MODELING CODE-SWITCHING SPEECH ON UNDER-RESOURCED LANGUAGES FOR LANGUAGE IDENTIFICATION

Koena Ronny Mabokela, Madimetja Jonas Manamela, Mabu Manaileng

Telkom Centre of Excellence for Speech Technology
Department of Computer Science, University of Limpopo (Turfloop Campus)
Private Bag, X1106, Sovenga, 0727, Polokwane, South Africa
{krmabokela, manailengmj}@gmail.com, jonas.manamela@ul.ac.za

ABSTRACT

This paper presents an integration of phonotactic information to perform language identification (LID) in a mixed-language speech. A single-pass front-end recognition system is employed to convert the spoken utterances into a statistical occurrence of phone sequences. To process such phone sequences, a hidden Markov model (HMM) is utilized to build robust acoustic models that can handle multiple languages within an utterance. A supervised Support Vector Machine (SVM) learns the language transition of the phonotactic information given the recognized phone sequences. The back-end SVM-based decision classifies language identity given the likelihood scores phone occurrences. The experiments are conducted on commonly mixed-language Northern Sotho and English speech utterances. We evaluate the system measuring the performance of the phone recognition and LID portions separately. We obtained a phone error rate of 15.7% when a data-driven phoneme mapping approach is modeled with 16 Gaussian mixtures per state. However, the proposed integrated LID system has achieved a considerable performance with an acceptable LID accuracy of 85.0% and average of 81% on code-switched speech and monolingual speech segments respectively.

Index Terms—Code-switching speech, under-resourced languages, phonotactic information, acoustic models, language model

1. INTRODUCTION

Currently, most multilingual speakers tend to show a tendency for engaging in code-switching – a mixed-language phenomenon that is referred to as the usage of more than one language within an utterance. This style of speaking appears to be preferred commonly in multilingual societies [1]. South Africa is a multilingual society with eleven official languages where code-switching appears to be an acceptable modern-day style of communication. However, the African reality in many communication episodes is that, English is frequently mixed with indigenous under-resourced official languages (e.g., Radio Broadcast, TV Shows) [2]. For most African indigenous under-resourced language speakers, numeric data such as *date items, personal identification codes, time representations*, are not usually spoken in their first language but the preference is the English language [3]. Although code-switching speech is commonly more spoken than formally written, it is necessary to find large orthographically well-constructed text data sets for building suitable language models [1]. In addition, multilingual language models allow the exploitation of multilingual pronunciation dictionaries. Code-switching requires large amount speech corpus to develop suitable context-dependent acoustic models. Code-switching speech has significant influence on automatic speech recognition (ASR) systems. As a consequence, the acoustic models, pronunciation models and language models need to be redesigned in order to accommodate words from different languages [2]. It is a far greater challenge for ASR system to accurately handle code-switched utterances without robust acoustic models and pronunciation models, and language models. Therefore, it is highly plausible to classify code-switched speech itself into the same category as under-resourced languages due lack of speech technology resources for developing accurate ASR systems [1, 3].

The language identification (LID) system is an enabling technology for a wide range of multilingual speech processing applications, such as routing telephone calls to human operators, more particular for handling emergency calls [4, 5]. It is a relevant study since it is common in South Africa for more than one language to be spoken in the same region. In this paper we propose LID system that is integrated with speech recognizer to identify code-switched speech between Northern Sotho (also known as Sepedi) and English. Although we report only on experiments conducted using two official languages of South Africa, the same procedure can still be followed on other under-resourced

languages. To date, there are two approaches to handle a code-switched speech recognition system [1, 6].

a) The first approach employs two monolingual speech recognition systems and a LID module. The LID module extract the input code-switched utterances then decide on the identity of each of the speech segments before passing them into their respective monolingual ASR system.

b) The second approach use a multilingual ASR system comprising of a multilingual acoustic model of the languages concerned, a multilingual pronunciation dictionary which combines the words from targeted languages, and a multilingual language model that allows mixing of different language units.

