SIGNAL BASED ETHIOPIAN LANGUAGES IDENTIFICATION USING GAUSSIAN MIXTURE MODEL

Mikias Wondimu and Menore Tekeba
Emails: mikowond@gmail.com, menore.tekeba@aait.edu.et
School of Electrical and Computer Engineering, Addis Ababa Institute of Technology, AAU

ABSTRACT

A language identification (LID) system is an approach in which machines can determine the language and identify it from relatively brief audio spoken samples. Very few attempts have been made on LID Systems for Ethiopian languages. The importance of LID is increasing due to the development of telecommunication infrastructures. Using an LID, service calls from customers can be forwarded to a person who knows the language. Therefore, an LID system involving four Ethiopian languages (Amharic, Oromiffa, Guragegna and Tigregna) is done using Gaussian mixture models (GMM). A dataset consisted of recordings from seven different speakers of each language was prepared and after preprocessing the data, the features are extracted using Mel frequency cepstral coefficients (MFCC) and classification is done using GMM. The performance of the LID system was tested with scenarios by taking two, three and four languages at a time. The LID system is also tested for utterance and speaker dependence performances. The average accuracy of utterance dependent LID test for the four languages was about 93%, the utterance independent test for the four languages was about 70% while the speaker independent test, being tested on utterance dependent scenario only, for the four languages was nearly 91%.

Keywords: *Accuracy, GMM, LID, Language Identification System, MFCC, Utterance*

INTRODUCTION

Language is a means used for human communication either in the form of speech or text. Speech is primarily intended to convey some message. The speech signal contains not only the intended message, but also the characteristics of the utterance speaker and the language of communication. The language is conveyed through the sequence of sound units. The present work focuses on signal form of speech. With the growth of global partnership, the demand for communication across the languages is increasing. This has given rise to new challenges for automatic language recognition, followed by speech recognition system before the machine can understand the meaning of the utterance [1].

Automatic LID is the task of automatically recognizing a language from a given spoken utterance. With increasing interest in multi-lingual speech systems, such as international telephone-based information access, there has been a great deal of research in LID techniques over the last decades. The importance of having an efficient LID system dealing with large databases of languages is to allow for further processing to be carried out on the hypothesized languages [2].

There was no research in LID up to 1970. Even though it was started in early 1970's, there was no momentum in this area for nearly 20 years. Afterwards, much progress was made taking the advantage of the openly available multilingual corpus of speech [3]. The LID systems implemented till date mainly vary in their methods for modeling languages.

All the LID systems can be broadly classified into two groups, namely, text-based (Explicit) and signal-based (Implicit) LID systems. All the existing LID systems use some amount of language specific information [3]. These two approaches differ only in the extent of information used for the LID task. The performance evaluation of an LID system comprises of accuracy and complexity.

It mainly depends on the amount of linguistic information given to the system. In the training stage some systems require only the speech signal and the true identity of the language [1]. In this type of systems language models are created from the speech signals alone, which are given to the system at the time of training. The text-based LID systems may require segmented and labeled speech corpus of all languages under consideration to create the language models during training [1]. Even though the performance of text based LID systems is better than signal-based systems, inserting a new language into such system is a difficult task. If the number of languages under consideration is large, it is obvious to make a choice between the performance and simplicity [1].

Signal (speech) based LID has various applications where one application could be a telephone based front whose main work is to route the call to the corresponding operator who is knowledgeable to that language. Other application of LID would be in the speech-to-speech translation, shopping, airports and other commercial areas.

The main goal of this research is to develop and test LID system for the selected four languages spoken in Ethiopia. The system development consists of three important steps: it starts by recording and preparing the raw speech data and followed by the training stage and finally to determine the effectiveness of the system the evaluation stage proceeded.

LITERATURE REVIEW

The paper only made a review on speech signal-based LID systems as thework is signal based from spoken utterances.

Spectral similarity approach for LID is used which concentrates on the differences in spectral content among languages [4]. The main aim of an acoustic feature based LID is to capture the fundamental differences between the languages. These can be captured by modeling the distribution of spectral features which can be done by extracting a language independent set of spectral features from segments of speech. The differences in phoneme inventory, variations in the frequency of occurrence of phonemes and acoustic realization of similar phonemes cause the languages to differ from each other in their short time acoustic features [3].

Calvin Nkadimeng demonstrated an LID system for African languages that is based on simple stochastic models and implementations of various approaches. The use of GMMs in various configurations and using various MFCC-based parameterizations were evaluated. It was found that increasing mixtures led to a general improvement, but leveled out above 300 mixtures. For single GMM systems MFCC gave the best performance.

