Noname manuscript No.
(will be inserted by the editor)

Collecting and evaluating speech recognition corpora for 11 South African languages

Jaco Badenhorst · Charl van Heerden · Marelie Davel · Etienne Barnard

Received: date / Accepted: date

Abstract We describe the Lwazi corpus for automatic speech recognition (ASR), a new telephone speech corpus which contains data from the eleven official languages of South Africa. 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. We find that a surprisingly small number of speakers (fewer than 50) and around 10 to 20 hours of speech per language are sufficient for the purposes of acceptable phone-based recognition.

Keywords speech recognition · Lwazi corpus · resource-scarce languages · South African languages

1 Introduction

There is a widespread belief that spoken dialog systems (SDSs) will have a significant impact in the developing countries of Africa [Tucker & Shalonova, 2004], where the avail-

Jaco Badenhorst
Human Language Technology Competency Area, CSIR Meraka Institute, Meiring Naude Road, Pretoria, South Africa,
Tel.: +27-12-8413028, Fax: +27-12-8414720,
E-mail: jbadenhorst@csir.co.za

Charl van Heerden
Human Language Technology Competency Area, CSIR Meraka Institute,
E-mail: cvheerden@csir.co.za

Marelie Davel
Human Language Technology Competency Area, CSIR Meraka Institute,
E-mail: mdavel@csir.co.za

Etienne Barnard
Multilingual Speech Technologies, North-West University, Vanderbijlpark 1900, South Africa
E-mail: etienne.barnard@nwu.ac.za

ability 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.

SDSs 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., 2007; Sharma et al., 2009] and government services [Barnard et al., 2003]. Recent years have seen extensive research on the application of speech technology in the developing world - for a recent review, see [Patel et al., 2010]. 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., 2000; Seid & Gambäck, 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 USA, 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.

In this paper we describe and evaluate a new speech corpus of South African languages recently developed (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 (that is, the language that the speaker identifies with most strongly), over the telephone, in an efficient and affordable manner. Commissioned by the South African Department of Arts and Culture, the activities of the first stage of this project (2006-2009) included 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 that were developed include phoneme sets, electronic pronunciation dictionaries and the speech and text corpora required to develop ASR and TTS systems for all eleven official languages of South Africa. The usability of these resources were demonstrated during a national pilot in 2009. All outputs from the project have since been released as open source software and open content [Meraka-Institute, 2009].

Resources were developed for all eleven languages that are recognised as official languages in South Africa (SA) and contribute to the available HLT components [Grover et al., this volume]. These languages are:

1. isiZulu (ISO 639-3: zul) and isiXhosa (ISO 639-3: 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 (ISO 639-3: nso), Setswana (ISO 639-3: tsn), Sesotho (ISO 639-3: sot), together the home language of 26% of the SA population.
3. Afrikaans (ISO 639-3: afr), a Germanic language, which is the home language of approximately 13% of the SA population.
4. South English (ISO 639-3: eng), the home language of only 8% of the population, but widely spoken as an additional language.
5. The two Nguni languages less widely spoken in SA: Siswati (ISO 639-3: ssw) and isiNdebele (ISO 639-3: nbl), together the home language of 4% of the SA population.
6. Xitsonga (ISO 639-3: tso) and Tshivenda (ISO 639-3: ven), the home languages of 4% and 2% of the SA population, respectively [Lehohla, 2003].

For all these languages, new pronunciation dictionaries, text and speech corpora were 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). Table 1 provides a summary of the amount of speech for the different languages.

Table 1 *The official languages of South Africa, their ISO 639-3:2007 language codes, and the amount of speech contained in the Lwazi corpus [van Heerden et al., 2009].*

Language	code	# total minutes	# speech minutes	# distinct phonemes
isiZulu	zul	525	407	46
isiXhosa	xho	470	370	52
Afrikaans	afr	213	182	37
Sepedi	nso	394	301	45
Setswana	tsn	379	295	34
Sesotho	sot	387	313	44
SA English	eng	304	255	44
Xitsonga	tso	378	316	54
siSwati	ssw	603	479	39
Tshivenda	ven	354	286	38
isiNdebele	nbl	564	465	46

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 (Section 3.1) and ASR corpus design (Section 3.2). In Section 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 & Barnard, 2004; Kominek & Black, 2006; Maskey et al., 2004].

Small audio corpora can be used efficiently by utilising techniques that share data across languages, 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 & Waibel, 2001; Byrne et al., 2000]. Both [Wheatley et al., 1994] and [Schultz & 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.

Nagroski et al. [2003] use Principle Component Analysis (PCA) 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. Interestingly, the effect of utterance length has also been analysed as a significant source of variation [Riccardi & Hakkani-Tur, 2003].

