INTEGRATION OF PHONOTACTIC FEATURES FOR LANGUAGE IDENTIFICATION ON CODE-SWITCHED SPEECH

Koena Mabokela

Technopreneurship Centre, School of Consumer Intelligence and Information Systems, Department of Applied Information Systems, University of Johannesburg, South Africa

ABSTRACT

In this paper, phoneme sequences are used as language information to perform code-switched language identification (LID). With the one-pass recognition system, the spoken sounds are converted into phonetically arranged sequences of sounds. The acoustic models are robust enough to handle multiple languages when emulating multiple hidden Markov models (HMMs). To determine the phoneme similarity among our target languages, we reported two methods of phoneme mapping. Statistical phoneme-based bigram language models (LM) are integrated into speech decoding to eliminate possible phone mismatches. The supervised support vector machine (SVM) is used to learn to recognize the phonetic information of mixed-language speech based on recognized phone sequences. As the back-end decision is taken by an SVM, the likelihood scores of segments with monolingual phone occurrence are used to classify language identity. The speech corpus was tested on Sepedi and English languages that are often mixed. Our system is evaluated by measuring both the ASR performance and the LID performance separately. The systems have obtained a promising ASR accuracy with data-driven phone merging approach modelled using 16 Gaussian mixtures per state. In code-switched speech and monolingual speech segments respectively, the proposed systems achieved an acceptable ASR and LID accuracy.

KEYWORDS

Language identification, phonotactics, acoustics, language model, ASR, LID, code-switch speech.

1. INTRODUCTION

It is common for multilingual speakers to engage in code-switching or switching between more than one language in an utterance, a phenomenon known as mixed-language usage [1]. In multilingual societies, it seems to be commonly preferred. According to the constitution established by the national legislation of South Africa, chapter 1 - of the Bill of Rights - section 6, it states that: “Pan South African Language Board (PanSALB) has the right to promote the use all eleven official languages.” The result is that South Africa is a multilingual nation, with eleven official languages. Most native speakers are likely to speak more than one official language in their daily, everyday conversation. In general, South African languages are represented by a mixed mode of usage (e.g., in radio and television dramas, news broadcasts, religious worship services, and interviews and presentations). Ethnologue's database notes that of over 6909 natural languages spoken worldwide, English is traditionally used for global communication [2]. English is often mingled with indigenous languages that are under- resourced in many African communication episodes. Native speakers of African languages utilize the English language to express numerical digits, times, and codes. In South Africa, it is common to hear more than one language being spoken in the same area. This research was therefore relevant.

Normally code-switched speech consists of two or more words or sentences from another language. The embedded language is also known as the primary language. There is no formal

written form of code-switched speech. In this sense, code-switched speech falls into the same category as under-resourced languages [1]. Modern style of communication, characterized by the mixing of two or more official languages, presents greater challenges for speech-enabled technologies. As the state of the art of human language technology (HLT) advances, it is now focusing on automating systems that allow multilingual individuals to interact easily with smart computers. As a result, the HLT sector benefits a wide range of multicultural societies - from the less-literate ruralites in remote regions who need help obtaining relevant medical information over a cell-phone line, to the sophisticated industry researchers who need assistance solving commercial problems with computer equipment [3].

In recent years, much research has been done on proven systems for spoken language processing which are based on pattern recognition, one of the most challenging computational problems. An artificial intelligence system is defined as a system that is capable of recognizing, identifying, or classifying speech patterns [4, 6]. Known as spoken language identification (LID), spoken language identification refers to an automatic process that can accurately determine the language spoken in every sampled speech utterance. There are many multilingual speech processing applications that can be enabled by LID systems, including multilingual information retrieval [5], spoken language translation [4] and telephone call routing systems [6]. A major trend in today's speech technologies is the ability to support multiple input and output languages, especially when the applications are targeted at global markets and linguistically diverse communities [4].

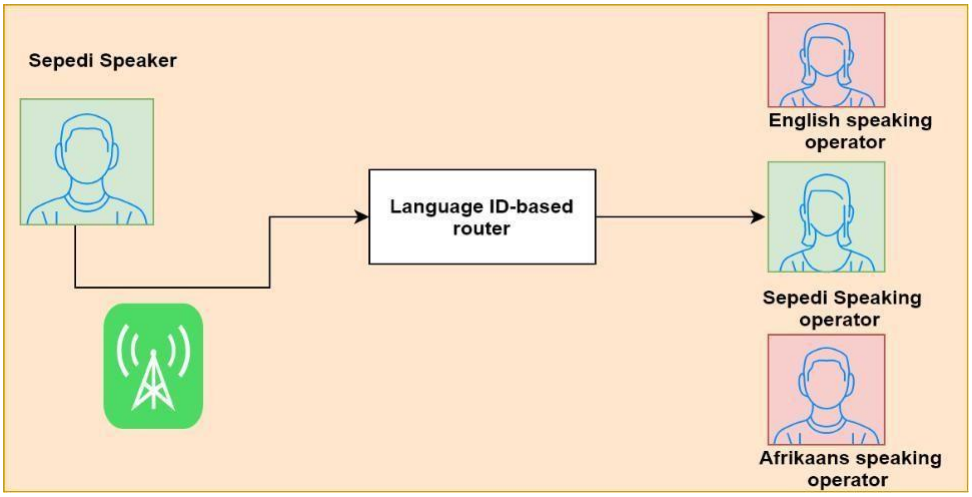

Figure 1. Example of an LID based call routing system.