However, when they train the system using a Universal Background Model (UBM), small further improvements are achieved by also including acceleration coefficients. Using full covariance did not improve with use of diagonal covariance when the number of parameters increased [2].

Pinki Roy and Pradip K. Das presented the efficiency of a speech dependent LID system for four Indian languages namely Indian English, Hindi, Assamese and Bengali. The evaluation of languages is done on standard recorded databases where the features are extracted using MFCC and classification is done using GMM. The results show that the accuracy of LID is best for all languages in mixture order of 1024. The accuracy of LID is very good lowest up to 93% for Assamese and highest up to 100% for Bengali, Hindi and English [5].

David Martinez, Lukas Burget, Luciana Ferrer and Nicolas Scheffer [7]: an automatic language recognition system that extracts prosody information from speech and makes decisions about the language with a generative classifier based on iVectors is built. The system is tested on the NIST LRE09 dataset. The prosodic system (2048 Gaussians, 400- dimension iVectors) and the fusion of both systems improve performance in all conditions. The relative improvements obtained over the acoustic system are: 10.93% for 3 seconds; 15.24% for 10 seconds; and 9.39% for 30 seconds [7].

L.F. Lamel and J.L. Gauvain [8]: Demonstrated phone-based acoustic likelihoods to the problem of LID using laboratory quality speech. With 2 sec of speech the LID performed around 99% accuracy on average. On spontaneous telephone speech from the Oregon Graduate Institute (OGI) corpus, the language can be identified as French or English with 82% accuracy with 10s of speech. The 10 LID rate using the OGI corpus is 59.7% with 10s of signal [8].

Koena R. Mabokela, Madimetja J. D. Manamela and Nalson Gasela [9]: This paper presents an approach to the development of the automatic LID system on mixed-language speech. Speech corpus to be used for simulation involves mixed utterances of Northern Sotho (aka Sepedi) and English. Language boundary detection methods are used to identify multiple languages within an utterance. Overall, the research work aims at ultimately enhancing the performance of a general-purpose speech recognizer with automatic LID capabilities [9].

From the review of the signal-based approaches (acoustic-phonetics and prosody approaches) the research focuses on the acoustic approaches of an LID system. An acoustic-phonetic approach is advantageous over other signal-based LID systems in that it doesn't require language specific knowledge. Because of this, the development and insertion of new language into the system is not a difficult task. Therefore, thisresearch has used acoustic-phonetic approach for the LID implementation of the four Ethiopian languages selected.

Methodology of LiD System and Data Processing

The paper uses different methods for the implementation of LID system for Ethiopian languages. The methods used and the descriptions of algorithms used for data collection, processing and final classification is given in the following sections from 3.1 to 3.5.

Preparation of Database Mono Channel and Sampling

For preparation of the database mono channel recording was done for Amharic, Guragegna, Oromiffa and Tigregna languages in relatively closed and quiet noise-free room. Seven male native speakers in the age group of 20-30 for each language have been selected. They are made to utter one paragraph and one sentence which were used for training and testing as shown in Table 1 below.

Table 1: Dataset description of utterance dependent and independent system

Language	Amharic	Guragegna	Tigrigna	Oromiffa	Total
No. of Speaker (N)	7	7	7	7	28
No. of times each speaker utters	15	15	15	15	60
Total training sample(N*10)	70	70	70	70	280
Utterance dependent sample(1 min long each)	35	35	35	35	140
Utterance Independent sample(3-5sec long each)	35	35	35	35	140
Sample Frequency	16kHz				

Each speaker is instructed to utter the same paragraph (~1min long) for 15 times and the same sentence ((3-5) sec long) for 5 times. Out of 15 recorded one minute long utterance database, 10 were used for training the system and the remaining five utterances were used for testing the accuracy of the LID system.

All (3-5) sec long speech database was used to test the accuracy of the model for the utterance independent system. The database mono channel was prepared with the help of Adobe Audition 3.0.

The LID System Model

The LID system comprises of pre-processing, feature extraction and classification. First the raw speech of known language is given to pre-processing. Once the speech is pre-processed the next step is extracting important features from the speech signal. The important spectral features are next given to GMM for training and creating a model for each language. Once the system creates the unique model for each language and it will be compared to the speech of the unknown language. After choosing the best and the most likelihood languages from all models the system gives its decision. Figure 1 shows the whole processes that have been carried on for LID task.

