Human and automatic accent identification of Nguni and Sotho Black South African English

Febe de Wet[a*], Philippa Louw[a] and Thomas Niesler[b]

It is well established that accent can have a detrimental effect on the performance of automatic speech recognition (ASR) systems. Whereas accents can be labelled in terms of a speaker's mother tongue, it remains to be determined if and when this distinction is appropriate for the development of ASR technology. This study compares the varieties of South African English produced by mother-tongue speakers of the Nguni and Sotho languages, who account for over 70% of the country's population. The aim of the investigation was to determine whether these two accent groups should be treated as a single variety by ASR systems, or whether it is better to consider them separately. To this end, two sets of experiments were carried out. First, a perceptual experiment was performed in which human listeners were required to classify different English accents. Subsequently, automatic speech recognition experiments were conducted to determine how the accuracy of an automatic accent identification system compares with these perceptual results, and whether the acoustic models benefit from the incorporation of Nguni/Sotho accent classifications. The results of the perceptual experiment indicated that most listeners could not correctly identify a speaker's mother tongue based on their English accent. This finding was supported by the results of the automatic accent identification and speech recognition experiments.

Introduction

South Africa is a multilingual society, with a total of eleven officially-recognized languages and an even greater number used in practice. South African English, which serves as the lingua franca, is therefore characterized 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.[1–3] The performance of ASR systems operating in multi-accent settings can be improved by the integration of accent-specific modelling. However, a prerequisite to the development of such systems is the identification of appropriate accent groups.

The accents of South African English include English spoken by English, Afrikaans, coloured (mixed-race), Indian and black mother-tongue speakers. In this study, we focus on Black South African English (BSAE)[‡] and as a benchmark also Standard South African English (SSAE). In particular, we try to determine whether BSAE, the variety of English spoken by second-language speakers whose mother tongue is one of the indigenous Bantu languages of South Africa, 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[6] 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 occur at the suprasegmental level.[7] 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'.[5]

Contrasting claims have also been made concerning differences at the 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 that investigated the types of labels given to BSAE speech showed that, when listeners tried to identify a person's mother tongue based on the speaker's English accent, they were not able to do so accurately.[4]

The study reported here investigated the relevance of these claims for the development of accent-robust ASR technology by means of both perceptual and ASR experiments. In our study, we have involved two variants of BSAE: English as spoken by mother-tongue speakers of a Sotho language (Northern Sotho, Southern Sotho, Tswana), and by mother-tongue speakers of an Nguni language (Zulu, Xhosa, Swati, Ndebele). This grouping is based on the observation that the Sotho languages share a system of seven vowels, whereas the Nguni languages have a common five-vowel system. Nguni and Sotho languages are widely spoken, and are the mother tongue to 45.7% and 25.5% of the South African population, respectively. Although our research focuses on the two BSAE varieties, English mother-tongue speech was also included to serve as a benchmark. This variety is referred to as Standard South African English and is known to differ strongly from BSAE.[5,9] We will focus only on the pronunciation differences associated with the chosen accent groups, and will not be concerned with accent-related differences in grammar and vocabulary.

In our perceptual experiment, listeners were asked to classify a person's English accent in terms of his/her mother tongue from a recording of speech. The ASR experiments determined whether it was possible to automatically classify a speaker's mother tongue correctly from a recorded English utterance, and also whether or not the speech recognition performance can be improved by keeping data from the two accent groups apart during acoustic modelling. The data for both the perceptual and the ASR experiments were taken from the African Speech Technology database of telephone speech.[10]

[‡] Assigning appropriate and commonly accepted labels to the different varieties of English spoken in South Africa remains a contentious issue among linguistic scholars.[4] Based on the arguments presented by De Klerk,[5] we have decided to use the terms Standard South African English and Black South African English in this paper. The former is a variety used by English mother-tongue speakers, whereas the latter refers to the English spoken by second-language speakers whose mother tongue is one of the indigenous Bantu languages of South Africa. For the purposes of this study, English as spoken in other African countries with indigenous Bantu languages, such as Zimbabwean English, is not included.

[a] Centre for Language and Speech Technology, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa.
[b] Department of Electrical and Electronic Engineering, University of Stellenbosch
*Author for correspondence. E-mail: fdw@sun.ac.za

The AST database

The African Speech Technology (AST) project was funded by the Department of Science and Technology from 2000 to 2003.[10] 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. To make provision for these known varieties, five English and three Afrikaans databases were collected.

