DEVELOPMENT OF A SPOKEN LANGUAGE IDENTIFICATION SYSTEM FOR SOUTH AFRICAN LANGUAGES

M. Peché[*] **, M.H. Davel**[†] **and E. Barnard** [‡]

[*] *Department of Electrical, Electronic and Computer Engineering, University of Pretoria. E-mail: mariuspeche@gmail.com*

[†] *HLT Research Group, Meraka Institute, CSIR. E-mail: mdavel@csir.co.za*

[‡] *HLT Research Group, Meraka Institute, CSIR. E-mail: ebarnard@csir.co.za*

Abstract: This article introduces the first Spoken Language Identification system developed to distinguish among all eleven of South Africa's official languages. The PPR-LM (Parallel Phoneme Recognition followed by Language Modeling) architecture is implemented, and techniques such as phoneme frequency filtering, which aims to utilize the available training data to maximum efficiency, are utilized. The system performs reasonably well, achieving an overall accuracy of 71.72% on test samples of three to ten seconds in length. This accuracy improves when the predicted results are combined into language families, reaching an overall accuracy of 82.39%

Key words: Spoken Language Identification, Parallel Phoneme Recognition followed by Language Modeling, South African languages.

1. INTRODUCTION

Spoken Language Identification (S-LID) is a process whereby the most probable language of a segment of audio speech is determined. This choice is made from a set of possible target languages, be it a closed set where all possibilities are known or an open set with unknown languages included in the test corpora as well. S-LID is a difficult task since the process to identify and extract meaningful tokens (from the audio) upon which a decision can be made is itself prone to errors.

This article describes the process followed to create an S-LID system that is able to distinguish among all eleven official languages of South Africa. The accuracy of the system, as well as techniques used to improve on initial baseline performance are reported. In addition, the effect of clearly defined language families on the accuracy of the system is investigated.

The article is structured as follows: Section 2 provides an overview of the background to the S-LID task and the specific challenges encountered in the South African environment. A short description of the general system design and data sets used follows in Section 3. Section 4 discusses the research approach, Section 5 details the results obtained on a per language basis and Section 6 investigates the role of language families. The article is concluded in Section 7 with some suggestions for further optimization.

2. BACKGROUND

In this section, the background to the S-LID task is discussed in greater detail, specifically focusing on the following areas: Section 2.1 provides a brief overview of the S-LID task; variables that influence the accuracy of S-LID systems is examined in Section 2.2 with the formulas to measure the system performance given in Section 2.3; and Section 2.4 takes a more in-depth look at the Parallel Phoneme Recognition followed by Language Modeling (PPR-LM) approach to S-LID. Furthermore, background is also provided on the official languages of South Africa (in Section 2.5) and possibilities for a South African S-LID system are examined in Section 2.6.

2.1 S-LID task overview

A segment of audio speech has several features that can differ from language to language. These may be used in different S-LID system designs, each with varying complexities and results. Prosodic information, which includes factors such as rhythm and intonation, was an early focus of S-LID research [3], as was spectral information [11], which characterizes an utterance in terms of its spectral content. Other approaches include reference sounds [5] and other raw waveform features [4]. These features represent a low level of linguistic knowledge (limited or no language-specific information is required), and the systems utilizing these features are typically simple in design.

In addition, systems that use language-specific information (such as syntax or semantics) have been developed, and it has been conjectured that there is a correlation with the level of knowledge presented within the extracted features and the results of the system utilizing those features [9]. However, systems with a higher linguistic knowledge representation have also proven more difficult to design, requiring a more complex architecture and greater computational power. This in turn implies that more accurate systems require more labeled training data. Therefore most researchers prefer to utilize acoustic resources [12], especially phonetic tokens, as is the case with the PPR-LM approach, which is discussed later. Variations of the PPR-LM approach provide state-of-the-art accuracies and require no additional linguistic knowledge apart from what can be deduced from large labeled audio corpora.

2.2 Variables that influence S-LID accuracy

Related work has identified a number of factors which play an important role in the classification of a complex syntax that has been constructed from a set of distinct tokens, such as natural language. These factors include the following:

- The number of tokens available in a sample used for testing [12].
- The amount of available training data [8].
- The classification algorithm [7].
- The level of similarity of the target languages [2].