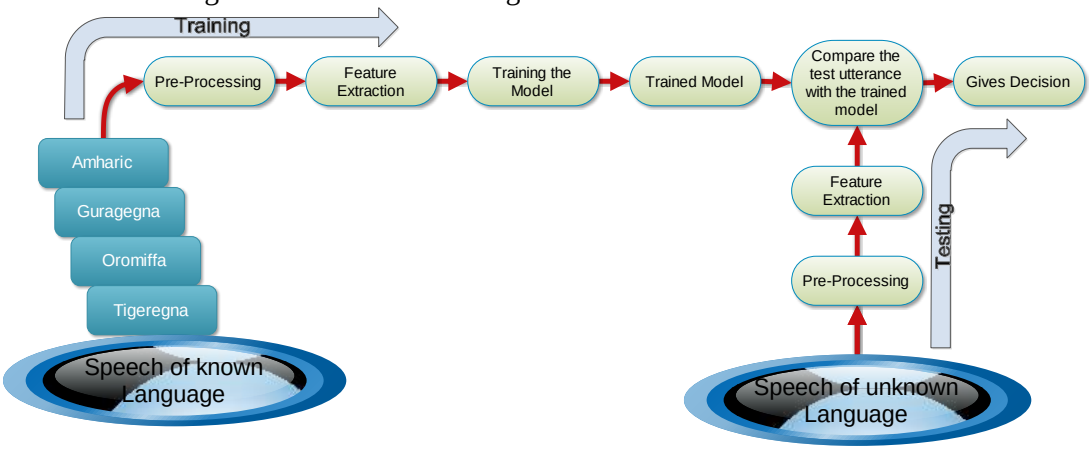

Figure 1: The LID System Model

Speech Data Pre- Processing

The first step in LID system is processing of the row speech data to be compatible with the feature extraction used for the study. Pre-processing speech is converting of the analog representations (first air pressure, and then analog electrical signals in a microphone) into a digital signal. The process of analog-to-digital conversation of speech signal has two steps: sampling and quantization.

A signal is sampled by measuring its amplitude at a particular time and to have a better sampling it is necessary to have at least two samples in each cycle (i.e. one measuring the positive part of the wave and one measuring the negative part of the wave). According to the Nyquist-Shannon sampling theorem a time-continuous signal x(t) that is bandlimited to a certain finite frequency fmax needs to be sampled with a sampling frequency of at least 2fmax [10].

Most information in human speech is in frequencies below 10 KHz; thus, taking Nyquist theorem into account a 20 KHz (2fmax) sampling rate would be necessary for complete accuracy. A 16 KHz sampling rate (sometimes called wideband) is often used for microphone speech. Since the speech is recorded using microphone the appropriate sampling for the present work is 16 KHz sampling. The sampled digital signal usually stored as integers, either 8 bit or 16 bit. This process of representing real-valued numbers as integers is called quantization.The process of analog to digital conversion of the wave form is done using software called wavesurfer 1.8.8p4.

Feature Extraction

The speech signal cannot directly given to the LID system (i.e. weaker signal has to be amplified, longer silences have to be removed and speech with background noise is to be extracted for further processing).

A feature vector should emphasize the important information regarding the specific task and suppress all other information which is not required. The speaker dependent characteristics, the characteristics of the environment and recording equipment should be suppressed because these characteristics do not contain any information about the linguistic message. Furthermore, the feature extraction should reduce the dimensionality of the data to reduce the computation time and the number of training samples [11].

The feature analysis component of the LID system plays a crucial role in the overall performance of the system. Many feature extraction techniques are available which include:

- Linear Predictive Coding (LPC)
- Perceptual Linear Predictive Coefficients (PLP)
- Dynamic Time Warping (DTW)
- Relative spectral filtering of log domain coefficients (RASTA)
- Mel-frequency cepstral coefficients (MFCC)

Linear Predictive Coding (LPC)

The idea behind the Linear predictive coding analysis is that a speech sample can be approximated as a linear combination of past speech samples. LPC is a frame based analysis of the speech signal which is performed to provide observation vectors of speech [16]. To compute LPC features, initially the digitized speech signal is put through a low order digital system. The output of the pre-empesizer network is blocked into frames of N samples. After frame blocking, the next step is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame. The next step is to auto correlate each frame of windowed signal. Finally, the LPC analysis that converts each frame of autocorrelations into LPC parameter set by Durbin's method [17].

Perceptual Linear Predictive Coefficients (PLP)

The perceptual linear prediction model developed by Hermansky. The PLP speech analysis technique is based on the short-term spectrum of speech and it provides the human speech based on the concept of the psychophysics of hearing. PLP discards irrelevant information of the speech and thus improves speech recognition rate. It is identical to LPC except that its spectral characteristics have been transformed to match the characteristics of the human auditory system [18].

