Nguni and Sotho varieties of South African English - distant cousins or twins?

Febe de Wet[†], *Thomas Niesler*[‡] *& Philippa Louw*[†]

[†]Centre for Language and Speech Technology, Stellenbosch University, South Africa
{fdw,phlouw}@sun.ac.za

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

Abstract

It is well established that accent can have a detrimental effect on the performance of automatic speech recognition (ASR) systems. While accents are usually classified in terms of a speaker's mother tongue, it remains to be determined if and when this linguistic classification is appropriate for the development of ASR technology. This study focuses on South African English as produced by mother tongue speakers of Nguni and Sotho languages, which account for over 70% of the country's population. The aim of the investigation is to determine whether these two accent groups should be treated as a single variety, or whether it is better to treat them separately. We begin with a perceptual experiment in which human listeners classify different English accents. Subsequently, speech recognition experiments are conducted to determine whether the acoustic models benefit from the incorporation of Nguni/Sotho accent classifications. The results of the perceptual experiment indicate that most listeners cannot correctly identify a speaker's mother tongue based on their English accent. This finding is supported by the results of the recognition experiments.

1. Introduction

South Africa is a multilingual society, with a total of eleven languages recognised officially by the constitution and many more used in practice. South African English, which serves as the lingua franca, is therefore characterised by a large variety of accents. This has important implications for the development of automatic speech recognition (ASR) systems, because their performance is known to deteriorate for non-native speech [4, 1]. Accent modelling can be integrated into ASR systems to improve their performance with non-mother tongue speakers. A first step in the development of accent-specific recognition systems is the identification of appropriate accent groups.

The accents of South African English include English spoken by English, Afrikaans, Coloured, Indian and Black mother-tongue speakers. In this study, we focus on Black South African English (BSAE)[1]. In particular, we want to determine whether BSAE should be treated as a homogeneous accent group, or whether sub-groups should be defined in terms of differences in mother-tongue.

There does not appear to be consensus on this issue in the linguistic literature. Some authors [10] have argued that, because of the similar diphthong/tense vowel-structure of the Bantu languages, BSAE is a fairly coherent variety of English within which there is little variation that can be ascribed to different mother tongues. It has even been proposed that any perceivable differences between the English accents of speakers of different Bantu languages will most probably be on the suprasegmental level [11]. In contrast, other authors maintain that "the idea of a single uniform variety of BSAE would thus seem to be an optimistic figment of the linguistic imagination" [3].

Contrasting claims have also been made concerning differences at a perceptual level. For example, in a study regarding the comprehensibility of South African English varieties, it was reported that Black language teachers claimed to be able to distinguish between the English spoken by Xhosa and Zulu mother tongue speakers [8]. However, a different study investigating the types of labels given to BSAE speech showed that, when listeners tried to pin point a person's mother tongue based on the speaker's English accent, they were not able to do so accurately [2].

This study investigates the relevance of these claims for the development of accent-robust ASR technology by means of both perceptual and ASR experiments. We focus on two variants of BSAE: English as spoken by mother-tongue speakers of a Sotho language (Northern Sotho, Southern Sotho, Tswana) and English as spoken by mother-tongue speakers of a Nguni language (Zulu, Xhosa, Swati, Ndebele). These two language groups constitute 25.5 and 45.7% of the South African population, respectively. In the perceptual experiment, listeners were asked to classify a person's English accent in terms of his/her mother tongue. The ASR experiments aimed to determine whether or not the quality of the acoustic models can be improved by keeping data from the two accent groups separate. The data for both experiments was taken from the African Speech Technology database of telephone speech.

2. The AST database

The African Speech Technology (AST) project was funded by the Department of Science and Technology of the South African national government from 2000 to 2003 [7]. During the project, telephone speech databases in five of South Africa's eleven official languages were compiled, namely Xhosa, Southern Sotho, Zulu, South African English and Afrikaans. The variation in spoken South African English and Afrikaans is considerable and, in many instances, culturally-bound. In order to make provision for these known varieties, eleven databases based on the five languages were developed.

For the English database-group, English mother-tongue

[1]Assigning appropriate and commonly accepted labels to the different varieties of English spoken in South Africa is still a bone of contention amongst linguistic scholars (e.g. [2]). Based on the arguments presented in [3], we have decided to use the term Black South African English in this paper.

(EE) speakers as well as Black, Coloured, Asian and Afrikaans non-mother-tongue speakers were targeted. Within the Black speaker group, speakers having any one of Xhosa, Zulu, Southern Sotho (Sesotho), Tswana (Setswana) or Northern Sotho (Sepedi) as their mother tongue were included. The data that was used to conduct the experiments in this study was selected from the resulting BSAE corpus.

