Accent identification in the presence of code-mixing

Thomas Niesler[†] *& Febe de Wet*[‡]

[†]Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
trn@dsp.sun.ac.za

[‡]Centre for Language and Speech Technology, Stellenbosch University, South Africa
fdw@sun.ac.za

Abstract

We investigate whether automatic accent identification is more effective for English utterances embedded in a different language as part of a mixed code than for English utterances that are part of a monolingual dialogue. Our focus is on Xhosa and Zulu, two South African languages for which code mixing with English is very common. In order to carry out our investigation, we extract English utterances from mixed-code Xhosa and Zulu speech corpora, as well as comparable utterances from an English-only corpus by Xhosa and Zulu mother-tongue speakers. Experiments show that accent identification is substantially more accurate for the utterances originating from the mixed-code speech. We conclude that accent identification is more successful for these utterances because accents are more pronounced for English embedded in mother-tongue speech than for English spoken as part of a monolingual dialogue by non-native speakers.

1. Introduction

The South African constitution officially recognises eleven languages and, in practice, many more are spoken by the country's population. As is often the case in such a multilingual society, English is usually the lingua franca. South African English is consequently characterised by a large variety of accents. Automatic language and accent identification as well as multilingual and accent specific speech recognition are therefore important aspects of the development of ASR technology in the region.

Since most citizens are multilingual, it is common for South African speakers to switch among different languages both between utterances (code-switching) and within an utterance (code-mixing). In modern Xhosa and Zulu, for example, it is very common to revert to English when citing numbers, dates and money amounts.

The research presented in this paper investigates the effect of such code-switching and code-mixing on the accuracy of an automatic accent-identification system. In particular, we determine how accurately the mother tongue of Xhosa and of Zulu speakers can be determined when the utterance is part of a mixed code, and when it is not.

The work was motivated by the findings of two previous studies. In the first, the performance of language identification systems were investigated for Afrikaans, English, Xhosa and Zulu [1]. Here it was established that even in mixed-code utterances consisting predominantly of English words, the language could be correctly classified with approximately 70% accuracy. The second study focussed on the identification of accents in English spoken by mother-tongue speakers of an African language, and for whom English is a second or third language [2]. Instead of considering individual languages, a distinction was in this case made between the Nguni (Zulu, Xhosa, Swati, Ndebele) and Sotho (Northern Sotho, Southern Sotho, Tswana) language families. These are the two largest language families in South Africa, and each shares a similar vowel system. It was shown by means of both perceptual tests involving human subjects as well as automatic accent identification systems that neither humans nor machines were able to distinguish reliably between the Nguni and Sotho English accents.

These two findings were seemingly contradictory. On the one hand, reliable accent identification was not possible between Nguni and Sotho accent groups. On the other hand, Xhosa and Zulu accents could be identified with fair accuracy automatically, although both are Nguni languages and hence strongly related. There was, however, a difference in the type of English utterances that had been used for testing. For the Xhosa/Zulu experiments, the English words were embedded in the respective African mother tongues as part of mixed codes. For the Nguni/Sotho experiments, data was drawn from a corpus in which all words were English. Hence in this second case, code mixing and switching did not occur.

In this paper we attempt to establish experimentally whether or not the performance of automatic accent-identification systems is indeed improved for mixed-code utterances produced by mother-tongue speakers Xhosa and of Zulu.

2. Speech databases

Our experiments are based on the African Speech Technology (AST) corpora, which consist of recorded and annotated South African speech collected over mobile as well as fixed telephone networks [3]. For the compilation of these corpora, speakers were recruited from targeted language groups and given a unique datasheet with items designed to elicit a phonetically diverse mix of read and spontaneous speech. The datasheets included read items such as isolated digits, as well as digit strings, money amounts, dates, times, spellings and also phonetically-rich words and sentences. Spontaneous items included references to gender, age, mother tongue, place of residence and level of education.

