Comparative phonetic analysis and phoneme recognition for Afrikaans, English and Xhosa using the African Speech Technology telephone speech databases

T Niesler[a] P Louw[b]

[a]*Department of Electronic Engineering, University of Stellenbosch, Stellenbosch, South Africa,* trn@dsp.sun.ac.za
[b]*Research Unit for Experimental Phonology, University of Stellenbosch, Stellenbosch, South Africa,* phlouw@sun.ac.za

Abstract

This paper concerns the Afrikaans, English and Xhosa speech databases recently developed as part of the African Speech Technology project. The three corpora are analysed and compared in terms of their phonetic content, diversity and mutual overlap. Connected phoneme recognition systems are subsequently developed and tested in each language.

Keywords: *Multilingual speech recognition, phoneme recognition*
Computing Review Categories: *I.2.7*

1 Introduction

The African Speech Technology (AST) project is pioneering the technological development of the indigenous languages of South Africa. One of the outcomes of this project is the development of speech corpora in five of the country's eleven official languages, namely South African English, Xhosa, Afrikaans, Zulu and Southern Sotho. This data is aimed at the development of multilingual speech technology within the South African context. The work presented here describes the development of a set of benchmark speech recognition systems using the Afrikaans, English and Xhosa databases.

2 The AST speech databases

The AST telephone speech databases have been collected over both mobile and fixed networks and have been designed in the first instance to support general information retrieval and transaction teleservices [7]. A total of 400 speakers were recruited from each of the 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 phonetically-rich words and sentences. Spontaneous items included gender, age, mother-tongue, place of residence and level of education.

Databases for the following language groups were used for experimentation in this work:

EE : English spoken by English mother-tongue speakers.

AA : Afrikaans spoken by Afrikaans mother-tongue speakers.

XX : Xhosa spoken by Xhosa mother-tongue speakers.

Further databases are still under development within the AST project and will be used in future research as they become available. In particular, databases of English and Afrikaans spoken by non-mother-tongue speakers as well as Zulu and Southern Sotho were not considered.

2.1 Database contents

Each completed AST speech database includes:

- 8-bit a-law encoded speech waveforms of each utterance obtained at a sampling rate of 8kHz.
- Orthographic (word-level) transcriptions of each utterance.
- Phoneme-level transcriptions of each utterance.

The orthographic transcriptions were produced and validated by human transcribers. Initial phoneme transcriptions were obtained from the orthography using grapheme-to-phoneme rules except for English, where a pronunciation dictionary was used instead. These were subsequently corrected and validated manually by human experts. Since the work presented here focuses on phoneme recognition, we employ mainly the audio data and phoneme transcriptions in our experiments.

2.2 Phoneme set

A global set of 154 phonemes, referred to in the following as **ASTBET**, has been used to transcribe the AST databases. As for several other multilingual speech databases [5], [9], [10] these phonemes were based on the definitions of the International Phonetic Alphabet. The appendix presents a table of the phonemes that occur in the AA, EE and XX databases. Both the IPA as well as the corresponding ASTBET strings have been listed.

ASTBET has been chosen to cover all the sounds occurring in the AST corpora. Besides spanning five languages and various non-mother-tongue variations, it includes labels for diphthongs and tripthongs that occur across word boundaries in spontaneous speech. Furthermore, several combinations of diacritic symbols and consonants or vowels have been considered to be unique phonemes. The diacritics in question include ejection, aspiration, syllabification, voicing, devoicing and duration.

2.2.1 Vowels

A total of 29 vowels are represented in ASTBET. For all vowels except lax vowels a longer-duration counterpart is included. For example, for the phoneme `/i/` in the Afrikaans word "*siek*" there is a phoneme `/i_long/` as found in "*vier*". Duration in English and Afrikaans is only transcribed when a sound is distinctively longer or when it is phonemic[1]. Both lax and tense vowels are included in the phoneme set because lax vowels occur frequently in English but not at all in Afrikaans or the African languages. Examples of lax vowels are those found in the words "*him*" and "*put*".

2.2.2 Diphthongs and tripthongs

ASTBET includes 31 unique diphthongs, of which 10 occur in English and 14 in Afrikaans. The English and Afrikaans diphthongs are similar, but no diphthongs are shared between the two languages mainly due to the use of lax vowels in English. Diphthongs occur only during code-switching[2] in the African languages and are not intrinsic to the languages themselves.

A number of diphthongs that occur across word boundaries in spontaneous speech are also included in the ASTBET phoneme set, for example in the Afrikaans words "*drie-en-twintig*". Finally, ASTBET also includes 5 tripthongs, most of which occur across word boundaries in spontaneous speech. Examples in this case are the Afrikaans words "*hy is*" or the Afrikaans mother-tongue pronunciation of the English word "*about*".

2.2.3 Affricates

ASTBET contains 14 unique affricates of which 12 are unique to the African languages and 9 occur in Xhosa.

