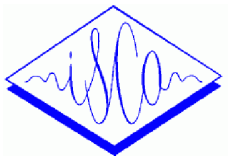

Third Workshop on Spoken Language
Technologies for Under-resourced Languages
Cape Town, South Africa
May 7-9, 2012

# MEDIUM-VOCABULARY SPEECH RECOGNITION FOR UNDER-RESOURCED LANGUAGES

*Charl van Heerden, Marelie H. Davel and Etienne Barnard*

North-West University, Vanderbijlpark, South Africa

## ABSTRACT

We report on the development of speech-recognition systems that are able to perform accurate recognition on medium-vocabulary tasks (i.e. tasks that require distinctions between approximately 200 different terms). We are able to achieve error rates of less than 5% (our design goal) on four under-resourced languages as well as English, by using training corpora that contain 70–100 hours of speech per language. The majority of the errors stem from words such as abbreviations, foreign words or names, which do not adhere to the standard orthography of the target language. We also find that recognition accuracy does not depend strongly on the number of occurrences of a term in the training set or the length of the term to be recognized, and that a few problematic speakers are responsible for a disproportionate number of errors.

***Index Terms***— Speech recognition, under-resourced languages, multilingual speech processing

## 1. INTRODUCTION

The potential of speech-recognition systems to play a meaningful role in developing-world applications is widely understood [1, 2, 3], but the development of such systems for the under-resourced languages which are ubiquitous in the developing world is a major obstacle to the fulfilment of this potential. One way to address this issue is to understand how well recognizers can function if limited resources are used in their development. With this in mind, we have recently shown that flexible and accurate small-vocabulary recognizers can be trained with relatively small amounts of speech [4]. However, recent developments in smartphone-based data collection have significantly reduced the complexity of creating sizeable speech corpora in under-resourced languages, thus making it feasible to attempt more ambitious recognition tasks in such languages. We therefore investigate the issues that occur when developing a set of recognizers for medium-sized vocabularies (around a few hundred words per grammar) for a group of under-resourced languages, and compare the resulting performance with that achieved with an English recognizer developed under comparable circumstances.

Our first goal is to investigate how recognizers trained on 70–100 hours of speech per language perform on realistic medium-vocabulary recognition tasks for information-access applications. Such tasks are characterized by moderately high perplexities (we study perplexities from 50 to 200), variable-length recognition terms and a significant prevalence of proper names and words taken from foreign languages. We then investigate how each of these characteristics affects the performance of the trained recognizers, and also characterize the speaker differences that we observe when this task is performed.

In Section 2, we summarize some of the relevant work that has been done on speech recognition for under-resourced languages, including a discussion on the developments in corpus collection that have enabled us to expand our corpora efficiently. Section 3 details the corpus that was used for our experiments, as well as the experimental protocol followed. Section 4 contains our experimental results, and in Section 5 we discuss the implications of this work.

## 2. BACKGROUND

Since data scarcity is the basic problem for the development of ASR systems in under-resourced languages, most of the research on such systems has involved some form of data sharing, either with a well-resourced “donor” language, or by pooling data from two or more related under-resourced languages. Several methods using such data sharing are summarized in [5]. Although each of those methods – ranging from phoneme mapping for recognition with “pure” donor-language acoustic models to complete retraining with appropriately pooled data – shows some improvement from data sharing, it is fair to say that the amount of target-language speech remains the most important determinant of the recognition accuracy achieved. That is, none of the pooling strategies is able to deliver comparable accuracy with that which is achieved with sufficient training data from the target language.

Against this background, recent developments in ASR data collection tools are extremely encouraging. In particular, Hughes *et al.* [6] described a smartphone-based application, that greatly streamlines the process of ASR data collection by combining the basic components of such collection in a single, easy-to-use and portable package. The application is loaded with a collection of prompts which are to be

recorded by speakers in the target population; each respondent is then handed a smartphone, which displays a sequence of prompts to the speaker and records the resulting speech. This application inspired the development of Woefzela, an open-source application for Android smartphones, which is specifically geared towards the practicalities of developing-world data collection – thus, it supports the logistics of field workers who may need to perform data collection in rural or otherwise challenging environments, performs some quality control while the collection is taking place, and does not rely on the constant availability of Internet connectivity.