For the five English databases, English mother-tongue speakers (SSAE) as well as black (BSAE), coloured, Asian and Afrikaans non-mother-tongue speakers were targeted. The BSAE database, in turn, includes speakers whose mother tongue is Xhosa, Zulu, Southern Sotho (Sesotho), Tswana (Setswana) and Northern Sotho (Sepedi). We used the AST SSAE and BSAE corpora to conduct the experiments in this study.

Each AST database contains between 38 and 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, words or phrases relevant to spoken dialogue systems, as well as 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 were recorded digitally using a Dialogic D/300-SC Primary Rate ISDN Interface.

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 roughly half of the speakers calling a toll-free number from a mobile (cellular) phone and the other half from a landline phone. Each speaker was presented with a unique data sheet containing the items to be read. The final SSAE and BSAE databases contain 303 and 236 phone calls, respectively, corresponding to between six and seven hours of speech per database.

Perceptual experiments

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, first, by means of a perceptual experiment. We report later on similar investigations using ASR systems.

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 (the final year of secondary school). Although some of the speakers had a higher, university qualification, they were not considered to have reached the acrolectal level.[7] Table 1 shows the distribution of mother tongues in the speaker population.

Of the 119 speakers, 37 were first-language speakers of a Sotho language, 35 of a Nguni language, and 47 of English. Overall, 62 speakers were female and 57 male, with a similar male/female ratio within the Nguni, Sotho and English groups. The average age of the Sotho and Nguni speakers was 30, with a standard deviation of 9 years, and the average age of the English speakers was 44 with a standard deviation of 12 years.

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

Mother tongue of speaker	Number of speakers		
	Male	Female	Total
Total English	**22**	**25**	**47**
Southern Sotho	9	4	13
Tswana	9	15	24
Total Sotho	**18**	**19**	**37**
Xhosa	8	8	16
Zulu	8	10	18
Ndebele	1	0	1
Total Nguni	**17**	**18**	**35**

Stimuli

We used a total of 180 stimuli, consisting of 30 single words and 30 phrases pronounced by native speakers of English, a Sotho language and an Nguni language, as set out in Table 1. We will refer to these three varieties respectively as Standard South African English (SSAE), 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 judgment. We would also not be able to determine which specific sounds influenced the listeners' judgment if only sentences were used as stimuli.

The phonetic content of the stimuli was selected according to established descriptions of BSAE.[6,7,9,12] 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 because they do not occur in the AST databases. Table 2 gives an overview of the contexts and realizations of the BSAE sounds as they occurred in the stimuli. All stimuli contained only English words, hence code-mixing did not occur in our data.

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

Listeners

A total of 36 participants (none of whom partook in the AST project) were recruited on campus. The mother-tongue distribution among the listeners is shown in Table 3 and indicates an equal representation of the Nguni and Sotho language groups. Most listeners were enrolled for an undergraduate course at the university, but the group also included a few postgraduate students. The female/male ratio of the group was 20:16.

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 buttons, representing the options

[‡] The terms basilect, mesolect, and acrolect are used in creole studies to describe the variation observed between different speakers' command of English. The continuum has been adapted to characterize the range of forms observed in so-called New Englishes.[11] According to the modified definition, basilect English is spoken by people who have little contact with L1 English and who have received little or no formal education, e.g. labourers, domestic servants, etc. Acrolect English is usually internationally intelligible and is used by highly educated speakers such as university lecturers, politicians and medical specialists, and whose English differs only slightly from the L1 English spoken in the region. Mesolect English accounts for everything in between and is usually nationally intelligible and used informally by educated people such as students, teachers, and nurses. In this study, we adhere to the definition[7] that 'mesolect speakers are characterized by phonetic and phonological differences from the standard variety, whereas an acrolect speaker differs as far as phonetic properties are concerned, but not phonological properties'.

available to choose from, that is, 'Sotho', 'Nguni', 'English' and 'I don't know'. Each stimulus was played only once; as soon as the participant had made his/her choice by selecting one of the four options, the next stimulus was played. The participants did not have advance knowledge of the words and sentences used as stimuli.

