LANGUAGE IDENTIFICATION AND MULTILINGUAL SPEECH RECOGNITION USING DISCRIMINATIVELY TRAINED ACOUSTIC MODELS

Thomas Niesler[†], *Daniel Willett*[‡]

[†]Department of Electrical and Electronic Engineering, University of Stellenbosch, South Africa
trn@dsp.sun.ac.za

[‡]Harman/Becker Automotive Systems, Ulm, Germany
dwillett@harmanbecker.com

Abstract

We perform language identification experiments for four prominent South-African languages using a multilingual speech recognition system. Specifically, we show how successfully Afrikaans, English, Xhosa and Zulu may be identified using a single set of HMMs and a single recognition pass. We further demonstrate the effect of language identification-specific discriminative acoustic model training on both the per-language recognition accuracy as well as the accuracy of the language identification process. Experiments indicate that discriminative training leads to a small overall improvement in language identification accuracy while not affecting the speech recognition performance strongly. Furthermore, language identification is found to be more error prone and discriminative training less effective for code-mixed utterances, indicating that these may require special treatment within a multilingual speech recognition system.

1. Introduction

Language identification and multilingual speech recognition are key to the development of spoken dialogue systems that can function in multilingual environments. In South Africa, which officially recognises eleven languages and whose citizens almost without exception speak more than one of these languages fluently, the development of such systems is an especially relevant challenge. In this paper, we develop language identification systems based on continuous speech recognition, and evaluate these for four languages, Afrikaans, English, Xhosa and Zulu, which are spoken by 13.3%, 8.2%, 17.6% and 23.8% of the South African population respectively [11].

Several approaches to language identification (LID) have been proposed in the literature [15]. Systems based on Gaussian mixture models (GMMs) classify the speech feature vectors independently, and do not make explicit use of phonotactic or other higher-level information. Neural networks and support vector machines can also be used in this way and have been demonstrated to be promising alternatives to GMMs [10, 3]. All of these approaches have low computational requirements when compared with alternatives based on speech recognition. Furthermore, they do not require orthographically or phonetically annotated training material, which is an advantage when dealing with languages for which such resources are not available.

A more sophisticated class of language identification systems makes use of continuous speech recognition algorithms. These systems can in turn be divided into two broad groups: those based on phone recognition, and those based on word recognition.

The basic phone-based LID system is often referred to as *Phone Recognition followed by Language Modelling* (PRLM) [2, 13, 15] and is illustrated in Figure 1. A single phone recogniser, employing either unilingual or multilingual phone units, "tokenises" the acoustic input into a sequence of phone labels. This sequence is fed to a bank of parallel n-gram language models, one for each language to be identified. Finally, a classifier determines the language of the input speech based on the language model score.

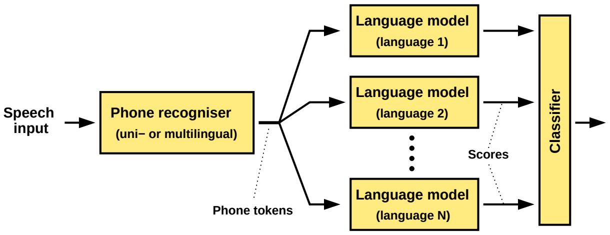
Speech input
Phone recogniser (uni– or multilingual)
Phone tokens
Language model (language 1)
Language model (language 2)
Language model (language N)
Scores
Classifier

Figure 1: Phone recognition followed by language modelling (PRLM).

A slightly more elaborate version of PRLM known as *Parallel Phone Recognition followed by Language Modelling* (PPRLM) [2, 15] is illustrated in Figure 2.

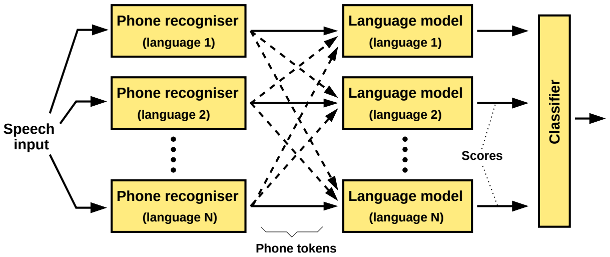
Speech input
Phone recogniser (language 1)
Phone recogniser (language 2)
Phone recogniser (language N)
Phone tokens
Language model (language 1)
Language model (language 2)
Language model (language N)
Scores
Classifier

Figure 2: Parallel phone recognition followed by language modelling (PPRLM).