Using these applications and a collection of, say, ten or fifteen relatively inexpensive smartphones, it is a simple matter to collect speech from a few hundred mother-tongue speakers of a target language in a week or less. As a consequence, the basic assumption that developing-world ASR implies small speech corpora is overturned. Of course, many other challenges remain for ASR in under-resourced languages – the creation of suitable text corpora and pronunciation lexicons as well as user-interface issues and the like remain as formidable obstacles to be overcome in the quest for high-impact speech technology for the developing world.

We can distinguish three overlapping categories of applications which could conceivably contribute to such impact [7, 8]:

- Small-vocabulary recognizers, with vocabularies from 2 to around 20 words, are useful for tasks such as menu traversal and the confirmation of choices or preferences. One can think of such systems as touchtone replacements, which is a useful capability in light of the difficulties that many developing-world users experience with touchtone interfaces [3].

- Medium-vocabulary systems, which typically operate with vocabularies up to several hundred words, can also be used for selecting items from longer lists (e.g. to choose destinations in a travel application) or to search for information in limited domains.

- Systems with practically unrestricted vocabularies represent the state of the art in speech recognition; these are used in applications such as voice search, dictation, navigation and numerous others.

We have previously pointed out that useable small-vocabulary systems (i.e. with speaker-independent accuracies in the order of 95 % or better) can be developed for under-resourced languages with around 5 – 10 hours of orthographically transcribed speech in the target language. The main goal of the current contribution is to investigate whether similar performance is achievable for medium-vocabulary tasks, given the substantially larger corpora that are rendered feasible with tools such as the application of Hughes *et al.* and Woefzela.

## 3. MATERIALS AND METHOD

Our investigation involves five of the official languages of South Africa, as shown in Table 1. Amongst these, the South African dialect of English is the only well-resourced language (and even SAE does not have the benefit of large publicly-available dialect-specific speech corpora or pronunciation lexicons). We employed preliminary versions of speech corpora in those languages that are being developed at the Meraka Institute; statistics of these corpora are also provided in Table 1.

**Table 1**: Summary of corpora used in speech-recognition experiments. 8 speakers were held out for future evaluations while the rest of the corpus was used for experiments by performing 4-fold cross-validation.

| Language | # speakers | # utts. | vocab. | #hours |
|---|---|---|---|---|
| Afrikaans | 190 | 99620 | 3835 | 82 |
| English | 214 | 111374 | 7619 | 80 |
| isiNdebele | 209 | 90297 | 17666 | 112 |
| isiZulu | 166 | 82552 | 2793 | 95 |
| Sesotho | 199 | 106550 | 2660 | 88 |

Since the corpora all consist of short prompts (mostly 1 – 5 words in length) read by variable numbers of speakers, we constructed our test grammars by randomly selecting 200 prompts from each language as the target utterances. This allows us to investigate the influence of factors such as training-vocabulary dependence and the length of the terms to be recognized. (Due to slightly variable collection protocols employed in the different languages, we had to limit the number of prompts that were spoken by only one or two speakers in some languages. In those languages, we ordered the prompts by their occurrence frequencies in the corpus, and modified our procedure to ensure approximately uniform coverage of the various occurrence frequencies.) For our investigation into the effect of variable perplexity on recognition accuracy, random subsets of these 200 target utterances were drawn.

Our protocol does not include a manual verification process of the recordings – hence, reading and other errors are known to be present in both the training and test sets. We are currently developing other tools to detect such errors automatically [9], but the present work does not utilize any of those tools. Hence, our error-rate estimates are pessimistic; we return to this matter in the conclusion below.