Corpora were compiled in five different languages, namely Afrikaans, English, Southern Sotho, Xhosa and Zulu. Furthermore, the accents of South African English spoken by English, Afrikaans, Coloured, Indian and Black mother-tongue speakers were gathered separately, resulting in five accent-specific English corpora. In the work presented here, we have made use of the Xhosa and Zulu mother-tongue corpora, as well as the Black English corpus. The latter consists of English spoken chiefly by mother-tongue speakers of Xhosa, Zulu, Southern Sotho (Sesotho), and Tswana (Setswana), as set out in Table 1. The former two corpora, on the other hand, each consist of speech uttered exclusively by Xhosa and by Zulu mother-tongue speakers respectively. Furthermore, although most words were uttered in Xhosa or in Zulu, mixed codes were very common in these two corpora, especially for dates, times, numbers, money amounts and spelled words. By identifying such predominantly English utterances in the Xhosa and the Zulu data, and considering comparable utterances by Xhosa and Zulu speakers in the Black English corpus, the effect of code-mixing on accent-identification can be studied.

Mother tongue	% of speakers
Xhosa	23
Zulu	18
Sesotho	23
Tswana	32
Other	4

Table 1: Mother tongues of the speakers making up the Black English AST corpus.

Together with the recorded speech waveforms in each corpus, both orthographic (word-level) and phonetic (phone-level) transcriptions were available for each utterance. The orthographic transcriptions were produced and validated by human transcribers. Initial phonetic transcriptions were obtained from the orthography using grapheme-to-phoneme rules for Xhosa and Zulu, and using a pronunciation dictionary for English. These transcriptions were subsequently corrected and validated manually by human experts.

3. Automatic accent identification

The structure of our automatic accent identification system in is shown in Figure 1. This architecture, referred to as *Parallel Phone Recognition followed by Language Modelling* (PPRLM), uses a parallel set of phone recognisers for explicit accent classification [2, 4, 5].

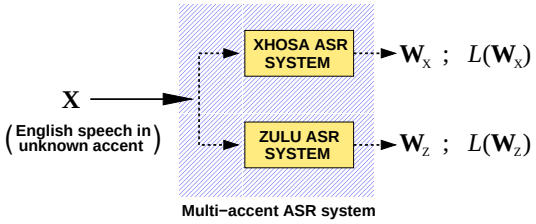

Figure 1: Accent-identification using parallel recognition systems for Xhosa and Zulu.

In Figure 1, each recognition system includes accent-specific acoustic and language models. Since we would like to determine the accent (Xhosa or Zulu) of English input speech, the acoustic models should ideally be trained using only English utterances by speakers of the appropriate mother tongue. However, due to the very limited size of the corpora at our disposal, and since approximately 45% of the words in the Xhosa and Zulu corpora are in fact English due to code mixing and switching, these two corpora have been used to train the acoustic models instead. Using the system depicted in Figure 1, each recognition system provides a transcription $\mathbf{W}$ of the speech utterance $\mathbf{X}$ in terms of its phone inventory and a corresponding language-specific language model. In addition to this transcription, each recogniser provides a likelihood $L(\mathbf{W})$, which is used to identify the accent of the input speech.

3.1. Data preparation

In the following, we will refer to the Xhosa, Zulu and Black English corpora described in Section 2 by the abbreviations XX, ZZ and BE respectively. Because the mother tongue of each speaker in the BE corpus is known, it was possible to extract sub-portions of this corpus uttered by Xhosa and Zulu speakers. These two corpora (XBE and ZBE respectively), were large enough to provide testing material free of code mixing and switching for the purposes of our evaluation. However, as already pointed out in the previous section, the XBE and ZBE corpora were not large enough to provide training data for the acoustic models. Instead, these were trained using data obtained from the XX and ZZ corpora, each containing approximately 7 hours of audio data, as indicated in Table 2.

Corpus name	Speech (h)	No. of utts.	No. of spkrs.	Phone tokens	Word tokens
XX	6.98	8 538	219	177 843	36 676
ZZ	7.03	8 295	203	187 249	35 568

Table 2: Training sets for the Xhosa and the Zulu (XX and ZZ) corpora, as used to train acoustic models.

Independent test sets were also prepared for Xhosa and Zulu, each containing approximately 25 minutes of speech. These two test sets, as well as the XBE and ZBE test sets, are described in Table 3. There was no speaker-overlap between the training and any of the test sets, and each contained a balance of male and female speakers.

Corpus name	Speech (min)	No. of utts.	No. of spkrs.	Phone tokens	Word tokens
XX	26.8	609	17	10 925	2 480
ZZ	27.1	583	16	11 008	2 385
XBE	23.8	614	17	9 112	2 709
ZBE	25.8	643	16	9 841	2 856

Table 3: Test sets prepared for the Xhosa and Zulu (XX and ZZ) corpora as well as the Xhosa and Zulu subsets (XBE and ZBE) of the BE corpus.