In this method, the audio data is presented to a parallel bank of phone recognisers, each for a different language. In some cases a phone recogniser is present for each language to be identified, and in some cases a subset or even a different set is used.

As before, the phone recogniser outputs are fed to a bank of parallel language models, and a classifier again selects the language with the highest score. PPRLM is computationally more complex due to the need for parallel recognisers but has been shown to outperform PRLM-based systems.

Finally, language identification can also be achieved as a by-product of word-based large vocabulary speech recognition, as illustrated in Figure 3.

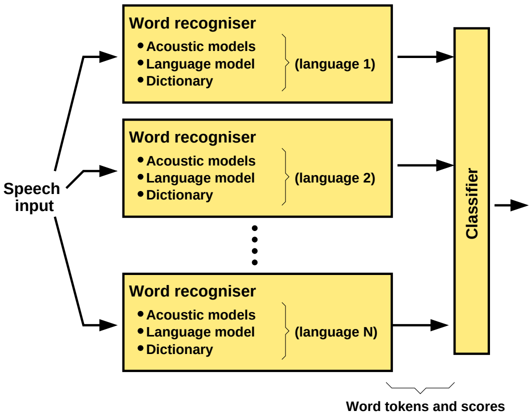

Figure 3: Parallel word recognition and language modelling (PWRLM).

This approach is similar to PPRLM, with one word-based speech recogniser per language operating in parallel, and will therefore be referred to as *Parallel Word Recognition followed by Language Modelling* (PWRLM) [4, 5, 9]. From the results quoted in the literature, PWRLM systems reduce the LID error rate by between 10 and 50% relative to PPRLM systems.

We propose a variant of a PPRLM/PWRLM system in which the acoustic models of the individual phone or word recognisers are merged into a single set of models, as illustrated in Figure 4.

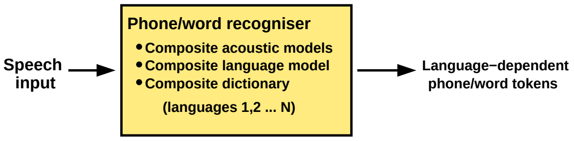
Speech input
Phone/word recogniser
• Composite acoustic models
• Composite language model
• Composite dictionary
(languages 1,2 ... N)
Language−dependent phone/word tokens

Figure 4: Language identification by combined acoustic and language modelling.

Since language dependence is maintained, this does not represent an IPA-based merging of phones, but rather a single set of acoustic models consisting of decoupled sub-sets of phone models for each language. The merged set of acoustic models can then be used in a single recognition pass with an appropriately defined language model. Hence parallel competing paths in each language are allowed during the recognition search as in PPRLM. However, since decoding occurs within a single pass, alternative paths belonging to unfavourable languages may be expected to be pruned from the set of recognition paths quickly. Hence this method represents a computational saving in comparison with PPRLM.

The same approach can be applied to word recognition. In this way both the spoken word sequence as well as the language are identified in a single pass without explicitly decoding the utterance for each language in parallel, as would be done in PWRLM. This can facilitate the development of multilingual dialogue systems, which share a dialogue structure and branch logic. It will in particular allow the language of discourse to alter during the dialogue, a phenomenon known as code mixing that is prevalent among South African speakers.

This paper deals with the phone-based variant of the approach illustrated in Figure 4. Our aim is to determine whether the language identification performance of such a composite set of acoustic models can be improved by means of discriminative training, and how this affects speech recognition accuracy.

2. Speech databases

We have based our experiments on the African Speech Technology (AST) databases, which consist of recorded and annotated speech collected over both mobile and fixed telephone networks [7]. For the compilation of these databases, speakers were recruited from targeted language groups and given a unique datasheet with items designed to elicit a phonetically diverse mix of read and spontaneous speech. The datasheets included read items such as isolated digits, as well as digit strings, money amounts, dates, times, spellings and also phonetically-rich words and sentences. Spontaneous items included references to gender, age, mother tongue, place of residence and level of education.

