Collecting and evaluating speech recognition corpora for nine Southern Bantu languages

Jaco Badenhorst, Charl van Heerden, Marelie Davel and Etienne Barnard
HLT Research Group, Meraka Institute, CSIR, South Africa
jbadenhorst@csir.co.za, mdavel@csir.co.za
cvheerden@csir.co.za, ebarnard@csir.co.za

Abstract

We describe the Lwazi corpus for automatic speech recognition (ASR), a new telephone speech corpus which includes data from nine Southern Bantu languages. Because of practical constraints, the amount of speech per language is relatively small compared to major corpora in world languages, and we report on our investigation of the stability of the ASR models derived from the corpus. We also report on phoneme distance measures across languages, and describe initial phone recognisers that were developed using this data.

1 Introduction

There is a widespread belief that spoken dialog systems (SDSs) will have a significant impact in the developing countries of Africa (Tucker and Shalonova, 2004), where the availability of alternative information sources is often low. Traditional computer infrastructure is scarce in Africa, but telephone networks (especially cellular networks) are spreading rapidly. In addition, speech-based access to information may empower illiterate or semi-literate people, 98% of whom live in the developing world.

Spoken dialog systems can play a useful role in a wide range of applications. Of particular importance in Africa are applications such as education, using speech-enabled learning software or kiosks and information dissemination through media such as telephone-based information systems. Significant benefits can be envisioned if information is provided in domains such as agriculture (Nasfors, 2007), health care (Sherwani et al., ; Sharma et al., 2009) and government services (Barnard et al., 2003). In order to make SDSs a reality in Africa, technology components such as text-to-speech (TTS) systems and automatic speech recognition (ASR) systems are required. The latter category of technologies is the focus of the current contribution.

Speech recognition systems exist for only a handful of African languages (Roux et al., ; Seid and Gambck, 2005; Abdillahi et al., 2006), and to our knowledge no service available to the general public currently uses ASR in an indigenous African language. A significant reason for this state of affairs is the lack of sufficient linguistic resources in the African languages. Most importantly, modern speech recognition systems use statistical models which are trained on corpora of relevant speech (i.e. appropriate for the recognition task in terms of the language used, the profile of the speakers, speaking style, etc.) This speech generally needs to be curated and transcribed prior to the development of ASR systems, and for most applications speech from a large number of speakers is required in order to achieve acceptable system performance. On the African continent, where infrastructure such as computer networks is less developed than in countries such as America, Japan and the European countries, the development of such speech corpora is a significant hurdle to the development of ASR systems.

The complexity of speech corpus development is strongly correlated with the amount of data that is required, since the number of speakers that need to be canvassed and the amount of speech that must be curated and transcribed are major factors in determining the feasibility of such development. In order to minimise this complexity, it is important to have tools and guidelines that can be used to assist in designing the smallest corpora that will be sufficient for typical applications of ASR systems. As minimal corpora can be extended by sharing data across languages, tools are also required to indicate when data sharing will be beneficial and when detrimental.

Proceedings of the EACL 2009 Workshop on Language Technologies for African Languages – AfLaT 2009, pages 1–8,
Athens, Greece, 31 March 2009.

In this paper we describe and evaluate a new speech corpus of South African languages currently under development (the Lwazi corpus) and evaluate the extent in which computational analysis tools can provide further guidelines for ASR corpus design in resource-scarce languages.

2 Project Lwazi

The goal of Project Lwazi is to provide South African citizens with information and information services in their home language, over the telephone, in an efficient and affordable manner. Commissioned by the South African Department of Arts and Culture, the activities of this three year project (2006-2009) include the development of core language technology resources and components for all the official languages of South Africa, where, for the majority of these, no prior language technology components were available.

The core linguistic resources being developed include phoneme sets, electronic pronunciation dictionaries and the speech and text corpora required to develop automated speech recognition (ASR) and text-to-speech (TTS) systems for all eleven official languages of South Africa. The usability of these resources will be demonstrated during a national pilot planned for the third quarter of 2009. All outputs from the project are being released as open source software and open content(Meraka-Institute, 2009).