2.2.4 Stop sounds

In the African languages ejection and aspiration are phonemic and therefore the ejective and aspirated version of a stop sound are always included in the phoneme set. However, aspiration is not phonemic and ejective sounds do not occur in Afrikaans or in English. Therefore basic[3] versions of the stop sounds are also included in ASTBET.

As a result the phoneme set includes at least three versions of each stop sound e.g. `/p/`, `/p_asp/` and `/p_edj/`, and in some cases also voiceless and implosive versions, e.g. `/b_vcls/` and `/b_imp/`.

2.2.5 Clicks

Xhosa uses dental, lateral and palatal click sounds. These three basic clicks are extended to a total of fifteen by including combinations of aspiration, nasalisation and voicing, which can be phonemic [3].

2.3 Training- and test-sets

Each database has been split into a training set used for development of the acoustic models, and a test set for subsequent evaluation, in the ratio of approximately 95:5. No speaker occurring in the test set occurs also in the corresponding training set. Table 1 shows the amount of data (time) as well as number of speakers and number of phonemes in each training set.

Database	Speech (hours)	#Speakers	#Phonemes
AA	6.82	249	195,145
EE	6.31	258	178,738
XX	6.61	205	168,827

Table 1: AST speech databases: training sets.

Each test set was further divided into a development and an evaluation test set (dev-test and eval-test respectively). The former was used to optimise various parameters used during acoustic model training as well as recognition, while the latter was reserved for final testing. There was no speaker-overlap between the development- and evaluation-sets, and each contained both male and female speakers. Table 2 summarises the division of the test sets for each database.

Database	Speech (mins)		#Speakers		#Phonemes	
	dev	eval	dev	eval	dev	eval
AA	6.5	19.5	4	12	3,125	9,299
EE	6.4	16.4	4	11	2,880	7,579
XX	5.2	16.9	4	9	2,398	7,482

Table 2: AST speech databases: test sets.

Response-rates[4] differed among the various speaker target groups, leading to the variation in the quantity of data that was collected for each database and as is reported in tables 1 and 2.

From table 1 we also observe that while for English and Afrikaans the training data contains 8.0 and 7.9 phonemes per second of recorded speech respectively, for Xhosa this figure is just 7.1. This may be attributed to the fact that, as a general rule, the penultimate syllable in Xhosa is lengthened. This applies to single words, and to the last word of a phrase or sentence. Is was also found that spelled words and digit strings were generally pronounced more slowly than they were for

[1] We understand duration to be phonemic when it affects the meaning of the word.

[2] Code-switching occurs when a speaker changes from one language to another during discourse.

[3] Neither ejective nor aspirated.

[4] The proportion of the 400 datasheets (as set out in section 2) that were responded to successfully.

the other two languages. In cases where penultimate syllable lengthening occurred, the vowel in this syllable was transcribed with the appropriate diacritic to indicate length.

2.4 Noise and silence models

In addition to phonemes representing speech sounds, the following non-speech sounds were labeled in all databases.

`sil` : Silence model.

`spk` : Speaker noise. This includes various sounds and noises made by the speaker that are not part of the prompted text. Examples include: lip smack, cough, throat clear, tongue click, loud breath, laugh and loud sigh.

`sta` : Stationary noise. This category contains background noise that is not intermittent and has a more or less stable amplitude spectrum. Examples include car noise, road noise, voice babble and street noise.

`int` : Intermittent noise. This category contains noises of an intermittent nature. These noises typically have pauses between them, or change their colour over time. Examples include music, background speech, baby crying, phone ringing, door bell and paper rustle.

`ext` : External noises are typically abrupt, like a door slam. Often they occur between words without influencing intelligibility. However when an external noise appears while a word is being spoken, the utterance is in some cases considered to be unintelligible and marked accordingly.

All the above non-speech sounds were transcribed at the place of occurrence, using reserved symbols. Furthermore, when `sta`, `int` or `ext` overlap with speech, this overlap is indicated by means of start and end markers in the transcription text.

3 Corpus analysis

This section presents a comparative analysis of the phonetic content and character of each of the three databases under study.

3.1 Phoneme statistics

Table 3 shows the degree to which phonemes in one language's phoneme set are also present in those of the others'. These figures exclude the noise- and silence phonemes described in section 2.4. It is evident that the number of different phonemes in Xhosa is significantly larger than for either English or Afrikaans. Furthermore, each language contains phonemes not present in any of the others. For example, 19% of phonemes in the Afrikaans phoneme set are not covered by the Xhosa phoneme set. The particular phonemes in question may be identified from the table in the appendix.

Phoneme set	Number of phonemes	% Phonemes not in phoneme set		
		AA	EE	XX
AA	83	0%	25%	19%
EE	72	14%	0%	14%
XX	110	39%	44%	0%

Table 3: Phoneme set coverage between languages.