To ascertain the language of code-switched utterances, we propose integrating a multilingual One-Pass ASR system followed by a language identification system. Using a multilingual One-Pass ASR system, we would be able to decode multiple languages simultaneously within the same utterance [1, 7]. Our experimental methods employed context dependent HMMs based on multilingual acoustic models and phone LMs. For language identification on code-switched speech, we used nearly the same technique as parallel phoneme recognition followed by language modelling (PPRLM). Even though we report only on experiments that were conducted in South Africa in two official languages, we believe that the same process can be applied to other under-resourced languages as well [23]. However, the question then is does ASR performance influence

the performance of the LID system? We perform experiments to show the answer is possible and positive.
Our main contributions can be summarized as follows:

- We formalise the framework of multilingual One-Pass ASR-LID system built with different open-source tools.
- We show that a significant improvement in the performance of the ASR system directly affect the performance of the LID system. We provide a detailed analysis of ASR-LID improvements.
- We show that integration of phone merging strategies and the use of mixed corpora's within the ASR system can handle code-mixed or code-switched utterances.
- We evaluate the ASR-LID task on TCoE4ST-LWAZI corpus and create acoustic models and LMs which capture sufficient phonotactic information which we leverage to show improvements on phone error rates and LID accuracy.

This paper is organized as follows: Section 2 discusses the related works of ASR and LID on code-switched speech and/or multilingual setting. Section 3 provide an explanation of the system architecture. Section 4 describes the mixed speech corpus used for the experiments. Section 5 discusses the experimental setup and results. Lastly, the conclusion and future directions are provided.

2. RELATED WORK

In Singapore, Mandarin and English are often mixed in spoken conversations [1], in Hong Kong a code-switching between Cantonese and English is used on many occasions [8] and in Taiwan, a Mandarin-Taiwanese code-switching speech was reported [9]. Others reported a mixed- language speech found in India between Hindi and English [10]. Similarly, code-switching is observed in South Africa, where code-switching between two indigenous South African languages such as Xhosa and Zulu has been studied for multilingual speech recognition [11]. Recently, Modipa et al. [12] reported a context-dependent modelling technique of English vowels in Sepedi code-switched speech where the process of obtaining a phone mapping from embedded language to the matrix language was investigated.

In the repertoire of code-switched speech, only a few approaches have been reported. In order to detect different languages in code-switch speech utterances, multiple cues such as acoustics, prosody, and phonetics are integrated [8]. In order to detect more than one language within an utterance, a Language Boundary Detection (LBD) method is used [9]. In code-switched utterances, the other method is used to separate languages like English, Mandarin and Taiwanese using Delta-Bayesian information criteria (Delta-BIC) and Latent semantic analysis (LSA) [9]. To jointly segment and identify utterances of a mixed language, an estimation approach that utilizes maximum posteriori estimation was used [13]. LID modules that incorporate LBD modules are used in the above-mentioned approaches. It is generally not preferred to use LID systems incorporating the LBD module because the incorrect assumption is that the code-switched speech segments are independent and as a result, errors in the LID module cannot be recovered [1]. As a result, in this case, if the LBD module cannot reach 100%, the LID module will also experience limitations, thereby limiting the performance of the speech recognition module [1, 10].

An alternative multilingual approach could handle code-switched speech that incorporates multilingual acoustic models and multilingual pronunciation dictionaries, as well as multilingual language models that share models across multiple language units [1, 10]. However, a multilingual ASR approach does not require an additional language identification module

because language information becomes part of the system directly [1, 24]. The first technique involves using linguistic knowledge to establish a multilingual map of phonetic features or clustering of them that is based on the same training data [7]. The International Phonetic Alphabet (IPA), the Assessment Methods Phonetic Alphabet (SAMPA), and the Wordbet are three common examples [15]. Using a data- driven approach, such as clustering specific phones according to the distance between similar acoustic models, is another technique for mapping language-dependent phones. These methods take into account spectral characteristics and include Confusion Matrix, Bhattacharyya Distances, and Kullback-Leibler Divergent [15].

According to Lyu et al. [16], code-switching speech within an utterance could be distinguished by a word-based lexical model LID system. With the help of a large vocabulary continuous speech recognition (LVCSR) system, a two-stage scheme system is employed. Using recognized word sequences, a trained word-based lexical model is applied to identify languages. There are several approaches to the LID system, including PPRLM, phoneme recognition followed by language modelling (PRLM), and parallel phoneme recognition vector space modelling (PPR-VSM). Several phoneme recognizers are used to tokenize the speech waveform into sequences of phonemes in the PPRLM approach. After determining the most probable language from the target languages, the resulting sequence of phonemes is fed into an n-gram LM [6, 18]. Supervised SVMs are the most effective classifiers [16]. Using a similar model, South Africa distinguishes the eleven official languages [17]. Using a PPRLM architecture and phoneme frequency filtering technique, SVM-based classifiers are used to classify languages at the back end. In test samples with a length of 3-10 seconds, the SVM classifier achieved an average LID rate of 71.78%; and when clustering similar language families was considered, the LID rate reached 82.39%.