However, the XX and ZZ test sets both contain a significant deal of code mixing and switching. Figure 2 shows the fraction of utterances in the XX and ZZ test sets that contain at least a certain threshold proportion of English words. This threshold is indicated on the horizontal axis, so that it can be seen, for example, that approximately 25% of the utterances in both test sets consist entirely of English words.

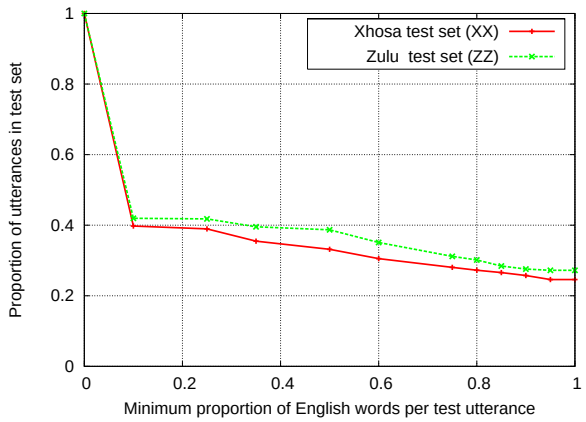

Figure 2: Fraction of Xhosa and Zulu (XX and ZZ) test set utterances that contain a certain proportion of English words.

We will refer to the subset of utterances in the XX and ZZ test sets that consist entirely of English words as XXE and ZZE respectively. These subsets of the full XX and ZZ test sets are described in Table 4. The same table also shows the subsets of XBE and ZBE that consist of the same English words found in the XXE and ZZE test sets. These subsets of the full XBE and ZBE test sets will be referred to as XXBE and ZZBE respectively. The composition of these various test sets is illustrated in Figure 3.

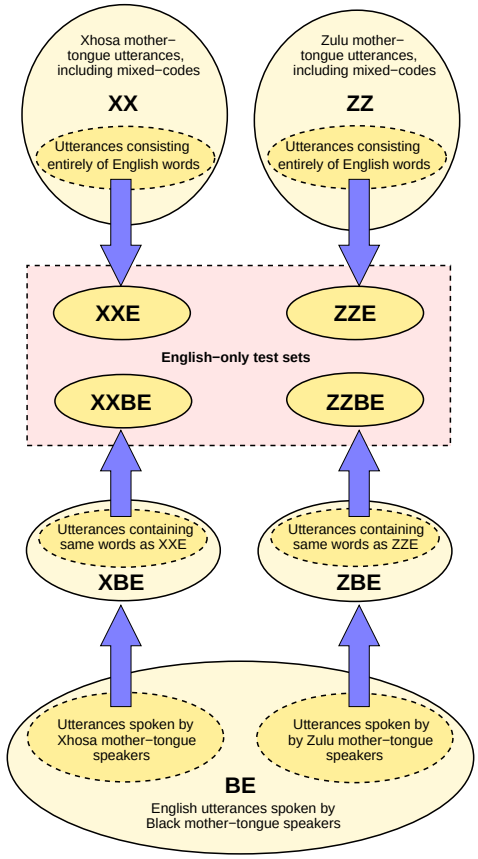

Figure 3: Composition of the test sets described in Table 3 and in Table 4.

Corpus name	Speech (min)	No. of utts.	No. of spkrs.	Phone tokens	Word tokens
XXE	10.0	148	17	2 873	1 332
ZZE	10.5	158	16	3 139	1 272
XXBE	12.2	237	17	4 007	1 752
ZZBE	10.2	214	16	3 472	1 456

Table 4: English-only subsets of each test set in Table 3.

The utterances in the XXE and ZZE test sets are therefore quite similar to those comprising XXBE and ZZBE, because they are made up of the same English

words that occur often as part of Xhosa and Zulu mixed codes. XXE and ZZE differ from XXBE and ZZBE in that the former two have been extracted from the XX and ZZ corpora respectively, while XXBE and ZZBE both originate from the BE corpus. Hence XXE and ZZE are drawn from mixed codes, while XXBE and ZZBE are drawn from monolingual speech.