The much larger number of phonemes in Xhosa is due to not only to the intrinsically greater variety of sounds (e.g. clicks and affricates), but also to the large number of Afrikaans and English words used during code-switching. This issue is explored further in section 3.2.

Table 3 illustrates the overlap between phoneme sets but does not take the relative frequencies of each phoneme into account. Table 4 indicates to what degree the phoneme set of each language covers the phonemes in each of the other languages' training sets. For example, the table shows that 1.3% of the phonemes in the Afrikaans training set are not covered by the Xhosa phoneme set.

Training set	Number of phonemes	% phonemes not in phoneme set		
		AA	EE	XX
AA	195,145	0.0%	1.5%	1.3%
EE	178,738	0.6%	0.0%	7.2%
XX	168,827	10.9%	11.6%	0.0%

Table 4: Phoneme set coverage between training databases.

Since the values in table 4 are all lower than the corresponding values in table 3, we conclude that the phonemes not covered between phoneme sets are also the less-frequent phonemes in each language. For example, even though 25% of the phonemes in the AA phoneme set do not appear in the EE phoneme set, these account for only 1.5% of the Afrikaans training data.

The phoneme set for Xhosa covers 98.7% of the phonemes in the Afrikaans data, and 92.8% of those in the English data. On the other hand, the Afrikaans and English phoneme sets respectively cover just 89.1% and 88.4% of the phonemes found in the Xhosa data. The Afrikaans phoneme set covers 99.4% of the phonemes in the English data, while the English phoneme set covers 98.5% of those in the Afrikaans data. Hence we see that the Xhosa phoneme set covers Afrikaans and English to a greater extent than either of the latter two cover Xhosa. However Afrikaans and English exhibit high mutual phoneme coverage. These issues are explored further in the following section.

3.2 Borrowed phonemes and words

For the purposes of this analysis, a borrowed word will refer to one imported from another language during code-switching. There is an important distinction between borrowed words and *loanwords*. Loanwords are foreign words that have been phonologised and thus more fully integrated into the phonetic structure of the language. For example, the Xhosa word "*idesika*" has been derived from the English word "*desk*". This would be referred to as a loanword since the phonology of the original English word has been altered to conform to the morphological structure of Xhosa. On the other hand, in the Xhosa sentence "*Ezi zinto zidla i-twenty five Rand*", meaning literally: "*These items cost twenty five Rand*", the words "*twenty five Rand*" are borrowed words because they have not been phonologised. According to our definition, code-switching occurs when borrowed words but not when loanwords are spoken.

South African speakers of one of the three languages studied in this paper will most often also speak at least one of the remaining two. Hence it is interesting to consider the proportion of each language's phoneme set used to pronounce borrowed words.

During orthographic transcription of each database, words considered foreign to the language in question were marked accordingly. For example, when the Afrikaans exclamation "*ag*" was uttered in the EE database, it was marked, as were English numbers in the XX database. Hence it was possible to determine a list of the borrowed words present in each of the three databases. Table 5 presents these as percentages of both the vocabulary as well as the training set. The vocabulary is considered to be the set of all unique words occurring in the orthographic transcription of the training set (including borrowed words).

Training	Borrowed words as % of		Examples of
set	vocabulary	training set	borrowed words
AA	5.8%	0.3%	phone, sorry, Brixton
EE	2.8%	0.3%	kloof, ja, Dalsig
XX	14.8%	40.1%	ten, o'clock, Durban

Table 5: Prevalence of borrowed words in eachlanguage.

Table 5 shows that all languages have a significant proportion of borrowed words in their vocabularies, and that this proportion is highest for Xhosa. In particular, while over 40% of the Xhosa training set consists of borrowed words, this figure is just 0.3% for both English and Afrikaans. Borrowed words in English and in Afrikaans were found to consist mostly of proper nouns. Xhosa borrowed words, on the other hand, consisted not only of proper nouns, but most often of English numbers and names of months.

Code-switching occurs often in modern Xhosa, and leads to the high proportion of borrowed words. In particular, it is accepted practice in some African languages to cite digits, numbers and amounts in either the mother-tongue or in English (and sometimes Afrikaans). In Xhosa, for instance, the item "2353" is often read simply as "*Two thousand three hundred and fifty three*". However it could also be read as "*Amawaku amabini namakhulu amathathu namashumi amahlanu nantathu*", meaning literally: "*Thousands-that-are-two and hundreds-that-are-three and tens-that-are-five and three*". Code-switching is also likely to appear in the spontaneous citing of dates and times. For example, a Xhosa-speaking person might cite the time as "*Ixhesha ngoku ifive past ten*", meaning literally: "*The time now is five past ten*" [7].

Once the set of borrowed words for each language had been determined, the set of phonemes used to pronounce words intrinsic to each of the three languages (i.e. not borrowed) could be determined. From these we could determine which phonemes in the phoneme set of each language are used exclusively to pronounce borrowed words. Table 6 expresses these borrowed phonemes as percentages of both the complete phoneme set as well as of the total number of phonemes in the training set.

