PHONE CLUSTERING METHODS FOR MULTILINGUAL LANGUAGE IDENTIFICATION

Ronny Mabokela

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

ABSTRACT

This paper proposes phoneme clustering methods for multilingual language identification (LID) on a mixed-language corpus. A one-pass multilingual automated speech recognition (ASR) system converts spoken utterances into occurrences of phone sequences. Hidden Markov models were employed to train multilingual acoustic models that handle multiple languages within an utterance. Two phoneme clustering methods were explored to derive the most appropriate phoneme similarities between the target languages. Ultimately a supervised machine learning technique was employed to learn the language transition of the phonotactic information and engage the support vector machine (SVM) models to classify phoneme occurrences. The system performance was evaluated on mixed-language speech corpus for two South African languages (Sepedi and English) using the phone error rate (PER) and LID classification accuracy separately. We show that multilingual ASR which fed directly to the LID system has a direct impact on LID accuracy. Our proposed system has achieved an acceptable phone recognition and classification accuracy in mixed-language speech and monolingual speech (i.e. either Sepedi or English). Data-driven, and knowledge-driven phoneme clustering methods improve ASR and LID for code-switched speech. The data-driven method obtained the PER of 5.1% and LID classification accuracy of 94.5% when the acoustic models are trained with 64 Gaussian mixtures per state.

Keywords

Code-switching, Phone clustering, Multilingual speech recognition, Mixed-language, Language identification

1. INTRODUCTION

Most multilingual societies are capable of code-switching in their daily conversations. This appears to be an acceptable modern-day style of communication, usually preferred in multilingual societies [1], [2]. Code-switching speech is commonly more spoken than formally written, and a large textual dataset is required to build a suitable language model which is necessary for developing a multilingual ASR system [3], [4]. However, the African reality in many communication episodes is that English is frequently mixed with indigenous under-resourced official languages.

Code-switching speech has a significant impact on existing ASR systems and a large speech corpus is required to develop suitable context-dependent acoustic models [3]. The existing monolingual ASR systems are not accurate enough to handle code-switched speech utterances. Consequently, acoustic, pronunciation and language models need to be redesigned to accommodate foreign or unknown words from different languages [5]. A multilingual ASR

David C. Wyld et al. (Eds): NLP, JSE, MLTEC, DMS, NeTIOT, ITCS, SIP, CST, ARIA - 2020
pp. 285-298, 2020. CS & IT - CSCP 2020 DOI: 10.5121/csit.2020.101421

system employs multilingual language models that allow the exploitation of multilingual pronunciation dictionaries. It is highly plausible to classify code-switched speech itself in the same category as under-resourced languages due to lack of speech technology resources for developing accurate ASR systems [2], [3]. ASR systems deployed in this environment should be able to process multilingual speech that includes such code-switching utterances. The LID system identifies language speech processing applications, such as telephone calls routing to human operators, particularly for handling emergency calls [3], [5]. In this paper, we propose an ASR system that is integrated with an LID system to classify code-switched speech for Sepedi and English. Only the experiments conducted using two official South African languages are reported on.

There are two ASR approaches that are reported to handle code-switched speech [1], [6], [7]. The first approach employs two monolingual ASR systems and an LID module. The LID module extracts the input code-switched utterances and then decides on the identity of each speech segment before passing them into their respective monolingual ASR systems. This approach is very simple because it applies acoustic and language modelling methods which achieve excellent monolingual performance. However, this approach is not preferred by many researchers due to LID error propagation which leads to poor ASR performance. The second approach employs a single-pass multilingual ASR system comprising 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. The approach needs a complete redesign of the acoustic and language models [6]. The major advantage of this approach is that it does not require the use of an LID system and it avoids the errors presented by the LID system.

In this research we propose a multilingual ASR system to perform LID on mixed-language corpora. We investigated whether the second approach can be adopted to achieve suitable multilingual acoustic modelling which can be used to handle Sepedi-English code-switching speech. We present the first study which relied solely on the mixed monolingual speech corpus of Sepedi and English but was evaluated using a code-switched speech corpus. Furthermore, we investigated which phoneme clustering method yields better ASR accuracy. We also examined how a multilingual acoustic model can impact LID classification accuracy in a mixed-language corpus. The novel approach proposed in this study is the first to offer a framework that integrates acoustic features and phonotactic information to achieve the LID system for mixed-language speech. This is a relevant study since it is common in South Africa for more than one language to be spoken in the same region.