The AST databases were collected in five different languages, as well as in a number of non-mother tongue variations. In this work we have used of the Afrikaans, English, Xhosa and Zulu mother tongue databases. Note that, due to the prevalence of code-mixing, each of these databases may contain words in other languages.

Together with the recorded speech waveforms, both orthographic (word-level) and phonetic (phone-level) transcriptions were available for each utterance. The orthographic transcriptions were produced and validated by human transcribers. Initial phonetic transcriptions were obtained from the orthography using grapheme-to-phoneme rules, except for English where a pronunciation dictionary was used instead. These were subsequently corrected and validated manually by human experts.

2.1. Training and test sets

Each database was divided into a training and two test sets. The four training sets each contain between six and eleven hours of audio data, as indicated in Table 1. Phone types refer to the number of different phones that occur in the data, while phone tokens indicates their total number. Note that a slightly lower speech rate was observed for Xhosa and Zulu compared with English and Afrikaans.

Database name	Speech (hours)	No. of speakers	Phone types	Phone tokens
Afrikaans	6.18	234	84	180,904
English	6.02	271	73	167,986
Xhosa	6.98	219	107	177 843
Zulu	10.87	203	101	285,501

Table 1: Training sets for each database.

Development and evaluation test sets were prepared for each language and contain approximately 15 and 25 minutes of speech respectively, as shown in Table 2. There was no speaker-overlap between the training and any of the test sets, and each contained a balance of male and female speakers.

Database	Development test		Evaluation test	
	Speech (mins)	No. of speakers	Speech (mins)	No. of speakers
Afrikaans	15.3	12	24.4	20
English	14.2	10	24.0	18
Xhosa	15.3	10	26.8	17
Zulu	16.8	10	27.1	16

Table 2: Development and evaluation test sets for each database.

The development test set was used for approximate optimisation of the word insertion penalty required during speech recognition, and also for the choice of appropriate learning rates during discriminative training. The evaluation test set was reserved for completely independent testing.

3. Baseline maximum-likelihood system

A set of baseline acoustic models was obtained using the HTK tools [14] and the AST speech corpora. The speech was parameterised as Mel-frequency cepstral coefficients (MFCCs) and their first and second differentials, with cepstral mean normalisation (CMN) applied on a per-utterance basis. Diagonal-covariance speaker-independent cross-word triphone models with three states per model and eight Gaussian mixtures per state were trained using the phonetically-labeled training sets by embedded Baum-Welsh re-estimation and decision-tree state clustering. Phones from different databases with the same IPA symbol were kept distinct by means of database-specific labels. We refer here to “databases” rather than “languages” to avoid confusion with respect to mixed-code utterances, which are especially common in Xhosa and Zulu, and will be given further attention in Section 6. In our experiments, all phones in a particular database were labelled with the language of the database (which corresponds to the speakers’ mother tongue), and not the language of the words themselves. Hence our triphones model code-mixing as it occurs in the training data, but do not allow switching between models of different databases. Silence and speaker noise models were shared among the four databases, and were permitted as left and right triphone contexts but not expanded to triphones themselves. In all experiments reported here, switches between models of different databases across the context-independent silence and speaker noise models were disallowed by appropriate design of the language model.

Test database	Phone error rate (%)	
	Dev test	Eval test
Afrikaans	45.74	41.70
English	46.40	44.21
Xhosa	43.89	49.68
Zulu	49.65	47.80
Average	46.42	45.85

Table 3: Phone error rates (%) of baseline acoustic models.

The resulting HMM model set has a total of 3840 clustered states. Speech recognition performance in terms of phoneme error rates is shown in Table 3. Phoneme loops are used as language models for each database respectively, and hence these figures represent the speech recognition performance achievable when language identification is perfect.

Language identification was performed using a language model consisting of a combined set of four decoupled phone loops. This allows competing recognition hypotheses using phone models from different databases, but prohibits mid-utterance switching. Tables 4 and 5 show the confusion matrices describing the performance of the baseline acoustic models on the development and evaluation test sets.