Resources are being developed for all nine Southern Bantu languages that are recognised as official languages in South Africa (SA). These languages are: (1) isiZulu (zul[1]) and isiXhosa (xho), the two Nguni languages most widely spoken in SA. Together these form the home language of 41% of the SA population. (2) The three Sotho languages: Sepedi (nso), Setswana (tsn), Sesotho (sot), together the home language of 26% of the SA population. (3) The two Nguni languages less widely spoken in SA: siSwati (ssw) and isiNdebele (nbl), together the home language of 4% of the SA population. (4) Xitsonga (tso) and Tshivenda (ven), the home languages of 4% and 2% of the SA population, respectively (Lehohla, 2003). (The other two official languages of South Africa are Germanic languages, namely English (eng) and Afrikaans (afr).)

For all these languages, new pronunciation dictionaries, text and speech corpora are being developed. ASR speech corpora consist of approximately 200 speakers per language, producing read and elicited speech, recorded over a telephone channel. Each speaker produced approximately 30 utterances, 16 of these were randomly selected from a phonetically balanced corpus and the remainder consist of short words and phrases: answers to open questions, answers to yes/no questions, spelt words, dates and numbers. The speaker population was selected to provide a balanced profile with regard to age, gender and type of telephone (cellphone or landline).

[1] After each language name, the ISO 639-3:2007 language code is provided in brackets.

3 Related work

Below, we review earlier work relevant to the development of speech recognisers for languages with limited resources. This includes both ASR system design (Sec. 3.1) and ASR corpus design (Sec. 3.2). In Sec. 3.3, we also review the analytical tools that we utilise in order to investigate corpus design systematically.

3.1 ASR for resource-scarce languages

The main linguistic resources required when developing ASR systems for telephone based systems are electronic pronunciation dictionaries, annotated audio corpora (used to construct acoustic models) and recognition grammars. An ASR audio corpus consists of recordings from multiple speakers, with each utterance carefully transcribed orthographically and markers used to indicate non-speech and other events important from an ASR perspective. Both the collection of appropriate speech from multiple speakers and the accurate annotation of this speech are resource-intensive processes, and therefore corpora for resource-scarce languages tend to be very small (1 to 10 hours of audio) when compared to the speech corpora used to build commercial systems for world languages (hundreds to thousands of hours per language).

Different approaches have been used to best utilise limited audio resources when developing ASR systems. Bootstrapping has been shown to be a very efficient technique for the rapid development of pronunciation dictionaries, even when utilising linguistic assistants with limited phonetic training (Davel and Barnard, 2004).

Small audio corpora can be used efficiently by utilising techniques that share data across lan-

guages, either by developing multilingual ASR systems (a single system that simultaneously recognises different languages), or by using additional source data to supplement the training data that exists in the target language. Various data sharing techniques for language-dependant acoustic modelling have been studied, including cross-language transfer, data pooling, language adaptation and bootstrapping (Wheatley et al., 1994; Schultz and Waibel, 2001; Byrne et al., 2000). Both (Wheatley et al., 1994) and (Schultz and Waibel, 2001) found that useful gains could be obtained by sharing data across languages with the size of the benefit dependent on the similarity of the sound systems of the languages combined. In the only cross-lingual adaptation study using African languages (Niesler, 2007), similar gains have not yet been observed.

3.2 ASR corpus design

Corpus design techniques for ASR are generally aimed at specifying or selecting the most appropriate subset of data from a larger domain in order to optimise recognition accuracy, often while explicitly minimising the size of the selected corpus. This is achieved through various techniques that aim to include as much variability in the data as possible, while simultaneously ensuring that the corpus matches the intended operating environment as accurately as possible.

Three directions are primarily employed: (1) explicit specification of phonotactic, speaker and channel variability during corpus development, (2) automated selection of informative subsets of data from larger corpora, with the smaller subset yielding comparable results, and (3) the use of active learning to optimise existing speech recognition systems. All three techniques provide a perspective on the sources of variation inherent in a speech corpus, and the effect of this variation on speech recognition accuracy.