Training	Borrowed phonemes as % of	
set	phoneme set	training set
AA	29%	1.6%
EE	31%	0.3%
XX	34%	10.3%

Table 6: Phonemes borrowed from other languages.

The figures in table 6 indicate that each language devotes an approximately equal proportion of its phoneme set to borrowed words. However for Xhosa these phonemes represent a much larger proportion of the training set than for Afrikaans and for English. We conclude that Xhosa is by far the most phonetically rich of the three languages. Furthermore, the variety of sounds produced by Afrikaans mother-tongue speakers has been expanded more by the presence of the other languages than for English mother-tongue speakers. This is true to an even greater extent for Xhosa mother-tongue speakers.

3.3 Language models

In order to obtain some insight into the diversity of the three phoneme sets, unigram language models were obtained from the training set phoneme transcriptions of each database. Perplexities were calculated on the evaluation-test sets and are shown in table 7. Perplexity is a measure of the predictability of a phoneme sequence [2]. A higher perplexity indicates that the next phoneme in a sequence is harder to predict.

Training set	Number of unigrams	Unigram perplexity (eval-test)
AA	83	29.6
EE	72	34.3
XX	110	36.9

Table 7: Phoneme unigram language models.

Afrikaans has a lower unigram phoneme perplexity than English, even though it has a larger number of phonemes. This implies that a larger proportion of Afrikaans phonemes are used less frequently than in English. This is confirmed by the graph in figure 1, which shows the fraction of the phonemes in each training set covered by the *n* most-frequent phonemes, were *n* is varied on the horizontal axis.

The relatively large number of infrequent phonemes in Afrikaans is due to the prevalence of words imported from English and other languages, as already pointed out in section 3.2. This effect is even more pronounced for Xhosa, which has approximately the same unigram perplexity as English although it has a substantially larger phoneme set.

From figure 1 we see that for Afrikaans and English more than 99% of the phonemes in the training set are covered by the most-frequent 34 and 40 phonemes respectively, while for Xhosa 68 phonemes must be retained for the same coverage.

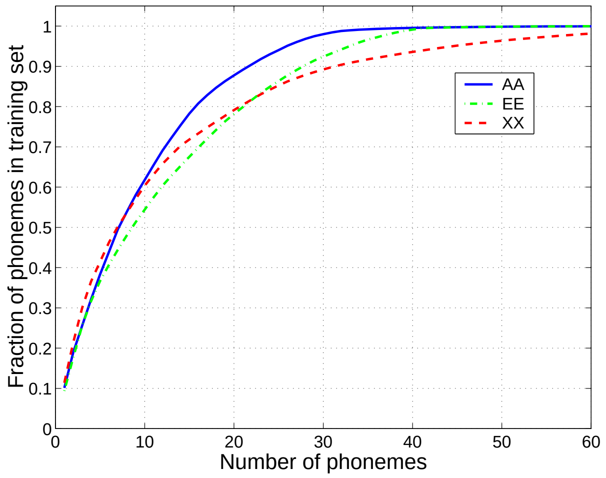

Figure 1: Training set coverage by most-frequent phonemes.

The less-frequent phonemes in Xhosa occur relatively more-frequently than their counterparts in English and Afrikaans, due to the much higher prevalence of borrowed words. This implies that Xhosa requires a substantially larger phoneme set for accurate phonetic modelling than the other two languages.

A backoff bigram language model was obtained from the training set phoneme transcriptions of each database [4]. Table 8 lists the test set perplexity of each bigram language model. Absolute discounting was used for the estimation of language model probabilities [8].

Training set	Number of bigrams	Bigram perplexity (eval-test)
AA	1567	13.0
EE	2032	14.2
XX	2715	14.2

Table 8: Phoneme bigram language models.

It is interesting to note that Xhosa shows the same bigram perplexity as English even though it has a higher unigram perplexity and significantly larger phoneme set. This indicates that there is a stronger sequential relationship between consecutive phonemes for Xhosa that there is for English. We believe this to be because the Xhosa language conforms to the phonological /CV/ or /CwV/ structure where a consonant is always followed by a vowel or the glide /w/. As discussed in section 3.2, loanwords are phonologised so that they conform to the /CV/ structure of Xhosa. For example the Xhosa word "*itafile*", meaning "*table*", has a /VCVCVCV/ structure and has been derived from the Afrikaans word "*tafel*".

4 Phoneme recognition

Phoneme recognition experiments have been performed due to the lack of language model training data as well as the widely varying vocabularies. Phoneme recognition experiments allow the quality of the acoustic models to be evaluated independently of the lexical constraints that are imposed by word recognition. This is especially relevant for Xhosa for which the concept of a word is less clearly defined than it is for Afrikaans and English.

A set of baseline HMM acoustic models was trained for each language listed in table 1 using the HTK tools [12]. The following sections describe the development and evaluation of these models.