2. RELATED WORK

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

There are few reported approaches in code-switched speech. One approach is to integrate multiple cues such as acoustics, prosodics and phonetics to distinguish between languages in a code-switched speech utterance [8]. A language boundary detection (LBD) method is applied to detect multiple languages within an utterance [9]. The second approach, such as the delta-Bayesian information criterion (Delta-BIC) and latent semantics analysis (LSA), has been used to separate English, Mandarin and Taiwanese in code-switched utterances [9]. Lastly, an approach that uses maximum a posteriori-based estimation was used to jointly segment and identify utterances of a mixed language [13]. The above approaches use an LID module that incorporates an LBD module. The LID systems that incorporate an LBD module are usually not preferred due to incorrect assumptions that code-switched speech segments are independent of each other and as a result, errors in the LID module cannot be recovered [1]. Therefore, if the LBD module cannot achieve 100%, it will directly influence the performance of the LID module, thereby limiting the performance of the speech recognition module [1], [10].

On the other hand, a multilingual ASR approach can handle code-switched speech. It comprises a multilingual acoustic model, a multilingual pronunciation dictionary and a multilingual language model that allows the mixing or sharing of models across different language units [1], [10]. A multilingual ASR approach does not need an additional LID module to identify speech segments since language information is incorporated directly into the system [1]. One technique is to use a linguistic knowledge-based method to establish a multilingual phone set mapping or clustering of similar phonetic features that share the training data [7]. Common examples are the International Phonetic Alphabet (IPA), Speech Assessment Methods Phonetic Alphabet (SAMPA) and Wordbet [15]. Another technique is to map language-dependent phones using a data-driven approach such as clustering specific phones according to distance measured between similar acoustic models. Examples of data-driven methods are the Confusion Matrix, Bhattacharyya Distances and Kullback-Leibler Divergent which takes spectral characteristics into consideration [15].

Lyu et al. [16] propose a word-based lexical model LID system which uses the lexicon information to distinguish between code-switching speech within an utterance. A two-stage scheme system is used with a large vocabulary continuous speech recognition (LVCSR) system. Then a trained word-based lexical model is applied to identify languages via recognised word sequences. The approaches such as Parallel Phone Recognition and Language Modelling (P-PRLM) [4, 18] and parallel phoneme recognition vector space modelling (PPR-VSM) [17] are some of the most popular approaches to the LID system. The P-PRLM approach employs multiple phoneme recognisers that tokenise the speech waveform into sequences of phonemes. The resulting sequence of phonemes is then passed to the n-gram language model which determines the most probable language from the target languages [6], [18]. The supervised support vector machine (SVM) model has proven to be the best classifier [18]. A similar approach was used to distinguish between 11 officially spoken languages of South Africa [18]. It was implemented using P-PRLM architecture and techniques such as phoneme frequency filtering - where an SVM-based classifier is used to classify languages at the back end. The SVM classifier was able to achieve an average LID rate of 71.78% on test samples of 3-10 seconds long and an LID rate of 82.39% when clustering similar language families [18]. The diagram below shows the P-PRLM system employed for LID in South African languages. The Figure 1 below shows the PPRLM system employed for LID in South African languages. This work is similar to this current research work, but was employed for single-language identification.

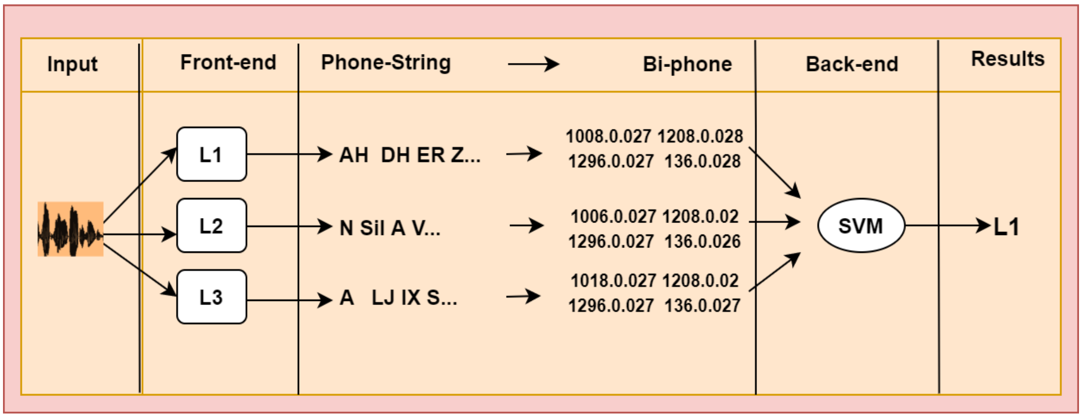

Figure 1. LID system based on P-PRLM scheme (adopted from[18]).

3. MIXED-LANGUAGE CORPUS

The speech corpus in this study contained two monolingual speech sets of data for Sepedi and English. The third speech corpus was Sepedi-English code-switched speech data which was recorded during the speech data collection phase. The speech corpora were divided into training and testing datasets. The amount of mixed-language speech data that was used for the training and testing of the system is summarised in the sections below.