In (Nagroski et al., 2003), Principle Component Analysis (PCA) is used to cluster data acoustically. These clusters then serve as a starting point for selecting the optimal utterances from a training database. As a consequence of the clustering technique, it is possible to characterise some of the acoustic properties of the data being analysed, and to obtain an understanding of the major sources of variation, such as different speakers and genders (Riccardi and Hakkani-Tur, 2003).

Active and unsupervised learning methods can be combined to circumvent the need for transcribing massive amounts of data (Riccardi and Hakkani-Tur, 2003). The most informative untranscribed data is selected for a human to label, based on acoustic evidence of a partially and iteratively trained ASR system. From such work, it soon becomes evident that the optimisation of the amount of variation inherent to training data is needed, since randomly selected additional data does not necessarily improve recognition accuracy. By focusing on the selection (based on existing transcriptions) of a uniform distribution across different speech units such as words and phonemes, improvements are obtained (Wu et al., 2007).

In our focus on resource-scarce languages, the main aim is to understand the amount of data that needs to be collected in order to achieve acceptable accuracy. This is achieved through the use of analytic measures of data variability, which we describe next.

3.3 Evaluating phoneme stability

In (Badenhorst and Davel, 2008) a technique is developed that estimates how stable a specific phoneme is, given a specific set of training data. This statistical measure provides an indication of the effect that additional training data will have on recognition accuracy: the higher the stability, the less the benefit of additional speech data.

The model stability measure utilises the Bhattacharyya bound (Fukunaga, 1990), a widely-used upper bound of the Bayes error. If P_i and $p_i(X)$ denote the prior probability and class-conditional density function for class i, respectively, the Bhattacharyya bound ϵ is calculated as:

$$\epsilon = \sqrt{P_1 P_2} \int \sqrt{p_1(X) p_2(X)} dX \quad (1)$$

When both density functions are Gaussian with mean μ_i and covariance matrix Σ_i, integration of ϵ leads to a closed-form expression for ϵ:

$$\epsilon = \sqrt{P_1 P_2} e^{-\mu(1/2)} \quad (2)$$

where

$$\mu(1/2) = \frac{1}{8}(\mu_2 - \mu_1)^T \left[\frac{\Sigma_1 + \Sigma_2}{2}\right]^{-1} (\mu_2 - \mu_1) + \frac{1}{2} ln \frac{\left|\frac{\Sigma_1 + \Sigma_2}{2}\right|}{\sqrt{|\Sigma_1||\Sigma_2|}} \quad (3)$$

is referred to as the Bhattacharyya distance.

In order to estimate the stability of an acoustic model, the training data for that model is separated into a number of disjoint subsets. All subsets are selected to be mutually exclusive with respect to the speakers they contain. For each subset, a separate acoustic model is trained, and the Bhattacharyya bound between each pair of models calculated. By calculating both the mean of this bound and the standard deviation of this measure across the various model pairs, a statistically sound measure of model estimation stability is obtained.

4 Computational analysis of the Lwazi corpus

We now report on our analysis of the Lwazi speech corpus, using the stability measure described above. Here, we focus on four languages (isiNdebele, siSwati, isiZulu and Tshivenda) for reasons of space; later, we shall see that the other languages behave quite similarly.

4.1 Experimental design

For each phoneme in each of our target languages, we extract all the phoneme occurrences from the 150 speakers with the most utterances per phoneme. We utilise the technique described in Sec. 3.3 to estimate the Bhattacharyya bound both when evaluating phoneme variability and model distance. In both cases we separate the data for each phoneme into 5 disjoint subsets. We calculate the mean of the 10 distances obtained between the various intra-phoneme model pairs when measuring phoneme stability, and the mean of the 25 distances obtained between the various inter-phoneme model pairs when measuring phoneme distance.

In order to be able to control the number of phoneme observations used to train our acoustic models, we first train a speech recognition system and then use forced alignment to label all of the utterances using the systems described in Sec. 5. Mel-frequency cepstral coefficients (MFCCs) with cepstral mean and variance normalisation are used as features, as described in Sec. 5.