Active and unsupervised learning methods can be combined to circumvent the need for transcribing massive amounts of data [Riccardi & 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 & Davel, 2008; Badenhorst, 2009] a technique is developed that estimates how stable a specific phoneme model 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. The Bayes error provides an indication of the separability between two probability distributions. If a probability distribution is calclulated for two phonemes, say /a/ and /e/, then the ease with which a new audio sample can be classified as being an /a/ or an /e/ depends on how separable ('different') the two distributions are. The more similar the distributions, the more miss-classifications are expected, and the higher the minimum expected miss-classification rate or Bayes error. When two distributions are identical, it becomes impossible to determine to which of the two classes a new sample should belong, apart from guessing. The expected miss-classification rate then becomes 0.5 (50%).

By determining how close to identical the probability distributions are of the same phoneme calculated using different sections of the training corpus, it is possible to determine whether the developed models are stable. If the probability distribution of phoneme /a/ trained on one section of the corpus is quite different to the probability distribution of the same phoneme /a/ trained on another section of the corpus, then the training subset is still too small to produce stable acoustic models. When these probability distributions are very similar (and the Bayes error approaches 0.5) then the training data were sufficient, and the acoustic models are stable: adding additional data will not influence the estimated probability density significantly.

Since the Bayes error itself cannot always be calculated analytically (depending on the complexity of the probability distributions being compared) an upper bound provides a 'close-enough' estimate of the value itself. The Bhat bound is such an estimate, and provides the assurance that the true error will never be larger than the bound calculated.

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:

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

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:

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

where

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

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 in section 3.3. 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 Section 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 Section 5.1. Mel-frequency cepstral coefficients (MFCCs) with cepstral mean and variance normalisation are used as features, as described in Section 5.1.

4.2 Analysis of phoneme variability

In an earlier analysis of phoneme variability of an English corpus [Badenhorst & Davel, 2008], it was observed that similar trends are observed when utilising different numbers of mixtures in a Gaussian mixture model. (That is, a model with a limited number of mixtures is a good predictor of the behaviour of a more complex model.) Similarly, it was found that context dependent and context independent models also produced comparable behaviour. (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 and the phone model topology can thus be viewed as single-state HMMs.

As we also observe similar trends for phonemes within the same broad categories, we report on a couple of examples from several broad categories which occur in most of our target languages. Using X-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.

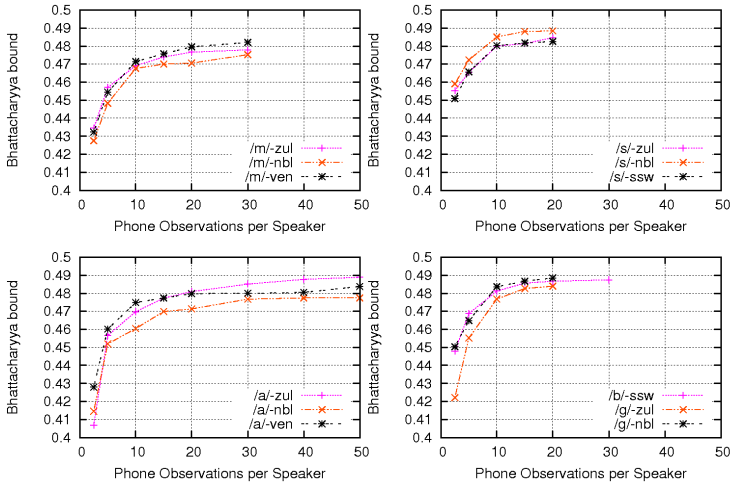


Fig. 1 *Effect of number of phoneme utterances per speaker on mean of Bhattacharyya bound for different phoneme groups using data from 30 speakers*

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, since a value of 0.5 corresponds to indistinguishable distributions (with 50% two-class error rates). In Figure 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.

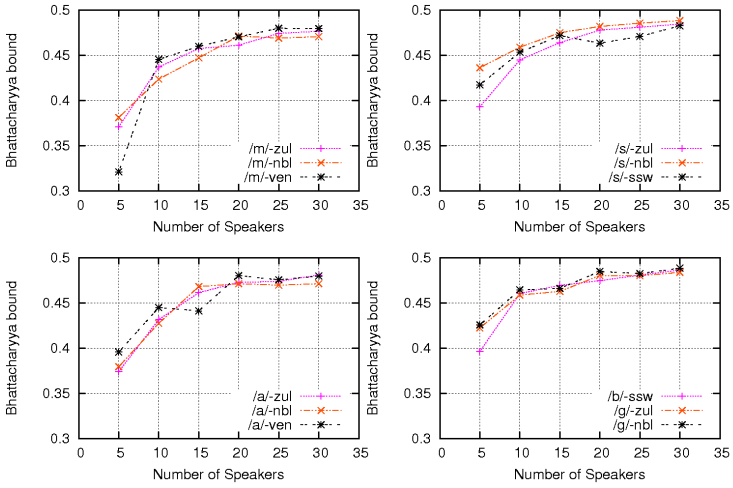


Fig. 2 *Effect of number of speakers on mean of Bhattacharyya bound for different phoneme groups using 20 utterances per speaker*

For the results in Figure 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 Figure 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, ranging from less than 10 for the voiceless fricatives to 30 or more for vowels, liquids and nasals. We return to these observations in section 5.