In this paper we are interested in the second approach since it is the most typical approach to deal with code-switched utterances. The first approach has many disadvantages which lead to it not being preferred by many researchers [6]. As for LID systems, identifying segmented speech is also a challenge. For this reason, a single-pass ASR system is employed to decode code-switched speech utterances. We trained the front-end ASR system on both Northern Sotho and English speech data and model code-switching at the pronunciation level. We explicitly model the English words within the code-switched utterances by adopting a linguistically-motivated and data-driven phoneme mapping methods to develop suitable acoustic models of the target languages. The pronunciation modeling provides a convenient tactic of dealing with out-of-language words [2]. The LID module that we used is the back-end component which gets the decoded phoneme strings from the front-end ASR system to perform language identification [4, 7]. Our acoustic models are trained on the training data set, and the SVM-based classifier is evaluated on the phoneme strings recognized from the same set. We engaged a supervised SVM-based classifier that learns the language transitions of the phonotactic information given the recognized phone sequences. The back-end SVM-based decision classifies language identity given the likelihood scores phone occurrences. The proposed integrated approach is more like parallel phone recognition followed by language model (P-PRLM) approach and has been shown to reliably perform well [4, 5, 7].

This paper is organized as follows: the next section describes the integrated LID system. Section 3 describes the phoneme mapping approaches used to derive suitable phoneme sets. In section 4, we describe a mixed-language speech corpus which is used for experimentation. Section 5 presents the baseline experimental setup. The experimental results are discussed in Section 6 and conclusions are finally drawn in Section 7.

2. AN INTEGRATED LID SYSTEM

This section describes our proposed language identification system, which integrate the acoustic units and phonotactic information to perform language identification on mixed speech utterances. Our integrated LID system is intended to identify only two languages, i.e., Northern Sotho and English, on code-switched speech utterances. In this section we first describe the phone recognition system utilizing acoustic features and describe the language classification based on the phonotactic information. Although we discuss the system phases separately, the overall proposed integrated LID system forms one unit.

2.1. Front-end Phone Recognition System

Fig. 1 shows the front-end of the phone recognition system designed to decode mixed-language speech utterances. A phone recognition system takes speech waveform and output the corresponding phone sequences. This is done when a phone recognition system estimates the likelihood score of the optimal phone sequences given the acoustic features extracted from the speech utterance waveform. We assume that the speech waveform can be segmented into a sequence of phones. To achieve this, a phone *n*-gram language model is employed to estimate the likelihood score of the *n*th phone given the (*n*-1) of the preceding phones.

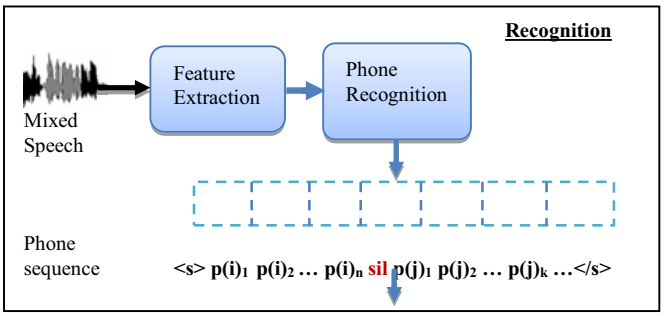

Fig. 1: Front-end phone-based recognition on code-switched speech

A *Baum-Welch* iteration algorithm is used during training of acoustic models to perform HMM-based parameter re-estimation. For the recognition purpose, the acoustic features are compared with the HMM-based acoustic models as well as the phone language model. The sequences of phone strings are decoded by the *Viterbi* decoding algorithm which searches the optimal sequence of the phones using the combined likelihood scores from the acoustic model and phone language model.

2.2. Back-end SVM-based Classification System

The SVM-based classifier is used to identify only two class feature samples, languages outside the targeted range will

not be classified. For each phone sequence generated from the phone recognition, the bi-phone occurrences are extracted from the phone sequences and converted into a suitable SVM format with a unique numerical representation. This approach is similar to vector space modeling [4]. The LID is performed by using SVM-based classifier to score the phoneme sequence of a test utterance. The language model with the highest log likelihood score is chosen to be the most likely sample for classification.

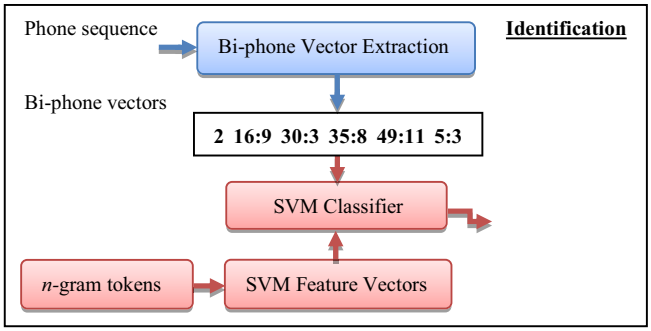

Fig. 2: Back-end scheme designed for language identification on code-switched speech.