4.2 Analysis of phoneme variability

In an earlier analysis of phoneme variability of an English corpus (Badenhorst and Davel, 2008), it was observed that similar trends are observed when utilising different numbers of mixtures in a Gaussian mixture model. For both context dependent and context independent models similar trends are also observed. (Asymptotes occur later, but trends remain similar.) Because of the limited size of the Lwazi corpus, we therefore only report on single-mixture context-independent models in the current section.

As we also observe similar trends for phonemes within the same broad categories, we report on one or two examples from several broad categories which occur in most of our target languages. Using SAMPA notation, the following phonemes are selected: /a/ (vowels), /m/ (nasals), /b/ and /g/ (voiced plosives) and /s/ (unvoiced fricatives), after verifying that these phonemes are indeed representative of the larger groups.

Figures 1 and 2 demonstrate the effects of variable numbers of phonemes and speakers, respectively, on the value of the mean Bhattacharyya bound. This value should approach 0.5 for a model fully trained on a sufficiently representative set of data. In Fig. 1 we see that the various broad categories of sounds approach the asymptotic bound in different ways. The vowels and nasals require the largest number of phoneme occurrences to reach a given level, whereas the fricatives and plosives converge quite rapidly (With 10 observations per speaker, both the fricatives and plosives achieve values of 0.48 or better for all languages, in contrast to the vowels and nasals which require 30 observations to reach similar stability). Note that we employed 30 speakers per phoneme group, since that is the largest number achievable with our protocol.

For the results in Fig. 2, we keep the number of phoneme occurrences per speaker fixed at 20 (this ensures that we have sufficient data for all phonemes, and corresponds with reasonable convergence in Fig. 1). It is clear that additional speakers would still improve the modelling accuracy for especially the vowels and nasals. We observe that the voiced plosives and fricatives quickly achieve high values for the bound (close to the ideal 0.5).

Figures 1 and 2 – as well as similar figures for the other phoneme classes and languages we have studied – suggest that all phoneme categories require at least 20 training speakers to achieve reasonable levels of convergence (bound levels of 0.48 or better). The number of phoneme observations required per speaker is more variable, rang-