3.1. Training Dataset

The corpus used for training the acoustic model included recorded speech data and the respective transcriptions of locally produced Sepedi developed within the Telkom Centre of Excellence for Speech Technology (TCoE4ST) and freely available LWAZI South African English speech data. The TCoE4ST locally produced Sepedi speech corpus had 3 465 utterances. From the LWAZI English speech corpus, 1 680 recorded speech data and the respective sentences from utterances were selected. The training dataset contained a total of 5 505 speech utterances and was 5.5 hours long.

3.2. Testing Dataset

The speech corpus used for testing the phone recognition contained 660 speech utterances which were not part of the training speech data. Code-switched speech is generally spoken but not formally written. However, it is not easy to find code-switched speech data. It is for this reason that simple finite loop grammar was used to generate 60 artificially code-switched sentences that were syntactically correct. These texts were recorded by 10 speakers to produce 660 utterances, 300 of which were used as a testing set and the remaining 300 were included in the training set. The quality of the utterances was manually improved by removing dysfluencies such as long pauses, laughs and hiccups. The average ratio of code-switched English words within each utterance was not more than 0.5. The testing utterances were 1 hour long.

4. PHONEME MAPPING METHODS

In this research project, two different phone mapping strategies are proposed to determine the similarity between the target languages. The first mapping strategy is based on an IPA-based scheme which requires linguistic experts, and the other strategy is a data-driven method derived by measuring a confusion matrix.

4.1. Linguistic-Knowledge Phoneme Mapping

The methods which are used to deal with similar phoneme inventories have been studied [1], [8], [9]. A single-pass speech recogniser on two languages with a multilingual acoustic modelling technique for the available speech corpus was proposed in this study. The multilingual acoustic model was developed by mapping the English phonemes to the Sepedi phonemes. This approach was motivated by the occurrence of similar phonemes from the target languages and also aimed to reduce a larger number of phonemes. The criteria that were used to construct the linguistic-knowledge strategy are described below.

Phoneme Mapping Criteria

C-1. If the IPA classification is like a Sepedi phoneme, then the English phoneme is mapped directly to the Sepedi phoneme.

C-2. Each English phoneme is mapped to its closest matching Sepedi phoneme.

C-3. If no closely matching phoneme is found, then the English phoneme that occurs most frequently in the phoneme inventory is extended to the phoneme set.

C-4. If none of the above criteria are applicable, then each phoneme is mapped to the Sepedi phoneme that it is mostly confused with, according to a confusion matrix.

The above criteria resulted in the phoneme mapping list indicated in Table 1 below. Phonemes such as ***/au/*** and ***/e@/***from the LWAZI dictionary were decoupled to a single phoneme ***/a/, /u/***and ***/e/, /@/*** respectively.

Table 1. Examples of the phoneme list achieved with linguistically motivated method.

Phoneme mapping list							
from	**to**	**from**	**to**	**from**	**to**	**from**	**to**
{	**E**	Oi	**O i**	i:	**i**	g	**k_>**
3:	**E**	p	**p_h**	i@	**i @**	@i	**@ i**
a:	**a**	Q	**O**	k	**k_h**	u:	**u**
au	**a i**	r\	**r**	O:	**O**	u@	**u**
ai	**a u**	t	**t_h**	@:	**a**	Z	**d_0Z**
d	**l’**	T	**f**	U	**u**	h_b	**h**
D	**l’**	tS	**tS_h**	v	**B**	**Additions**	
e@	**E @**	@u	**O**	z	**s**	@	**b**

In our case, the diphthongs of the English language were separated into vowels using the traditional IPA-based strategy in the mixed phoneme set. Each phonemic vowel was mapped toits equivalent Sepedi phoneme directly using C-2 criteria.

4.2. Data-driven Phoneme Mapping

The data-driven mapping which is based on the confusion matrix was built by including all the Sepedi and English phonemes [10, 11]. The confusion matrix was generated when the acoustic models of the source language were applied to the speech utterances of the target language. The recognised phoneme sequences of the source phoneme candidates were then mapped to phoneme sequences of the target phoneme candidates as indicated for Table 2. This mapping method consists of the counts of the confusion pairs 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 a linguistic expert, which can be time-consuming. For each phoneme of the

English language, the most often confused phoneme with the Sepedi language was selected for mapping. For each phoneme P_{L1} from the target language $\boldsymbol{P}_{L2}$, the best respective source candidate phoneme was matched. The similarities were measured by selecting the number of phoneme confusions as **C** ($\boldsymbol{P}_{L1}$, $\boldsymbol{P}_{L2}$). The target phoneme was matched as follows:

$$P_{Ln} = MaxC(P_{L1}, P_{L2}) \tag{1}$$

where ***L1*** denotes the target language (Sepedi) and ***L2*** the source language (English). Thus, for each target phoneme, a source candidate phoneme with the highest number of confusions was determined. If the same number of confusions occurred on two or more source candidate phonemes, the decision on the choice of the target phoneme was made by a knowledge expert. The same strategy was followed even when there were no confusions found between target and source candidate phonemes.

