HMM)-based one in our case, and q' identified during a free phone decode.

GOP was developed for phone-level analysis. In [10] a word-level version of GOP was defined, with two variants – frame-based and phone-based – depending on how duration normalisation is applied. In the same study, it was found that triphones provide more accurate results than monophones for word-level analysis. (Monophones are more typical of GOP scores used for phone-level pronunciation assessment). Word-level GOP, using triphones and frame-level normalisation, forms the basis for the analysis peformed here.

III. Approach

In code-switched speech, vowels are typically produced in one of two ways: the true embedded language pronunciation is produced or at least approximated fairly closely, or the target phone is substituted for a counterpart from the matrix language. As there is always the possibility of producing the true pronunciation, modeling both these possibilities requires that variants be introduced to the pronunciation lexicon: both 'embedded language' and 'matrix language' versions for each code-switched word are therefore required. Examples of such pronunciation variants are shown in Table I, using SAMPA notation[1].

TABLE I
EXAMPLES OF EMBEDDED AND MATRIX PRONUNCIATIONS

word	embedded language (English)	matrix language (Sepedi)
pressure	/p r\E S @ r\/	/ p r E S a /
fifteen	/ f @ f t i: n /	/ f i f t i n /

Our goal is then to determine which features influence vowel substitutions when they *do* occur. That is, we would like to predict which vowel substitutions can be expected in the matrix pronunciation, specifically. Given the two examples in Table I, we therefore would like to predict that $/@/ \rightarrow /a/$ in one case, and $/@/ \rightarrow /i/$ in the other.

As introduced earlier, we first identify which vowel substitutions occur by auto-tagging a speech corpus based on acoustic characteristics. In order to determine the accuracy of the auto-tagger, we manually create a small labeled test set and evaluate the accuracy of the auto-tags against the manual labels. Once the auto-tagging process has been verified, we then tag a much larger corpus and determine the predictability of this auto-tagged corpus, using non-acoustic features.

In the remainder of this section, we discuss our approach to the auto-tagging process (Section III-A), the development of the manually labeled test set (Section III-B) and the feature selection and classification process (Section III-C). All experiments are performed using the SPCS corpus, introduced in section II.

A. Auto-tagging process

The SPCS corpus is first partitioned into a training and test set, and the training set is used to develop a standard Hidden Markov Model (HMM) based ASR system. The system is implemented using the HTK toolkit [12]. 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. Cepstral Mean and Variance Normalization (CMVN) preprocessing is used and semi-tied transforms are applied.

During training, a pronunciation lexicon is used that retains the English schwa as a unit. Every word in this lexicon has a single variant: Sepedi words are modelled using Sepedi pronunciation models and English words using English pronunciation models. Sepedi pronunciations are obtained directly from G2P models (default-and-refine models [13] trained on the NCHLT *in-lang* dictionaries [14]). English

[1]The 'Speech Assessment Methods Phonetic Alphabet' is a standard computer-readable notation for phoneme descriptions. See [11].

words are first predicted using English G2P models[15], then manually reviewed. (As English words form the focus of the analysis, accurate pronunciations are expected to be important for the rest of the study.)

Once the ASR system has been trained, the same training data is realigned, using five different options: each replacing the schwa in the pronunciation lexicon with a different Sepedi vowel (/a/, /E/, /i/, /O/ or /u/). The resulting alignments then produce both timing information, as well as the likelihood of each vowel, given the data. These aligments are referred to as '$spcs\_@$', '$spcs\_a$', '$spcs\_E$', etc. depending on which vowel was used during alignment.

Time alignments were manually verified to be accurate before proceeding. Accurate alignments were only obtained when training and aligning on the same corpus. Note that the lexicons are only changed during alignment, no re-training is performed. This ensures that all the data being studied are combined in the schwa-model, and not incorporated into any of the other vowel-models.

Once the likelihoods have been obtained, word-level GOP scores are extracted for each alignment option. Word-level GOP scores are used, as timing information otherwise introduces unnecessary variability. Word beginning and end times stay fairly consistent, while phone beginning and end times may differ significantly, especially if there is a mismatch between the vowel actually produced and the alignment candidate. For the same reason, frame-based (rather than phone-based) GOP scores are extracted, as defined in [10].