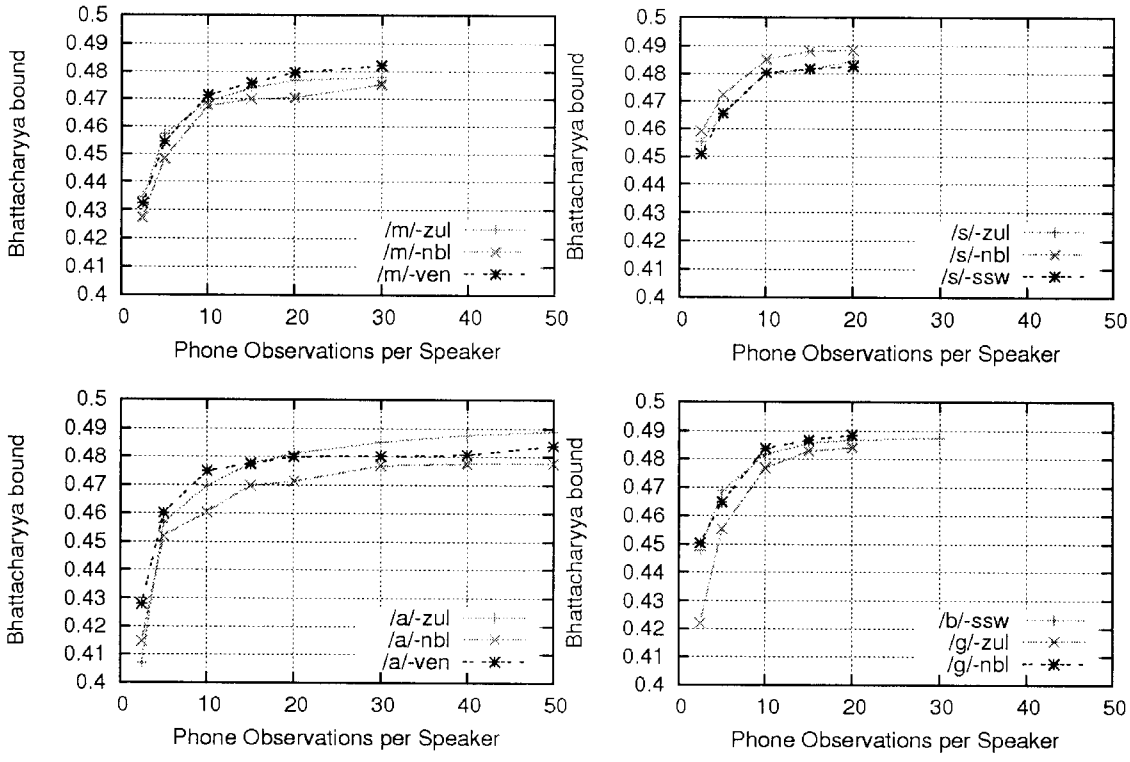

Figure 1: *Effect of number of phoneme utterances per speaker on mean of Bhattacharyya bound for different phoneme groups using data from 30 speakers*

ing from less than 10 for the voiceless fricatives to 30 or more for vowels, liquids and nasals. We return to these observations below.

4.3 Distances between languages

In Sec. 3.1 it was pointed out that the similarities between the same phonemes in different languages are important predictors of the benefit achievable from pooling the data from those languages. Armed with the knowledge that stable models can be estimated with 30 speakers per phoneme and between 10 and 30 phonemes occurrences per speaker, we now turn to the task of measuring distances between phonemes in various languages.

We again use the mean Bhattacharyya bound to compare phonemes, and obtain values between all possible combinations of phonemes. Results are shown for the isiNdebele phonemes /n/ and /a/ in Fig. 3. As expected, similar phonemes from the different languages are closer to one another than different phonemes of the same language. However, the details of the distances are quite revealing: for /a/, siSwati is closest to the isiNdebele model, as would be expected given their close linguistic relationship, but for /n/, the Tshivenda model is found to be closer than either of the other Nguni languages. For comparative purposes, we have included one non-Bantu language (Afrikaans), and we see that its models are indeed significantly more dissimilar from the isiNdebele model than any of the Bantu languages. In fact, the Afrikaans /n/ is about as distant from isiNdebele /n/ as isiNdebele and isiZulu /l/ are!

5 Initial ASR results

In order to verify the usability of the Lwazi corpus for speech recognition, we develop initial ASR systems for all 11 official South African languages. A summary of the data statistics for the Bantu languages investigated is shown in Tab. 1, and recognition accuracies achieved are summarised in Tab. 2. For these tests, data from 30 speakers per language were used as test data, with the remaining data being used for training.

Although the Southern Bantu languages are tone languages, our systems do not encode tonal