The bi-phone frequencies are then used as an input to the back-end SVM-based classifier. The bi-phone feature vectors have the following numerical attributes, a *label* is the class label in a numerical representation, a *feature index* represents ordered feature indexes - that is, the location of that particular bi-phone feature, usually, integer representation, and in our case, a *feature value* represents the frequency count or occurrences of each bi-phone feature attribute. The SVM classification model is used to separate vectors in a binary classification and hypothesize the maximuim likelihood score of the bi-phone frequencies of each language [5].

3. MULTILINGUAL PHONEME MAPPINGS

We adopted two different phoneme mapping strategies to determine the phone similarities among the target languages. The first mapping technique is linguistically motivated phoneme mapping which require linguistic expertise while the other technique is a data-driven phoneme mapping.

3.1. Linguistically motivated mapping

We used linguistic knowledge-based method to construct a combined phone set for the target languages [8, 9]. To achieve this, the language-dependent speech units are defined based on the characteristics of their phonemic properties as represented on the International Phonetic Alphabet (IPA) scheme [10]. We used knowledge-based IPA method to create linguistically motivated phonetic pairwise mappings. Although no serious expert knowledge was involved but at some point, we relied on extensive language documentations.

The multilingual acoustic model is built by mapping the English phonemes to the Northern Sotho phonemes. This approach is motivated by the occurrence of similar phonemes from our target languages, an observation that leads to a reduced number of phonemes. The criterion to construct linguistically-motivated mappings is obtained as follows:

- If the IPA classification is similar to the one of the Northern Sotho phoneme then the English phonemes are mapped directly.
- Each English phoneme is mapped to its closest matching Northern Sotho phoneme based on the IPA scheme.
- If there is no close match to be found, then an English phoneme that occurs most frequently, the phoneme inventory is extended with that English phoneme.
- If none from the above criterion is applicable, then each phoneme is mapped to Northern Sotho phoneme that is mostly confused with as according to a confusion matrix.

In our case, the diphthongs of English were separated into vowels when applying an IPA-based method. Each phonemic vowel was then mapped to its equivalent phoneme of the target language.

3.2. Data-driven mapping

We performed the same procedure which was followed in section 3.1 but this time, we defined the data-driven phoneme mapping of English to Northern Sotho using the confusion matrix. The data-driven mapping which is based on the confusion matrix is built by including Northern Sotho language and English phonemes [10]. This mapping method consists of the counts of the confusion pairs that existed when aligning the speech recognition output and transcriptions of the speech data. The advantage of this approach is that it is fully data-driven and does not require much of some form of linguistic expertise [8].

For each phoneme of the English language, the most often confusable phoneme to the matrix language is selected for mapping. The phoneme mapping is obtained as follows: For each phoneme $Ø_{L1}$ from the target language, the best respective source candidate phoneme $Ø_{L2}$ is matched. We now measure the similarity by selecting the number of phoneme confusions as $\mathbf{c}(Ø_{L1}, Ø_{L2})$. The target phoneme $Ø_{L1}$ is matched as follows:

$$Ø_{L1} = \textbf{\textit{max}}\ \mathbf{c}(Ø_{L1}, Ø_{L2}) \quad (1)$$

Thus, for each target phoneme $Ø_{L1}$ source candidate phoneme $Ø_{L2}$ with the highest number of confusions is determined. If the same number of confusions occurs on two or more source candidate phonemes, the decision on the choice of the target phoneme $Ø_{L1}$ is made by expert. The same procedure is employed even when there are no confusions between target and source candidate phonemes.

4. MIXED-LANGUAGE SPEECH CORPUS

The state-of-the-art LID system requires a large amount of training speech data [11]. Under this condition, a large portion of the mixed speech corpus was attained by combining two monolingual speech data corpus. The corpus used for training of acoustic model include recorded speech data and their respective transcriptions of locally-produced primary Northern Sotho developed within Telkom Centre of Excellence for Speech Technology (TCoE4ST) and freely available LWAZI (third party corpus) South African English speech data often used for speech technology experiments [12]. The speech corpus is divided into two components; training and testing data sets. The TCoE4ST locally-produced Northern Sotho speech corpus had an amount of 3465 utterances. From the LWAZI English speech corpus, we selected 1840 speech utterances and their respective sentential form utterances that were used as training speech data set of the integrated system. Each speaker produced approximately 30 utterances that were phonetically balanced. The speech data were recorded over a telephone channel at 8 kHz sampling rate. The two speech corpora were combined together to form a large vocabulary of sizable mixed-language speech corpus for training the overall integrated system. Table 1 shows the summarized amount of speech data of the mixed-language speech corpus with two sub-sets.