4.1 Speech recognition engine

Speech recognition experiments were performed with the HTK time-synchronous Viterbi decoder [12]. This hidden Markov model (HMM) based speech recogniser performs a time-synchronous beam-search using the Token-Passing procedure [13]. The word insertion penalty and language model scaling factors were adjusted to optimise recognition accuracy measured on the development test set.

4.2 Acoustic parameterisation

The 8kHz sampled speech waveform was divided into overlapping frames each containing 256 samples at a frame-rate of 100Hz. A 39-dimensional feature vector was calculated for each speech frame, consisting of 12 MFCCs, energy, and their first and second differentials. This parameterisation has been found to perform very well by a number of authors [1], [11]. Cepstral mean normalisation (CMN) was not applied since these models are targeted for use in real-time telephone dialogue systems. The utterances to be processed by these systems are expected often to be too short for straightforward application of CMN.

4.3 Monophone models

Diagonal-covariance speaker-independent monophone models with three states per model and one Gaussian mixture per state were trained using the phonetically-labeled training sets by embedded Baum-Welsh re-estimation. The recognition accuracy of these acoustic models measured on the evaluation test set are shown in table 9. The bigram language model described in section 3.3 was used during decoding.

Training set	Phoneme recognition accuracy (%)
AA	42.4
EE	44.5
XX	40.1

Table 9: Phoneme recognition accuracy of monophone models.

In this work the monophone models serve only as interim models in the process of triphone model development. Hence they will not be considered in any further detail.

4.4 Triphone models

Cross-word triphone models were obtained from the monophone models by decision-tree state clustering [11]. This process groups together triphones with the same base phoneme on grounds of the acoustic similarity of their left- and right-context phonemes. This is done in order to reduce the enormous number of possible triphones to a number for which the parameters can be more reliably estimated from the limited amount of acoustic training data. The five silence and noise models described in section 2.4 were permitted as left- and right-contexts for the remaining triphones in the clustering process. However these models were not expanded to triphones themselves. Table 10 shows the effect of clustering on each language's set of acoustic models. The second column indicates the total possible number of distinct triphones for each language. Due to the inclusion of the silence and noise models as valid triphone contexts, this figure is somewhat larger than N_p^3, where N_p is the number of phonemes in the language's phoneme set as indicated in table 3. The third column of table 10 indicates the total number of distinct triphones present in the training set. Notice that this is much smaller than the corresponding number in the second column, indicating that training data is available for only a small faction of the set of possible triphones. The fourth and fifth columns indicate the number of distinct models and states remaining after clustering. While clustering proceeds at a state-level, often all three states of a triphone model will be clustered in the same way as those of a different model. In this case the two triphones themselves will be considered to be clustered. Since this is not always the case, however, the number of clustered triphones is smaller than three times the number of clustered states.

Training set	#Possible triphones	#Distinct triphones	#Clustered triphones	#Clustered states
AA	642,757	10,020	5,562	2,623
EE	444,137	13,852	7,741	2,546
XX	1,454,755	12,289	7,879	2,858

Table 10: State-clustering process for triphone models.

The HMM model sets for the three languages as described in table 10 contain approximately the same number of clustered states and therefore also approximately the same number of free parameters.

Finally the number of Gaussian mixtures per state of each set of triphone models was gradually increased. Each such increase was followed by four iterations of embedded re-estimation in order to update the HMM parameters. A total of 8 mixtures per state achieve approximately optimum phoneme recognition performance for all three languages measured on the development test sets. Increasing the number of mixtures even further led to eventual deterioration in recognition accuracy. Table 11 shows the corresponding evaluation test set phoneme recognition accuracies at several stages of this process. The table also indicates the recognition accuracies for the 8-mixture models when using a unigram instead of the bigram language model (LM).

From table 11 we see that the EE models outperform the AA models by a relative 7.3% margin, and these in turn outperform the XX models by a relative 3.1% margin. English has the smallest number of phonemes, followed by Afrikaans and then Xhosa, so this result is not unexpected. However figure 1 and tables 7 and 8 indicate that Afrikaans has the lowest unigram and bigram perplexity and also exhibits the best training set coverage for a reduced number of phonemes. From this point of view one might expect better recognition results to be achievable for Afrikaans than for English.

Number	Phoneme recognition accuracy (%)					
of	Bigram LM			Unigram LM		
mixtures	AA	EE	XX	AA	EE	XX
1	60.4	66.0	56.3	-	-	-
2	64.7	69.8	60.0	-	-	-
4	65.9	73.0	63.0	-	-	-
6	67.4	74.0	63.8	-	-	-
8	67.4	74.7	64.3	60.2	68.8	57.5

Table 11: Phoneme recognition accuracy of triphone models.