Test database	Percentage classified as			
	Afrikaans	English	Xhosa	Zulu
Afrikaans	82.3	9.0	4.2	4.6
English	8.7	84.8	3.0	3.5
Xhosa	0.9	3.1	82.3	13.7
Zulu	1.0	4.1	18.7	76.2
Average	81.43%			

Table 4: Language identification accuracy (%) of baseline models on the development test set.

Test database	Percentage classified as			
	Afrikaans	English	Xhosa	Zulu
Afrikaans	87.3	4.4	4.4	3.9
English	8.0	84.5	3.4	4.0
Xhosa	3.3	3.6	78.7	14.4
Zulu	3.4	5.5	25.9	65.2
Average	79.69%			

Table 5: Language identification accuracy (%) of baseline models on the evaluation test set.

From these tables it is evident that Afrikaans and English are identified with the highest accuracy, and that greater confusion occurs between Xhosa and Zulu. This is not surprising since the phonetic composition of these two indigenous languages are known to be strongly related [6].

4. Discriminative training for LID

The acoustic model parameters $\phi_{\mathbf{ml}}$ of the speech recogniser described in the previous section have been optimised by maximizing the posterior probability of the training data, given the reference phonetic or orthographic transcription, as shown in Equation (1).

$$\phi_{\mathbf{ml}} \quad = \quad \operatorname*{argmax}_{\phi\in\Phi} \prod_{u=1}^{U} p_{\phi}(X_u \mid W_u) \tag{1}$$

Here Φ represents the set of all possible HMM parameter values, ϕ a particular set of values for these parameters, U the total number of training utterances, X_u the acoustic observations for the u_{th} utterance, and W_u the known correct transcription for X_u. This process, known as maximum likelihood (ML) parameter estimation, can be performed successfully by means of the Baum-Welch algorithm at a relatively low computational cost.

It is well known that more discriminative objective measures can yield better parameter estimates with respect to recognition performance, for which discrimination among models is more important than model accuracy in the ML sense [12, 8]. However, discriminative training is computationally costly and improvements cannot be guaranteed even with the well-established Extended Baum-Welch algorithm.

We investigate the applicability of discriminative training approaches to the optimisation of acoustic models used for language identification. If we assume that the languages occur with equal probability, and that no language model is used, the discriminative training process for language identification can be viewed as the determination of the optimal HMM parameter values $\phi_{\textbf{disc}}$ according to Equation (2).

$$\begin{aligned} \phi_{\textbf{disc}} &= \operatorname*{argmax}_{\phi\in\Phi} \prod_{u=1}^{U} \frac{p_\phi(X_u \mid L_u)}{p_\phi(X_u)} \\ &= \operatorname*{argmax}_{\phi\in\Phi} \prod_{u=1}^{U} \frac{p_\phi(X_u \mid L_u)}{\sum\limits_{L} p_\phi(X_u \mid L)} \\ &= \operatorname*{argmax}_{\phi\in\Phi} \prod_{u=1}^{U} \frac{p_\phi(X_u \mid L_u)}{\sum\limits_{L\neq L_u} p_\phi(X_u \mid L)} \end{aligned} \qquad (2)$$

Here L_u represents the known correct language for X_u, and $P_\phi(X_u)$ the overall likelihood of the acoustic observations X_u given the current HMM model parameter values ϕ. The quantity $\phi_{\textbf{disc}}$ is the likelihood ratio between the posterior probability of the training data utterances X_u given the labeled language L_u and the unconditional likelihood of X_u. Note that, as indicated in Equation (2), the denominator can be approximated by a sum over all incorrect languages without affecting the optimal parameter values $\phi_{\textbf{disc}}$.