Table 2. Examples of the phoneme list achieved with data-driven method.

Phoneme mapping list			
From	**to**	**from**	**to**
{	a	Oi	**E**
3:	E	p	**p_>**
a:	**a**	Q	**O**
au	**i**	r\	**r**
ai	**u**	t	**t_h**
d	**l’**	T	**F**
D	**l’**	tS	**tS_h**
e@	**E**	@u	**O**
G	**G**	u:	**U**
@i	**E**	u@	**O**
i:	**E**	U	**U**
i@	**a**	v	**B**
k	**k_>**	z	**S**
O:	**O**	Z	**d_0Z**
@	**a**	b	**B**
h_b	**h**	@:	**a**

5. PROPOSED MULTILINGUAL ASR-LID SYSTEM

The multilingual ASR-LID system is targeted to identify only two languages. A multilingual recognition system takes speech waveform and outputs the corresponding phone sequences. This is done when an ASR system estimates the likelihood score of the optimal phone sequences given the acoustic features extracted from the speech utterance waveform. To achieve this, a multilingual acoustic and language model was employed to estimate the likelihood scores for the spoken utterance. The phoneme mapping technique was applied to generate the shared phoneme set for robust multilingual acoustic modelling. Figure 2 shows the proposed multilingual ASR system.

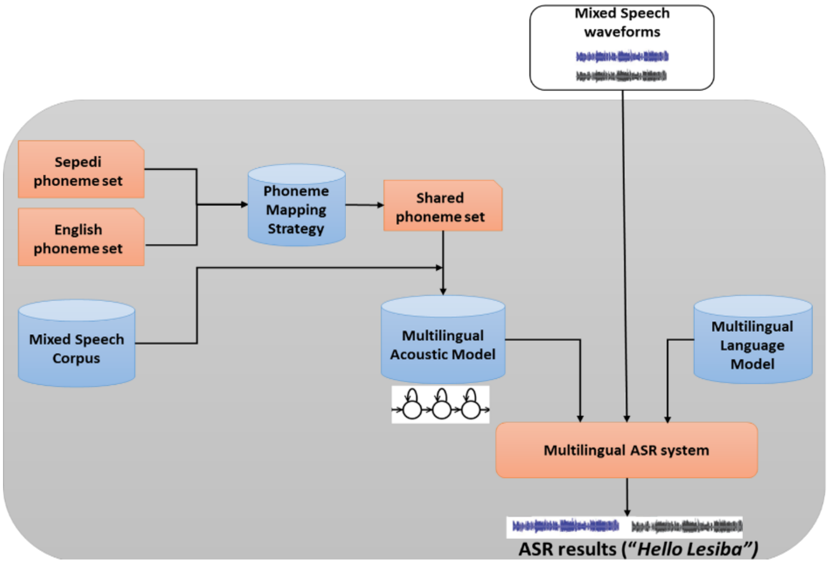

Figure 2. Multilingual automatic speech recognition system.

Multilingual acoustic models are used to perform HMM-based parameter re-estimation. For recognition purposes, the multilingual acoustic features were compared with the HMM-based multilingual acoustic models as well as the language model. The sequences of phone strings were decoded by the Viterbi decoding algorithm, which searches the optimal sequence of the phones using the combined likelihood scores from the multilingual acoustic model and language model.

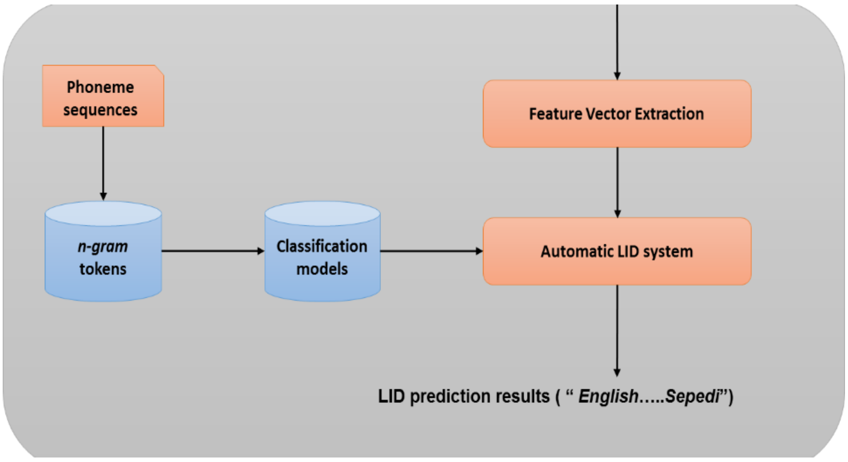
Phoneme sequences
Feature Vector Extraction
n-gram tokens
Classification models
Automatic LID system
LID prediction results (" *English.....Sepedi*")

Figure 3. The language identification system for mixed-language speech.