3. SYSTEM ARCHITECTURE

This section explains the proposed approach for integrating the phonotactic features in the LID systems. The approach is an integrated system targeted to identify multiple languages, namely, Sepedi and English on code-switched speech utterances.

Recognition. Figure 2 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. Assume the speech waveform is segmented into a sequence of phones. To achieve this, a phone n-gram LM is employed to estimate the likelihood score of the nth phone given the (**n-1**) th of the preceding phones..

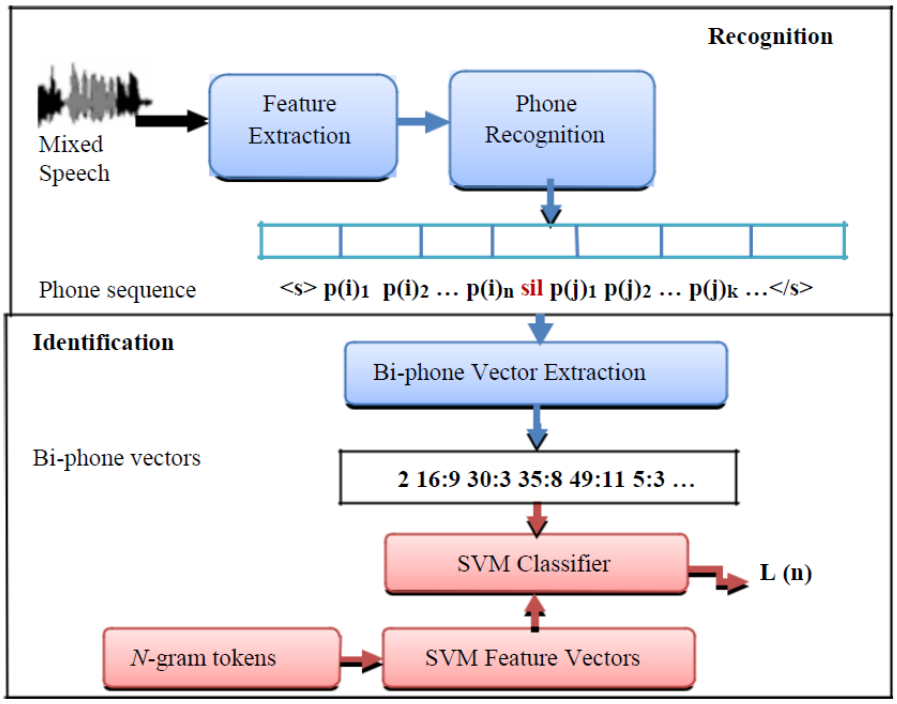

Figure 2. A multilingual One-Pass ASR-LID for code-switched speech.

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 LM 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 LM.

Identification/classification. The SVM-based classifier is used to identify 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 like vector space modeling [16]. Each word has its unique pronunciation representation. For example, the word [ABILITY] in the dictionary is phonetically represented as: [a p_> i l i t_h i]. The possible n-gram feature vectors from the given phoneme string are generated. Then LID is performed by using SVM classifier to score the phoneme sequence of a test utterance. The most likely sample for classification is chosen from the LM with the highest log-likelihood score.

The bi-phone frequencies are then used as an input to the backend SVM-based classifier. The numerical attributes of the bi-phone feature vectors are as follows: A label represents a class name in numerical representation, a feature index represents a specific location for that bi-phone feature - normally, an integer representation, and a feature value represents the frequency count of each bi-phone feature attribute. In binary classification, the SVM classification model is used to separate the vectors and to hypothesize the maximum likelihood scores of the bi-phone frequencies for each language [17].

4. SPEECH CORPUS

Research has shown that the amount of training data strongly influences the accuracy and robustness of ASR systems [8–12]. The multilingual ASR system requires a large amount of training speech data. Under these conditions, most of the mixed speech corpus was derived from the combination of two monolingual data sets [10, 12]. The training and testing parts of a speech corpus were separated. Although, a code-switched speech dataset is difficult to collect and annotate, this process is often time-consuming and requires more sophisticated linguistic experts [22, 25]. This section describes the speech corpus used for training of multilingual acoustic model. Furthermore, we discuss how the mapping of the phone set, and creation of multilingual pronunciation dictionary was achieved.

4.1. Data Description

Data pre-processing. A speech corpus used to train the model included recordings and transcriptions of Sepedi from the Telkom Centre of Excellence for Speech Technology (TCoE4ST) and LWAZI South African English. These data are often used for experiments of speech technology [18, 24, 25]. Apart from the mixed-language speech data, TCoE4ST-LWAZI corpus consists of foreign words from local languages other than Sepedi and English. We have also found misspelled words and borrowed words such as dlala (i.e., play), gymnasium, television, and ambulance. Such words are labelled correctly in the Sepedi version of pronunciation. A few utterances contained non-speech sounds like laughing, breathing, etc. Since our focus in this work is to investigate acoustic models and LMs for code-switched speech in a multilingual setting, we excluded low-quality utterances.