Both the numerator and denominator of Equation (2) should ideally be estimated by summing likelihoods over all possible model sequences **W**, as indicated in Equation (3), where L_W denotes the language of the model sequence W.

$$p_\phi(X_u|L_u) \;=\; \sum_{\substack{W\in\mathbf{W};\\ L_W=L_u}} p_\phi(X_u|L_u,W) \qquad (3)$$

Since this is impractical, the numerator and denominator must be approximated appropriately. Lattices and N-best lists have successfully been used for this purpose in discriminative training for large vocabulary speech recognition [12, 8]. We have followed a different and computationally much simpler approach by approximating numerator and denominator in Equation (2) with a single model sequence. For the numerator, we simply employ the reference model sequence W_u used during ML training.

$$p_\phi(X_u \mid L_u) \;\approx\; p_\phi(X_u \mid W_u) \qquad (4)$$

The denominator of Equation (2) is approximated with the best fitting model sequence, independent of whether it is in the correct language or not, as determined by Viterbi decoding.

$$p_\phi(X_u) \;\approx\; \max_{W} p_\phi(X_u \mid W) \qquad (5)$$

Alternatively, the denominator can be approximated with the best model sequence from an incorrect language, as indicated in Equation (6). This best incorrect model sequence can again be computed by Viterbi decoding using a recognition grammar that simply lacks the respective correct language.

$$p_\phi(X_u) \;\approx\; \max_{\substack{W;\\ L_W\neq L_u}} p_\phi(X_u \mid W) \qquad (6)$$

The baseline maximum-likelihood trained acoustic models described in the previous section were further refined by two iterations of discriminative training using the approximations described above. In both cases, a third iteration yielded no additional improvements on the development test set. Optimization was carried out with the Extended Baum-Welch algorithm as described in [8]. The state dependent tuning parameters (D_s for means and variances and C_s for the mixture weights) were set as proposed in [8] to ensure positive variances and a positive denominator in the mixture weight reestimation. The learning rate h [8] was set conservatively during a small number of preliminary experiments using the development set. The language identification and speech recognition performance respectively of the resulting models is shown in Tables 6 and 7. Baseline performance figures are repeated from Tables 4 and 5.

Model set	Configuration	Train	Dev test	Eval test
HML	Baseline	87.13	81.49	79.69
HD1	Eq.4 & Eq.5	88.58	81.61	79.31
HD2	Eq.4 & Eq.6	87.22	82.68	79.92

Table 6: Average language identification accuracy (%) of baseline (HML) and discriminatively trained (HD1 and HD2) acoustic models on training, development and evaluation sets.

Model	Phone error rate (%)	
set	Dev test	Eval test
HML	46.42	45.85
HD1	46.62	46.08
HD2	46.17	45.97

Table 7: Speech recognition performance of baseline (HML) and discriminatively trained (HD1 and HD2) acoustic models on development and evaluation test sets.

The second line of Tables 6 and 7 shows the performance resulting when the baseline model set is subjected to two iterations of discriminative training using Equations (4) and (5) to approximate numerator and denominator respectively. The last line in the two tables indicates the corresponding performance when approximating numerator and denominator using Equations (4) and (6).

For model set HD1 it is apparent that a considerable improvement in language identification accuracy has been achieved for the training data, but that this improvement does not generalise on the test data sets, where we even observe some performance degradation. However, for model set HD2 we see improvements on both development and evaluation test sets, although the latter is small. The following two sections investigate the results presented in Table 6 in greater detail.

5. Length of test utterances

The average lengths of utterances in the evaluation test sets are given in Table 8. Histograms indicating the distribution of lengths are shown in Figure 5.

Database	Average Length (sec)
Afrikaans	2.0
English	2.1
Xhosa	2.6
Zulu	2.6

Table 8: Average length (in seconds) of test utterances for each language.

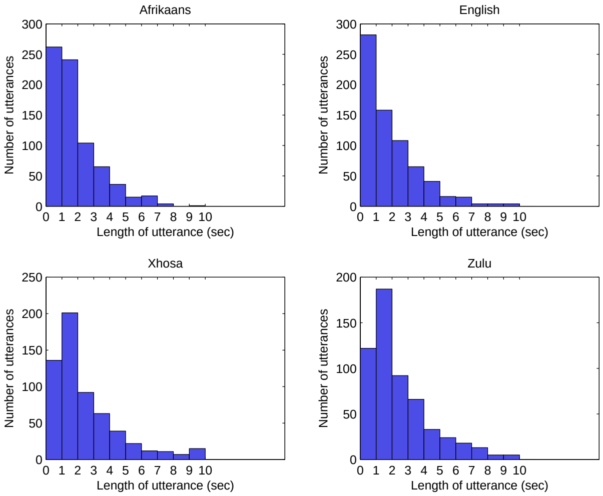

Figure 5: Distribution of test utterance lengths (in seconds) for each language.