For each phone sequence generated from the ASR system, the bi-phone occurrences were extracted from the phone sequences and converted into a suitable SVM format with a unique numerical representation. This approach is like vector space modelling [5]. As a final step (see Figure 3), the SVM-based classifier was used to identify only two class feature samples; languages outside the targeted range were not classified. For each phoneme sequence generated from the ASR system, the phoneme occurrences were extracted from the phoneme sequences and converted into a suitable SVM format with a unique numerical representation. The classification model with the highest log-likelihood score was chosen to be the most likely sample for classification. The bi-phone vectors were then used as an input to the SVM-based classifier to build the classification model. The phoneme 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 phoneme feature, usually integer representation, and in our case, a feature value represents the frequency count or occurrences of each phoneme feature attribute. The SVM classification model was used to separate vectors in a binary classification and hypothesise the maximum likelihood score of the bi-phone frequencies of each language [18].

6. MODEL DEVELOPMENT AND SYSTEM SETUP

This section describes the experimental setup, the tools used to develop the ASR systems, LID system and the configurations of each system setup. All the multilingual ASR systems were developed, and experiments were performed using the hidden Markov Model Toolkit (HTK) [12]. The experimental results obtained from these ASR and LID systems are later analysed and discussed.

6.1. Baseline Acoustic Models

To build multilingual acoustic models, we applied a Hamming window of 25ms length with an overlapping window frame length of 10ms. Acoustic features were 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 were applied to the hidden Markov Models (HMM). The CMVN was used to overcome the undesired variations across the channels and distortion. Figure 4 shows the configuration file that was used for extracting speech features. The HMM-based ASR system was created with a widely used standard HTK [12]. The acoustic model used a three-state left-to-right HMM. The HMM-based system consisted of the tied-state triphones clustered by a decision tree technique. Each HMM state distribution was modelled by eight Gaussian mixture models (GMM) 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 phones during speech decoding.

```
CEPLIFTER     =   22
ENORMALISE  =   FALSE
NUMCEPS       =   12
NUMCHANS     =   26
PREEMCOEF   =   0.97
SAVECOMPRESSED =   TRUE
SAVEWITHCRC     =   FALSE
SOURCEFORMAT    =   WAVE
TARGETKIND      =   MFCC_0_D_A_Z
TARGETRATE      =   100000.0
USEHAMMING      =   TRUE
WINDOWSIZE      =   250000.0
ZMEANSOURCE     =   TRUE
LOFREQ          =      150
HIFREQ          =     4000
```

Figure 4. The configurations used for mfcc feature extraction

6.2. Language Modelling

To build the multilingual language model, many data texts had to be collected and normalised. The word coverage of the multilingual speech was improved by applying language-aware context-based text normalisation. Thus, for example, digits, temperature, time, currency amount, percentage, etc., were converted to appropriate words. A phone language model was incorporated in the speech recogniser for the purpose of robust speech decoding. The training transcriptions, together with the generated code-switched texts, were formatted into phone transcriptions and were used to develop the multilingual language model. The combined vocabulary that was used to train the language model consisted of over 85 000 unique word tokens for both Sepedi and English. The language model that was used for multilingual ASR experiments was implemented using the Stanford Research Institute language model toolkit [13]. It was trained independently with discount interpolation. The interpolation weights were optimised using the training set perplexity as a performance measure.

6.3. Multilingual Dictionary and Phoneme Set

The experimental bilingual pronunciation dictionary used was achieved by merging several monolingual pronunciation word lexicons without retaining duplicate words. For the primary Sepedi language, we used a limited vocabulary of a freely available Sepedi pronunciation dictionary that was locally produced within the TCoE4ST and LWAZI. For the English language, we used a freely available LWAZI English pronunciation dictionary often used for speech technology research tasks. All the words in the pronunciation dictionary were manually verified and checked for redundant phone representation. There were 7 176 Sepedi and 78 722 English words in the bilingual dictionaries. The dictionaries were further redesigned and rectified where necessary.