Table 11 also shows that the largest improvement in recognition accuracy when moving from a unigram to a bigram language model is achieved for the AA models (7.2% relative increase), followed by XX (6.8%) and then EE (5.9%). This agrees with the earlier observation that Afrikaans has the lowest unigram and bigram perplexity and that Xhosa shows a larger perplexity improvement than English when moving from a unigram to a bigram language model. We conclude that the constraints imposed by the /CV/ phonetic structure in Xhosa have been captured by the bigram language model and have aided recognition.

Finally, table 11 indicates that the EE models exhibit the largest relative gain (8.7%) in performance when increasing the number of mixtures from 1 to 8, followed by XX (8.0%) and then AA (7.0%).

Our phoneme recognition rates for EE compare well with those obtained by others for American English using the TIMIT corpus [6]. The different recording conditions of the AST and TIMIT corpora as well as the smaller number of phonemes used in the latter (61) do not allow a strict comparison of these figures. However it does appear to indicate general agreement with related published work.

5 Summary and conclusions

This paper has presented a comparative analysis at the phonetic level as well as phoneme recognition experiments for the Afrikaans, English and Xhosa speech databases developed as part of the AST project. These are the first set of benchmark evaluations performed with this data and will serve as a basis for further work.

Best performance in terms of phoneme recognition accuracy was achieved for English, followed by Afrikaans and then Xhosa. Analysis of the phoneme sets and phoneme transcriptions of each language showed Xhosa to be substantially more phonetically diverse than the other languages. From this point of view phoneme recognition may be expected to be intrinsically more difficult for Xhosa. However the analyses also show Afrikaans to have the most compact phoneme set and predictable sequential phonetic structure of the three languages. Hence there appears to be room for improvement in recognition accuracy for Afrikaans.

6 Acknowledgments

The AST project was supported by the Innovation Fund (project number 21213) of the Department of Arts, Culture, Science and Technology (DACST) of the South African National Government. The authors would like to thank Prof. J.C. Roux of the Research Unit for Experimental Phonology at Stellenbosch University for his valuable input, as well as the reviewers for their constructive comments.

References

[1] R. Haeb-Umbach and M. Loog. An investigation of cepstral parameterisations for large vocabulary speech recognition. In *Proc. Eurospeech*, pages 1323–1326, Budapest, Hungary, 1999.

[2] F. Jelinek, R.L. Mercer, and L.R. Bahl. A maximum likelihood approach to continuous speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 5:179–190, March 1983.

[3] M. Jessen and J.C. Roux. Voice quality differences associated with stops and clicks in Xhosa. *Journal of Phonetics*, 30:1–52, 2002.

[4] 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*, 35(3):400–401, March 1987.

[5] J. Köhler. Multi-lingual phoneme recognition exploiting acoustic-phonetic similarities of sounds. In *Proc. ICASSP*, pages 2195–2198, Atlanta, 1996.

[6] L.F. Lamel and J.L. Gauvain. High performance speaker-independent phone recognition using CDHMM. In *Proc. Eurospeech*, pages 121–124, Berlin, 1993.

[7] P.H. Louw, J.C. Roux, and E.C. Botha. African speech technology (AST) telephone speech databases: corpus design and contents. In *Proc. Eurospeech*, Aalborg, Denmark, September 2001.

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

[9] T. Schultz. Globalphone: a multilingual speech and text database developed at Karlsruhe University. In *Proc. ICSLP*, pages 345–348, Denver, 2002.

[10] T. Schultz and A. Waibel. Language independent and language adaptive large vocabulary speech recognition. In *Proc. ICSLP*, Sydney, 1998.

[11] P.C. Woodland, J.J. Odell, V. Valtchev, and S.J. Young. Large vocabulary continuous speech recognition using HTK. In *Proc. ICASSP*, pages 125–128, Adelaide, 1994.

[12] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland. *The HTK book, version 2.2*. Entropic, 1999.

[13] S.J. Young. *Token passing, a simple conceptual model for connected speech recognition systems*. Technical Report TR38, Cambridge University, 1989.

7 Appendix

Table 12 presents a list of the ASTBET phonemes used to transcribe the Afrikaans (AA), English (EE) and Xhosa (XX) databases. An example of a word in which the phoneme occurs in each language is given. Absence of such an example indicates that, in our data, the phoneme did not occur in that language. Furthermore, if the example word appears in italics, it indicates that it is a borrowed word, as defined in section 3.2. Noise and silence phonemes, as detailed in section 2.4, are not included in the table.