Table 1: *The overall statistics of the mixed-language speech corpus*

	Train set	Test set	Total
# Speakers	143	5	148
Duration (hours)	5.5	1.0	6.5
# Utterances	5305	660	5965

The speech data set used for testing was not part of training data set. Code-switched speech is generally spoken but not formally written. However, it is not easy to find a code-switched speech data corpus [1]. It is for this reason that a simple finite loop grammar was used to generate about 300 artificially code-switched texts that are syntactically correct. The generated texts were recorded and included in a test set. We manually improved the quality of the utterances by removing dysfluencies such as long pauses, laughs and hiccups. Within the code-switched speech corpus, the calculated ratio of Northern Sotho words to English words is approximately 3:1. The average ratio of code-switched English words within each utterance was not more than 0.5. We have extended the test data set by adding 360 monolingual utterances.

5. BASELINE EXPERIMENTAL SETUP

We developed the front-end phone recognition and back-end classification portion of the baseline integrated LID system using the mixed-language speech corpus discussed in section 4. The experimental bilingual pronunciation dictionary used was achieved by combining several monolingual pronunciation dictionaries without retaining duplicate words. For the primary Northern Sotho language we used a limited vocabulary of Northern Sotho pronunciation dictionary that was locally-produced within the TCoE4ST and LWAZI, a freely available Northern Sotho pronunciation dictionary. For the secondary English language, we used a freely available LWAZI English pronunciation dictionary [12]. All the words in the pronunciation dictionary were manually verified and correctly checked for redundant phone representation. The compiled bilingual dictionary contained 85891 unique word tokens. The representation used in the bilingual pronunciation dictionary followed the Speech Assessment Method Phonetic Alphabet (SAMPA) notations and also taking into consideration of pronunciation rules [10].

A phone language model was incorporated in the phone recognizer for the purpose of speech decoding. The training transcriptions together with the generated code-switched texts were formatted into phone transcriptions and were used to develop the phone language model. A suitable bigram phone language model with discount interpolation was independently trained using a freely available SRI language model (SRILM) toolkit [13]. The resultant best phone language model had a perplexity of 13.8 without reporting out-of-vocabulary (OOV) rate.

For speech feature extraction, we apply a Hamming window of 25ms length with an overlapping window frame length of 10ms and the pre-emphasis coefficient of 0.97. Acoustic features are obtained using 39-dimensional static Mel-frequency Cepstral Coefficients (MFCCs) with 13 deltas and 13 acceleration coefficients. The Cepstral Mean and Variance Normalization (CMVN) pre-processing and semi-tied transformations are applied to the HMM-based acoustic models. The CMVN is used to overcome the undesired speech variations across the channels and distortion [9]. The HMM-based phone recognition system was created with a widely used standard Hidden Markov Model Toolkit (HTK) [14]. The acoustic model uses a 3-state left-to-right HMMs. The HMMs consist of the tied-state triphones clustered by a decision tree technique. Each HMM state distribution is modeled by 8 Gaussian mixtures

with a diagonal covariance matrix. Furthermore, the optimal phone insertion penalties and language scaling factors were properly tuned to balance the number of inserted and deleted phone during speech decoding.

To classify the targeted languages, the SVM-based classifier was implemented using a freely downloadable library for SVM (LIBSVM) toolkit - an integrated package for training SVM classifier [15]. This SVM program is a suitable package for classifying numerical attributes. The phone sequences used to train the classifier resulted with 3201 support vectors from models. The training process was also aimed at maximizing the margin as well as minimizing the training errors. The bi-phone vector attributes were then scaled in the range of [0, 1]. We used the same scaling values on training and testing data sets. The benefit of scaling data sets is to speed up training and classification process in order to obtain the best model performance and to avoid numerical differences that could lead to over-fitting if the training data attributes are in a large range [16]. A grid search is a simple search technique which was used to estimate the SVM parameters such as C, gamma, margin-error trade-off parameter and kernel width before training the classifier [4, 7]. The best kernel that we used was Radial Basis Function (RBF) training the classifier. We obtained the optimal parameter for RBF kernel and applied 5-fold cross-validation on the training set as well as estimating each grid point for the accuracy of the classifier.