Data distribution. We construct training and testing sets from the pre-processed TCoE4ST-LWAZI corpus. It was obtained by using the TCoE4ST locally produced Sepedi speech data contains an amount of 3749 utterances together with the LWAZI English speech data, we selected 1556 clean speech data and their respective sentential transcriptions that were used as training speech data set in these experiments. The two speech corpora were combined to form a large vocabulary of sizable mixed-language speech corpus for model training. Table 1 shows detailed statistics of each split. The out-of-vocabulary (OOV) rates on the test sets are 3.4%. The speech data which was used for system testing was not part of training data set.

Table 1: The statistics of the mixed corpus

	Train set	**Test set**	**Total**
# Speakers	143	6	149
Duration (hours)	5.5	1.2	6.7
# Utterances	5305	715	6020

Additional data. As code-switched speech is generally spoken but not formally written, it is not easy to find code-switched speech data [1, 25]. It is for this reason that a simple finite loop grammar was used to generate 60 artificially code-switched sentences that are semantically and syntactically correct. The sentences that were not syntactically and semantically were fewer which some were marked for either correction or deletion. The generated sentences were verified by linguistic experts and then recorded by selected native speakers under strict protocols. The total number of code-switched speech utterance which was recorded were 300 after cleaning and pre-processing. These additional speech data was included as part of the test set. Furthermore, we manually reassured the quality of the utterances by removing disfluencies such as long pauses, laughs and hiccups. Within code-switched speech data, the percentage of Sepedi words is 74.2%

and English words are 25.8% excluding silences. We estimated an average ratio of code-switches within each utterance to be between 0.4-0.5 when counting only switches that exists within the target languages. We extended the test set by adding 415 randomly selected monolingual utterances from both Sepedi and English corpus making the total test set of 715 utterances.

4.2. Dictionary and Phone Set

To obtain the pronunciation dictionary, the dictionaries for Sepedi and English are modified using the phone set. The multilingual pronunciation dictionary used in this study was created by combining several monolingual pronunciation lexicons without keeping duplicate words. To create the Sepedi pronunciation dictionary, we used both TCoE4ST and LWAZI, a freely available dictionary of Sepedi. For the English language, we used freely available LWAZI English pronunciation dictionary often used for speech technology research tasks [18]. It is checked manually to ensure that each word has been translated accurately and there are no redundant phonetic entries in the pronunciation dictionary. There were 85.9k unique words in the combined bilingual dictionary. A bilingual pronunciation dictionary based on the Speech Assessment Method Phonetic Alphabet (SAMPA) notation used by the International Phonetic Association (IPA) and taking into account pronunciation rules [7, 10, 25]. In this case, phones with similar phonetic features were mapped into a single best phone candidate representation to reduce confusion within the combined phone set. Several English vowel phones were left unmapped since they did not match any Sepedi vowel phone.

5. PHONE MERGING METHODS

In this paper, we adopted two different phone mapping strategies to determine the phone similarities among the target languages. The first mapping technique is linguistically motivated phoneme mapping which requires a linguistic expert while the other technique is a data-driven phoneme mapping.

5.1. Manually merging similar phones

We employ a manually merging similar phones (MMSP) method. We adopted this method to build the merged phone set for the matrix and embedded languages using linguistic knowledge [7, 9, 25]. To achieve this, the language-dependent speech units are defined based on the characteristics of their phonemic properties as represented on the IPA-based scheme [7]. By merging pairwise phonetic phonetics using IPA, we were able to create linguistically justified phonetic pairwise merges. Initially, we merged English and Sepedi phones to see if merging had an impact. To build this multilingual acoustic model, the English phonemes are combined with the Sepedi phonemes. Our goal is to reduce as many phonemes as possible by following the occurrence of similar phonemes in our target languages.

The criteria for constructing linguistically motivated mappings are as follows [7]: (i) If the IPA classification resembles one of the Sepedi phones, the English phones are directly merged. (ii) The IPA is used to calculate the matching phone between each English phone and its closest Sepedi match. (iii) The phone inventory is expanded with the most frequently occurring English phone if no close match matches are found. (iv) Those phones that do not meet the above criteria are mapped to the Sepedi phone that is most often confused with according to the confusion matrix.

When using an IPA-based method, the diphthongs of English were separated into vowels. Next, the phone vowels were merged to their equivalents in the target language.