Instructions were given verbally to the participants before the experiment started. A short pre-test, consisting of three utterances, was carried out to demonstrate the procedure as well as to ensure that the participants could hear the stimuli clearly over a set of headphones. The test stimuli were presented in three sets of 60 and participants were allowed to take a short break between sets. On average, the participants required 20 minutes to complete the perceptual test. All listeners who participated in the experiment received a monetary reward for their contribution.

Results

Overall, 10% of all responses were 'I don't know'. These were mostly (77%) responses to the single-word stimuli. The percentage correct results, calculated after removal of all 'I don't know' responses, are summarized in Table 4. The table shows that 71.3% of the SSAE stimuli were correctly identified. This result indicates that listeners were able to distinguish between the SSAE and BSAE accents. Table 4 also shows that the identification accuracy for the NE and SE stimuli was much lower than for the SSAE stimuli.

Table 5 illustrates the results for the NE and SE stimuli, with responses to SSAE stimuli removed from the dataset. According to the data in this table, on average, only 53.4% of the stimuli was correctly identified as NE or SE. Listeners performed slightly better on the sentence stimuli (54.2%) than on the isolated words (52.4%), but the difference was found not to be significant according to an ANOVA significance test. This result seems to 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.

It was also observed that the listeners' responses showed a bias towards their own mother tongue (L1). The Nguni L1 listeners classified 65% of the BE sentence stimuli and 70% of the BSAE word stimuli as NE. A similar but weaker trend was observed for the Sotho L1 listeners' responses, who classified 56% of the BSAE sentence stimuli and 54% of the BSAE word stimuli as SE.

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' judgments.

Automatic accent identification

The previous section has demonstrated that human subjects cannot reliably distinguish between the two BSAE accents under study. In this section, we determine whether greater success can be achieved by means of the automatic classification system shown in Fig. 1. This architecture, referred to as *Parallel Phone Recognition followed by Language Modelling* (PPRLM), uses a parallel set of accent-specific automatic speech recognisers for explicit accent classification and has been used successfully for the purpose of language identification.[13,14] Each accent-specific speech recognition system includes an accent-specific language model. We now evaluate its ability to distinguish between SSAE,

Table 2. Realization of sounds in BSAE and example words.

Realization of sounds in BSAE	Example words
Vowels	
Neutralization of tense/lax vowels or long/short vowels	list, task, half, information, into, mini
Avoidance of central vowels	sentence, matric
[æ] replaced by [ɛ]	help
Schwa in open syllables replaced with [a]	operator
Diphthongs	
Narrower diphthongs realized as single monophthongs	telephone, make, day, change, only, over
Consonants	
Affricate [tʃ] becomes fricative [ʃ]	opportunities
Trilled [r] as allophone for liquid	directory
Velar plosive devoicing	goodbye
General	
cancel, check, continue, development, eleven, goodbye, perhaps, repeat	

Table 3. Mother-tongue distribution of the listeners who participated in the perceptual experiment.

Mother tongue of listener	Number of listeners		
	Male	Female	Total
Northern Sotho	3	4	7
Southern Sotho	4	3	7
Tswana	2	2	4
Total Sotho	**9**	**9**	**18**
Xhosa	3	6	9
Zulu	4	5	9
Total Nguni	**7**	**11**	**18**

Table 4. Accent identification accuracy (%) for SSAE, NE and SE.

Test database	Classified as (%)		
	SSAE	NE	SE
SSAE	**71.3**	15.0	13.7
NE	6.8	**55.3**	37.9
SE	8.3	48.2	**43.5**
Average correct		56.7	

Table 5. Accent identification accuracy (%) for NE and SE.

Test database	Classified as (%)	
	NE	SE
NE	**59.3**	40.7
SE	52.5	**47.5**
Average correct	53.4	

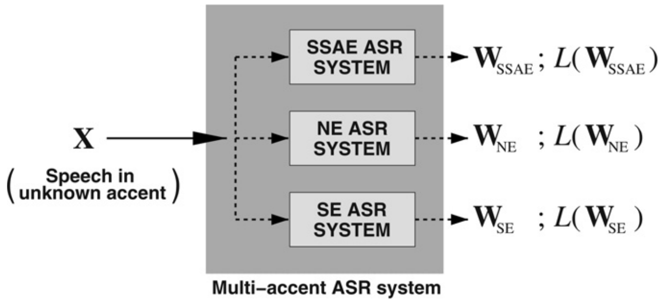

Fig. 1. Accent identification using parallel recognition systems for SSAE, NE and SE.