From Table 8 and Figure 5 we see that the utterances in our test set are mostly fairly short. In fact, the average utterance length is considerably shorter than that employed in the NIST language identification evaluation, where different experiments use 3s, 10s and 45s speech segments [10].

To investigate the effect the utterance length has on the language identification classification accuracy, we divided the test utterances into two sets according to whether each utterance exceeded the average length or not. The language identification performance for each group is shown in Table 9.

HMM	Shorter utterances		Longer utterances	
set	Dev test	Eval test	Dev test	Eval test
HML	79.42	79.36	84.51	80.19
HD1	79.73	78.79	84.51	80.09
HD2	81.29	79.55	84.98	80.28

Table 9: Language identification accuracy (%) of baseline and discriminatively trained models for long and short utterances.

The figures in Table 9 confirm that, on average, language identification is more accurate for longer utterances than for shorter utterances. Furthermore, the acoustic model HD2 shows consistent improvements over the baseline HML for short as well as long utterances, over both development and evaluation test sets. Finally, it appears as if the discriminative training has narrowed the performance gap between short and long utterances in the case of HD2. This is consistent with the aim of discriminative training to focus on the most confusable utterances, in this case the shorter ones.

6. Code mixing in Xhosa and Zulu

It is accepted practice in several African languages, including modern Xhosa and Zulu, to cite numbers, dates and amounts in either the mother tongue or in English. This occurs because the English alternative is often much shorter. In Xhosa, for instance, the item "2353" is often read simply as "*Two thousand three hundred and fifty three*". However, it could also be read as "*Amawaku amabini namakhulu amathathu namashumi amahlanu nantathu*", meaning literally: "*Thousands-that-are-two and hundreds-that-are-three and tens-that-are-five and three*". Code-mixing is also likely to appear in the spontaneous citing of dates and times. For example, a Zulu-speaking person might cite the time as "*Isikhathi manje u-five past ten*", meaning literally: "*The time now is five past ten*" [7].

Code-mixing occurs frequently in the Xhosa and the Zulu databases. In addition, recent research has shown that the English accents of Xhosa and Zulu mother-tongue speakers are very similar. We may therefore expect the language identification performance to deteriorate for these languages when code switching occurs. To test whether this is indeed so, we divided the Xhosa and Zulu test-sets into two subsets: one consisting of all utterances containing 75% or more English words (Table 10) and the other of the remainder.

Database	% code-mixed test utterances.
Xhosa	27.9%
Zulu	31.2%

Table 10: Percentage of utterances in the Xhosa and Zulu test sets containing 75% or more English words.

In order to investigate whether Xhosa and Zulu utterances with a high proportion of English words are more difficult to classify, we divided their test sets according to Table 10. The language identification accuracies for these partitions of the test sets are shown in Table 11.

HMM	Code-mixed utterances		Remaining utterances	
set	Dev test	Eval test	Dev test	Eval test
HML	77.78	70.57	81.91	81.08
HD1	77.78	69.43	82.13	80.82
HD2	80.42	69.70	83.05	81.39

Table 11: Language identification accuracy (%) of baseline and discriminatively trained models for code-mixed utterances.

The figures in Table 11 confirm that the language identification is in general less accurate for code-mixed utterances than for utterances containing few or no English words. Furthermore, for HD2 discriminative training improved language identification accuracy for both development and evaluation test sets for utterances containing at least a few mother tongue words, while it leads to performance deterioration on the evaluation set for code-mixed utterances. We believe that this is due to the similarity with which Xhosa and Zulu speakers pronounce English words [1]. It is often not even possible for human subjects to differentiate these accents successfully. Hence, when dealing with code-mixed training utterances, the discriminative training approach may attempt to differentiate these utterances on the grounds of speaker or channel differences, which will consequently not generalise to the test sets.