The AST contents specification totals 38 to 40 utterances per speaker comprising a mixture of spontaneous and read speech. The types of read utterances include isolated digit items, natural numbers, dates, times, money amounts, application/domain specific words or phrases, and phonetically rich words and sentences. Spontaneous responses were gathered by asking the speakers to say their age, home language, date of birth and to answer yes/no questions. The data was recorded digitally using a Dialogic D/300-SC board which interfaced directly to an ISDN Primary Rate Interface channel.

In total, 300 to 400 speakers between the ages of 20 and 60 were recruited for each database. An approximately equal male/female balance was achieved, with 50% of the speakers calling a toll-free number from a land-line phone and the other 50% from a mobile phone. Each speaker was presented with a unique data sheet containing the items to be read. Of the almost 6000 calls recorded, 41% were classified as empty or unusable. This data loss was anticipated for by distributing 400 datasheets per database, while aiming for only 300 usable calls. The final BSAE database contains 235 phone calls, corresponding to approximately 6 hours of speech.

3. Perceptual experiment

Mother-tongue speakers of African languages often claim that they can determine the mother tongue of other African-language speakers from their English accent. We investigated this claim by means of a perceptual experiment.

3.1. Speakers

To ensure that only the speech of mesolect speakers was used as stimuli, the minimum level of education of the speakers was grade 12 (matric). Although some of the speakers had a higher university qualification, they were not considered to have reached the acrolectal level as described in [11]. Table 1 shows the distribution of mother tongues in the speaker population.

Mother tongue	Number of speakers
Southern Sotho	13
Tswana	24
Xhosa	16
Zulu	18
Ndebele	1

Table 1: Mother tongue distribution of the *speakers* who participated in the perceptual experiment.

Of the 72 speakers, 37 were female and 35 male. The male/female ratio within the Nguni and Sotho groups were similar to the overall distribution.

3.2. Stimuli

We used a total of 180 stimuli, consisting of 30 single words and 30 phrases pronounced by native speakers of each of the three language groups native English (EE), Sotho English (SE) and Nguni English (NE). The idea behind the single word stimuli was that listeners would be able to focus on a limited number of sounds in a limited context. On the other hand, the phrase stimuli were intended to provide listeners with a variety of sounds as well as prosodic cues which may influence accent judgement. We would also not be able to determine which specific sounds influenced the listeners' judgement if only sentences were used as stimuli.

The phonetic content of the stimuli was selected according the descriptions of BSAE in [5, 10, 11, 9]. We attempted to represent as many of the relevant phonetic/phonological BSAE phenomena as possible. However, the exact example words given by these authors could not be used since they do not occur in the AST databases.

Almost all the single words were selected from utterances in which they occurred in phrases. In this way, each word could be presented both in isolation and within a phrase. As far as possible, the words and phrases were selected from the EE and BSAE databases in such a way that the contents were the same for each language group.

3.3. Listeners

A total of 22 participants (none of whom partook in the AST project) were recruited on campus. The mother tongue distribution[2] amongst the listeners is shown in Table 2.

Mother tongue	Number of listeners
Southern Sotho	4
Xhosa	9
Zulu	9

Table 2: Mother tongue distribution of the *listeners* who participated in the perceptual experiment.

The majority of the listeners were enrolled for an undergraduate course at the university, but the group also included a few postgraduate students. The female/male ratio in the group was 14/8.

3.4. Test administration

The perceptual experiment was set up using the Praat software package (www.praat.org). The 180 stimuli were played in a random sequence, but the same sequence was used for all participants. The question "Can you identify the language group to which this speaker belongs?" was displayed on the computer screen together with four clickable buttons, representing the options available to choose from i.e. "Sotho", "Nguni", "English" and "I don't know". Each stimulus was played only once and as soon as the participant had made his/her choice by clicking on one of the four buttons, the next stimulus was played.

Instructions were given verbally to the participants before the experiment started. A short pre-test, consisting of three test utterances, was carried out to demonstrate the procedure as well as to ensure that the participants could hear the stimuli clearly. The participants listened to the stimuli using earphones. The test stimuli were presented in three sets of 60 and participants were allowed to take a short break between sets. On average,

[2]Because Stellenbosch University is in the Western Cape province, it was much more difficult to recruit Sotho than Nguni participants for our experiments. We decided against trying to recruit a larger group of Sotho listeners once it became apparent how evenly matched the results were between the Nguni and Sotho subjects.

the participants required 20 minutes to complete the perception test[3].