The data in each language was split into four partitions with no speaker overlap – each partition thus contained approximately 12 – 15 hours of speech, from approximately 50 speakers. These partitions were then used in cross validation: we repeatedly trained acoustic models on three of the partitions; recognition accuracy was then measured on the remaining portion in each case.

A standard 3-state left to right HMM architecture is used to model context-dependent triphones in each language, using HTK [10]. As acoustic features, 39-dimensional Mel Frequency Cepstral Coefficients are used: 13 static coefficients with cepstral mean normalisation applied, 13 delta and 13 double delta coefficients. Triphones are tied at the state level using decision tree clustering, and each tied-state triphone is estimated with 8 Gaussian mixtures per state. Semi-tied transforms are also employed throughout.

## 4. RESULTS

Fig. 1 shows the error rates that were achieved for the 200-term task across the five languages, as well as the accuracies achieved on various subsets of the data. We see that the fundamental goal of 5% error rate on "generic" terms was achieved for all languages, with error rates on those terms ranging between 1.1% for English to 3.7% for isiNdebele. On the other categories of terms, however, performance was highly variable – for isiZulu, English and Afrikaans, good performance was achieved across all classes, whereas isiNdebele performance on all the non-standard items was much worse. These differences stem from at least two causes. On the one hand, the sophistication of resources (such as pronunciation dictionaries and pre-processing rules to place text in canonical form) differs significantly across the various languages. Another factor is the difference in recording protocols followed for the various languages: these early versions of the NCHLT corpus have significantly different repetition frequencies of the various prompts by different speakers. We return to these factors in Section 5 below.

We next investigate the vocabulary independence of our recognizers, by studying how recognition accuracy of each term depends on the number of occurrences of that term in the training set. Thus, we keep track of the number of instances of each term in a given training fold, and tally the resulting test-fold performance of that term against this training-set count. This process is repeated across all folds. The resulting trends of error rate against the number of training occurrences are shown in Fig. 2. As before, there seem to be significant differences between the different languages – however, there are not strong training-count dependencies in any of the languages, suggesting that the accuracies achieved are, to a large extent, vocabulary independent.

Fig. 3 summarizes our findings on another important issue related to the performance of our system, namely the relationship between the accuracy achieved and the length of the term to be recognized. One would generally expect longer terms to be recognized more accurately, since such terms are more distinct from one another. For English and perhaps Sesotho there is evidence of such a trend; however, those trends are generally fairly weak, suggesting that utterance length is not a strong determinant of confusability within the parameters of our investigation.

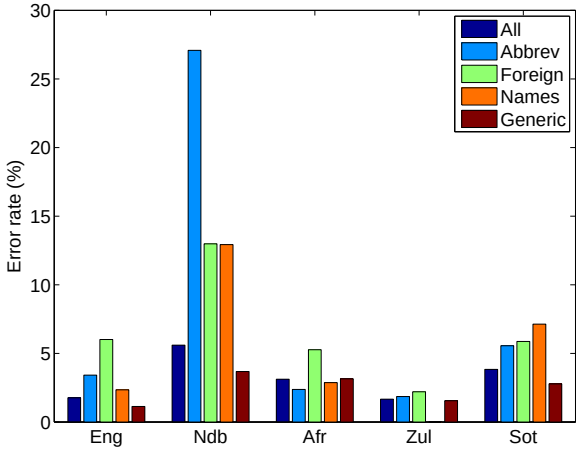


**Fig. 1**: Error rates achieved when performing 200-term recognition in several languages. "Abbrev", "Foreign" and "Names" refer to error rates on subsets of the terms containing one or more words in those categories, whereas "generic" refers to error rates achieved on terms that do not contain words in any of those categories.