7. Discussion and conclusions

In this paper, we have demonstrated the performance of integrated speech recognition and language identification systems for the South African languages Afrikaans, English, Xhosa and Zulu. Furthermore we have demonstrated that language identification performance can be improved by means of discriminative training without strongly affecting speech recognition performance. However, overall, the gain from discriminative training is small and considering its enormous computational cost, it is not clear whether it is worthwhile.

Nevertheless, several factors still remain to be investigated before this conclusion can be reached with finality. Firstly, we have not made use of language models in our experiments. Discriminative training may be more effective when language models are present, since training could then focus to a greater degree on confusable phones or words that are likely to compete during the Viterbi search in different languages, rather than attempt to discriminate between all such units simultaneously. Further experimentation with the inclusion of language models is therefore warranted. Secondly, due to computational and time constraints, all experiments have been carried out using a fairly narrow beam during Viterbi decoding. The effect of this on the success of the discriminative training steps is not clear at this stage. Finally, more accurate approximations to the numerator and denominator of Equation 2 may result in more substantial performance gains.

8. Acknowledgements

This work was supported by the National Research Foundation (NRF) under grant number FA2005022300010.

9. References

[1] F. De Wet, T.R. Niesler, and P.H. Louw. Nguni and Sotho varieties of South African English - distant cousins or twins? In *ISCA ITRW on Multilingual Speech Processing (MULTILING)*, Stellenbosch, South Africa, 2006.

[2] F. Fernández, R. de Córdoba, J. Ferreiros, V. Sama, L.F. D'Haro, and J. Macías-Guarasa. Language identification techniques based on full recognition in an air traffic control task. In *Proc. ICSLP*, Jeju, Korea, 2004.

[3] S. Herry, B. Gas, C. Sedogbo, and J-L. Zarader. Language detection by neural discrimination. In *Proc. ICSLP*, Jeju, Korea, 2004.

[4] J.L. Hieronymus and S. Kadambe. Robust spoken language identification using large vocabulary speech recognition. In *Proc. ICASSP*, pages 779–782, Munich, Germany, 1997.

[5] S. Mendoza, L. Gillick, Y. Ito, S. Lowe, and M. Newmann. Automatic language identification using large vocabulary continuous speech recognition. In *Proc. ICASSP*, Atlanta, USA, 1996.

[6] T.R. Niesler, P.H. Louw, and J.C. Roux. Phonetic analysis of Afrikaans, English, Xhosa and Zulu using South African speech databases. *Southern African Linguistics and Applied Language Studies*, 23(4):459–474, 2005.

[7] J.C. Roux, P.H. Louw, and T.R. Niesler. The African Speech Technology project: An assessment. In *Proc. LREC*, Lisbon, Portugal, 2004.

[8] R. Schlüter, W. Macherey, B. Müller, and H. Ney. Comparison of discriminative training criteria and optimization methods for speech recognition. *Speech Communication*, 34(3):287–310, 2001.

[9] T Schultz, Rogiana I., and Waibel A. LVCSR-based language identification. In *Proc. ICASSP*, Atlanta, USA, 1996.

[10] E Singer, P.A. Torres-Carrasquillo, T.P. Gleason, W.M. Campbell, and D.A. Reynolds. Acoustic, phonetic, and discriminative approaches to automatic language identification. In *Proc. Eurospeech*, pages 1345–1348, Geneva, Switzerland, 2003.

[11] Statistics South Africa, editor. *Census 2001: Primary tables South Africa: Census 1996 and 2001 compared*. Statistics South Africa, 2004.

[12] V. Valtchev, J.J. Odell, P.C. Woodland, and S.J. Young. MMIE training of large vocabulary recognition systems. *Speech Communication*, 22(4):303–314, 1997.

[13] K-k. Wong and M-h. Siu. Automatic language identification using discrete hidden markov model. In *Proc. ICSLP*, Jeju, Korea, 2004.

[14] S. Young, G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland. *The HTK book, version 3.2.1*. Cambridge University Engineering Department, 2002.

[15] M.A. Zissman and K.M. Berkling. Automatic language identification. *Speech Communication*, 35:115–124, 2001.