The combined bilingual dictionary contained 85 898 unique words. The representation used in the bilingual pronunciation dictionary followed the SAMPA notations based on IPA rules and also taking into consideration the pronunciation rules [8], [14]. Some 67 phones were combined into a mixed phone set, attained by combining Sepedi and English phones directly without silent phonemes. In this case, phones with similar phonetic features were mapped into one best phone candidate representation to lower confusion within the combined phone set. Some of the English vowel phones were left unmapped since they did not match any Sepedi vowel phones.

The combined mixed phone set included all phonemes of Sepedi and English that were used during the training phase without performing phoneme mapping. We used a knowledge-based IPA method to construct linguistically motivated phonetic pairwise mappings. The IPA-based phoneme set, and data-driven phoneme set contained 38 phonemes, excluding the silent phonemes. In this case, to train our multilingual acoustic model that effectively handled Sepedi-English code-switched speech, we adopted the technique used by Biswas et al. [6], Shan et al. [7] and Bhuvanagiri and Kopparapu [8]. Lastly, problematic words of Sepedi or English origin were manually reviewed for correct pronunciation prior to training the HMMs.

6.4. Language Identification Classifier

The SVM-based classifier based on bi-phone frequencies as an output was used to classify only two class feature samples; languages outside the targeted range were not classified. For each phone sequence generated from the phone recognition, each bi-phone occurrence was extracted from the phone sequences and converted into a suitable SVM format with a unique numerical representation called a bi-phone feature vector. Therefore, the phoneme features were calculated for every utterance. The bi-phone frequency vectors were then used as input to the back-end SVM classifier. The SVM classification model was used to separate vectors in a binary classification and hypothesise the maximum likelihood score of the bi-phone frequencies of each language. The bigram phones were trained to create the classification model.

The SVM-based classifier was implemented using LIBSVM library [15]. The SVM training dataset size was 12 147 KB of phone tokens. The training dataset that was used for training the SVM-based classifier was extracted from the phone-based transcriptions. The phone sequences were used to train the SVM classifier, which resulted in support vectors from models. The training process was also aimed at maximising the margin as well as minimising the training errors. The bi-phone vector attributes for both testing and training were scaled in the range of [0, 1]. The benefit of scaling datasets is to speed up the 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 was used to estimate the SVM parameters such as ***C***, ***gamma***, **margin error**, **trade-off parameter** and **kernel** width before training the classifier [5], [17]. The Radial Basis Function (RBF) kernel was used for training the classifier. We obtained the optimal parameter for this kernel and applied five-fold cross-validation to the training set and estimated each grid point for the accuracy of the classifier.

7. RESULTS AND DISCUSSION

For experimentation, three ASR and LID systems were developed. The baseline ASR system was achieved by directly combining monolingual ASR systems for Sepedi and English into a multilingual ASR system. The baseline (i.e. directly mixed) ASR-LID system was evaluated and compared with the multilingual ASR-LID systems that were developed using the two phoneme mapping techniques. No specific phoneme mapping was performed in the phoneme set. The phoneme set size was large with 67 phonemes. The HMM-based acoustic models trained on these systems contained eight Gaussian mixtures per state.

In this experiment, we modified a mixed recognition system by applying two different phoneme mapping techniques. We trained the front-end acoustic models on both Sepedi and English speech data and performed modelling of code-switching at the pronunciation dictionary level. Linguistically motivated and data-driven phoneme mapping methods were applied to determine the similarity between the phonemes of our target languages. We first applied a linguistically motivated phoneme mapping method using an IPA-based scheme. Our experiments were performed on the same speech data which was used for developing the directly mixed LID system. As a result of phoneme mapping, the number of phonemes in the directly combined phoneme set were reduced. The results obtained from the three systems using mixed-language speech are shown in Table 3.

The experimental results presented in Table 3 show that phone error rate (PER) and LID classification accuracy improved when the phoneme clustering methods were applied. The quality of the correct phoneme recognition output was typically captured by the PER metric formulated as:

$$PER = \frac{(S+I+D)}{N} \tag{2}$$

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.

In Table 3, the SVM classifier yielded a promising and acceptable LID accuracy rate of 95.2% on directly mixed speech utterances with a total of 3 201 support vectors. In this case, mixed speech utterances included both monolingual and code-switched utterances. The experimental results of the SVM-based LID classifier were also obtained using RBF kernel. The SVM-based classifier was trained using a five-fold cross-validation and RBF kernel which yielded an estimation rate of 99.75% on the trained classification models. Both phoneme mapping approaches achieved a significant improvement over the baseline system results. The data-driven approach outperformed the baseline directly mixed system and the IPA-based system. The IPA-based approach performed better with a PER of 4.5% but LID classification accuracy was about 9% lower. The data-driven approach performed better with a PER of 14.5% as well as a LID classification accuracy of 2.3%. These methods allow sharing of the parameters in the HMM-based acoustic models of the target languages.