3.5. Results

Overall, 14% of all responses were "I don't know". These were mostly (77%) responses to the single word stimuli and were removed from the dataset before the percentage correct results were calculated. Of the EE stimuli, 70.6% were correctly identified as "English". This result shows that listeners were able to distinguish between the EE and BSAE accents.

Figure 1 illustrates the results for the BSAE stimuli (responses to EE stimuli removed from the dataset). According to the data in the figure, only 47.8% of the stimuli were correctly identified as NE or SE. Listeners performed only slightly better when judged only on their responses to sentences. These results indicate that the listeners could not reliably determine whether a speaker's mother tongue was from the Nguni or Sotho language group, irrespective of whether supra-segmental information was present or not.

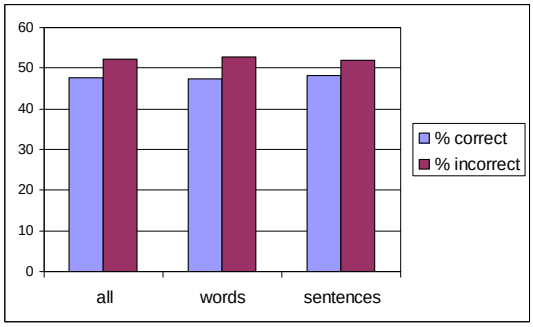

Figure 1: Percentage correct and incorrect listener responses to NE and SE stimuli.

It was also observed that the listeners' responses showed a bias towards their own mother tongue. Nguni mother tongue listeners classified the majority of the stimuli as Nguni, irrespective of the language group to which the speakers actually belong. The Sotho listeners' responses showed the same bias towards Sotho.

Since almost every stimulus was produced by a different speaker, we did not attempt to determine whether there were any speaker specific attributes that may have influenced the listeners' judgements.

4. Automatic speech recognition

For optimal speech recognition performance, the character of the training- and test-sets should match as closely as possible. This implies that, when accents are distinct, the training data should be drawn from the same accent group as the test data. If the character of the Nguni and Sotho varieties of South African English differ, we should therefore find that the best speech recognition performance for each variety is achieved when the system's training data stems from the same variety. In this section we will determine experimentally whether such a difference in performance can be found.

[3]All speakers and listeners who participated in the experiment received a monetary reward for their contribution.

4.1. Data

The AST BSAE database (as described in Section 2) was used for all the ASR experiments. Manually produced and checked phonetic as well as orthographic transcriptions of this data were available. Furthermore, the mother tongue and level of education of each speaker was recorded. Hence it was possible to extract sub-portions of this corpus uttered by mesolect Nguni and Sotho speakers. These two databases, which will henceforth be referred to as "NE" and "SE", were each further subdivided into a training and a test set, as shown in Table 3.

	Training set			Test set		
Data-Base	No.of spkrs	Size (h)	Phone tokens	No.of spkrs	Size (min)	Phone tokens
NE	88	2.57	62,351	10	12.7	5,112
SE	92	2.55	61,519	10	13.6	5,571

Table 3: Nguni- and Sotho-English databases.

Both NE and SE test-sets were designed to have 50:50 male/female as well as cell/landline ratios. Finally, separate development sets, consisting of approximately 6 minutes of speech from 4 speakers, were prepared for the NE and SE databases. These were used only for the optimisation of recognition parameters, before final evaluation on the test-set. There is no overlap between the development set and either the test or training sets.

4.2. Acoustic models

Since our aim was to determine whether it is better to have distinct Nguni- and Sotho-English recognisers or to have a single, general Black English recogniser, we further subdivided the NE and SE training sets as shown in Figure 2.

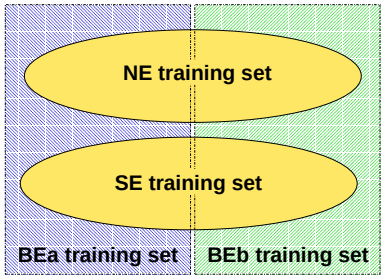

Figure 2: Division of NE and SE sets into BEa and BEb.

Both the NE and SE training sets were divided in half, taking care to maintain the male/female and cellphone/landline balance. Two new training sets, BEa and BEb, were then formed by pooling an NE and a SE subset. Hence BEa and BEb are accent-neutral with respect to the Nguni/Sotho distinction. Furthermore, since BEa and BEb contain approximately the same amount of data as the NE and SE training sets, the performance of speech recognition systems trained on this data can be compared.