For large-vocabulary systems, requiring context-dependent modeling, the picture is unfortunately much more complicated. In that case, one also has to consider the impact of state tieing, and the trade-off between the number of clusters and the amount of training data per cluster becomes an important issue.

4.3 Distances between languages

In Section 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

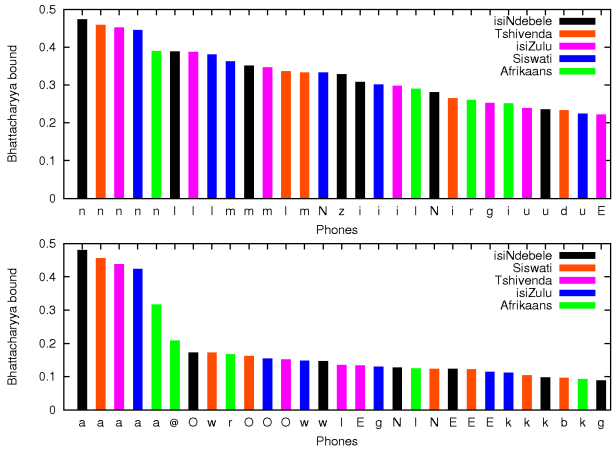


Fig. 3 *Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/n/-nbl top and /a/-nbl bottom) and each of its closest matches within the set of phonemes investigated*

languages. Armed with the knowledge that stable models can be estimated with 30 speakers per phoneme and between 10 and 30 phoneme 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 Figure 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 /l/ and isiZulu /l/ are.

Analysis of another isiNdebele vowel /i/ and nasal /m/ are shown in Figure 4. Interestingly all distances are found to be diminished and we conclude that the models for /i/ are indeed very similar across the investigated language borders. For /m/ all the Bantu languages are also even more closely related than for the model /n/.

To complete the picture, an isiNdebele fricative /s/ and plosive /g/ are also investigated. It can be seen in Figure 5 that immediately all of the closest matches are non-vowel sounds. We also find that models are no longer that dissimilar between the different plosive phonemes. For the isiNdebele /g/ it can be seen that /k/ and the Afrikaans /d/ models are actually closer than models for isiZulu /g/ or the Siswati /b/.

The similarity of the fricative /s/ also proves interesting across language borders. Although the Siswati model is found to be closest, the model for Afrikaans is very similar. All of the fricative models are, however, distinct from the other investigated phones of the languages.

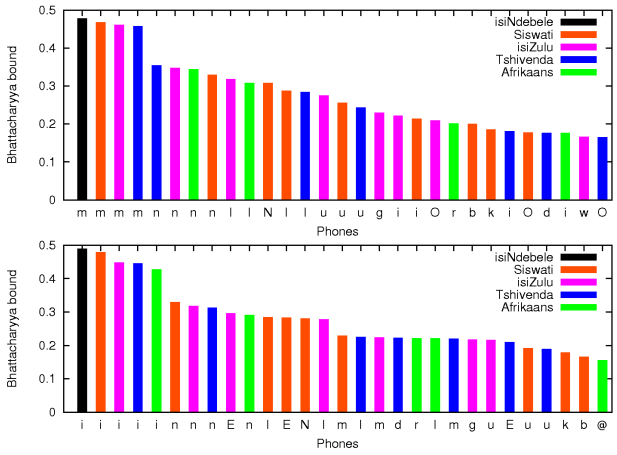


Fig. 4 *Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/m/-nbl top and /i/-nbl bottom) and each of its closest matches within the set of phonemes investigated*

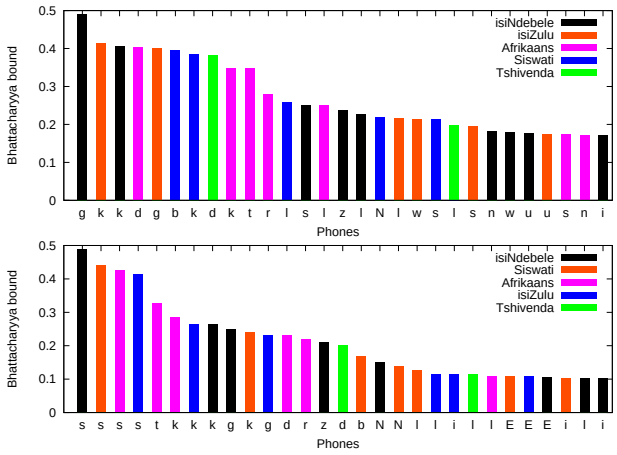


Fig. 5 *Effective distances in terms of the mean of the Bhattacharyya bound between a single phone (/g/-nbl top and /s/-nbl bottom) and each of its closest matches within the set of phonemes investigated*

5 Recognition results

In this section, we aim to confirm the measurements reported in Section 4 with several ASR measures. Baseline accuracies for both phone recognition and small vocabulary word recognition are established in Sections 5.1 and 5.2 respectively. We then measure ASR accuracy with varying amounts of data, based both on the number of speakers and the number of phones in Section 5.3. In Section 5.4, we use these results together with a heuristic relationship [Schuurmans, 1997] to get a rough estimate of the amount of data required to achieve a particular phone recognition accuracy.