Two main results are obtained from the tagging process, for each schwa observed:

- The most likely vowel candidate (apart from schwa).
- Whether the phone matches the broad schwa category best, or is better matched to one of the Sepedi vowels.

These results form the basis of the analysis described in Section IV.

B. Manual tagging

In order to evaluate the effectiveness of the auto-tagging process, we created a manually labelled verification set. Two subjects were asked to listen to the words in question independently, and manually label each schwa with the most probable phone produced. Subjects were encouraged to only select 'schwa' if a true schwa was produced, and to otherwise select the closest vowel candidate. Both subjects were bilingual English/Afrikaans speakers, with subject B a trained linguist with exposure to Northern Sotho.

C. Classification process

Once the entire corpus has been labelled (using the auto-tagging procedure from section III-A), we review the results and identify possible features that influence vowel prediction. We use Naive Bayes classification to obtain an indication of whether these features are applicable. This is discussed further in Section IV.

IV. Results and Analysis

We first analyse the validity of the manual tagging process (Section IV-A) before evaluating the accuracy of the auto-tagging process (Section IV-B). The influence of various possible non-acoustic features is evaluated in Section IV-D, before considering the overall predictability of vowel substitutions in Section IV-E.

A. Manual labels: inter-subject agreement

Any labels that were not true examples of code-switched schwas were removed from the evaluation set. (For example, where 'pressure' is only produced as / p r\E S / without articulating the last phoneme at all.) The number of remaining labels and inter-subject agreement measured using these labels, are shown in Table II. While inter-subject agreement is medium to fair at 70.0%, a review of the data shows that disagreement is mainly due to schwa boundaries being drawn in different places by the two subjects. If only labels that are not considered to be very close to schwa by either of the subjects are considered, inter-subject agreement increases to 85.7%.

TABLE II
Inter-subject agreement during manual tagging.

	#observations	#agreement	%agreement
All labels	50	35	70.00
Schwa excluded	35	30	85.71

B. Accuracy of the auto-tagger

As the auto-tagger is restricted to always select a possible alternative pronunciation (excluding schwa), it only makes sense to evaluate tagging accuracy against non-schwa labels. For the sake of completeness, results on both the full set and the non-schwa set are included in Table III.

TABLE III
Accuracy of the auto-tagger, when measured against different manually labeled test sets.

	#observations	#agreement	%agreement
All labels: subject A	53	23	43.40
All labels: subject B	53	28	52.83
Schwa excluded: subject A	38	23	60.53
Schwa excluded: subject B	33	28	84.85
Schwa excluded: overall	71	51	71.83

From this analysis, it is clear that the auto-tagging process produces usable results, but agrees more strongly with subject B. While overall agreement is only at 71.8%, agreement between the auto-tagging process and subject B reaches 84.9%.

While the discrepancies may simply be due to greater transcription skill exhibited by subject B, it is worth understanding where human error occurs. For this reason, a formant figure was created as shown in Fig. 1 and Fig. 2.