Acoustic models were trained using the HTK tools [12] and the SE, NE, BEa and BEb corpora. 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 triphone models 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 600 clustered states for each set of acoustic models.

4.3. Recognition results

Speech recognition was accomplished using the HTK decoder and a bigram language model obtained from the reference transcriptions. Because the amount of training data was very limited, phoneme recognition was performed. All recognition systems employed a common set of 90 phones, including silence and speaker noise. The performance of the triphone recognition system is shown in Table 4.

Model set	Test-set PER (%)		
	NE	SE	Average
NE	51.4	51.2	-
SE	51.2	49.9	-
NE/SE	-	-	50.7
BE	51.5	49.8	50.7

Table 4: Phone error rates (%) for triphone recognition experiments evaluated on the NE and SE test sets.

The entry labelled "NE/SE" indicates the average performance of the NE acoustic models when tested on the NE test-set and the SE models tested on the SE test-set. This represents the ideal matched condition in terms of NE/SE accents. The model set labelled BE represents the average performance of the two recognition systems trained on BEa and BEb, respectively. This was done to avoid any bias which may have resulted from the particular way in which NE and SE were split into BEa and BEb.

The results show that, in all cases, the performance of the matched recognition systems NE and SE is approximately the same as the general BE system. In particular, the NE/SE average performance is indistinguishable from the average BE performance. We therefore conclude that there is no merit in separating English as spoken by Nguni and Sotho mother tongue speakers when developing automatic speech recognition systems.

5. Discussion and conclusions

The results obtained in the perceptual experiment do not support the claim made by some mother tongue speakers of the Bantu languages that they can determine a speaker's mother tongue from his/her English accent. This finding is supported by the speech recognition experiments, which showed no discernible difference in performance between accent-specific and accent-neutral systems.

However, it should be noted that the AST data is not ideal for conducting perceptual experiments. For example, the data was recorded over telephone lines as read prompts and it may not be possible to perceive the distinctive suprasegmental features of the different BSAE varieties in this kind of data. Although the experimental design did take level of education into account, there was still much variation in the English proficiency of the speakers, and different degrees of accentedness could have had an influence on the listeners' opinions. A number of these shortcomings have been addressed in a follow-up study, which has recently been submitted for publication [6].

6. Acknowledgements

This work was supported by the National Research Foundation (NRF) under grants FA2005022300010 and TTK2005072700022.

7. References

[1] S. Aalburg and H. Hoege. Approaches to foreign-accented speaker-independent speech recognition. In *Proceedings of Eurospeech 2003*, pages 1489–1492, Geneva, Switzerland, 2003.

[2] S. Coetzee-Van Rooy and B. van Rooy. South African English: labels, comprehensibility and status. *World Englishes*, 24(1):1–19, 2005.

[3] V. de Klerk. Towards a norm in South African Englishes: the case for Xhosa English. *World Englishes*, 22(4):463–481, 2003.

[4] S. Goronzy, M. Sahakyan, and W. Wokurek. Is non-native pronunciation modelling necessary? In *Proceedings of Eurospeech 2001*, pages 309–312, Aalborg, Denmark, 2001.

[5] D. Gough. Black English in South Africa. In V. de Klerk, editor, *Focus on Africa*, Varieties of English around the world. G15, pages 53–77. John Benjamins Publishing Company, Amsterdam, 1996.

[6] P. H. Louw and F. de Wet. The perception and identification of accent in spoken Black South African English. *Southern African Linguistics and Applied Language Studies*, Under review, 2005.

[7] J. C. Roux, P. H. Louw, and T. R. Niesler. The African Speech Technology project: An assessment. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages I:93–96, Lisbon, Portugal, 2004.

[8] C. van der Walt. The international comprehensibility of varieties of South African English. *World Englishes*, 19(2):139–153, 2000.

[9] B. van Rooy. Black South African English: phonology. In E. Schneider, K. Burridge, B. Kortmann, R. Mesthrie, and C. Upton, editors, *Handbook of Varieties of English I*, pages 943–952. Mouton, Berlin, 2004.

[10] B. van Rooy and G. van Huyssteen. The vowels of BSAE: current knowledge and future prospects. *South African Journal of Linguistics*, Supplement 38:15–33, 2000.

[11] D. Wissing. Black South African English: A new English? Some observations from a phonetic viewpoint. *World Englishes*, 21(1):129–144, 2002.

[12] 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.