The composition of the target languages is of particular importance in this article, as it is much harder to distinguish between related languages such as isiZulu and isiXhosa than between two unrelated languages such as isiZulu and English [2]. The number of target languages also has a significant influence on overall system accuracy, as researchers have been able to achieve much better results using language-pair recognition than trying to recognize between several languages at once [11]. The use of an open test set instead of a closed test set also has a negative effect on system accuracy, and complicates the design of the system as a whole.

Current benchmark results are established by the National Institute of Standards and Technology (NIST) Language Recognition Evaluation (LRE) [12]. Initially started in 1996, the next evaluation was in 2003, after which the evaluation has been repeated every two years. The NIST-LRE measures system achievements based on pair-wise language recognition performance. A score according to the probability that the system incorrectly classifies an audio segment is calculated for each target-non-target language pair. The average of these scores (C_{AVG}) then represents the final system performance.

Figure 1 expresses the historic scores achieved across three different test sample lengths. Note how the longer segments outperform those of a much shorter length.

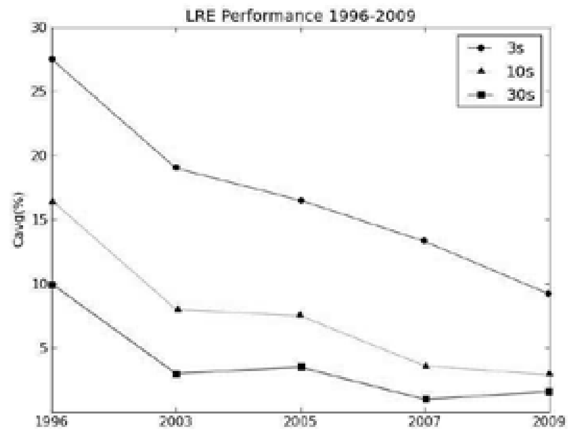

Figure 1: Historic scores for the 1996 - 2009 NIST LRE in the general language recognition task. Image has been reproduced from [12].

2.3 Measuring the system performance

The system developed for this article does not function on a language-pair basis, and therefore this specific measurement will not be used for the remainder of the article. Instead, the performance of the system as a whole is measured by evaluating both the front-end and the back-end separately.

The Automatic Speech Recognition (ASR) systems used for phoneme recognition in the front-end are evaluated using both the phoneme recognition accuracy and phoneme correctness. Accuracy and correctness are defined as follows:

$$accuracy = \frac{N-D-S-I}{N} * 100\% \quad (1)$$

$$correctness = \frac{N-D-S}{N} * 100\% \quad (2)$$

where N is the total number of labels, D is the number of deletion errors, S is the number of substitution errors, and I is the number of insertion errors.

As for the back-end classifiers, the overall accuracy of the S-LID system, as well as the precision and recall for each language is reported. The overall accuracy is simply the percentage of all utterances correctly identified by the classifier. Precision and recall scores of a specific language l are defined as follows:

$$precision = \left(\frac{l_{correct}}{l_{correct}+l_{incorrect}}\right) * 100\% \quad (3)$$

$$recall = \left(\frac{l_{correct}}{l_{correct}+O_{incorrect}}\right) * 100\% \quad (4)$$

where $l_{correct}$ is the number of utterances correctly classified, and $l_{incorrect}$ and $O_{incorrect}$ represents the number of false accepts and false rejects respectively.

2.4 The PPR-LM approach to S-LID

PPR-LM is an S-LID system configuration whereby the audio data is first processed into a phoneme string before the actual classification is performed [8]. The first step (referred to as the front-end) extracts phonotactic information in the form of phonemic tokens from the audio signal. The resulting string of tokens is then passed to a back-end where some form of language modeling (for example n-grams) is used to determine the most probable language from the set of target languages [6].

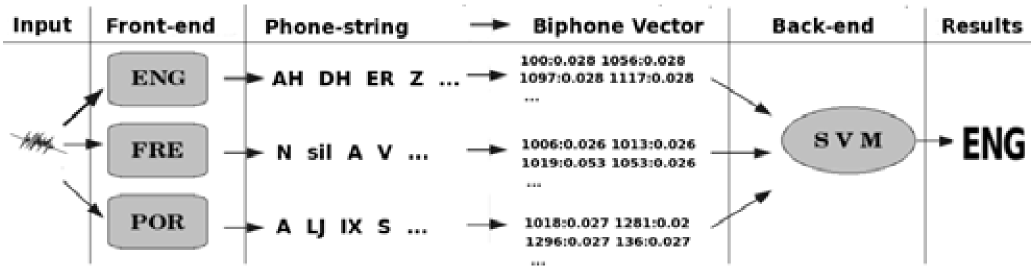

Figure 2: Visual representation of the PPR-LM architecture.