5.1 Phone recognition with the Lwazi corpus

The recognisers we employ are standard HMM-based systems. We use HTK 3.4 to build a context-dependent cross-word HMM-based phone recogniser with triphone models. Each model has 3 emitting states with 7 mixtures per state. (These parameter choices were determined to be optimal for phone-recognition accuracy with the complete corpora during pilot experiments.) 39 features are used: 13 MFCCs together with their first and second order derivatives. Cepstral Mean Normalisation (CMN) as well as Cepstral Variance Normalisation (CVN) are used to perform speaker-specific normalisation. A diagonal covariance matrix is used; to partially compensate for the implicit assumption of feature independence, semi-tied transforms are applied. A flat phone-based language model is employed for phone recognition.

The optimal values of parameters such as the number of mixtures and the insertion penalty (during language modelling) will in general depend on the amount of training data available. Since our values are optimised for the full corpus, our reported accuracies for reduced corpora are underestimates. Although we have not exhaustively evaluated all parameter options, we have verified that the dependencies are quite weak, and that the overall trends reported below are also observed when the parameters are adjusted.

As the initial pronunciation dictionaries were developed to provide good coverage of each 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 & Barnard, 2008] and used to generate missing pronunciations. For the reason cited above, tone is not modelled in the system.

For phone recognition, we divided the data into a test set, which consists of 30 randomly selected speakers in each language, and a training set (the remaining speakers, approximately 170 per language). The recogniser for each language was built using all the training data for that language, using the recognition architecture as described above. These recognisers were then evaluated by performing a Viterbi search (an efficient search technique [Viterbi, 1967]) with a language model that allows unrestricted transitions between any pair of phonemes. Dynamic programming was used to match the resulting phoneme strings against the strings that result from automatic phonemic transcription of the orthographic transcriptions of the test utterances. The resulting accuracies are summarised in Table 2. Phone-recognition correctness refers to the percentage of correctly recognised phone labels with regard to the total number of expected phone labels, while the dynamic programming used to calculate the accuracy values take into account phone label insertions/deletions as well. The table also lists the phonotactic perplexity of each language – that is, the perplexity that is measured if a bigram model is used to model the phoneme sequences that occur in the training set. Lastly, the table contains word-recognition results, which are discussed in Section 5.2 below.

Interestingly, the correctness and accuracy of all other languages are higher than that of English, despite the fact that most languages have more phonemes than English. One possible explanation for this observation is the fact that English has fewer phonotactic constraints than any of the other languages, as can be deduced from the perplexity values in Table 2. (The Southern Bantu languages employ CV (consonant-vowel) or V syllable structures predominantly.) Overall, however, phonotactic perplexity does not correlate well with correctness or accuracy in our results, so other explanations for the relative accuracies should also be investigated. Finally, the relatively high recognition accuracies obtained with these small corpora confirm the observations summarized in Figures 1 and 2.

Table 2 *Phone and Word-recognition results. Phone-recognition correctness ("Corr") and accuracy ("Acc") achieved are listed for each of the languages in the Lwazi corpus. "Ave # phones" refers to the average number of occurrences of each phone for each speaker, and the final column lists the phonotactic perplexity of each language in our corpus. NTIMIT results from [Morales et al., 2008] are provided for comparative purposes. Small vocabulary word recognition accuracies are given for* 10 *languages. Each system is required to distinguish between ten different semantic categories, with each category represented by one to three different lexical items.*

Language	Phone recognition				Word recognition			
	% Corr	% Acc	Ave # phones	Phone ppl	Lwazi models	English Recognisers Lwazi	Ntimit	WSJ
Afrikaans	71.76	63.14	16.55	14.45	96.11	90.35	60.36	79.15
SA English	62.51	54.26	14.61	15.80	91.94	91.94	82.86	81.95
isiNdebele	74.21	65.41	28.66	10.26				
isiXhosa	69.25	57.24	17.79	10.67	95.29	77.78	34.34	61.28
isiZulu	71.18	60.95	23.42	11.20	90.53	80.00	37.57	69.19
Tshivenda	76.37	66.78	19.53	9.99	97.74	66.37	57.56	52.14
Sepedi	66.44	55.19	16.45	11.54	89.49	83.72	54.91	43.41
Sesotho	68.17	54.79	18.57	10.40	97.14	79.48	30.65	50.65
Setswana	69.00	56.19	20.85	11.15	87.66	76.95	39.02	52.09
Siswati	74.19	64.46	30.66	10.38	96.62	77.01	46.46	61.09
Xitsonga	70.32	59.41	14.35	10.34	97.90	77.58	54.99	60.60
NTIMIT	64.07	55.73						

5.2 Small-vocabulary speech recognition with the Lwazi corpus

Phone recognition is a useful benchmark to employ for recognition in new languages, since extensive intuition exists on phone-recognition accuracies achieved on standard corpora. However, initial applications of ASR in the developing world will in practice require accurate small-vocabulary recognition (as described in Section 1). We therefore describe experiments aimed at estimating our performance on such tasks next.