5.2. Data-driven phone merging

Another potential method to merge similar phone pairs is a data-driven phone merging (DDPM). In the ASR decoder, the errors made in the alignment can be used to identify the phone pairs to merge. The phone merging of English to Sepedi was defined by using the confusion matrix that was derived when the ASR system was trained by directly merging the phonemes of the target languages. With the help of the confusion matrix, data-driven merging is built by recognizing speech utterances from the target language with the aid of language source models [7, 10]. This merging technique consists of calculating the number of confusion pairs that exist between the speech recognition outputs and transcriptions. In addition to being fully data-driven, this approach does not require linguistic expertise [14]. We then determine the phoneme sequences by parsing the decoded lattice in the best possible way. Using the same acoustic model, we align the evaluation data and calculate the true phone sequences and durations corresponding to those sequences.

In this study, we choose pairs of expressed utterances with 80% or more overlap in duration in terms of aligned and decoded speech. The alignments are used to identify the English phone numbers that were mistakenly classified as Sepedi. Exchanges between languages are often referred to as cross-language swaps. Switching between the English and Sepedi phone pairs frequently indicates that the pairs should be merged. By observing this method, we see several merges that were missed in the manual merge. Whenever the same number of confusions occurs on more than one source candidate phoneme, a linguistic expert makes the decision on which phone to use as the target. Even when there is no confusion between target and source candidate phonemes, the same procedure is followed. More examples of phones are described in [23, 24].

6. EXPERIMENTS AND RESULTS

This section describes the experimental setup, the tools used to develop the system and the speech data used for testing of phone recognition system. Lastly, the experimental results are presented and discussed.

6.1. Acoustic Models

We used HTK toolkit to build all baseline ASR systems. To build a baseline ASR system, we applied a Hamming window of 25 ms length with an overlapping window frame length of 10 ms and the pre-emphasis coefficient. 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 HMMs. The CMVN is used to overcome the undesired variations across the channels and distortion [9]. The acoustic model uses a three state left-to-right HMM. The HMM-based consist of the tied-state triphones clustered by a decision tree technique. Each HMM state distribution is modelled by 8-Gaussian mixture models (GMM) with a diagonal covariance matrix. Furthermore, the optimal phone insertion penalties and language scaling factors are properly tuned to balance the number of inserted and deleted phone during speech decoding.

6.2. LMs and Perplexity Experiments

We used the SRILM toolkit [20] to build all our LMs. The baseline LM used for PPRLM system is a smoothed bigram LM estimated using the English and Sepedi monolingual texts. These LMs

are trained separately. Our proposed multilingual LM is a smoothed bigram LM estimated using the mixed-language texts which are referred to as mixed LM. We used a n-gram-count tool to compute language probabilities of the phone transcriptions. The training transcriptions together with the generated code-switched texts were formatted into phone transcriptions and were used to develop the phone LM. We train model bigram using the discounting. We enabled interpolated discount of order 2 by setting interpolation weights to the mid-point of [0, 1] range. The phone transcriptions were not part of the training data set. To evaluate the trained phone- based LM the was used where <-ppl> is an option for the determining the perplexity overall trained phone LM in given the test data set. A phone LM was incorporated in the phone recognizer for the purpose of speech decoding. The resultant best bigram phone LM had a perplexity value of 13.795 without reporting OOV rate. Below is a sample of an extract from the trained mixed LM file.

```
\data\
ngram 1=70
ngram 2=1838
\1-grams:
-0.8925686          </s>
-99          <s>    -4.006862
-2.198572           @:          -2.992164
-1.986697           @u          -3.171212
   ….
\2-grams:
-2.101489           <s> @u
-2.138814           <s> B
-3.502635           <s> BZ
….
-2.036361           {z
\end
```

6.3. Implementation Details

All the ASR systems were built using widely used Hidden Markov Model Toolkit (HTK) [19]. The supervised SVM classifier was implemented using a freely downloadable library for SVM (LIBSVM) toolkit - an integrated package for training SVM classifier [21]. This SVM program is a suitable package for classifying numerical attributes. A suitable bigram phone LM with discount interpolation was trained using a freely available Stanford Research Institute LM (SRILM) toolkit [20]. All these toolkits are integrated to develop the multilingual ASR-LID systems.

6.4. Evaluation Metrics

Classification is then performed on the testing data to compute the error rate. The quality of the correct phone recognition output is typically captured by the phone error rate (PER) metric. The accuracy and error rate are defined as:

$$\mathbf{PER} = \frac{S+I+D}{N} * 100 \qquad \text{..... (1.1)}$$

where (N) is the total number of labels, (D) is the number of phone deletion errors, (S) is the number of phone substitution errors, and (I) is the number of phone insertion errors. The language identification was done on per-utterance basis by estimating the sequence of phone strings that have the maximum likelihood of being selected. We defined accuracy of the LID as follows:

$$\textbf{Accuracy} = \left(\frac{P_{Correct}}{P_{all}}\right)\textbf{*100} \quad \textbf{..... (1.2)}$$

Where $P_{correct}$ is the number of utterances in the test data set that are accurately identified and while Pall the represents the total number of samples classified.