Dynamic Time Warping (DTW)

DTW is a time series alignment algorithm developed originally for speech recognition. It aims at aligning two sequences of feature vectors by warping the time axis iteratively until an optimal match between the two sequences is found. Dynamic time warping (DTW) is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed. This technique is also used to find the optimal alignment between two time series. If one time series may be "warped" non-linearly by stretching or shrinking it along its time axis. This warping between two time series can then be used to find the corresponding regions between the two time series or to determine the similarity between the two time series [16].

Relative Spectral Filtering log domain coefficients (RASTA)

The term RASTA comes from the word RelAtive SpecTrA. RASTA processing is studied in a spectral domain which is linear-like for small spectral value and logarithmic-like for a large spectral value. The rate of change of non-linguistic components in speech often lies outside the typical rate of change of vocal tract shape. RASTA filtering is often coupled with PLP for robust speech recognition.

It is a separate technique that applies a band-pass filter to the energy in each frequency sub band in order to smooth over short-term noise variations and to remove any constant offset resulting from static spectral coloration in the speech channel [19].

Mel-frequency cepstral coefficients (MFCC)

Till now, for conventional LID systems, features are extracted using Mel-frequency cepstral coefficients (MFCC) [1]. This paper focuses on this most popular, prevalent, widely used and efficient technique for featureextraction[1,2,19]. Humans have the ability of distinguishing languages without having a much knowledge of that language. The idea behind language identification is the representation of human capability into machine understanding.

The speech generated by humans is filtered by the shape of the vocal tract and representing this shape accurately is the main job of feature extraction techniques. The advantage of MFCC is that its ability to represent this shape in a more appropriate and accurate way. Therefore, MFCC have been used as a tool to represent the signal with its important spectral feature vectors. The block diagram [12] shown in the Figure 2 shows the main steps of MFCC.

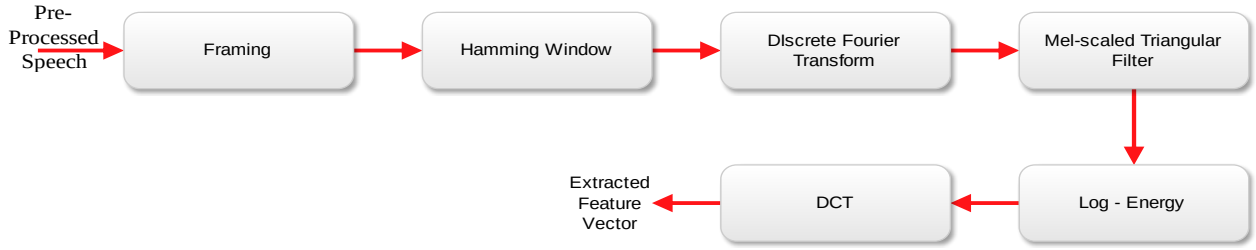

Figure 2: Block diagram showing MFCC main steps

Speech is a non-stationary signal and hence it is necessary to extract spectral features from a small window of speech that characterizes a particular sub-phone. The speech extracted from each window is called a frame.

The more common window used in MFCC extraction is the Hamming window, which shrinks the values of the signal near to zero at the window boundaries, avoiding discontinuities. The hamming window is described byequn. [12]:

$$w[n]= 0.54 - 0.46 \cos(2\pi n/L) \quad 0 \leq n \leq L-1 \quad [1]$$

Where w[n] is the value of the window at time n and L is the speech extracted from each window (frame)The next step is to extract spectral information for the windowed signal and to do so how much energy the signal contains at different frequency bands should be known. The extraction of spectral information for discrete frequency bands for a discrete time signal is the Discrete Fourier Transform (DFT).

The advantage of applying the mel-scale is that it approximates the nonlinear frequency resolution of the human ear [6]. The formula for converting from frequency to Mel scale is:

$$M(f) = 1125 \ln(1 + f\backslash 700) \quad [2]$$

Where $M(f)-$ Denotes the mel scale in frequency domain and $f-$ Denotes frequency Once the mel fiter bank energies are calculated, the nest step is to take the log of each of the mel spectrum values. In general, the human response to signal level is logarithmic. In addition, taking logarithms allows us to use cepstral mean subtraction, which is a channel normalization technique [12].Finally, a discrete cosine transform is applied to the log of the filter bank results in the raw MFCC vector. The highest cepstral coefficients are omitted to smooth the cepstral and minimize the influence of the pitch which are irrelevant to the LID process [12].