DESCRIPTION	**IPA**	**ASTBET**	**AA**	**EE**	**XX**
Voiceless Bilabial Plosive	p	p	pop	spit	*Cape*
Aspirated Bilabial Plosive	p^h	p_asp			phila
Ejective Bilabial Plosive	p’	p_edj			pasa
Voiced Bilabial Plosive	b	b	baba	baby	imbuzi
Devoiced Bilabial Plosive	b̥	b_vcls			bhala
Voiced Bilabial Implosive	ɓ	b_imp			ubawo
Voiceless Alveolar Plosive	t	t	tot	total	*Atlanta*
Aspirated Alveolar Plosive	t^h	t_asp			thatha
Ejective Alveolar Plosive	t’	t_edj			itakane
Ejective Alveolar Lateral Affricate	t͜ɬ’	t_lnklb_edj			intloko
Palatal Affricate	t͜ʃ	t_lnksh	tjokker	chocolate	tshixa
Aspirated Palatal Affricate	t͜ʃ^h	t_lnksh_asp			ukutshisa
Ejective Alveolar Affricate	t͜s’	t_lnks_edj			ukutsiba
Aspirated Alveolar Affricate	t͜s^h	t_lnks_asp			isithsaba
Voiced Alveolar Plosive	d	d	daar	death	indoda
Devoiced Alveolar Plosive	d̥	d_vcls			amadoda
Voiced Palatal Affricate	d͜ʒ	d_lnkzh	*John*	jug	inja
Voiced Alveolar Affricate	d͜z	d_lnkz	*George*		amanzi
Voiced Lateral Alveolar Affricate	d͜ɮ	d_lnklz			indlovu
Ejective Palatal Plosive	c’	c_edj			ukutya
Aspirated Palatal Plosive	c^h	c_asp			ityhefu
Voiced Palatal Plosive	ɟ	jb			indyevo
Voiceless Velar Plosive	k	k	koek	kick	*Brakpan*
Aspirated Velar Plosive	k^h	k_asp			yikha
Ejective Velar Plosive	k’	k_edj			kakubi
Ejective Velar Affricate	k͜x’	k_lnkx_edj			ikrele
Voiced Velar Plosive	g	g	berge	gun	ingubo
Glottal Stop	ʔ	gl	ver_as	co_operative	i_oyile
Bilabial Nasal	m	m	ma	man	mama
Syllabic Bilabial Nasal	m̩	m_syl			umfana
Alveolar Nasal	n	n	non	not	iyana
Palatal Nasal	ɲ	nlt	mandjie		inyama
Velar Nasal	ŋ	nj	lang	thing	ingubo
Alveolar Trill	r	r	roer	*Hartenbos*	isitrato
Uvular Trill	ʀ	rc	*roer*		isitrato
Alveolar Flap	ɾ	fh		better	
Voiceless Labiodental Fricative	f	f	vier	four	ukufa
Voiced Labiodental Fricative	v	v	water	vat	ukuvula
Voiceless Dental Fricative	θ	th	*Martha*	thing	*thousand*
Voiced Dental Fricative	ð	dh	*Northam*	this	*the*
Voiceless Alveolar Fricative	s	s	sies	some	sala
Voiced Alveolar Fricative	z	z	zoem	zero	ukuzama
Voiceless Palatal Fricative	ʃ	sh	sjoe	shine	kushushu

DESCRIPTION	**IPA**	**ASTBET**	**AA**	**EE**	**XX**
Voiced Palatal Fricative	ʒ	`zh`	genre	genre	*genre*
Voiceless Velar Fricative	x	`x`	gaan	*Gauteng*	irhafu
Voiceless Glottal Fricative	h	`h`	*Bethlehem*	hand	huhuza
Voiced Glottal Fricative	ɦ	`hht`	hand	*Johannes*	ihashe
Alveolar Approximant	ɹ	`rt`	*Brixton*	red	
Voiceless Alveolar Lateral Fricative	ɬ	`lb`	*Umhlanga*		hlala
Voiced Alveolar Lateral Fricative	ɮ	`lz`			dlala
Palatal Approximant	j	`j`	jas	yes	yima
Alveolar Approximant	l	`l`	lag	legs	lala
Voiced labio-velar Approximant	w	`w`	*William*	west	wela
Dental Click	ǀ	`cl1`			cinga
Nasalised Dental Click	ǀ̃	`cl1_nas`			nceda
Voiced Nasalised Dental Click	ǀ̬̃	`cl1_nas_vcd`			iingcango
Voiced Dental Click	ǀ̬	`cl1_vcd`			gcina
Aspirated Dental Click	ǀʰ	`cl1_asp`			chaza
Alveolar Lateral Click	ǁ	`cl4`			xela
Nasalised Alveolar Lateral Click	ǁ̃	`cl4_nas`			nxila
Voiced Nasalised Alveolar Lateral Click	ǁ̬̃	`cl4_nas_vcd`			ingxelo
Voiced Alveolar Lateral Click	ǁ̬	`cl4_vcd`			gxeka
Aspirated Alveolar Lateral Click	ǁʰ	`cl4_asp`			xhasa
Palatal Click	ǃ	`cl2`			qiqa
Nasalised Palatal Click	ǃ̃	`cl2_nas`			nqaba
Voiced Nasalised Palatal Click	ǃ̬̃	`cl2_nas_vcd`			ngciba
Voiced Palatal Click	ǃ̬	`cl2_vcd`			gquba
Aspirated Palatal Click	ǃʰ	`cl2_asp`			qhuba
High Front Vowel	i	`i`	siek	*Piet*	impilo
High Front Vowel with duration	iː	`i_long`	vier	keep	impilo
Lax Front Vowel	ɪ	`ic`	*Brixton*	him	*Cecil*
Rounded High Front Vowel	y	`y`	u		*u*
Rounded High Front Vowel with duration	yː	`y_long`	vuur		
High Back Vowel	u	`u`	boek	*Kapkaroord*	vulani
High Back Vowel with duration	uː	`u_long`	boer	blue	vula
Lax Back Vowel	ʊ	`hs`	*Woodstock*	push	*Newtown*
Mid-high Front Vowel	e	`e`	eweredig	*Senekal*	ndithengile
Mid-high Front Vowel with duration	eː	`e_long`	been	*Vrede*	ugqibeleni
Rounded Mid-high Front Vowel	ø	`phi`	neus		
Rounded Mid-high Front Vowel with duration	øː	`phi_long`	deur		
Rounded Mid-high Back Vowel	o	`o`	*Sibongile*	*Sibongile*	ukubonisa
Rounded Mid-high Back Vowel with duration	oː	`o_long`	boor		koloni
Mid-low Front Vowel	ɛ	`ep`	mes	nest	Themba
Mid-low Front Vowel with duration	ɛː	`ep_long`	lê	fairy	aneesenti
Rounded Mid-low Front Vowel	œ	`oe`	brug	nurse	
Rounded Mid-low Front Vowel with duration	œː	`oe_long`	brue	burst	
Central Vowel with duration	ɜː	`epr_long`	wîe	turn	*third*
Rounded Mid-low Back Vowel	ɔ	`ct`	mos	*Hartenbos*	molo
Rounded Mid-low Back Vowel with duration	ɔː	`ct_long`	môre	bore	anethoba
Low Back Vowel	ɒ	`ab`	*McDonald's*	hot	*box*
Lax Mid-low Vowel	ʌ	`vt`	*public*	hut	*hut*