6.5. Results and Discussion

We first evaluated the experimental results of the baseline systems and compared them with the results of the integrated LID system applied with two phone merging techniques for mapping the phones of the target languages. The baseline ASR-LID system was developed using directly combined phone set with no merging any phones in the set. The phone set size consist of 67 phones. The baseline system is then compared to PPRLM developed for English and Sepedi. The PPRLM system was developed using monolingual acoustic models and LMs from the respective speech corpus (i.e., described in section 3). We have built the PPRLM system following the approach described in [18]. However, we have used robust bigram LMs for each language to obtain comparable results on the monolingual corpus. The results obtained from the monolingual speech corpora and mixed speech test set are shown in Table 2 and Table 3. The phoneme recognition accuracies that were obtained from the respective phone recognizers were as follows; for Sepedi with a recognition accuracy of 53.13% and English with a recognition accuracy of 63.93%.

Table 2: The experimental results the of PPRLM and No Merge ASR-LID systems.

System Accuracies	**PPRLM**		**No Merge** (Mixed English & Sepedi)
	Sepedi	**English**	
(%) Word correctness	78.66	73.18	89.22
(%) Phone accuracy	53.13	63.93	66.79
(%) LID accuracy	**81.67**	**81.37**	**85.01**

The baseline ASR experiment employs acoustic models based on HMMs, with probability density functions of 8 Gaussians per state. The HMM-based acoustic models of the target languages can use these models to share parameters. We evaluated the systems based on the phone error rate (PER) and LID accuracy. Table 2 and Table 3 show experimental results for the ASR-LID system, the PPRLM system, and the phone set size. Based on the results, the combined phone merging strategies improve PER and LID accuracy. Due to the high number of confusable phones, it is clear from Table 2 that the No Merge ASR-LID system accuracy is low. On the other hand, there are promising results for the LID classification module. An average of 81.4% of monolingual utterances had language identification accuracy.

Table 3: The experimental results the of PPRLM and three (3) ASR-LID systems with standard 8-Gaussian mixtures per state.

Systems	**Phone set size**	**PER (%)**	**LID (%)**
PPRLM	81	44.47	82.48
No Merge	67	33.21	85.01
MMSP	38	**28.65**	**85.79**
DDPM	38	**19.22**	**87.33**

We trained a baseline SVM-based classifier using 5-fold cross validation which yielded an estimated accuracy of 97.5% on the trained classification models, and this method predicted a C value of 0.5 and an γ value of 0.5.0. Using Radial Basis Function (RBF) kernels, we have also obtained experimental results for the SVM-based LID classifier. The experimental results of the LID classifier were obtained using RBF kernel with log2 (C) = -1 and log2 (gamma) = -1 on the training dataset. Both phone merging approaches achieve a significant improvement over the baseline results. The DDPM system was able to outperform the baseline system and the MMSP system. The MMSP system was able to perform better with the PER of 5% and LID accuracy of 0.8%. The DDPM system was able to better the performance with the PER of 14.5% as well as the LID accuracy of 2.3%. A 10-fold cross-validation was used to further train the SVM-based classifier, and a RBF kernel was used to yield an accuracy of 99.75 percent on the trained classification models. The SVM-based classifier predicted C =2 and γ=0.5. The monolingual LID accuracy measured for Sepedi and English was 83.7% and 83.1%, respectively.

6.6. Error rates vs Gaussian mixtures

Figure 3 (a) and (b) represents the behaviour of the PER with an increasing Gaussian mixture per HMM state from 8 mixtures up to 64 Gaussian mixtures on each ASR system as well the LID performance. 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. After performing this procedure continuously until the models had 32 mixtures per state, the results of phoneme recognition on the test data set stopped improving significantly. In addition, it was further noticed that our context-dependent acoustic models with 16 - 64 Gaussian mixtures within a state tend to significantly improve the performance.

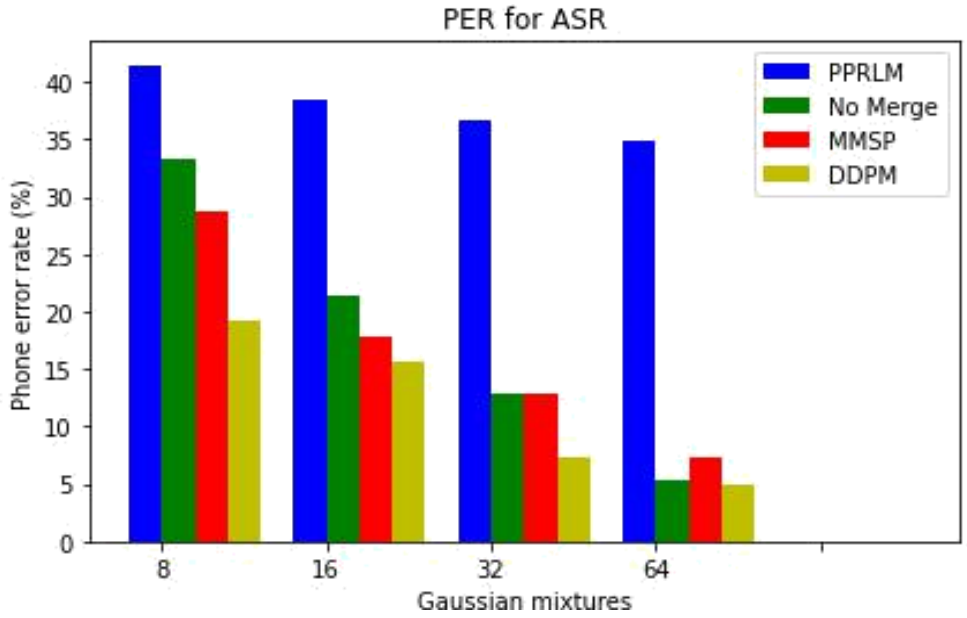

Figure 3 (a). The PER of the NoMerge, MMSP and DDPM ASR-LID system using 8, 16, 32, 64 Gaussian mixtures per HMM state.