During the collection of the Lwazi ASR corpus, callers were asked several questions, of which some resulted in only a small set of responses. These included the following:

- Are you married?
- Are you speaking on a landline or a cellphone?
- What is your gender?
- What is your mother tongue?
- Where do you live? / Where were you born?

Since these same questions were asked of all speakers across all languages, they form a suitable basis for small-vocabulary experiments. Mother tongue speakers were then asked to label all answers that were semantically equivalent. In this fashion, answers such "Egoli" (isiZulu name for Johannesburg, meaning "place of gold") and "Johannesburg" were considered equivalents.

This resulted in 10 distinct semantic concepts for each language, with approximately one to three different lexical items corresponding to the same concept in a language. Because of the similar questions, similar meanings are attached to the matching concepts in each language, except for minor variations because of cultural differences. (For example, the majority of English and Afrikaans speakers would simply answer "yes" or "no" to the first question. In contrast, the majority of Xitsonga, Sepedi and Tshivenda speakers would answer the question in different ways depending on their gender. A Xitsonga man would

for example say “ni tekile / a ni tekangi” (I have taken / I haven’t taken), whereas a woman would say “ni tekiwile / a ni tekiwangi” (I have been taken / I haven’t been taken).) Our small vocabulary task was constructed by removing all utterances that contain any of the phrases corresponding to any of these concepts from the training set, since such vocabulary-independent performance is the realistic goal for application in SDSs. For testing purposes, all utterances that contain only these phrases were employed; recognition was deemed correct if the phrase was placed into the correct semantic category. Because of the relatively small set of test utterances (five or fewer per speaker), we employed ten-fold cross validation to estimate recognition accuracy.

A vocabulary of ten words (actually, concepts) is a good test of typical recognition tasks in an SDS which is aimed at Interactive Voice Response (IVR) applications, where the dialogue is structured to contain mostly menu items and command words. Common tasks such as yes/no recognition require even smaller vocabularies, and larger tasks with highly distinctive vocabularies may in fact give comparable accuracies to those achieved with our artificially-constructed grammar.

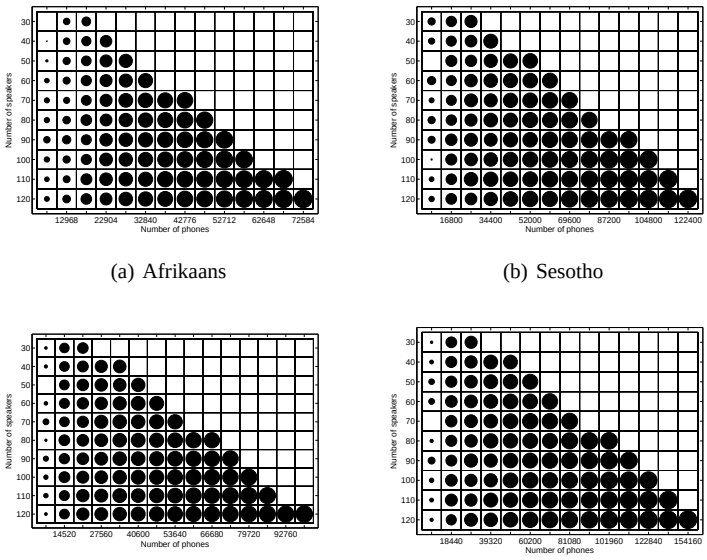
As a baseline for comparison, we have also measured the accuracies that can be achieved with the cross-language transfer procedure described, for example, in Sherwani et al. [2007, 2009]. That procedure, which is often a starting point for resource-scarce languages, utilises a well-trained recogniser in a world language such as English. All the words in the recognition task are transcribed using the phonemes of this well-trained recogniser, mapping the phonemes in the actual target language to the closest world-language phonemes where necessary. This cross-language dictionary is then used for recognition. Three English recognisers were investigated for our baseline, namely recognisers trained on the NTIMIT and Wall Street Journal corpora (the latter band-limited and downsampled to match our telephone corpus), and one trained on the English part of the Lwazi corpus.

Table 2 contains recognition results obtained with these baseline systems, as well as with our language-specific recognisers. (We were not able to carry out this experiment for isiNdebele, for lack of access to a mother-tongue speaker who could perform the semantic mappings.)