The front-end usually consists of the phonemic recognizer modules of a number of ASR systems, as the above-mentioned tokens are commonly language-dependent phonemes. Although PPR-LM systems can function with a tokenizer trained in only one of the target languages [11], tokenizers for several of the target languages that process the audio in parallel seem to be the most accurate system configuratio [7]. In such a case, the typical PPR-LM system usually utilizes one tokenizer for each of the target languages.

Once the audio signal has been processed by the front-end, the resulting token strings are scored by a classifie in the back-end and the language with the highest probability score is returned. Language models are usually employed to distinguish among languages, although the use of a Support Vector Machine (SVM) [6] has proved to be more successful [7]. The SVM requires that n-gram frequencies are extracted from the phoneme strings and represented as a point in a high-dimensional vector-space. Languages are then represented as groups of vectors.

Figure 2 provides a visual representation of the PPR-LM architecture. An utterance given as input to the system is passed to three ASR systems (English, French and Portuguese phoneme recognizers in the image) that together form the front-end of the system. These ASR systems produce phoneme strings which are then passed as a vector of biphone frequencies to the language model at the back-end (an SVM classifie in the image) which then predicts the language spoken in the utterance.

2.5 Languages of South Africa

Since 1994, South Africa has recognized eleven officia languages. These are listed in Table 1, along with each language's international ISO language code, the number of native speakers (in millions) and the language family to which it belongs. As can be seen from Table 1, several of South Africa's officia languages do not have a large speaker population, which makes the gathering of audio resources difficult

Of particular importance to this article is the language families. These families represent languages which exhibit similarities with regard to grammar, vocabulary and pronunciation. The Germanic languages are of Indo-European origin, but both the other two major families, the Nguni and the Sotho-Tswana families, represent two of the major branches of the Southern Bantu languages which originated in Central to Southern Africa. Tswa-Ronga and Venda are also classifie as being part of the Southern Bantu languages, though they fall into families of their own. [16]

Table 1: South Africa's eleven officia languages [16].

Language	ISO code	Native speakers	Language family
isiZulu	zul	10.7	Nguni
isiXhosa	xho	7.9	Nguni
Afrikaans	afr	6.0	Germanic
Sepedi	nso	4.2	Sotho-Tswana
Setswana	tsn	3.7	Sotho-Tswana
Sesotho	sot	3.6	Sotho-Tswana
SA English	eng	3.6	Germanic
Xitsonga	tso	2.0	Tswa-Ronga
Siswati	ssw	1.2	Nguni
Tshivenda	ven	1.0	Venda
isiNdebele	nbl	0.7	Nguni

It should be noted that though Afrikaans and English are both Germanic languages, they are subdivided into Low Franconian and Anglo-Frisian respectively, and will be treated as different families later in the article.

2.6 S-LID for South African languages

Currently, there is no existing S-LID system that can distinguish among all eleven of South Africa's officia languages. Prior work has been done in the related fiel of Textual Language Identificatio (T-LID) which has shown promise [2]. However, T-LID has already achieved impressive results on relative small test sets as early as 1994, and is therefore considered to be mostly a solved problem [13].

On the other hand, developing a S-LID system to distinguish among all South African languages will prove to be quite a challenge. Referring to Section 2.1, this environment poses the following two main difficulties

- Eleven target languages, and
- Closely related language families.

Another important obstacle which has to be overcome is the availability of high quality audio data. The linguistic resources available for South African languages (such as large annotated speech corpora) are extremely limited in comparison to the resources available for the major languages of the word. Section 3.2 gives more information on the corpus which is used for the development of the current system. Figure 3 should be given particular attention, as the very short lengths of the available utterances will also impact negatively on the system's performance.

3. SYSTEM DESIGN

This section describes the design of the South African S-LID system. A more detailed description of the system architecture follows in Section 3.1 and the corpus used to develop the ASR recognizers and the S-LID back-end is described in Section 3.2.