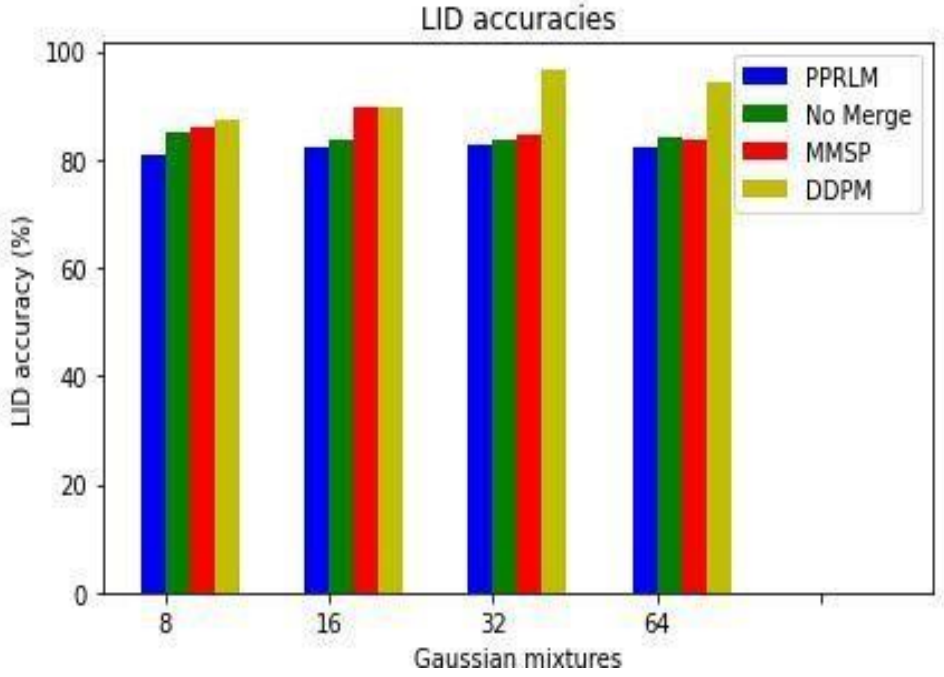

Figure 3 (b). The LID accuracies of the NoMerge, MMSP and DDPM ASR- LID system using 8, 16, 32, 64 Gaussian mixtures per HMM state.

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 expected, the DDPM system performed better even when the mixtures were increased. Both phone merging approaches give better results as compared to the baseline results. The most accurate results

were obtained when context-dependent acoustic models with 32 Gaussian mixture probability distributions were used unlike with the 64 mixtures. The DDPM system was able to achieve the PER of 7.7% outperforming even the MMSP system. Our analysis showed that there were quite few significant differences between our three systems and the SVM-based LID classifier since the accuracies of all three approaches ranged between 83.7% and 89.6% when 16 mixtures per HMM state were employed. The better LID accuracy was achieved by DDPM system. Based upon 32 Gaussian mixtures, three of the proposed systems achieved an acceptable LID accuracy ranging from 83.8% to 96.7%. The DDPM system was able to achieve the best LID accuracy of 96.7%. We have also observed that more speech data from the primary language also increase the chances better ASR-LID accuracies. Additionally, since we used the recorded code-switched speech data for testing of the systems, our systems will be able to perform significantly well for real multilingual annotated data. For instant the results show that ability of the multilingual acoustic model to recognise the testing set would have perform more or less the same on real multilingual annotated data and this is also demonstrated by the study in [12, 25]

7. CONCLUSIONS

This paper presents an incorporation of phonotactic information to perform multilingual ASR-LID on mixed-language speech. We proposed two phone mapping techniques to deal with code-switched or multiple language utterances. The MMSP approach is derived by manually merging the similar sounding phones while the DDPM system built by merging phones using ASR confusion matrix. Moreover, we investigated the behaviour of the PER using different number of Gaussian mixtures per state, which seem to show promising results. Our proposed MMSP and

DDPM systems have shown a significant improvement on both PER and LID accuracies. We observed that the DDPM system outperforms the MMSP system with an increase in the Gaussian mixtures per. We achieved a better PER of 7.7% with a DDPM system when the context-dependent acoustic models with 32 Gaussian mixtures per state were engaged. The increase of the GMM further lead to a decline of the LID classification accuracy. In future, we hope to train our systems with more real code-switched speech data to build even more robust acoustic models and LMs for further evaluation and performance analysis. This will be a clear proof to see if our model can handle a large amount of data. We plan to use other data- driven methods to perform phone similarity merging.