Table 6. English (SSAE), Nguni (NE) and Sotho-English databases.

Database	Training set			Test set		
	No. of speakers	Size (h)	Phone tokens	No. of speakers	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
SSAE	116	2.59	72 820	10	13.2	5 707

NE and SE accents. SSAE was included in these experiments primarily as a benchmark with which to confirm or refute the finding of the perceptual experiment that the SSAE and BSAE accents differ to a much greater degree than NE and SE.

In Fig. 1, each accent-specific recognition system provides a transcription **W** of the speech utterance **X** in terms of its phone inventory and a corresponding accent-specific language model. In addition to this transcription, each recognizer provides a likelihood $L(\mathbf{W})$. This probabilistic measure indicates how well the accent-specific phone models of each system are able to account for the speech **X**. In this framework, the accent of the input speech may be identified by simply identifying the transcription with the highest associated likelihood.

Data

The SSAE and BSAE databases described above were used for the automatic accent identification experiments. Manually produced and checked phonetic as well as orthographic transcriptions of these data were available. Because the mother tongue and the level of education of each speaker were known, it was possible to extract sub-portions of the BSAE corpus uttered by mesolect Nguni and Sotho speakers. These two databases (NE and SE, respectively), were each further subdivided into a training and a test set, as shown in Table 6.

The SSAE database indicated in Table 6 is a subset of the full AST SSAE database, and was designed to be of comparable size to the NE and SE databases. All test-sets were designed to have 50:50 male/female as well as cellular/landline ratios. Finally, separate development sets, consisting of approximately six minutes of speech from four speakers, were prepared for all three databases. These were used only for the optimization of recognition parameters, before final evaluation on the test-set. There was no overlap between the development set and either the test or training sets.

Acoustic models

Acoustic models were trained using the HTK tools[15] and the SSAE, SE and NE corpora. The SSAE recognition system used a set of 73 phones, including silence and speaker noise. The NE and SE recognition systems employed the same set of 90 phones, including silence and speaker noise, of which 70 were common to those used by the SSAE system. The 20 NE/SE phones not found in the SSAE data accounted for just 1.9% of the NE/SE phone tokens. Similarly, the 3 SSAE phones not found in the NE/SE data accounted for fewer than 0.1% of the SSAE phone tokens. Thus, although the phone sets used by the SSAE and the NE/SE systems were not exactly the same, there was a large degree of overlap.

The speech was parameterized as Mel-frequency cepstral coefficients (MFCCs) and their first and second differentials, with cepstral mean normalization (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 600 clustered states for each set of acoustic models.

Table 7. Accent identification accuracy (%) for SSAE, NE and SE.

Test database	Classified as (%)		
	SSAE	NE	SE
SSAE	**90.8**	3.6	5.6
NE	4.4	**43.9**	51.7
SE	7.5	38.3	**54.1**
Average correct		62.2	

Accent identification results

Table 7 shows the accuracy of the automatic accent identification system shown in Fig. 1, when presented with the test-sets listed in Table 6. The experimental results indicate that almost 91% of the SSAE test utterances were identified correctly, but that the classification accuracy for both NE and SE was substantially lower.

The results for a subsequent experiment in which only NE and SE were distinguished between are presented in Table 8. The equal percentages in the final column are incidental, and the figures do in fact differ after the first decimal.

Table 8 shows that, when distinguishing between the two accents of BSAE, the overall accuracy of the automatic identification system was 50.2%, which is almost chance. This agrees with the results of the perceptual experiments, indicating that it was not possible to discriminate reliably between NE and SE accents.

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 stem from the same variety. In this section, we will determine experimentally whether such a difference in performance can be found.

Acoustic models

Since our aim was to determine whether it is better to have distinct Nguni- and Sotho-English recognizers or to have a

Table 8. Accent identification accuracy (%) for NE and SE.

Test database	Classified as (%)	
	NE	SE
NE	**43.8**	56.2
SE	43.8	**56.2**
Average correct	50.2	

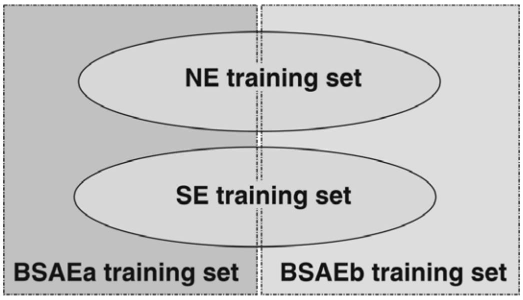

Fig. 2. Division of NE and SE sets into BSAEa and BSAEb.

single, general Black South African English recognizer, we further subdivided the NE and SE training sets as shown in Fig. 2.