DESCRIPTION	**IPA**	**ASTBET**	**AA**	**EE**	**XX**
Low Central Vowel	a	a	agter	*Garsfontein*	sala
Low Central Vowel with duration	aː	a_long	daar	*Klerksdorp*	apha
Low Back Vowel with duration	ɑː	as_long	*master's*	harp	
Central Vowel (Schwa)	ə	sw	niks	the	*degree*
Mid-low Front Vowel	æ	ae	eg, help	average	*camp*
Mid-low Front Vowel with duration	æː	ae_long	ver	dad	*camp*
Diphthongs					
	ɔ͜i	ct_lnki	rotjie		
	ə͜i	sw_lnki	allerlei	*Kleinmond*	*Spelled "H"*
	i͜u	i_lnku	geoloog		*Beaufort*
	i͜a	i_lnka	inisiatief		*financial*
	i͜ɔ	i_lnkct	Centurion	*Georgalli*	
	i͜œ	i_lnkoe	Pretorius		
	ɪ͜ə	ic_lnksw		here	*years*
	i͜ə	i_linksw		*hear*	*nearly*
	œ͜y	oe_lnky	luiaard		*Nelspruit*
	u͜i	u_lnki	moeiliker		
	ə͜u	sw_lnku	nou	*zero*	r*oad*
	ə͜ʊ	sw_lnkhs	*phone*	hope	r*oad*
	a͜i	a_lnki	poegaai	*Laaiplek*	*drive*
	aː͜i	a_long_lnki	baaie		*Laaiplek*
	a͜ɪ	a_lnkic		*Laaiplek*	*drive*
	ʌ͜ɪ	vt_lnkic	*Alpine*	fine	
	a͜ʊ	a_lnkhs	*Camperdown*	power	*south*
	a͜u	a_lnku	*Landsdown*	power	*south*
	o͜i	o_lnki	weggooi		point
	ɔ͜ɪ	ct_lnkic		boy	*point*
	ɔ͜u	ct_lnku			*road*
	e͜u	e_lnku	eeu		
	ɛ͜i	ep_lnki			*Cape*
	ɛ͜ɪ	ep_lnkic		waste	*Cape*
	e͜ɪ	e_lnkic	*gymnasium*	play	*Cape*
	ɛ͜ə	ep_lnksw		there	*there*
	ʊ͜ə	hs_lnksw		poor	*poor*
	u͜a	u_lnka	*Francois*		
	i͜ɛ	i_lnkep	twee en		
	i͜ə	i_lnksw	drie en		
	ʌ͜ə	vt_lnksw		*Bryan*	
Tripthongs					
	ɪ͜ɔ͜ɛ	ic_lnkct_ep		*Spelled "TOE"*	
	ə͜a͜ʊ	sw_lnka_lnkhs		*thousand*	
	ə͜i͜ə	sw_lnki_lnksw	hy is		

Table 12: Phonemes present in the AA, EE and XX databases.