6. EXPERIMENTAL RESULTS

We initiated our experiments with a baseline LID system. In subsequent experiments we first evaluated the experimental results of the baseline LID system before comparing them with the results of the integrated LID system applied with two phoneme similarity techniques for mapping the phonemes of our target languages. The baseline LID system was developed using a directly combined phoneme set. We did not perform any specific phoneme mapping in the phoneme set. The phoneme set size was large with 67 phonemes. The results evaluated on the mixed-language speech test set are shown in Table 2. Table 2 show the sizes of the phoneme set, phone error rates as well as the corresponding LID rates obtained.

Table 2: *The experimental results of the integrated LID systems and the phoneme set size.*

	Set size	PER (%)	LID rate (%)
Directly Mixed	67	33.7	85.0
IPA-based	39	28.7	85.8
Data-driven	**38**	**19.2**	**87.3**

The initial phoneme recognition experiments were as follows; context-dependent HMM-based acoustic models with 8 Gaussian probability density functions per state were engaged to obtain the experimental results in Table 2. In an IPA-based phoneme recognition system, we retained only two phonemes (such as /@/ and /b/) without specific mapping as there were no suitable phonemes which can be mapped to them. Approximately 45% of the English candidate phonemes were mapped to the Northern Sotho candidate phonemes using an IPA-based phoneme mapping approach and 48% of the English candidate phonemes were mapped to their highest confusable Northern Sotho phonemes using a data-driven phoneme mapping approach. Other phonemes were just directly mapped using both IPA-based and data-driven approach. These methods allow sharing of the parameters in the HMM-based acoustic models of the target languages [8, 9]. We report the phone error rate (PER) and LID accuracy. The results shows that PER and LID accuracy improves when the phoneme mapping is applied.

The baseline SVM-based classifier was trained using a 5-fold cross-validation which yielded an SVM accuracy of 97.5% on trained classification models and has predicted the best parameter value of C =0.5 and γ=0.5. The experimental results of the SVM-based LID classifier were also obtained using RBF kernel. In Table 2, both phoneme mapping approaches achieved a significant improvement over the baseline mixed LID results. The data-driven approach was able to outperform the baseline LID system and the IPA-based approach. The IPA-based approach was able to perform better with the PER of 5% and LID accuracy of 0.8%. The data-driven approach was able to better the performance with the PER of 14.5% as well as the LID accuracy of 2.3%. The average LID rate for monolingual utterances achieved was 81%. The SVM-based classifier was trained using a 5-fold cross-validation and RBF kernel which yielded an SVM accuracy of 99.75% on trained classification models and has predicted the best parameter value of C =2 and γ=0.5.

Fig. 3 represents the behavior of the PER with an increasing number of the Gaussian mixtures per state from 8 mixtures up to 32 mixtures.

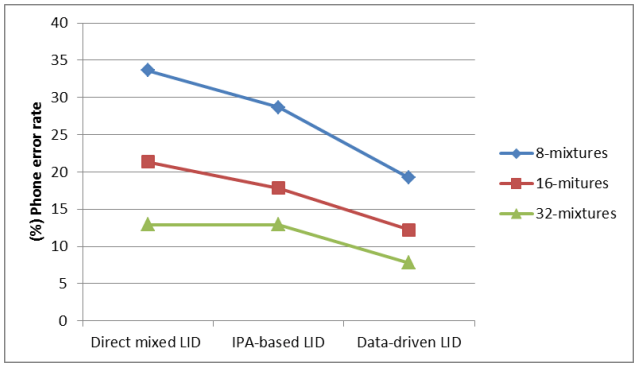

Fig. 3: The behavior of the PER of the directly mixed LID, IPA-based and data-driven LID system using 8 (blue), 16 (red), 32 (green) Gaussian mixtures per sate.