We see that accuracies above 90% are achieved in all languages except Sepedi and Setswana. With careful dialogue design [Cohen et al., 2004], this should be sufficient for a usable SDS. Of the three baseline systems that use phoneme mappings, the Lwazi English model is easily the most accurate. This is to be expected, since the acoustic conditions of NTIMIT and WSJ are somewhat dissimilar to those in Lwazi; however, the magnitude of the differences in accuracy is somewhat surprising. Even this best baseline system is, however, much less accurate than the language-specific acoustic models in most cases. Excluding English, the languages with the smallest absolute difference between baseline and trained models are Afrikaans, which is linguistically quite similar to English, and Sepedi, which also performed worst in the phone-recognition experiments (Section 5.1).

Given the similarities between the semantic categories in the different languages, it is interesting to compare the accuracies achieved in this task across languages (with error rates ranging between 2.1% and 12.3%). The quality of the phone recognisers partially explains these differences – in particular, the relatively poor performance of Sepedi and Setswana at both phone recognition and small-vocabulary word recognition is notable. However, Sesotho (with relatively accurate word recognition) and isiZulu (relatively accurate phone recognition) point to other relevant factors, such as the acoustic confusibility of words in semantically distinct classes that happens to occur in some languages but not others.

5.3 The number of training speakers



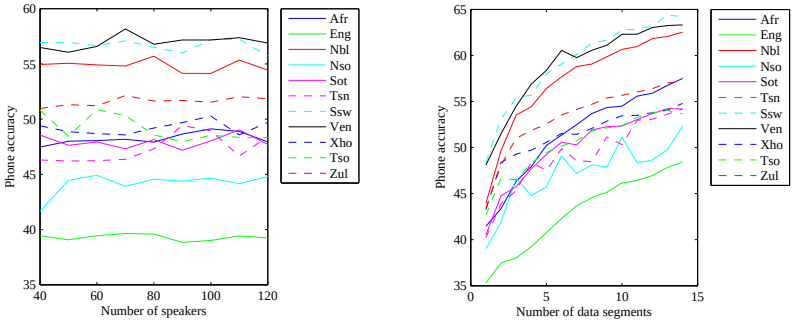
(a) Afrikaans (b) Sesotho

(c) isiZulu (d) isiNdebele

Fig. 6 *Phone accuracy as a function of both the number of speakers and the total amount of training data. The dots represent the measured accuracy, with bigger dots for higher accuracies and empty blocks (no dots) where no values are given.*

To analyse the influence of the number of training speakers on the recognition accuracy achieved, we investigate phone-recognition accuracy as a function of both the number of training speakers and the total number of phones used for training. (We use the number of phones rather than the number of words or utterances as measure of the amount of training data employed because of the significant differences in word and utterance lengths between the various languages – the phone count is therefore a better measure of the actual amount of speech employed.) The training sets are selected in such a way that the number of phones per speaker remains balanced.

Figure 6 shows typical results. (The empty blocks without dots in the upper right-hand corner of each figure represent experiments that could not be performed because sufficient data was not available for each individual speaker.) It is clear that the number of training speakers has little or no influence on the accuracy achieved, in the range that we have investigated. Whereas the figures show systematically increasing accuracy as the number of training phones is increased (from left to right), increasing the number of speakers contributing to a given set of training data has little effect (top to bottom). This same behaviour is observed for all eleven languages, and is confirmed by representations such as that shown in Figure 7(a) which shows the phone accuracy as a function of the number of training speakers, when about a quarter of the training data is used in each language. A similar insensitivity to the number of speakers was also observed for the training of context-dependent models [Barnard et al., 2009].



(a) (b)

Fig. 7 *Phone accuracy as a function of the number of speakers as well as a function of the amount of data in the training set, when data from all 120 training speakers is combined. In all cases, approximately* 25% *of the available training data is used to generate results for the number of speaker case. When all 120 training speakers is combined, the total amount of training data differs between the languages - the horizontal axis therefore indicates the number of segments used in each language, where the number of phone tokens per segment is (approximately) constant within a language, but different across languages.*

5.4 The amount of training data

Table 3 *Parameter values obtained by fitting measured phone-recognition rates. R^2 is the squared correlation between the estimated and actual values.*

Language	A	B	R^2
Afrikaans	64.94	549,900	0.9762
SA English	54.16	457,400	0.9650
isiNdebele	65.55	490,800	0.9722
Sepedi	55.35	380,200	0.2770
Sesotho	57.69	325,300	0.9201
Siswati	68.19	526,700	0.9757
Setswana	60.87	544,700	0.7975
Xitsonga	57.26	300,100	0.8839
Tshivenda	67.53	378,500	0.9616
isiXhosa	57.60	331,700	0.9710
isiZulu	59.96	352,100	0.9636