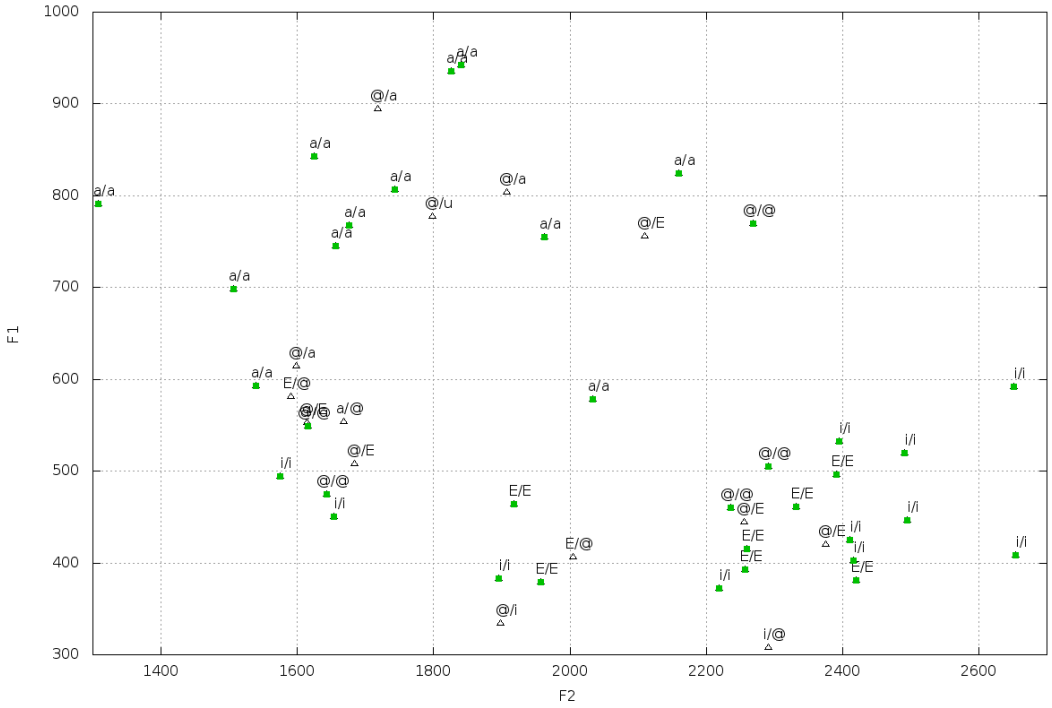

Fig. 1. F1/F2 positions of labels. Each A/B legend displays the tag provided by subject A and B,respectively.

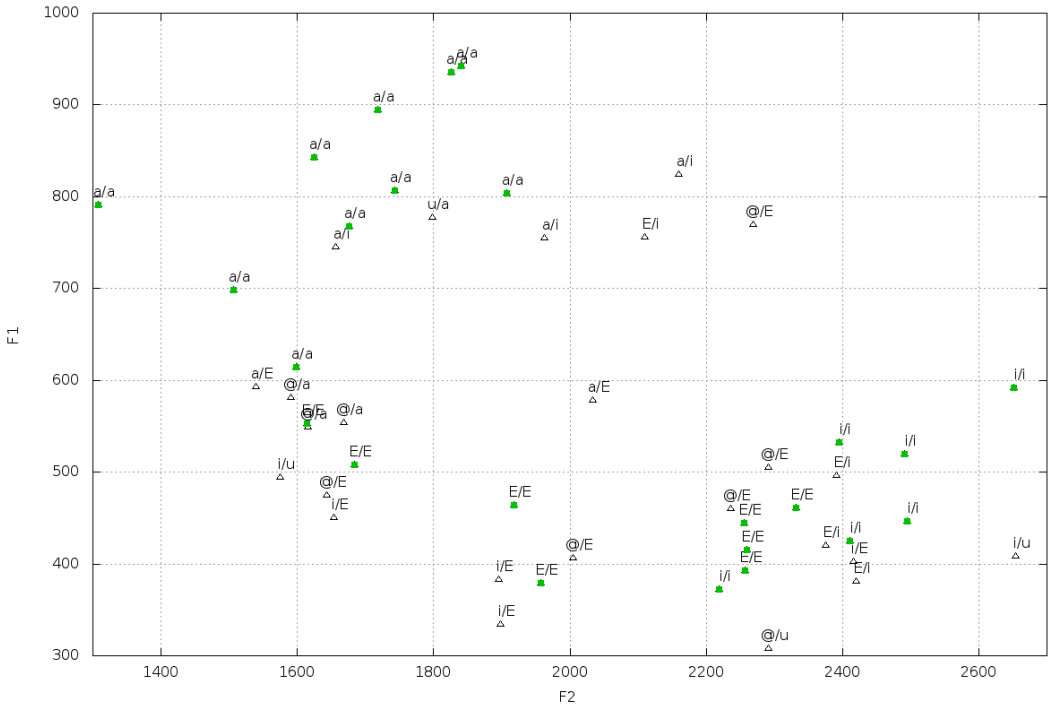

Fig. 2. F1/F2 positions of labels. Each B/T legend displays the tag provided by subject B and the auto-tagger,respectively.