3.1 System architecture

The South African S-LID system implements the popular PPR-LM architecture, as described in Section 2.3. Here as well, the phoneme recognizers of ASR systems are used as tokenizers to extract language-dependent phonemes from the audio signal.

These phoneme recognizers utilize context-dependent Hidden Markov Models (HMMs), which consist of three emitting states with Gaussian Mixture Models (GMMs) of seven mixtures within each state. The HMMs are trained using the Hidden Markov Model Toolkit (HTK) [10]. 13 Mel Frequency Cepstral Coefficients (MFCCs), their 13 delta and 13 acceleration values are used as features, resulting in a 39-dimensional feature vector. Cepstral Mean Normalization (CMN) as well as Cepstral Variance Normalization (CVN) are used as feature-domain channel normalization techniques. Semi-tied transforms are applied to the HMMs and a flat phone-based language model is employed for phone recognition [1]. Optimal insertion penalties are estimated by balancing insertions and deletions during recognition.

Per training sample, biphone frequencies are extracted from the phoneme strings of each phoneme recognizer and formatted into a vector, with each unique biphone representing a term. (Most of these terms would be zero.) The resulting vectors (one per tokenizer) are then concatenated one after the other to form a single vector for each utterance. This vector is used as an input to the SVM that serves as back-end classifier for the system. A grid search is used to optimize the main SVM parameters (margin-error trade-off parameter and kernel width) prior to classification.

3.2 Corpus statistics

The South African S-LID system is trained and tested with the Lwazi corpus, which consists of telephonic audio data as well as transcriptions, collected in all eleven of South Africa's official languages [1]. For the development of the South African S-LID system, all eleven languages in the corpus are utilized.

Table 2: Number of speakers (Spkrs) and utterances (Utts) in the training and test set of each language, as well as the duration of each set in hours.

Language	Set	Spkrs	Utts	Duration
Afrikaans	Train	170	4382	3.32
	Test	30	787	0.60
SA English	Train	175	4287	3.25
	Test	30	728	0.55
isiNdebele	Train	170	4878	3.69
	Test	30	846	0.64
isiZulu	Train	171	4852	3.23
	Test	29	685	0.52
isiXhosa	Train	180	4458	3.38
	Test	30	669	0.51
Sepedi	Train	169	3674	2.78
	Test	30	664	0.50
Sesotho	Train	170	4642	3.52
	Test	30	815	0.62
Setswana	Train	176	4257	3.10
	Test	30	686	0.54
Siswati	Train	178	4778	3.62
	Test	30	811	0.62
Tshivenda	Train	171	4414	3.34
	Test	30	770	0.58
Xitsonga	Train	168	4230	3.21
	Test	30	755	0.57

Table 2 provides statistics on the available data. The number of speakers per language is given, as well as the combined number of utterances and the length of the audio data in hours. Table 2 also provides the number of speakers and associated utterances in the training set, which is used to train both the ASR systems in the front-end and and the SVM at the back-end. As can be seen, a test set of about 15% the total size of a language's data is kept aside for validation purposes. Care has been taken to ensure that a speaker is not represented in both the train and test sets.

Figure 3 displays the distribution of the different utterance lengths across all eleven languages available in the corpus. Note that a large percentage of the utterances are extremely short (less than 6 seconds in duration).

4. GENERAL DISCUSSION OF THE APPROACH

The first step in the development of the proposed system is to establish baseline performance. This is done by developing a system which implements the PPR-LM architecture, as described previously. The baseline system utilizes phoneme recognizers from all eleven official languages of South Africa. The system results are given in Section 5.1

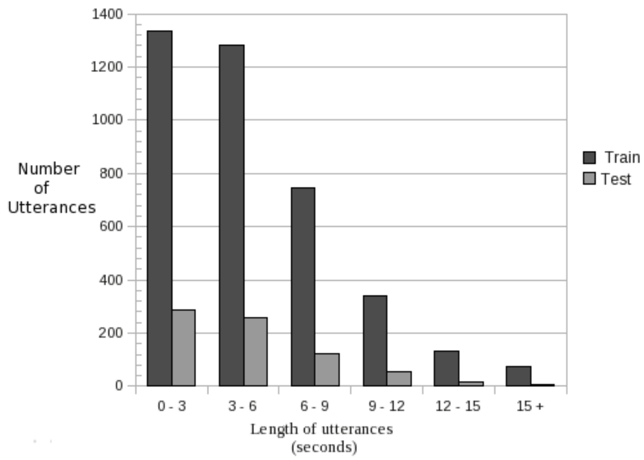

Figure 3: Distribution of utterance lengths in the Lwazi corpus. The number of training and test utterances are displayed separately.