3.2. Acoustic models

Acoustic models were trained using the HTK tools [6] and the XX and ZZ training sets described in Table 2. The Xhosa and Zulu recognition systems employ a common set of 90 phones, including silence and speaker noise. This phone set was verified to account for 99.5% of the phone tokens in the XBE and ZBE test sets. Thus, although the phones present in the XBE/ZBE and the XX/ZZ corpora are not exactly the same, there is an overwhelming degree of overlap.

The speech was parameterised as Mel-frequency cepstral coefficients (MFCCs) and their first and second differentials, with cepstral mean normalisation (CMN) applied on a per-utterance basis. Speaker-independent cross-word left-to-right triphone HMMs were trained by embedded Baum-Welsh re-estimation and decision-tree state clustering, using the phonetically-labelled training sets. Each model had three states, eight Gaussian mixtures per state and diagonal covariance matrices. Triphone clustering resulted in a total of approximately 1250 clustered states for each set of acoustic models.

3.3. Accent identification results

Table 5 shows the accuracy of the automatic accent identification system illustrated in Figure 1 when presented with the test sets listed in Table 4. In these experiments, the Xhosa and Zulu recognition systems depicted in Figure 1 each used an unweighted phone loop as a language model. Since the two systems share a common phone set, this means that discrimination between the two accents was based exclusively on acoustic differences.

Test corpus	Classified as (%)	
	Xhosa	Zulu
XXE	75.0	25.0
ZZE	29.1	70.9
Average correct	73.4%	
XXBE	48.9	51.1
ZZBE	37.4	62.6
Average correct	55.4%	

Table 5: Accent identification accuracy (%) for the XXE, ZZE, XXBE and ZZBE test sets using phone loop language model.

The experimental results in Table 5 indicate that while the accent of 73.4% of the utterances extracted from the XX and ZZ tests sets were correctly classified, this drops to 55.4% for a set of comparable utterances drawn from the XBE and ZBE test sets. Hence the accent is more difficult to classify when the speech is part of a monolingual English dialogue, and easier to classify when it forms part of a mixed code.

In order to gauge the effect which the proportion of English words in the test set has on accent identification, the accuracy was calculated for a series of points along the horizontal axis of Figure 2. These results, shown in Figure 4, indicate the effect which the degree of code mixing (proportion of English words in a Xhosa or Zulu sentence) has on the average accuracy of the identification system. The left-hand side of the graph corresponds to testing the system using the XX and ZZ test sets of Table 3, while the right-hand side of the graph corresponds to testing with the XXE and ZZE subsets described in Table 4. Figure 4 shows that, for utterances drawn from the XX and ZZ corpora, as the proportion of English words in the test utterances increases, identification accuracy deteriorates, but not drastically. This happens because a higher proportion of Xhosa and Zulu words in an utterance makes it easier to identify the mother tongue correctly.

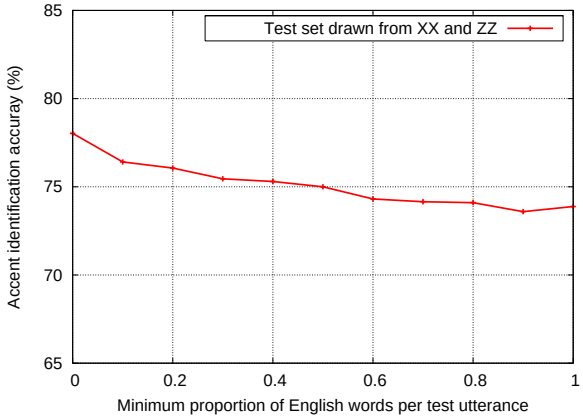

Figure 4: The effect of the degree of code-mixing in the test set on the accent identification accuracy.

Figure 5 shows the effect of restricting the test utterances drawn from the XBE and ZBE test sets to consist of the same words found in the XXE and ZZE test sets. Again, the left-hand side of the graph corresponds to testing the system using the full XBE and ZBE test sets of Table 3, while the right-hand side of the graph corresponds to testing with the XXBE and ZZBE subsets depicted in Table 4. The results in the figure show that the particular words present in the monolingual test sets drawn from the BE data do not affect the identification in a systematic way. This indicates that the strength of the Xhosa or Zulu

accent in an English word does not depend upon whether that word occurs frequently as part of a mixed code or not. Rather, when code mixing does not occur, the accent of the speaker cannot be determined accurately, irrespective of the vocabulary.