Both the NE and SE training sets were divided in half, taking care to maintain the male/female and cellular/landline balance. Two new training sets, BSAEa and BSAEb, were then formed by pooling an NE and an SE subset. Hence BSAEa and BSAEb were accent-neutral with respect to the Nguni/Sotho distinction. Furthermore, since BSAEa and BSAEb contained 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.

The same acoustic models used by the language identification system, and discussed in the section Acoustic models, were used to perform the ASR experiments. However, in this case additional models were trained using the BSAEa and BSAEb corpora, using the same set of 90 phones employed by the NE and SE systems. As before, each set of acoustic models consisted of 3-state left-to-right triphone HMMs with eight Gaussian mixtures per state, diagonal covariance matrices and a total of approximately 600 clustered states.

Speech recognition results

Because the amount of training data was very limited, phone recognition was performed. Recognition was accomplished using the HTK decoder and a bigram phone language model obtained from the training set transcriptions. The performance of the triphone recognition system is shown in Table 9. The model set labelled BSAE represents the average performance of the two recognition systems trained on BSAEa and BSAEb, respectively. This was done to avoid any bias which may have resulted from the particular way in which NE and SE were split into BSAEa and BSAEb.

As described under Acoustic models, the SSAE phone set was slightly different from that used by the other systems. Since the primary focus of our experiments is the treatment of accents within BSAE, we did not attempt to normalize the phone inventories by mapping phone labels to a smaller common set. The SSAE system was included as a baseline, and the differences in the phone sets affected fewer than 2% of the phones upon which the results are based.

Table 9. Phone accuracies for the triphone recognition experiments evaluated on the SSAE, NE and SE test sets.

Model set	Test-set accuracy (%)		
	SSAE	NE	SE
SSAE	71.3	27.9	32.2
NE	35.8	48.6	48.8
SE	36.1	48.8	50.1
BSAE	36.2	48.5	50.2

The results in Table 9 illustrate that, as anticipated, there was a definite difference between the accent of English mother-tongue speakers and that of Nguni or Sotho mother-tongue speakers. In particular, recognition performance of the SSAE system deteriorated strongly when presented with NE or SE utterances. Similarly, the recognition performance of the NE, SE and BSAE systems all deteriorated when tested using SSAE utterances. Hence we may conclude that there is a strong motivation for separating English as spoken by English mother-tongue speakers from English spoken by Sotho or Nguni mother-tongue speakers when developing automatic speech recognition systems. It is also evident from the values in Table 9 that the recognition accuracy of the SSAE system is much better than that of the BSAE, NE and SE systems. We believe that this is due to the higher fluency of the mother-tongue SSAE recordings relative to the three other non-mother-tongue datasets. For example, it has been shown elsewhere that the phone recognition accuracy of mother-tongue Xhosa speech is much closer to that achieved by an English mother-tongue system.[16]

The results in Table 9 also show that the NE, SE and BSAE systems all exhibit similar performance when tested on NE as well as SE utterances. From the same table, the average phone recognition accuracy of an ideally matched NE/SE system can be calculated to be 49.3%. Ideal matching in this case refers to the situation in which all NE utterances are processed by the NE recognition system, and all SE utterances by the SE system. This performance is indistinguishable from the average accuracy of the BSAE system, which is also 49.3%. From this we conclude that there is no merit in separating English as spoken by Nguni and Sotho mother-tongue speakers when developing automatic speech recognition systems.

Discussion and conclusions

The results obtained in the perceptual experiment do not support the claim made by some mother-tongue speakers of South Africa's Bantu languages that they can determine a speaker's mother tongue from his/her English accent. This finding is supported by the automatic accent identification system, which was also not able to distinguish between Nguni and Sotho varieties of BSAE. In contrast, both the perceptual and the automatic accent identification systems demonstrated that it is possible to distinguish between SSAE and BSAE.

All the above findings were corroborated by the speech recognition experiments. These showed no discernible performance difference between accent-specific and accent-neutral systems for the two varieties of BSAE. However, for optimal speech recognition performance separate accent-specific recognition systems should be maintained for SSAE and BSAE.