The first and second formants (F1 and F2) were extracted for each sample using Praat [16], and each sample plotted in the F1/F2 space. In Fig. 1, each sample is labeled with the tags from subject A and B. In Fig. 2, the sample is labeled with the tags of subject B and the auto-tagger. From the figures, it can be seen that most discrepancies among the three tags occur on the boundaries between classes. Tags in agreement are shown in green in both figures.

C. Corpus statistics

In total, 1 947 observations of schwa were auto-tagged, using the process described in Section III-A. The vowel /E/

was tagged most frequently, and the vowel /u/ least. The number of tags associated with each vowel is shown in Fig. 3.

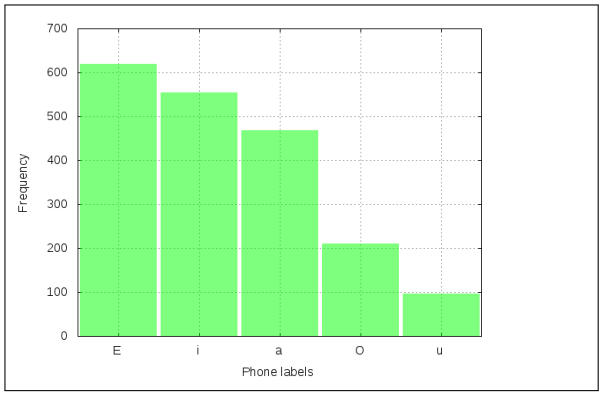

Fig. 3. Vowel distribution in the auto-tagged SPCS corpus.

D. Tag analysis

The tags were evaluated to determine how the observed vowel distribution is affected by non-acoustic factors. In Fig. 4 the vowel distribution is displayed per speaker. As is clear from this figure, speaker identity plays a minor role in determining which vowel is produced: per speaker, the distribution of vowels occurs approximately according to the same percentages as observed overall. This would immediately exclude other speaker-specific factors (such as gender or age). Speaker-specific features are therefore not considered further.

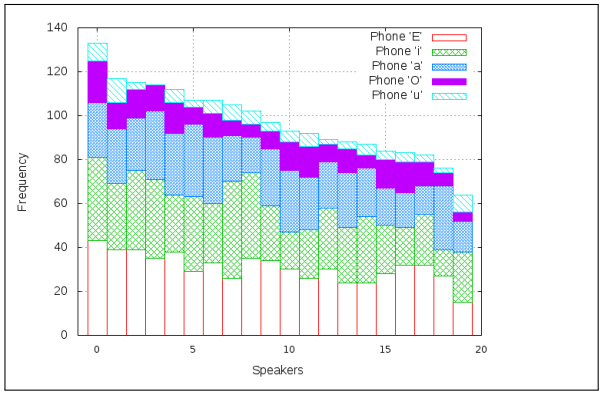

Fig. 4. The number of times each vowel was observed per speaker.

When analysing the tags per target word (all examples of the English word produced during code-switching are considered together), a clearer pattern emerges. The number of times a specific vowel is observed per word is shown in Fig. 5.

Based on the large role that word orthography plays, we next consider the graphemic string that produced the specific vowel (for example, *-a-, -i-, -io-* or *-ure-*). Note that the true prediction for each of these strings, in the specific context used, is 'schwa'. The data is first split into two parts to simplify analysis: where the schwa phone appears once in a word and where it appears multiple times. Since it is

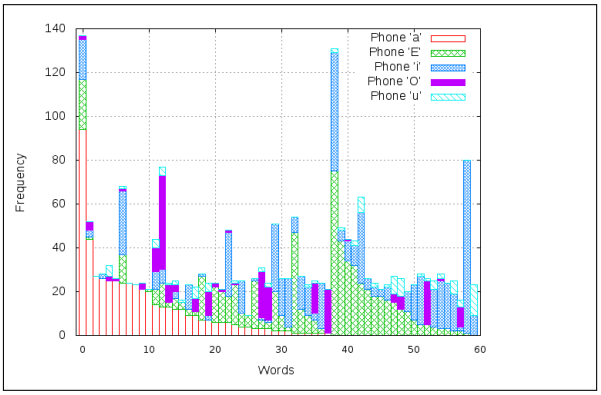

Fig. 5. The number of times each vowel was observed per unique word.