ACKNOWLEDGEMENTS

We acknowledge the National Research Foundation (NRF) for the Black Academics Advancement Programme (BAAP) award (REFERENCE NO: BAAP200225506825) in 2021

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. of SLTU 2012, Cape Town, South Africa, pp.1-4, May 2012.

[2] M. P. Lewis, G. F. Simons, and Charles D. Fenning (Eds). 2013. Ethonologue Languages of the World, Seventeenth edtion. Dallas, Taxas SIL International. Online version: http://www.ethonologue.com/world, Accessed on: 22 September 2021.

[3] HLT, http://www.meraka.org.za/humanLanguage.htm, Accessed on: 22 September 20212021.

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

[5] 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.

[6] 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.

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

[8] D. -C. Lyu and R. -Y. Lyu, "Language Identification on Code-Switching Utterances Using Multiple Cues," In Proceedings of INTERSPEECH, Brisbane, Australia, pp. 711-714, September 2008.

[9] C. -H. Wu, Y. -H. Chiu, C.-J. Shia and C. -Y. Lin, "Automatic Segmentation and Identification of Mixed-Language Speech Using Delta-BIC and LSA-Based GMMs," IEEE Transactions on Audio, Speech, and Language Processing, pp. 266-276, January 2006.

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

[11] T. Niesler and D. Willett, "Language identification and multilingual speech recognition using discriminatory trained acoustic models," In Multilingual Speech and Language Processing, 2006

[12] T. I. Modipa, M. H. Davel and F. de Wet, Context-dependent modelling of English vowels in Sepedi code-switched speech, In Proc. of PRASA 2012, Pretoria, South Africa, November 2012

[13] C. J. Shia, Y. H. Chiu, J. H. Hsieh and C. H. Wu, "Language Boundary Detection and Identification of Mixed-Language Speech Based on MAP Estimation," In Proc of ICASSP, pp. 381-384, 2004.

[14] C. –H Wu, H. –P. Shen and Y. –T. Yang, "Phone set construction based on context-sensitive articulatory attributes for code-switching recognition," In proc. ICASSP, pp.4865-4868, 2012

[15] D. -C. Lyu, R. -Y. Lyu, C. -L. Zhu and M.-T. Ko, "Language Identification In Code-switching Speech Using Words-based Lexical Model "Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium, Tainan, pp. 460-464, December 2010.

[16] W. M Campbell, JP Campbell, DA Reynolds, E Singer, "Support vector machines for speaker and language recognition", Computer Speech & Language, 20, pp.210-229, 2006.

[17] 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, December 2009.

[18] LWAZI, http://www.meraka.org.za/lwazi/, Accessed on: 06 September 2021.

[19] S. J. Young, G. Evermann, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, Cambridge University,2002. (For HTK Version 3.2.1), http://htk.eng.cam.ac.uk, Accessed on: 05 May 2013.
[20] Stolcke, "SRILM - An extensible Language modelling toolkit", In Proc. ICSLP, Denver, CO, pp. 901-904, November 2002
[21] C, -C. Chang and C. -J. Lin, LIBSVM-A library foe support vector machine, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm, Accessed on: 29 September 2021.
[22] A. Biswas, F. de Wet, E. van der Westhuizen, E. Yılmaz, and T. Niesler, "Multilingual Neural Network Acoustic Modelling for ASR of Under-Resourced English-isiZulu Code-Switched Speech," In Proc. Interspeech, 2018.
[23] K. R. Mabokela, M. J. Manamela, M. Manaileng, "Modeling code-switching speech on under-resourced languages for language identification", in: Spoken Language Technologies for Under-Resourced Languages, pp. 225-111, 2014.
[24] K. Mabokela, "Phone Clustering Methods for Multilingual Language Identification" In proc. Computer Science & Information Technology (CS & IT), pp. 285-298, 2020
[25] E. van der Westhuizen and T. Niesler, "A first South African corpus of multilingual code- switched soap opera speech," Proc. Language Resources and Evaluation Conference. pp. 2854 – 2859, 2018

AUTHOR

Mr. Ronny Mabokela is the Head of the Technopreneurship Centre (TPC) within the School of Consumer Intelligence and Information Systems & also a Lecturer in the Department of Applied information Systems at University of Johannesburg (UJ). He earned his MSc. degree in Computer Science from the University of Limpopo, where he obtained his undergrad degrees in Computer Science & Mathematics. Prior to joining the University of Johannesburg in 2019, he acquired a vast amount of industry experience, having worked for over 5 years in the Telecommunication sector as Solution design engineer and Business system Integration respectively. He has presented his research work on numerous platforms locally & internationally & has a keen research interest on Natural languages processing, Speech technologies, Multilingual sentiment analysis for under-resourced languages etc. He is currently pursuing his PhD studies in Computer Science at the University of Witwatersrand. He has received numerous awards in industry & academia, including NRF Black African Advancement Programme award in 2021. He is a member of the South African Institute for Computer Scientists & Information Technologists. He also serves on various boards.