In Figure 7(b) we show the trends of phone recognition accuracy as a function of the amount of training data, when all 120 speakers are used. Although the curves for some languages (especially Sepedi) are quite noisy, it seems clear that none of the languages is approaching asymptotic phone-recognition accuracy given the amount of training data available in our corpus. In order to obtain a rough estimate of the amount of training data required to approach such an asymptote, we employ a heuristic relationship that is expected to hold for a wide range of classifiers [Schuurmans, 1997]. This relationship states that the error rate will asymptotically depend on the number of training samples (N) through the relationship $A-(B/N)$, with A and B parameters corresponding to the asymptotic error rate and the number of training samples required to approach within 1% of that error rate, respectively. We have empirically determined that this relationship provides a reasonable fit to

our data for values of N greater than approximately 50,000; we have therefore used a linear least-squares fit to estimate A and B values for all our languages, including only measured accuracies for $N > 50,000$ in our analysis. Table 3 summarises the results obtained, and Figure 8 shows a typical fit obtained in this manner. We see that quite good fits are obtained for several languages ($R^2 > 0.96$), and that the B parameter, which is related to the number of training phones required for accurate training, ranges between approximately 300,000 and 550,000 for these languages. (For $N = B$, phone accuracies within 1% of the asymptotic value are predicted.) In our corpus, the average phone duration is approximately 150 ms - hence, corpora of approximately 750 to 1,400 minutes per language are suggested.

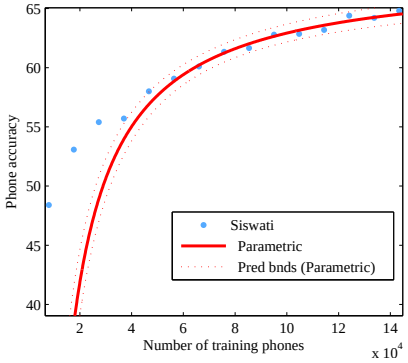


Fig. 8 *Example of parametric fit (for Siswati accuracies), with 95% confidence intervals computed from the fit.*

6 Conclusion

Collecting appropriate speech corpora for resource-scarce languages can be a challenging task, especially when financial resources are limited and speaker populations are small or geographically remote, with limited access to information and communication infrastructure. When collecting corpora from such environments, an understanding of the interplay between type and amount of data can be of great benefit, by ensuring that the collection effort is made as efficient as possible.

In this paper, we describe the Lwazi corpus for automatic speech recognition (ASR), a new telephone speech corpus for Sourth African languages. We analyse the data sufficiency of the corpus from both an analytical and a practical perspective: we measure the stability of ASR models derived from the corpus and evaluate phoneme recognition accuracy directly. We find that different phone classes tend to have different data requirements. Voiceless fricatives, for example, can be trained accurately with relatively few tokens per speaker, whereas nasals and vowels require more data per speaker for comparable convergence (stability) of the acoustic distributions. The number of speakers required for a given level of stability shows comparable, but not identical, trends.

Our investigation of the practical training of speech-recognition systems reveals that the number of training speakers is less of a constraint than the amount of data per speaker (under the circumstances investigated in this study). In particular, this investigation reveals

that systems of this nature can be trained successfully with around 40 to 50 training speakers; the total amount of speech to approach within 1% of asymptotic accuracy should be around 750 to 1,400 minutes per language. Clearly, more complicated recognition systems will benefit from more speakers and larger corpora; it is therefore important that work similar investigations should be carried out on larger multilingual corpora where such are available.

Another interesting avenue for future exploration follows from our findings that different phone classes have different data requirements. The data collection process could conceivably be made more efficient by biasing the recorded material towards the more "data-hungry" phonetic categories; it remains to be seen, however, whether that benefit can be obtained without making the recording protocol too unnatural.

References

Abdillahi, N., Nocera, P., & Bonastre, J.-F. (2006). Automatic transcription of Somali language. In *Proceedings of Interspeech*, (pp. 289–292). Pittsburgh, PA, USA.

Badenhorst, J. (2009). *Data Sufficiency Analysis for Automatic Speech Recognition*. Master's thesis, Potchefstroom Campus, North-West University.

Badenhorst, J., & Davel, M. (2008). Data Requirements for Speaker Independent Acoustic Models. In *Proceedings of the 19th Annual Symposium of the Pattern Recognition Association of South Africa*, (pp. 147–152). Cape Town, South Africa.

Barnard, E., Cloete, L., & Patel, H. (2003). Language and Technology Literacy Barriers to Accessing Government Services. *Lecture Notes in Computer Science*, *2739*, 37–42.

Barnard, E., Davel, M., & van Heerden, C. (2009). ASR Corpus Design for Resource-Scarce Languages. In *Proceedings of Interspeech*, (pp. 2847–2850). Brighton, U.K.

Byrne, W., Beyerlein, P., Huerta, J. M., Khudanpur, S., Marthi, B., Morgan, J., Peterek, N., Picone, J., Vergyri1, D., & Wang, W. (2000). Towards language independent acoustic modeling. In *Proceedings of the Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 2, (pp. 1029–1032). Istanbul, Turkey.