We have also studied how the vocabulary size (and thus, in our design, perplexity) influences recognition accuracy. To this end, we have selected subsets with, respectively, 50 and 100 terms from our original 200-term grammar, and performed the same recognition experiments as above with those target grammars. Fig. 4 shows that the dependency of error rate on perplexity is notable in all languages, but that the magnitude of this dependency again differs somewhat across languages. The weak trend for Afrikaans is easy to explain, given that most of the errors that occurred in the Afrikaans test were caused by a small number of speakers (2 or 3) who had various problems using our recording interface – these problems were so severe as to cause comparable error rates at all perplexities. The relatively steep curve for isiNdebele does not have such a clear-cut explanation, and deserves further study.

Finally, the error rates for the individual speakers are summarized in Fig. 5. In each language, the speakers have been ordered from highest to lowest error rate, and speakers who have contributed fewer than 5 test utterances have been excluded. We see similar trends across all languages, namely one or two speakers with extremely high error rates, followed by an intermediate group whose error rates are in the range 5% to 15%, and a long tail of speakers for whom recognition is perfect or nearly so. The small set of speakers with very high error rates produced poor recordings for a variety of reasons:

- A few speakers had difficulty with the recording hardware, apparently covering the recording microphone

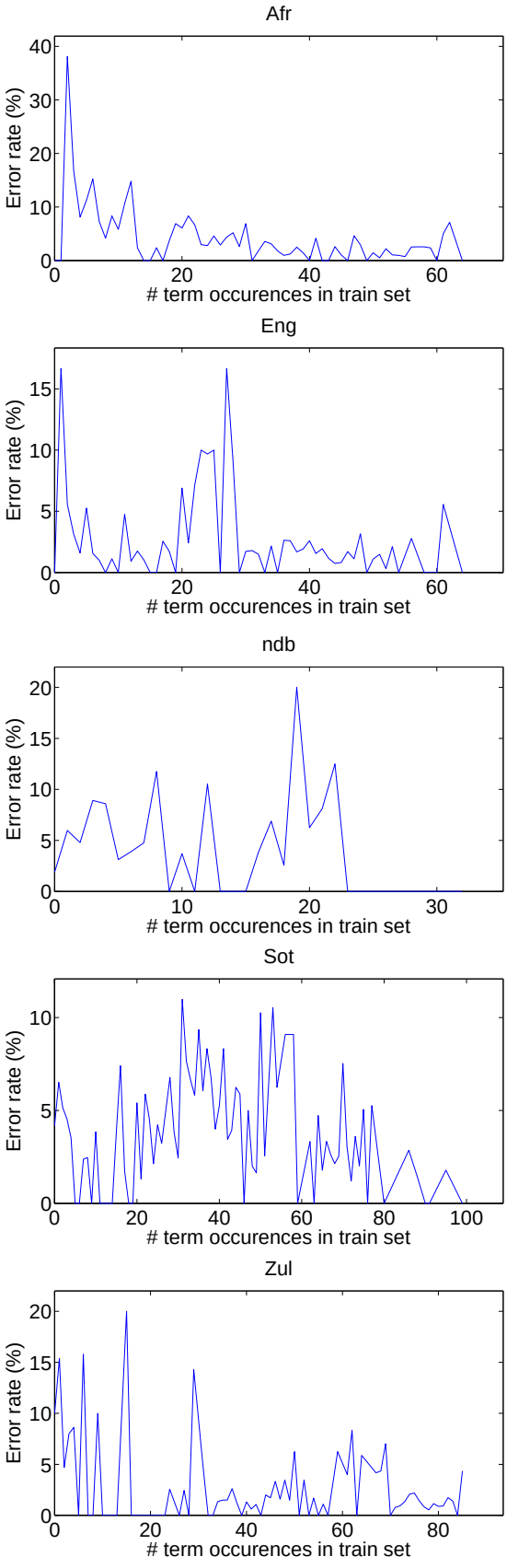


**Fig. 2**: Error rates as a function of the number of occurrences of a particular test term in the training set.

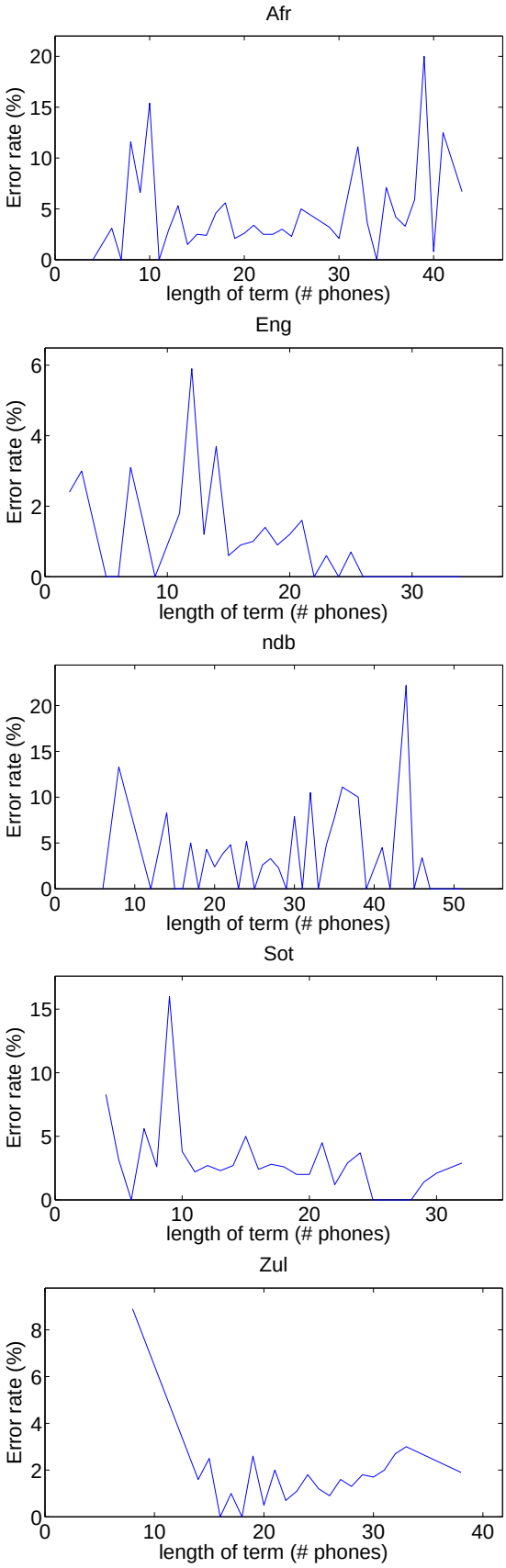


**Fig. 3**: Error rates as a function of the number of phones in the term to be recognized (for words with multiple pronunciations, we use the most common pronunciation).

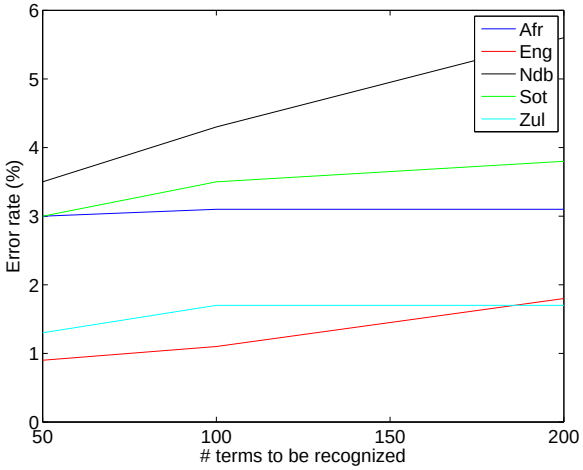


**Fig. 4**: Error rate as a function of the number of terms to be recognized.

while doing the recordings. The resulting audio files are significantly distorted (the high-frequency components are suppressed), but are generally still judged as intelligible by human listeners.

- Others had great difficulty in reading, often producing hesitant speech, partial repetitions and incorrect renderings of the prompts provided. Hence these should more properly be thought of as "reading errors" rather than "recognition errors".