The IPA-based LID system was able to achieve a better PER reduction of 4.5%, outperforming the directly combined mixed LID system. The best performance of the PER reduction of 19.2% was achieved by the data-driven LID system. The data-driven LID system achieved a PER difference of 9.4% compared to the IPA-based LID system. We also observed that the ASR system with a larger number of phonemes in the phoneme set performed badly compared to when a phoneme mapping method was engaged to reduce the phoneme set. The amount of quality training speech data can also improve the multilingual ASR system performance significantly.

Table 3. Experimental results showing the PER of the multilingual ASR and LID classification accuracy with 8, 16, 32 and 64 Gaussian mixtures per state.

Number of Gaussian Mixtures	**ASR-LID Systems**	**PER(%)**	**LID Accuracy (%)**
8	Directly mixed	33.2	95.2
	IPA-based	**28.7**	**85.8**
	Data-driven	**19.2**	**87.3**
16	Directly mixed	21.4	83.8
	IPA-based	**17.8**	**89.7**
	Data-driven	**15.6**	**89.5**
32	Directly mixed	12.9	83.8
	IPA-based	**12.9**	**84.7**
	Data-driven	**7.3**	**96.7**
64	Directly mixed	5.4	83.9
	IPA-based	**7.3**	**83.7**
	Data-driven	5.1	**94.5**

Table 3 shows the PER and LID accuracy or rate that was attained on the four LID systems when the HMM-based acoustic models contained 16 Gaussian mixtures per state. The use of context-dependent HMM-based acoustic models with increased Gaussian mixtures per state was adopted during training of the acoustic models as they tend to improve the performance of the phoneme recognition systems. The PER has a direct proportionate relationship to Gaussian mixtures per state. 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 repeated until the acoustic models had 32 Gaussian mixtures per state, after which the phoneme recognition results no longer improved significantly on the test set. The performance of the PER and LID classification accuracy was significantly better on both IPA-based and data-driven LID systems. Both systems were able to outperform the directly mixed LID system only in the LID classification accuracy. A slight difference of 0.12% in LID classification accuracy was observed between the IPA-based and data-driven LID systems. This was due to a larger number of the phoneme occurrences that were observed within the testing set.

Table 3 shows the PER and the LID classification accuracy attained on the four LID systems when the HMM-based acoustic models contained 32 Gaussian mixtures per state. The PER reduction was better, but the LID accuracy was a bit lower than expected for the data-driven LID system. We observed that a better performance of the PER does not necessarily result in a positive bias of the LID accuracy on both directly mixed and IPA-based LID systems, since the phoneme recognition systems are used only to obtain phoneme strings for SVM-based classifier training.

A slight reduction of the PER was achieved with eight Gaussian mixtures, but a better PER reduction was achieved with 32 Gaussian mixtures. This clearly shows that a data-driven LID system achieves a better PER with all Gaussian mixtures per HMM state represented in Table 3. The directly mixed LID and IPA-based LID system nearly achieved the same PER with 32Gaussian mixtures per HMM-based acoustic model state. However, a slight improvement of 0.3% was achieved. The phoneme recognition results in the experiment show that the application of phoneme mapping methods to our targeted languages and the increase of Gaussian mixture per shared HMM-based acoustic model significantly improve the performance of the phoneme recognition and LID system.

We also observed that by applying IPA-based and data-driven phoneme mapping techniques, these could yield extreme results such as increased sentence and phone correctness, phone recognition and LID accuracy of the proposed mixed-language integrated LID system, as well as reduced PER when 32 Gaussian mixtures per HMM state are considered. However, both the directly mixed ASR-LID system and data-driven system show a significant reduction of PER with 64 Gaussian mixtures but no further increase in the LID classification accuracy as indicated in Table 3.

8. CONCLUSIONS

This paper presents an integration of multilingual speech recognition into LID for code-switched speech using phonotactic features as language information. In this research work, two strategies are reported on to perform similar phoneme mapping of the target official languages. We have shown that existing monolingual corpora can handle code-switching utterances. The IPA-based approach is derived from linguistic knowledge, whereas the data-driven approach is based on the confusion matrix. Appropriate phoneme mapping approaches across the target languages offered robust context-dependent multilingual acoustic models which tended to produce acceptable ASR-LID system performance. Our proposed IPA-based and data-driven approaches have shown a significant improvement in both PER and LID classification accuracy. The data-driven method outperforms the IPA-based approach. An acceptable PER was achieved with the data-driven approach when multilingual acoustic models were employed that were trained with 32 Gaussian mixtures per state. Again, the 64 Gaussian mixtures do improve the PER, but has no impact on the performance of LID accuracy. In future, we hope to train the multilingual ASR-LID system with more robust context-dependent code-switched acoustic models for further evaluation and performance analysis. We also aim to collect more code-switched speech for ASR and LID research in future. As part of extending this research work, we aim to investigate more South African languages where speakers use code-switching in their daily conversations.