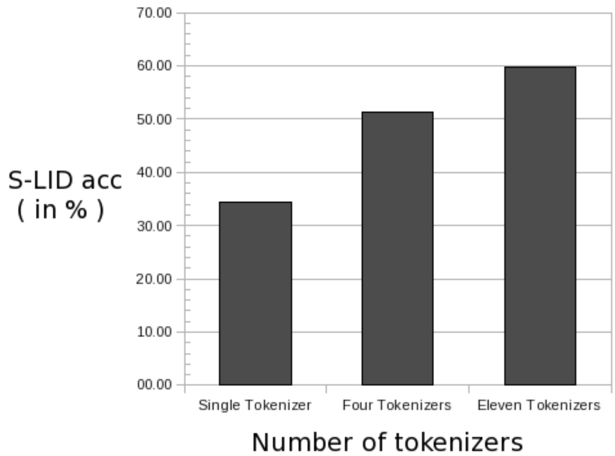

Figure 4: Overall SVM performance of all three configurations for the South African S-LID systems.

Two improvements are then implemented:

- The Lwazi corpus, as described in Section 3.2, consists of telephonic audio data which is not as clean as audio recorded under laboratory conditions. (Channel conditions, background noise and speaker disfluencies all affect the quality of the available audio.) Problematic utterances can have a negative influence on the overall accuracy of the system; therefore all utterances are automatically filtered in order to refine the system [14].
- Figure 3 displays the unbalanced nature of the Lwazi corpus when the lengths of the utterances are compared. This may also have a negative impact on the system as a whole, especially when it is considered that the length of the utterances is an important variable which influences the system accuracy (referring to Section 2.2). Therefore the training data is further refined, and all utterances shorter than three seconds are removed from the training set.

Both the baseline system and the refined system are tested on two test sets, one with all utterances three seconds and shorter, and the other containing only utterances longer then three seconds. Results are discussed in Section 5.2

5. INDIVIDUAL LANGUAGE CLASSIFICATION

5.1 S-LID baseline results

Section 4 describes the system which provides the baseline results against which the experiments in the rest of the article are compared. Figure 4 shows that an increasing number of tokenizers results in an increase in performance, as have previously been shown [15]. Therefore, even though it utilizes significant computational resources, the best performing system – with eleven tokenizers – serves as the baseline for the experiments in this article.

As predicted in Section 2.2, the longer utterances perform better than the shorter samples, achieving an overall accuracy of 69.89%, compared to the 56.60% achieved on the shorter utterances. Figure 5 displays the results graphically, with the system performance achieved for the test samples shorter and longer than three seconds plotted separately.

Table 3 summarizes the performance of the South African S-LID system on the test samples longer than three seconds in more detail. The accuracy and correctness of the phoneme recognizers as well as the precision and recall for each language are also given. Note that, since the length of the test sample does not influence the performance of the phoneme recognizers, the front-end results presented in Table 3 were generated on the entire test set.

5.2 Removing problematic utterances

In order to improve the South African S-LID system, the data used to train the SVM is filtered so that only utterances that have recorded a phoneme frequency of higher than three phonemes per second, and utterances longer than three seconds in length are selected. (The specific values of these variables were determined during prior experimentation.) Furthermore, the number of training utterances available from each language is limited to the same number in order to ensure that the SVM is not biased towards a particular language. A new, refined system is now developed, using this new subset of training data and the same training process as described above.