The triphone models were then improved by gradually increasing the number of Gaussian mixtures, and performing four iterations of embedded re-estimation after each increase. This procedure was continuously until the models had 32 mixtures per state, after which the phoneme recognition results no longer improved significantly on the test set. We further observed that our trained context-dependent acoustic models with 16 and 32 Gaussian mixtures per state as they tend to better the performance. In Fig. 3, the blue coloured line graph represents the calculated PER with an application of 8 Gaussian mixtures, red coloured line graph illustrates the PER with an application of 16 Gaussian mixtures and the green coloured line graph represents the PER with an application of 32 Gaussian mixtures per state. The results show that PER improves when context-dependent HMM-based acoustic models with 16 and 32 Gaussian probability density functions per state are engaged. As we expected, the data-driven approach performed better even when the Gaussian mixtures were increased to 16 and 32. Both phoneme mapping approaches give better results as compared to the baseline LID results. The highest performance was observed when our context-dependent acoustic models with 32 Gaussian mixtures per state were engaged. The data-driven approach was able to achieve the PER of 7.3% outperforming even the IPA-based approach. We also observed that not much significant differences from the back-end LID classifier has been achieved since the LID accuracies were found to be within the range of 84.7% and 89.5%.

7. CONCLUSION

In our attempt to present an integration of phonotactic information to perform language identification on mixed-language utterances, we proposed two successful phoneme mapping techniques to deal with code-switched utterances at the pronunciation level. The linguistic knowledge approach was derived on using IPA-based scheme while the data-driven approach was based on the confusion matrix. Moreover, we also investigated the performance of the PER on increased number of Gaussian mixtures per state. Our proposed IPA-based and data-driven approaches have shown a significant improvement on both PER and LID accuracies. We observed that the data-driven method outperforms the IPA-based approach. We achieved a better PER of 7.3% with a data-driven approach when the context-dependent acoustic models with 32 Gaussian mixtures per state were engaged. In future, we hope to train our system with real code-switched speech for further evaluation.

8. REFERENCES

[1] J. Weiner, N. T. Vu, D. Telaar, F. Metze, T. Schultz, D. –C. Lyu, E. Chng, and H. Li. “Integration of Language Identification into a Recognition System for Spoken Conversations Containing Code-Switches,” In Proc. SLTU, pp.61-64, 2012

[2] T. I. Modipa, M. H. Davel and F. de Wet, “Implications of Sepedi/English code switching for ASR systems”, In Proc. of PRASA, pp.65-69, 2013

[3] K. R. Mabokela and M. J. Manamela, “An Integrated Language Identification for Code-Switched Speech using Decoded-Phonemes and Support Vector Machine,” In Proc. SpeD, pp. 123-128, 2013

[4] H. Li, K. A. Lee and B. Ma, “Spoken Language Recognition: From fundamentals to practice,” in Proceedings of IEEE, Vol.101 (5), pp.1136-1159, 2013

[5] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language Identification: A Tutorial", In IEEE Circuits and Systems Magazine, Volume: 11, Issue 2, pp.82-108, 2011

[6] K. Bhuvanagiri and S. K. Kopparapu, “Mixed Language Speech Recognition without Explicit Identification”, American Journal of Signal Processing, Vol. 2, Issue 5, pp. 92-97, 2012

[7] M. Peche, M. Davel, and E. Barnard, “Development of a spoken language identification system for South African languages,” SAIEE Africa Research Journal, Vol. 100(4), pp. 97-105, 2009

[8] M. W. Christopher, S. Khundanphur, and J. K. Baker. “An investigation of acoustic models for multilingual code-switching”, In Proceedings of Interspeech, pp. 2691-2694, 2008

[9] W. Zhirong, U. Topkara, T. Schultz and A. Waibel, “Towards Universal Speech Recognition,” In Proc. ICMI 2002, Pittsburgh, 2002

[10] T. Modipa and M. H. Davel, “Pronunciation modelling of foreign words for Sepedi ASR,” in Proc. PRASA, pp. 185–189, 2010

[11] R. Tong, B. Ma, D. Zhu, H. Li, and E. S. Chng, “Integrating acoustic, prosodic and phonotactic features for spoken language identification,” In Proc. ICASSP, pp. 205–208, 2006.

[12] LWAZI, [Online] Available: http://www:meraka:org:za/lwazi

[13] A. Stolcke, “SRILM - An extensible language modeling toolkit,” In Proc. ICSLP, Denver, CO, pp. 901-904, 2002

[14] S. J. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, Cambridge University, “The HTK book Version 3.2.1,” 2002, [Online] Available: http://htk.eng.cam.ac.uk

[15] C. -C. Chang and C. -J. Lin, LIBSVM - A library for support vector machine, 2001, [Online] Available: http://www.csie.ntu.edu.tw/~cjlin/libsvm

[16] O. Giwa and M. H. Davel, “N-gram based Language Identification of Individual Words,” In Proc. of PRASA, pp.15-22, 2013