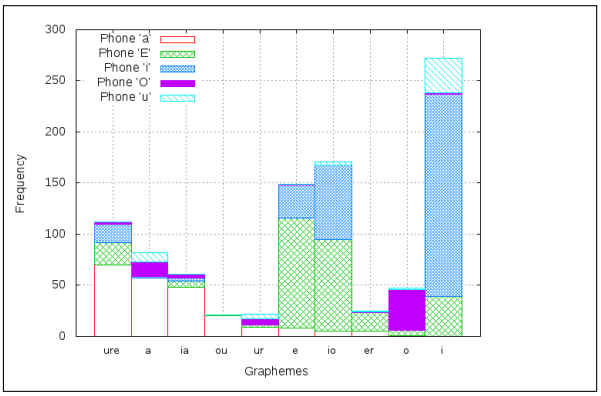

Fig. 6. The number of times each vowel was observed per unique grapheme string.

expected that one schwa may be realised in different ways in the same word, the single-schwa words (where the schwa phone appears once in the corresponding phone string of the word) are analysed. In Fig. 6 we display the number of times each vowel is observed per unique grapheme string, for the single-schwa subset. We see that the graphemes *-ure-* , *-a-, -ia-, -ou-* and *-ur-* were mostly (but not exclusively) realised as phone */a/* during code switching. There are three graphemes *-e-, -io-, -er-* which were mostly realised as the Sepedi phone */E/*. There was very little confusion, for the graphemes *-o-* and *-i-* as they were almost always realised as phones */O/ and /i/*. The most unnpredictable grapheme string was -io-, which was realised as either */i/* or */E/*, two phones that overlap on the vowel chart in Fig. 1.

E. Predictability

From the above analysis, we select triphone and grapheme as input for a simple Naive Bayes (NB) classifier, in order to obtain an initial indication of predictability. Only the single-schwa words are considered. Using 10-fold cross-validation, we train models using only these non-acoustic features, and evaluate using the ten test partitions. Results are shown in Table IV. The rows show the phone label tags (obtained from the auto-tagger) and the columns are the label predictions from the NB classifier. We show the agreement level on the diagonal, both in terms of counts and percentage accuracy. An overall classification accuracy of 67.36% is achieved.

TABLE IV
CONFUSION MATRIX WHEN PERFORMING 10-FOLD CROSS-VALIDATION WITH NON-ACOUSTIC FEATURES ONLY.

	E	O	a	i	u
E	**214(73.5)**	6	29	41	1
O	2	**47 (68.1)**	15	1	4
a	18	13	**189 (84.8)**	0	3
i	105	1	22	**198 (60.7)**	0
u	4	10	5	34	**0 (0)**

We see that the Sepedi phone labels *E* and *i* can be predicted fairly easily, although they are mutually highly confusable. On the other hand, the phone label *u* is never hypothesized. It also has the lowest occurance and predictability when analysed from a speaker-based, word-based, as well as grapheme-based perspective. It therefore seems better to not predict /u/ at all, but rather select either the second-most probable candidate, or not to introduce a variant for words where /u/ is predicted as the most probable realisation. Which of these two strategies is better, will require further experimentation. As the grapheme string seems to be the main predictor of the matrix pronunciation of code-switched speech, it is expected that using the auto-tagging process to generate training data in order to train alternative G2P models, will be a productive area for further research.

V. CONCLUSION

Two interlinked processes were demonstrated: (1) a technique for auto-tagging an acoustic corpus, and (2) an analysis of the auto-tags to determine useful non-acoustic features for pronunciation prediction. These would make it possible to generate a lexicon for code-switched speech, prior to having encountered acoustic data related to the specific vocabulary to be modeled.