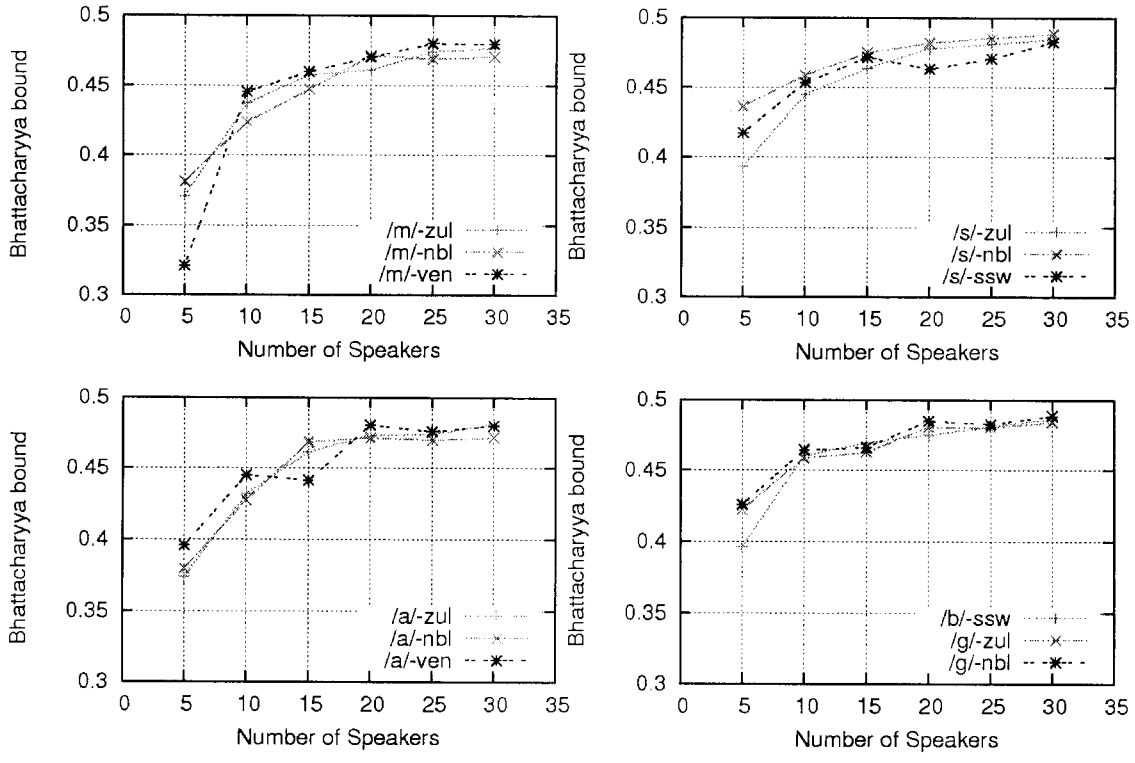

Figure 2: *Effect of number of speakers on mean of Bhattacharyya bound for different phoneme groups using 20 utterances per speaker*

Language	total # minutes	# speech minutes	# distinct phonemes
isiNdebele	564	465	46
isiXhosa	470	370	52
isiZulu	525	407	46
Tshivenda	354	286	38
Sepedi	394	301	45
Sesotho	387	313	44
Setswana	379	295	34
siSwati	603	479	39
Xitsonga	378	316	54
N-TIMIT	315	-	39

Table 1: *A summary of the Lwazi ASR corpus: Bantu languages.*

information, since tone is unlikely to be important for small-to-medium vocabulary applications (Zerbian and Barnard, 2008).

As the initial pronunciation dictionaries were developed to provide good coverage of the language in general, these dictionaries did not cover the entire ASR corpus. Grapheme-to-phoneme rules are therefore extracted from the general dictionaries using the Default&Refine algorithm (Davel and Barnard, 2008) and used to generate missing pronunciations.

We use HTK 3.4 to build a context-dependent cross-word HMM-based phoneme recogniser with triphone models. Each model had 3 emitting states with 7 mixtures per state. 39 features are used: 13 MFCCs together with their first and second order derivatives. Cepstral Mean Normalisation (CMN) as well as Cepstral Variance Normalisation (CMV) are used to perform speaker-independent normalisation. A diagonal covariance matrix is used; to partially compensate for this incorrect assumption of feature independence semi-tied transforms are applied. A flat phone-based language model is employed throughout.

As a rough benchmark of acceptable phoneme-recognition accuracy, recently reported results obtained by (Morales et al., 2008) on a similar-sized telephone corpus in American English (N-TIMIT) are also shown in Tab. 2. We see that the Lwazi results compare very well with this benchmark.