Cohen, M., Giangola, J., & Balogh, J. (2004). *Voice User Interface Design*. Addison-Wesley.

Davel, M., & Barnard, E. (2004). The Efficient Generation of Pronunciation Dictionaries: Human Factors during Bootstrapping. In *Proceedings of Interspeech*, (pp. 2797–2800). Jeju, Korea.

Davel, M., & Barnard, E. (2008). Pronunciation predication with Default&Refine. *Computer Speech and Language*, *22*(4), 374–393.

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

Kominek, J., & Black, A. W. (2006). Learning Pronunciation Dictionaries: Language Complexity and Word Selection Strategies. In *Proceedings of the Human Language Technology Conference of the NAACL*, (pp. 232–239). New York City, USA: Association for Computational Linguistics.

Lehohla, P. (2003). Census 2001: Census in brief. Report no. 03-02-03. Online: http://www.statssa.gov.za/census01/html/CIB2001.pdf.

Maskey, S., Black, A., & Tomokiyo, L. (2004). Boostrapping Phonetic Lexicons for New Languages. In *Proceedings of Interspeech*, (pp. 69–72). Jeju, Korea.

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

Morales, N., Tejedor, J., Garrido, J., Colas, J., & Toledano, D. (2008). STC-TIMIT: Generation of a Single-channel Telephone Corpus. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, (pp. 391–395). Marrakech, Morocco.

Nagroski, A., Boves, L., & Steeneken, H. (2003). In search of optimal data selection for training of automatic speech recognition systems. *Automatic Speech Recognition and Understanding, 2003. ASRU '03. 2003 IEEE Workshop on*, (pp. 67–72).

Nasfors, P. (2007). *Efficient Voice Information Services for Developing Countries*. Master's thesis, Department of Information Technology, Uppsala University, Sweden.

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

Patel, N., Chittamuru, D., Jain, A., Dave, P., & Parikh, T. S. (2010). Avaaj Otalo A Field Study of an Interactive Voice Forum for Small Farmers in Rural India. In *Proceedings of the 28th International Conference on Human Factors in Computing systems*, (pp. 733–742). Atlanta, GA, USA: ACM.

Riccardi, G., & Hakkani-Tur, D. (2003). Active and Unsupervised Learning for Automatic Speech Recognition. In *Proceedings of Eurospeech*, (pp. 1825–1828). Geneva, Switzerland.

Roux, J., Botha, E., & du Preez, J. (2000). Developing a Multilingual Telephone Based Information System in African languages. In *Second International Language Resources and Evaluation Conference*, (pp. 975–980). Athens, Greece.

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

Schuurmans, D. (1997). Characterizing Rational versus Exponential Learning Curve. *Journal of Computer and System Science*, *55*(1), 140–160.

Seid, H., & Gambäck, B. (2005). A Speaker Independent Continuous Speech Recognizer for Amharic. In *Proceedings of Interspeech*, (pp. 3349–3352). Lisboa, Portugal.

Sharma, A., Plauché, M., Kuun, C., & Barnard, E. (2009). HIV Health Information Access using Spoken Dialogue Systems: Touchtone vs. Speech. In *IEEE International Conference on Information and Communications Technologies and Development '09 (ICTD 09)*, (pp. 95–107). Doha, Qatar.

Sherwani, J., Ali, N., Mirza, S., Fatma, A., Memon, Y., Karim, M., Tongia, R., & Rosenfeld, R. (2007). Healthline: Speech-based Access to Health Information by low-literate users. In *Information and Communication Technologies and Development, International Conference on*, (pp. 131–139). Bangalore, India.

Sherwani, J., Palijo, S., Mirza, S., Ahmed, T., Ali, N., & Rosenfeld, R. (2009). Speech vs. touch-tone: Telephony interfaces for information access by low literate users. In *IEEE International Conference on Information and Communications Technologies and Development '09 (ICTD 09)*, (pp. 447–457). Doha, Qatar.

Tucker, R., & Shalonova, K. (2004). The Local Language Speech Technology Initiative. In *Proceedings of SCALLA Conference*. Nepal.

van Heerden, C., Barnard, E., & Davel, M. (2009). Basic Speech Recognition for Spoken Dialogues. In *Proceedings of Interspeech*, (pp. 3003–3006). Brighton, U.K.

Viterbi, A. (1967). Error bounds for convolutional codes and a asymptotically optimum decoding algorithm. *IEEE Transactions on Information Theory*, *13*, 260–269.

Wheatley, B., Kondo, K., Anderson, W., & Muthusumy, Y. (1994). An evaluation of cross-language adaptation for rapid HMM development in a new language. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, vol. 1, (pp. 237–240). Adelaide, SA, Australia.

Wu, Y., Zhang, R., & Rudnicky, A. (2007). Data selection for speech recognition. *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding*, (pp. 562–565).