We found that the vowel-tagging process is quite difficult for humans to do, showing medium to fair inter-subject agreement if subjects are allowed to select the embedded phone itself (schwa in this case) as an option. When only those samples were included where speakers were deemed to have produced a substitute from the matrix language, inter-subject agreement increased to 85.7%. The auto-tagger agreed more strongly with one of the subjects: achieving a matching accuracy of 60.5% agreement with one, but 84.9% agreement with the other.

The non-acoustic features found to be most useful for predicting phone labels were triphone and grapheme string. Using these features, a simple Naive Bayes classifier achieves an average of 67.4% classification accuracy, evaluated using 10-fold cross-validation.

From this analysis we conclude that the best matrix pronunciations for the English schwa phone are *E, i, O* and *a*, with specific substitutions occurring based on the larger grapheme string. A one-to-one vowel substitution is only possible if the grapheme representation is either *i* or *o*. Using the trained classifier, pronunciation dictionaries can now be developed by predicting the matrix pronunciations of English words in code-switched Sepedi speech, using only the vocabulary (no acoustic data) as input.

While the paper focused on the schwa found in English words embedded in Sepedi speech, the techniques are more generally applicable. (The English schwa is the foreign phone exhibiting the most variability, and was therefore selected as focus for this study.) Future work includes using a more sophisticated classifier, repeating this process for the full phone set, and conducting empirical ASR experiments using the proposed mappings.

VI. ACKNOWLEDGMENT

This work was partially supported by the National Research Foundation. Any opinion, findings and conclusions or recommendations expressed in this material are those of the author(s) and therefore the NRF do not accept any liability in regard thereto.

REFERENCES

[1] T. I. Modipa, M. H. Davel, and F. de Wet, "Pronunciation modelling of foreign words for Sepedi ASR," in *Proc. Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Johannesburg, South Africa, Dec. 2013, pp. 64–69.

[2] T. Modipa, M. H. Davel, and F. de Wet, "Context-dependent modelling of English vowels in Sepedi code-switched speech," in *Proc. Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Pretoria, South Africa, Nov. 2012, pp. 173–178.

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

[4] Y. Li, Y. Yu, and P. Fung, "A Mandarin-English code-switching corpus," in *Proc. LREC12*, 2012, pp. 2515–2519.

[5] D. Yu, L. Deng, P. Liu, J. Wu, Y. Gong, and A. Acero, "Cross-lingual speech recognition under runtime resource constraints," in *Proc. ICASSP*, 2009, pp. 4193–4196.

[6] V. B. Le, L. Besacier, and T. Schultz, "Acoustic-phonetic unit similarities for context dependent acoustic model portability," in *Proc. ICASSP*, 2006, pp. 1101–1104.

[7] C.-L. Huang and C.-H. Wu, "Phone set generation based on acoustic and contextual analysis for multilingual speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP)*, 2007, pp. 1017–1020.

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

[9] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2-3, pp. 95–108, 2000.

[10] M. H. Davel, C. van Heerden, and E. Barnard, "Validating smartphone-collected speech corpora," in *Proc. SLTU*, Cape Town, South Africa, May 2012, pp. 68–75.

[11] D. Gibbon, R. Moore, and R. Winski, *Handbook of standards and resources for spoken language systems*. Walter de Gruyter, 1997.

[12] 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.

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

[14] E. Barnard, M. H. Davel, C. J. V. Heerden, F. D. Wet, and J. Badenhorst, "The NCHLT speech corpus of the South African languages," in *Proc. SLTU*, 2014, pp. 194–200.

[15] L. Loots, M. Davel, E. Barnard, and T. Niesler, "Comparing manually-developed and data-driven rules for P2P learning," in *Proc. Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Stellenbosch, South Africa, Nov. 2009, pp. 35–40.

[16] P. Boersma and D. Weenink, "Praat: doing phonetics by computer (version 5.1.05) [computer program]," 2009, retrieved May 1, 2009, from http://www.praat.org/.