It may be questioned whether our use of telephone-bandwidth speech data might adversely affect the general correctness of our conclusions. Since no suitable wideband speech data were available for the development of automatic speech recognition and accent identification systems, we could not address this issue directly. However, the use of wideband and telephone-bandwidth speech has been compared in a similar set of perceptual experiments, and it was found that the restriction to telephone-bandwidth speech has very little effect on the accuracy with which participants are able to identify accent.[17]

Finally, the scarcity of suitable data with which to perform experiments motivated our use of the Nguni and Sotho language groupings. Further research is necessary to determine whether or not there are significant differences in accent within these two groups.

This article is a revised and extended version of a paper presented at the ISCA Tutorial and Research Workshop on Multilingual Speech and Language Processing (Stellenbosch, South Africa, 2006). Our work was supported by the National Research Foundation under grants FA2005022300010 and GUN2072874.

Received 22 September 2006. Accepted 16 April 2007.

1. Goronzy S., Sahakyan M. and Wokurek W. (2001). Is non-native pronunciation modelling necessary? In *Proceedings, Eurospeech 2001*, pp. 309–312. Aalborg, Denmark.
2. Aalburg S. and Hoege H. (2003). Approaches to foreign-accented speaker-independent speech recognition. In *Proceedings, Eurospeech 2003*, pp. 1489–1492. Geneva, Switzerland.
3. Kessens J. (2006). Non-native pronunciation modelling in a command and control recognition task: a comparison between acoustic and lexical modelling. In *Proceedings, First ISCA ITRW on Multilingual Speech and Language Processing (MULTILING)*, CDROM. Stellenbosch, South Africa.
4. Coetzee-Van Rooy S. and van Rooy B. (2005). South African English: labels, comprehensibility and status. *World Englishes* **24**(1), 1–19.
5. de Klerk V. (2003). Towards a norm in South African Englishes: the case for Xhosa English. *World Englishes* **22**(4), 463–481.
6. van Rooy B. and van Huyssteen G. (2000). The vowels of BSAE: current knowledge and future prospects. *S. Afr. J. Linguistics*, Suppl. **38**, 15–33.
7. Wissing D. (2002). Black South African English: A new English? Some observations from a phonetic viewpoint. *World Englishes* **21**(1), 129–144.
8. van der Walt C. (2000). The international comprehensibility of varieties of South African English. *World Englishes* **19**(2), 139–153.
9. van Rooy B. (2004). Black South African English: phonology. In *Handbook of Varieties of English*, vol. 1, eds E. Schneider, K. Burridge, B. Kortmann, R. Mesthrie and C. Upton, pp. 943–952. Mouton, Berlin.
10. Roux J.C., Louw P.H. and Niesler T.R. (2004). The African Speech Technology Project: an assessment. In *Proceedings, 4th International Conference on Language Resources and Evaluation*, Vol. 1, pp. 93–96. Lisbon, Portugal.
11. Meierkord C. (2005). Black South African Englishes – towards a variationist account. Online: http//webdoc.sub.gwdg.de/edoc/ia/eese/artic25/meierk/1_2005.html (accessed 16/05/2007).
12. Gough D. (1996). Black English in South Africa. In *Focus on Africa*, Varieties of English around the world, G15, pp. 53–77. John Benjamins, Amsterdam.
13. Zissman M.A. and Berkling K.M. (2001). Automatic language identification. *Speech Commun.* **35**, 115–124.
14. Niesler T.R. and Willett D. (2006). Language identification and multilingual speech recognition using discriminatively trained acoustic models. In *Proceedings, First ISCA ITRW on Multilingual Speech and Language Processing (MULTILING)*, CD-ROM. Stellenbosch, South Africa.
15. Young S., Evermann G., Hain T., Kershaw D., Moore G. Odell J., Ollason D., Povey D., Valtchev V. and Woodland P. (2002). *The HTK Book, version 3.2.1.* Cambridge University Engineering Department, Cambridge.
16. Niesler T.R. and Louw P.H. (2004). Comparative phonetic analysis and phoneme recognition for Afrikaans, English and Xhosa using the African Speech Technology telephone speech databases. *S. Afr. Computer J.* **32**, 3–12.
17. Louw P.H. and de Wet F. (in press). The perception and identification of accent in spoken Black South African English. *Southern African Linguistics and Applied Language Studies* **25**(1).