ACKNOWLEDGEMENTS

We acknowledge the team from Council for Scientific and Industrial Research (CSIR) that developed LWAZI speech data and the University of Limpopo for allowing us to use their Sepedi speech corpus. We also acknowledge Telkom SA for funding this project.

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, pp. 1–4.

[2] K. R. Mabokela, “A multilingual ASR of Sepedi-English code-switched speech for automatic language identification,” in International Multidisciplinary Information Technology and Engineering Conference (IMITEC), November 2019, pp. 430–437.

[3] K. Mabokela, M. Manamela, and M. Manaileng, “Modeling codeswitching speech on under-resourced languages for language identification,” in SLTU 2014, May, pp. 225–230.

[4]` L. Besacier, E. Barnard, A. Karpov, and T. Schultz, “Automatic speech recognition for under-resourced languages: A survey,” Speech Communication Journal, vol. 56, pp. 85–100, 2014.

[5] H. Li, B. Ma, and K. Lee, “Spoken language recognition: from fundamentals to practice,” in Proceedings of the IEEE, May 2014, pp. 1136– 1159.

[6] A. Biswas, F. de Wet, E. van der Westhuizen, E. Ylmaz, and T. Niesler, “Multilingual neural network acoustic modelling for ASR of under-resourced English-isiZulu code-switched speech,” in Proc. Interspeech 2018, pp. 2603–2607.

[7] C. Shan, C. Weng, G. Wang, D. Su, M. Luo, D. Yu, and L. Xie, “Investigating end-to-end speech recognition for Mandarin-English codeswitching,” in 2019 IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 6056– 6060.

[8] K. Bhuvanagiri and S. K. Kopparapu, “Mixed language speech recognition without explicit identification,” In American Journal of Signal Processing, 2012, pp. 92–97.

[9] F. Diehl, “Multilingual and cross-lingual acoustic modelling for automatic speech recognition,” PhD dissertation, Universitat Politecnica de Catalunya, Newark, 2007. [Online]. Available: http://mi.eng.cam.ac.uk/ fd257/publications/

[10] T. Modipa, M. Davel, and F. De Wet, “Pronunciation modelling of foreign words for Sepedi ASR,” in Proceedings of Pattern Recognition Association of South Africa, 2010, p. 185-189.

[11] T. Modipa and M. H. Davel, “Implications of Sepedi/English code switching for ASR systems,” in 24th Annual Symposium of the Pattern Recognition Association of South Africa, 2013, pp. 64–69.

[12] S. J. Young, A. D. K. G. Evermann, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, in Cambridge University (For HTK Version 3.2.1), 2013. [Online]. Available: http://htk.eng.cam.ac.uk

[13] A. Stolcke, “Srilm - an extensible language modelling toolkit,” in Proc. ICSLP, 2002, pp. 901–904.

[14] U. T. W. Zhirong, T. Schultz, and A. Waibel, “Towards universal speech recognition,” In Proc. ICMI 2002, 2002.

[15] C. C 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. Davel, “Language identification of individual words with joint sequence models,” in Proceedings of Interspeech 2014, 2014, pp. 1400 –1404.

[17] M. Peche, M. Davel, and E. Barnard, “Development of a spoken language identification system for South African languages,” in SAIEE Africa Research Journal, vol. 100(4), Dec 2009.

[18] E. Ambikairajah, H. Li, L. Wang, B. Yin, and V. Sethu, "Language identification: atutorial", In IEEE Circuits and Systems Magazine, 11(2), pp.82-108, 2011.

AUTHORS

Mr. Ronny Mabokela holds an MSc. degree in Computer Science from the University of Limpopo. Mr Mabokela has vast industry experience, having worked for Telkom and Vodacom South Africa. He was one of the remarkable team that established the formation of high-speed fibre-based internet. He contributed to the successful development of Vodacom internal systems, API integrations and business process automation. He received numerous awards, including being an exceptional performer in both Telkom & Vodacom and seconds prize award of research excellence at University of Limpopo (2013). He became a session chair and peer reviewer of the Southern Africa Telecommunication Networks and Applications Conference (SATNAC) and International Conference on Computing and Communications Technologies (ICCT) in 2014 and 2015, respectively. He has presented his work on numerous platforms including International Workshop on Spoken Language Technologies (SLTU) in Russia. He has a keen research interest to broadband network services, computational linguistics, natural language processing, speech technologies and machine learning.