The data set used for testing is also filtered in a similar fashion as described above, but then divided into two subsets: one containing utterances less than three seconds long, while the other contains all the utterances longer than three seconds. The performance of the refined system

Table 3: The performance of the ASR systems in the front-end as well as the SVM classifier in the back-end of the initial South African S-LID (baseline) system.

	afr	eng	nbl	nso	sot	tsn	ssw	ven	tso	xho	zul
	ASR Front-end										
% Correctness	70.49	58.72	73.02	68.00	67.76	70.64	74.04	75.76	68.53	68.54	69.81
% Accuracy	65.47	52.30	66.61	57.78	57.17	57.08	65.77	67.37	60.92	58.57	63.06
	SVM Back-end										
% Precision	75.90	82.59	64.47	64.09	57.76	66.12	63.15	69.45	68.27	61.42	50.90
% Recall	81.86	85.18	70.24	51.67	58.94	66.27	73.17	68.79	54.27	56.41	54.42
	Overall system accuracy : 69.89%										

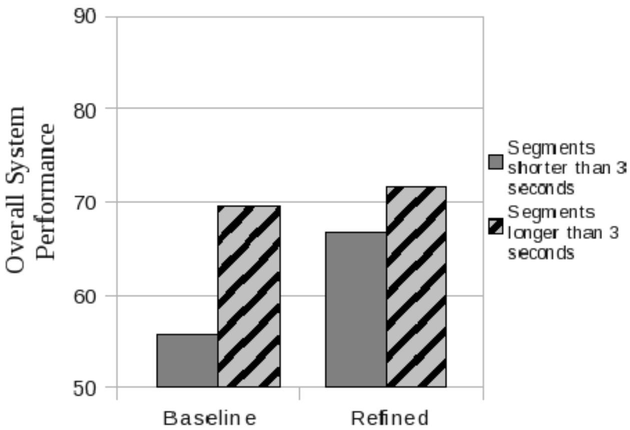

Figure 5: The results of both the baseline system and the refined system, with the overall system performance achieved for the test samples shorter and longer than three seconds plotted on separate graphs.

described above is then verified individually using both test sets. These new training and test sets are recognized by all eleven phoneme recognizers before the resulting phoneme strings are used to retrain and test the SVM classifier at the back-end.

Both sets of samples used to test the system report an increase in performance, achieving 67.29% and 71.72% for the shorter and longer segments respectively. Figure 5 also displays these results graphically, with the system performance achieved for the test samples shorter and longer than three seconds plotted on separate graphs.

Table 4 describes the performance of the South African S-LID system on the test samples longer than three seconds in more detail. The accuracy and correctness of the phoneme recognizers as well as the precision and recall for each language are given. The front-end results presented in Table 3 were again generated on the entire test set.

6. INVESTIGATING THE EFFECT OF LANGUAGE FAMILIES

The S-LID system developed in Section 5 distinguishes between eleven languages, some of which are closely related. In the light of the discussion in Section 2.2, we are interested to know how well the system performs when those distinctions are not attempted. (For practical applications, this may be acceptable, since several of the languages are mutually intelligible.) This was investigated by combining the results according to language families, instead of representing the languages individually. Doing this reduces the number of target groups the recognizer has to distinguish between, and simplifies the underlying relationships. (The languages with their associated families are listed in Table 1.)

When the ambiguity of closely related target languages is removed, the systems performance increases even more to an overall accuracy of 71.58% and 82.39% on test samples shorter and longer than three seconds respectively. Table 5 details the performance of the SVM classifier on the longer utterances for both the baseline system of Section 5.1 and the refined system of Section 5.2.

7. CONCLUSION

The South African context provides a challenging environment for Spoken Language Identification, for two main reasons: the large number of closely related target languages that occur, as well as the restricted availability of high quality linguistic resources.

This article has described the successful development of an S-LID system which performs reasonably well in differentiating among all the official languages of South Africa. Through the careful implementation of eleven different tokenizers, corpus filtering and corpus balancing a practically usable system was developed, The final system is capable of identifying the language family as well as the exact language spoken with an accuracy of 82% and 72%, respectively, for utterances of three seconds or longer. (With the longer utterances in the order of ten seconds in length.)

Further optimization of the system is currently being considered. Interesting areas of research include analyzing the effect of language family-specific tokenizers, experimenting with longer utterances and chunking and recombining utterances during SVM classification.