- A small number of speakers were obviously distracted during reading; their recordings are different from those in the previous category in that the readings are generally fluent, but inaccurate, containing extraneous speech. Prompts from these speakers are often cut off, as the recording application is not operated successfully by the speaker.

Utterances like these can be filtered using confidence measures, such as those described in [11], where a phone-based dynamic programming scoring technique is used to generate validation scores on a per utterance level.

## 5. CONCLUSION

We have shown that useful recognition accuracies can be achieved on medium-vocabulary recognition tasks in low-resource languages using corpora collected with the new generation of smartphone-based data-collection tools. Although details for the languages differ in various ways, a number of common themes emerge:

- The recognition accuracies for words and terms that do not correspond to the standard orthography of the lan-

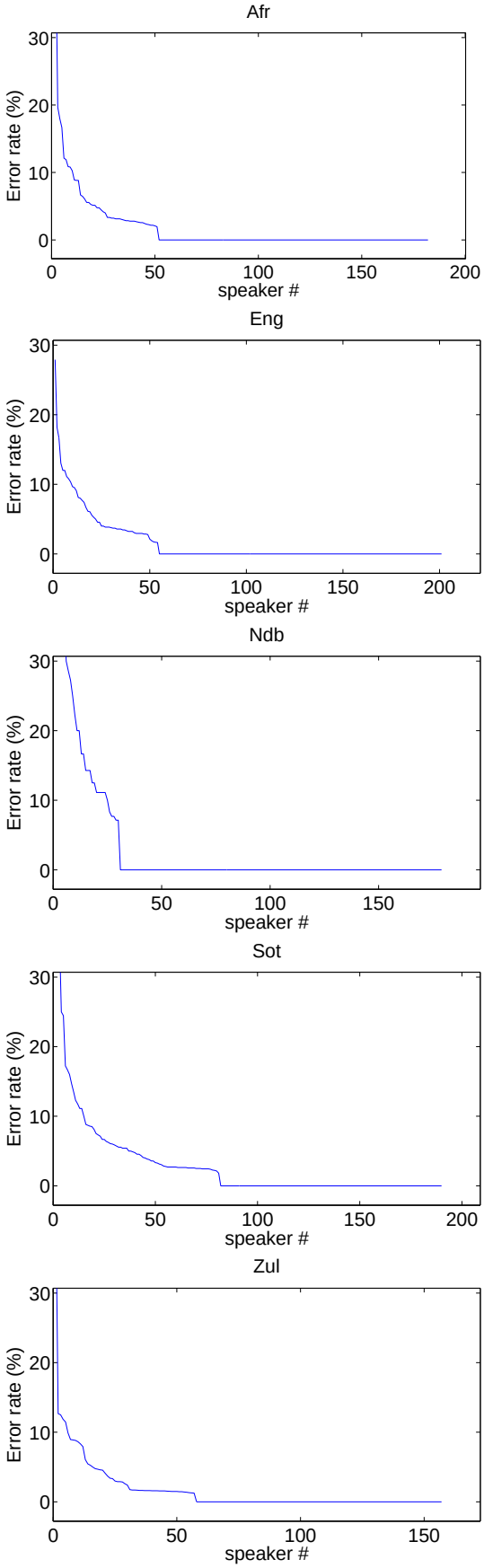


**Fig. 5**: Speaker-specific error rates, with speakers arranged from highest to lowest error rates.

guages (i.e. abbreviations, acronyms, names and foreign words) is generally much lower than that of standard terms.

- The accuracies achieved do not depend strongly on the amount of training data available for a particular term or the length of the term in phonemes.
- The perplexity of the recognition grammar predictably correlates quite well with the error rate observed.

Our analysis of the speaker-specific performance suggests that the corpus for each language contains a few "goats", who for some reason or another did not produce recognizable speech. These speakers should probably be excluded from the corpus, since their problems are typically interface issues rather than recognition failures. Excluding the three worst speakers in each corpus has a significant effect on the overall error rates, as we show in Table 2.