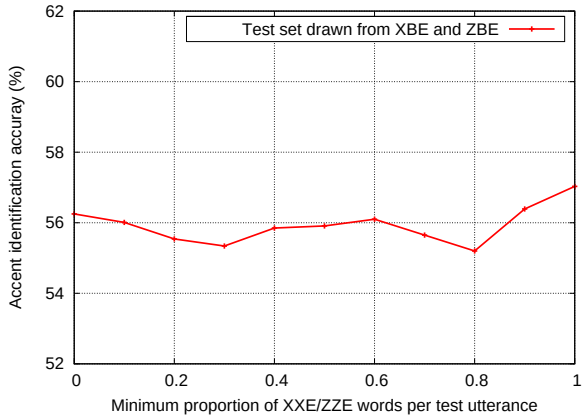

Figure 5: The effect of the similarity between the English words used in the monolingual and the mixed-code test sets on the accent identification accuracy.

Finally, in order to gauge the effect of the grammar on the accent identification accuracy, the experiments of Table 5 were repeated, but now with each of the two recognisers using a backoff bigram phone language model [7]. These were trained using the XX and ZZ training set transcriptions, and language model probabilities were estimated using absolute discounting [8]. The resulting accuracies are shown in Table 6.

Test corpus	Classified as (%)	
	Xhosa	Zulu
XXE	79.7	20.3
ZZE	29.7	70.3
Average correct	74.8%	
XXBE	56.5	43.5
ZZBE	42.1	57.9
Average correct	57.2%	

Table 6: Accent identification accuracy (%) for XXE, ZZE, XBE and ZBE when using a bigram language model.

It is apparent that the use of a bigram slightly improves identification accuracies for both code-mixed (XXE/ZZE) and monolingual (XXBE/ZZBE) test sets. Nevertheless, the accent identification accuracy for the utterances drawn from the monolingual BE corpus remains much lower than for the mixed-code utterances drawn from the XX and ZZ corpora.

4. Summary and conclusions

We have investigated the effect which code-mixing and code-switching have on the automatic identification of the English accent of Xhosa and Zulu mother-tongue speech. In particular, we have sought to determine whether the performance of an automatic accent identification system is different for English utterances that are embedded in Xhosa or Zulu as part of a mixed code, in relation to English utterances that are part of a monolingual dialogue. We have found that, in the latter case, it is not possible to distinguish between the two accents with good accuracy, while in the former case accuracies of above 70% were achieved. We therefore conclude that English which is embedded within a Xhosa or Zulu dialogue exhibits a more distinct accent than monolingual English produced by the same type of speakers. For the development of automatic speech recognition systems this appears to suggest that, while a common set of acoustic models is suitable for the recognition of monolingual English spoken by Xhosa and Zulu mother-tongue speakers, language-specific models will be more appropriate for the modelling of English words that are expected to be embedded in the respective African mother tongue as part of a mixed code.

5. Acknowledgements

This work was supported by the National Research Foundation (NRF) under grants FA2005022300010 and GUN2072874.

6. References

[1] T.R. Niesler and D. Willett, "Language identification and multilingual speech recognition using discriminatively trained acoustic models," in *First ISCA ITRW on Multilingual Speech and Language Processing (MULTILING)*, Stellenbosch, South Africa, 2006.

[2] F. de Wet, P.H. Louw, and T.R. Niesler, "Human and automatic accent identification of Nguni and Sotho Black South African English," *South African Journal of Science*, 2007.

[3] J.C. Roux, P.H. Louw, and T.R. Niesler, "The African Speech Technology project: An assessment," in *Proc. LREC*, Lisbon, Portugal, 2004.

[4] W. Shen and D. Reynolds, "Improving phonotactic language recognition with acoustic adaptation," in *Proc. Interspeech*, Antwerp, Belgium, 2007.

[5] M.A. Zissman and K.M. Berkling, "Automatic language identification," *Speech Communication*, vol. 35, pp. 115–124, 2001.

[6] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, *The HTK book, version 3.2.1*, Cambridge University Engineering Department, 2002.

[7] S. Katz, "Estimation of probabilities from sparse data for the language model component of a speech recogniser," *IEEE Transactions on Acoustics Speech and Signal Processing*, vol. 35, no. 3, pp. 400–401, March 1987.

[8] H. Ney, U. Essen, and R. Kneser, "On structuring probabilistic dependencies in stochastic language modelling," *Computer, Speech and Language*, vol. 8, pp. 1–38, 1994.