Table 4: The performance of the SVM classifier in the back-end of the refined South African S-LID system, when a test set containing only utterances longer than three seconds are used.

	afr	eng	nbl	nso	sot	tsn	ssw	ven	tso	xho	zul
% Precision	73.15	84.42	67.66	63.74	64.37	68.76	62.63	73.28	63.17	67.41	58.58
% Recall	85.80	88.40	73.66	66.35	54.45	69.20	73.96	71.22	57.75	59.60	53.37
	Overall system accuracy : 71.72%										

Table 5: The performance of the SVM classifier in the back-end of both the baseline and refined South African S-LID systems, when only language families are considered.

	Afrikaans	English	Nguni	Sotho-Tswana	Tswa-Ronga	Venda
	Baseline SA S-LID					
% Correctness	75.90	82.59	84.57	85.43	68.27	69.45
% Accuracy	81.86	85.18	90.38	80.97	54.27	68.79
	Overall system accuracy : 81.71%					
	Refined SA S-LID					
% Precision	73.15	84.42	85.92	85.81	67.41	73.28
% Recall	85.80	88.40	88,06	83.02	59.60	71.22
	Overall system accuracy : 82.39%					

8. REFERENCES

[1] Etienne Barnard, Marelie Davel, and Charl van Heerden: "ASR corpus design for resource-scarce languages", *Proceedings: Annual Conference of the International Speech Communication Association (Interspeech)*, Brighton, UK, pp. 2847-2850, Sep 2009.

[2] Gerrit R. Botha and Etienne Barnard: "Factors that affect the accuracy of text-based language identification", *Proceedings: Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Pietermaritzburg, South Africa, pp. 7-10, December 2007.

[3] J.T Foil: "Language identification using noisy speech", Proceedings: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New York, NY, pp.861-864, April 1986.

[4] S.C Kwasny, B.L. Kalman, W. Wu, and A.M. Engebretson: "Identifying language from speech: An example of high-level statistically-based feature extraction", *Proceedings: Annual Conference of the Cognitive Science Society*, Hillsdale, NJ, Vol. 14, pp. 53-57, August 1992.

[5] R.G Leonard and G.R. Doddington: "Automatic language identification", Technical report RADC-TR-74-200. *AirForce Rome Air Development Center*, Aug 1974.

[6] Haizhou Li and Bin Ma: "A Phonotactic Language Model for Spoken Language Identification", *Proceedings: Meeting of the Ass for Computational Linguistics,* Vol. 43, pp. 515-522, June 2007.

[7] Bin Ma and Haizhou Li: "A Comparative Study of Four Language Identification Systems", *Computational Linguistics and Chinese Language Processing*, Vol. 11 No. 2, pp. 80-985, Jan 2006.

[8] Yeshwant K. Muthusamy, Ethienne Barnard, and Ronald A. Cole: "Reviewing Automatic Language Identification", *IEEE Signal Processing Magazine*, October 1994.

[9] Rong Tong, Bin Ma, Donglai Zhu, Haizhou Li, and Eng Siong Chng: "Integrating Acoustic, Prosodic and Phonotactic Features for Spoken Language Identification", Proceedings: *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toulouse, France, pp. 205-208, May 2006.

[10] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, Valtcho Valtchev, and Phil Woodland: "The HTK book. Revised for HTK version 3.3", *Online: http://htk.eng.cam.ac.uk/*., 2005.

[11] M.A. Zissman: "Comparison of four approaches to automatic language identification of telephone speech", *IEEE Transactions on Speech and Audio Processing*, Vol. 4 No. 1, pp. 31-44, January 1996.

[12] NIST Website: "The 2009 NIST language recognition evaluation results", *http://www.nist.gov/speech/tests/lre/2009/ lre09 eval results vFINAL/index.html,* 2008.

[13] W. Cavnar and J. M. Trenkle: "N-gram based text categorization.", Proceedings: *Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, NV, Vol. 3, pp. 161-169, April 1994.

[14] Marius Pech´e, Marelie Davel, and Etienne Barnard: "Porting a spoken language identification system to a new environment", *Proceedings: Annual Symposium of the Pattern Recognition Association of South Africa (PRASA)*, Cape Town, South Africa, pp. 58-62, December 2008.

[15] M. Peché, M. Davel, and E. Barnard: "Phonotactic spoken language identification with limited training data", *Proceedings: Annual Conference of the International Speech Communication Association*, Antwerp, Belgium, pp. 1537-1540, Aug 2007.

[16] Pali Lehohla: "Census 2001", *Census in brief, Statistics South Africa*, 2003.