**Table 2**: Overall error rates achieved when all speakers are retained, and when the three speakers with the highest error rates are removed.

| | **Eng** | **Ndb** | **Afr** | **Zul** | **Sot** |
|---|---|---|---|---|---|
| All speakers | 1.76 | 5.60 | 3.11 | 1.66 | 3.83 |
| 3 speakers removed | 1.41 | 3.94 | 1.56 | 1.27 | 2.66 |

In general, though, the fraction of erroneous utterances in our corpus seems to be quite low, and we suspect that the quality of the trained acoustic models will not be affected much by our corpus-cleaning exercise (or even a manual verification of all utterances). Much more significant from an accuracy perspective would be a more systematic approach to pronunciation modelling for those items that do not correspond to the standard orthography – for example, by detecting foreign words and using alternative pronunciation rules for those words, and by doing more sophisticated normalization of abbreviations and acronyms. It is unlikely that we will be able to perform these tasks as successfully as can be done in well-resourced languages, but significant improvements from our current baseline approaches should be achievable.

In conclusion, we have shown that practically useable speech recognition for our target languages and vocabulary sizes can be achieved with the tools currently at our disposal. The most significant challenges for similar developments in new languages are (a) the initial collection of suitable text corpora and (b) the logistics of contacting and working with the first-language speakers who contribute the actual data. It will be interesting to see whether these tasks can be simplified further, and whether this can stimulate a widespread development of ASR systems with medium to large vocabularies for new languages.

## 6. REFERENCES


[1] N. Patel, D. Chittamuru, A. Jain, P. Dave, and T.S. Parikh, "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 (CHI 2010)*, Atlanta, GA, USA, April 2010, ACM, pp. 733–742.

[2] E. Barnard, M.H. Davel, and G.B. van Huyssteen, "Speech Technology for Information Access: a South African case study," in *Proceedings of the AAAI Spring Symposium on Artificial Intelligence for Development (AI-D)*, Palo Alto, CA, March 2010, pp. 8–13.

[3] J. Sherwani, S. Palijo, S. Mirza, T. Ahmed, N. Ali, and R. Rosenfeld, "Speech vs. Touch-tone: Telephony Interfaces for Information Access by Low Literate Users," in *Proc. IEEE Int. Conf. on ICTD*, Doha, Qatar, April 2009, pp. 447–457.

[4] C. van Heerden, E. Barnard, and M. Davel, "Basic speech recognition for spoken dialogues," in *Proc. Interspeech*, Brighton, UK, September 2009, pp. 3003–3006.

[5] C. van Heerden, N. Kleynhans, E. Barnard, and M. Davel, "Pooling ASR Data for Closely Related Languages," in *Proc. SLTU*, Penang, Malaysia, May 2010, pp. 17–23.

[6] T. Hughes, K. Nakajima, L. Ha, P. Moreno, and M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," in *Proc. Interspeech*, Makuhari, Japan, September 2010, pp. 1914–1917.

[7] E. Barnard, M. Plauché, and M. Davel, "The Utility of Spoken dialog systems," in *Proc. SLT*, Goa, India, December 2008, IEEE, pp. 13–16.

[8] E. Barnard, J. Schalkwyk, C. van Heerden, and P.J. Moreno, "Voice Search for Development," in *Proc. Interspeech*, Makuhari, Japan, September 2010, pp. 282–285.

[9] M.H. Davel, C. van Heerden, N. Kleynhans, and E. Barnard, "Efficient Harvesting of Internet Audio for Resource-Scarce ASR," in *Proc. Interspeech*, Florence, Italy, August 2011, pp. 3153–3156.

[10] S. Young, G. Evermann, M. Gaels, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, "The HTK book (for HTK version 3.4," March 2009.

[11] Marelie H. Davel, Charl J. van Heerden, and Etienne Barnard, "Validating smartphone-collected speech corpora," in *Proc. SLTU*, Cape Town, South Africa, May 2012, pp. 68–75.