An important issue in ASR corpus design is

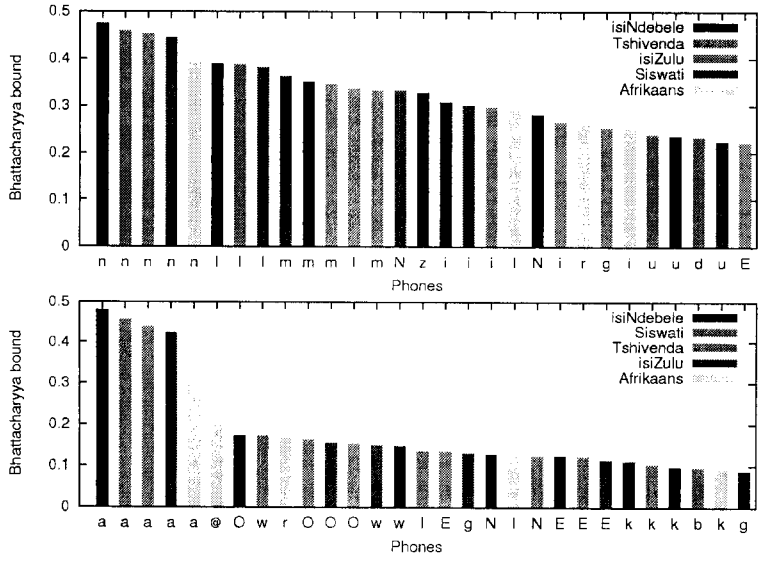

Figure 3: *Effective distances in terms of the mean of the Bhattacharyya bound between a single phoneme (/n/-nbl top and /a/-nbl bottom) and each of its closest matches within the set of phonemes investigated.*

Language	% corr	% acc	avg # phons	total # speakers
isiNdebele	74.21	65.41	28.66	200
isiXhosa	69.25	57.24	17.79	210
isiZulu	71.18	60.95	23.42	201
Tshivenda	76.37	66.78	19.53	201
Sepedi	66.44	55.19	16.45	199
Sesotho	68.17	54.79	18.57	200
Setswana	69.00	56.19	20.85	207
siSwati	74.19	64.46	30.66	208
Xitsonga	70.32	59.41	14.35	199
N-TIMIT	64.07	55.73	-	-

Table 2: *Initial results for South African ASR systems. The column labelled "avg # phonemes" lists the average number of phoneme occurrences for each phoneme for each speaker.*

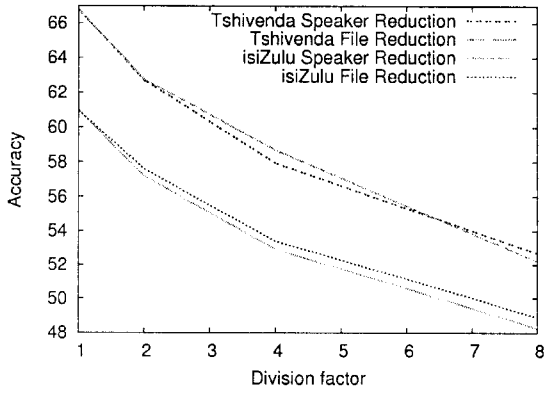

Figure 4: *The influence of a reduction in training corpus size on phone recognition accuracy.*

the trade-off between the number of speakers and the amount of data per speaker (Wheatley et al., 1994). The figures in Sec. 4.2 are not conclusive on this trade-off, so we have also investigated the effect of reducing either the number of speakers or the amount of data per speaker when training the isiZulu and Tshivenda recognisers. As shown in Fig. 4, the impact of both forms of reduction is comparable across languages and different degrees of reduction, in agreement with the results of Sec. 4.2.

These results indicate that we now have a firm baseline to investigate data-efficient training methods such as those described in Sec. 3.1.

6 Conclusion

In this paper we have introduced a new telephone speech corpus which contains data from nine Southern Bantu languages. Our stability analysis shows that the speaker variety as well as the amount of speech per speaker is sufficient to achieve acceptable model stability, and this conclusion is confirmed by the successful training of phone recognisers in all the languages. We confirm the observation in (Badenhorst and Davel, 2008) that different phone classes have different

data requirements, but even for the more demanding classes (vowels, nasals, liquids) our amount of data seems sufficient. Our results suggest that similar accuracies may be achievable by using more speech from fewer speakers – a finding that may be useful for the further development of speech corpora in resource-scarce languages.

Based on the proven stability of our models, we have performed some preliminary measurements of the distances between the phones in the different languages; such distance measurements are likely to be important for the sharing of data across languages in order to further improve ASR accuracy. The development of real-world applications using this data is currently an active topic of research; for that purpose, we are continuing to investigate additional methods to improve recognition accuracy with such relatively small corpora, including cross-language data sharing and efficient adaptation methods.

References

Nimaan Abdillahi, Pascal Nocera, and Jean-Franois Bonastre. 2006. Automatic transcription of Somali language. In *Interspeech*, pages 289–292, Pittsburgh, PA.

J.A.C. Badenhorst and M.H. Davel. 2008. Data requirements for speaker independent acoustic models. In *PRASA*, pages 147–152.

E. Barnard, L. Cloete, and H. Patel. 2003. Language and technology literacy barriers to accessing government services. *Lecture Notes in Computer Science*, 2739:37–42.

W. Byrne, P. Beyerlein, J. M. Huerta, S. Khudanpur, B. Marthi, J. Morgan, N. Peterek, J. Picone, D. Vergyri1, and W. Wang. 2000. Towards language independent acoustic modeling. In *ICASSP*, volume 2, pages 1029–1032, Istanbul, Turkey.

M. Davel and E. Barnard. 2004. The efficient creation of pronunication dictionaries: human factors in bootstrapping. In *Interspeech*, pages 2797–2800, Jeju, Korea, Oct.

M. Davel and E. Barnard. 2008. Pronunciation predication with Default&Refine. *Computer Speech and Language*, 22:374–393, Oct.

K. Fukunaga. 1990. *Introduction to Statistical Pattern Recognition*. Academic Press, Inc., 2nd edition.

Pali Lehohla. 2003. *Census 2001: Census in brief*. Statistics South Africa.

Meraka-Institute. 2009. Lwazi ASR corpus. Online: http://www.meraka.org.za/lwazi.

N. Morales, J. Tejedor, J. Garrido, J. Colas, and D.T. Toledano. 2008. STC-TIMIT: Generation of a single-channel telephone corpus. In *LREC*, pages 391–395, Marrakech, Morocco.

A. Nagroski, L. Boves, and H. Steeneken. 2003. In search of optimal data selection for training of automatic speech recognition systems. *ASRU workshop*, pages 67–72, Nov.

P. Nasfors. 2007. Efficient voice information services for developing countries. Master's thesis, Department of Information Technology, Uppsala University.

T. Niesler. 2007. Language-dependent state clustering for multilingual acoustic modeling. *Speech Communication*, 49:453–463.

G. Riccardi and D. Hakkani-Tur. 2003. Active and unsupervised learning for automatic speech recognition. In *Eurospeech*, pages 1825–1828, Geneva, Switzerland.

J.C. Roux, E.C. Botha, and J.A. du Preez. Developing a multilingual telephone based information system in african languages. In *LREC*, pages 975–980, Athens, Greece.

T. Schultz and A. Waibel. 2001. Language-independent and language-adaptive acoustic modeling for speech recognition. *Speech Communication*, 35:31–51, Aug.

Hussien Seid and Bjrn Gambck. 2005. A speaker independent continuous speech recognizer for Amharic. In *Interspeech*, pages 3349–3352, Lisboa, Portugal, Oct.

A. Sharma, M. Plauche, C. Kuun, and E. Barnard. 2009. HIV health information access using spoken dialogue systems: Touchtone vs. speech. Accepted at IEEE Int. Conf. on ICTD.

J. Sherwani, N. Ali, S. Mirza, A. Fatma, Y. Memon, M. Karim, R. Tongia, and R. Rosenfeld. Healthline: Speech-based access to health information by low-literate users. In *IEEE Int. Conf. on ICTD*, pages 131–139.

R. Tucker and K. Shalonova. 2004. The Local Language Speech Technology Initiative. In *SCALLA Conf.*, Nepal.

B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusumy. 1994. An evaluation of cross-language adaptation for rapid HMM development in a new language. In *ICASSP*, pages 237–240, Adelaide.

Y. Wu, R. Zhang, and A. Rudnicky. 2007. Data selection for speech recognition. *ASRU workshop*, pages 562–565, Dec.

S. Zerbian and E. Barnard. 2008. Phonetics of intonation in South African Bantu languages. *Southern African Linguistics and Applied Language Studies*, 26(2):235–254.