Language Classification

From literatures, the most popular classification techniques in the area of language identification are Deep Neural Network (DNN) and GMM. A DNN is a feed-forward, artificial neural network that has more than one layer of hidden units between its inputs and its outputs. DNNs are more accurate classifiers. DNN training is extremely time consuming, even with the aid of a graphical processing unit (GPU) or lots of CPU. The DNN training algorithm is very complicated because it's not guaranteed to converge to an optimal point in short period of time and requires huge computational resources [12].

Compared to DNNs, GMM is faster to compute and easier to learn. GMM training is reliable and if the data is clean enough , it is guaranteed to be trained a good system. The GMM approach to classification has been widely used in a variety of language processing applications. The system structure of a GMM based LID system is very simple. Accordingly, the computational requirements for processing are low. This simplicity advantage also extends to the development phase for such a system. The GMM LID system has significant potential advantage over other LID systems since they do not require orthographically or phonetically transcribed speech and are far more computationally efficient [7].

In GMM, language classification is performed according to the likelihood score calculated by the language-GMMs against a given feature vector. To determine the language in LID testing, multiple feature vectors are used. That is, likelihood scores are accumulated for each language and the decision making is delayed until the entire feature vectors are processed [13].

In a GMM model, the probability distribution of the observed data takes the form given by the following equation [1],

$$\rho(\bar{x}|\lambda) = \sum_{i=1}^{M} \rho_i b_i(\bar{x}) \qquad [3]$$

Where, M is the number of component densities, $\bar{x}$ is a D dimensional observed data, $b_i(\bar{x})$ is the component density as given in equation 4 below and ρ_i is the mixture weight for $i = 1, \dots, M$ as shown in Figure 3 below [18].

$$b_i(\bar{x}) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} \exp\{-\frac{1}{2}(\bar{x} - \bar{\mu_i})'\sum_i^{-1}(\bar{x}-\bar{\mu_i})\} \qquad [4]$$

Each component density $b_i(\bar{x})$ denotes a D-dimensional normal distribution with mean vector $\bar{\mu_i}$ and covariance matrix $\sum_i$. The mixture weights satisfy the condition $\sum_{i=1}^{M} \rho_i = 1$ and therefore represent positive scalar values. These parameters can be collectively represented $\lambda = \{\rho_i, \bar{\mu_i}, \sum_i\}$ for $i = 1, \dots, M$. Each language in a LID system can be represented by one distinct GMM and is referred by the language models λ_i, for $i = 1, 2, 3 \dots N$, where N is the number of languages.

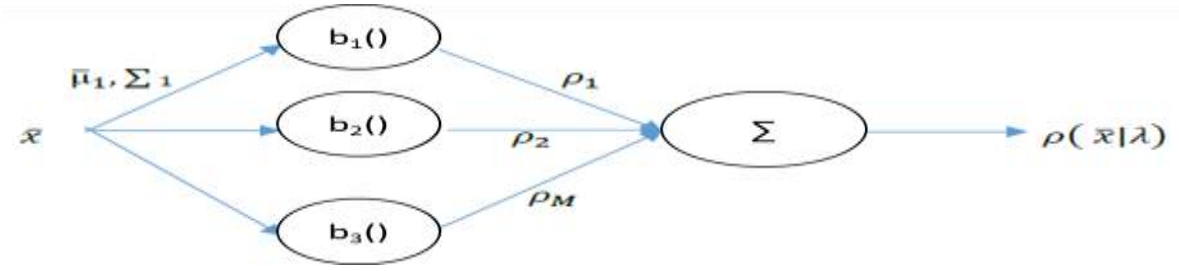

Figure 3: Diagram of Gaussian Mixture Model

Training GMM Classifier

In the training phase, a multivariate GMM for the spectral or cepstral feature vectors is created for each language. In the recognition phase, the likelihood of the test utterance feature vectors is computed given each of the training models. The language of the model having the maximum likelihood is anticipated as the language of the utterance.

The Expectation Maximization (EM) Algorithm is an iterative optimization of the means, variances and mixture weights of the M basis distributions of a Gaussian mixture model. The aim is to optimize the likelihood that the given data points are generated by the mixture of Gaussians. The EM algorithm alternates between performing an expectation (E) step, and a maximization (M) step [2].

- E - Computes an expectation of the likelihood by including the latent variables as if they were observed variables.
- M - Estimates the parameters by maximizing the expected likelihood found in the E step.

This technique is commonly referred to as the EM algorithm as given in [2]. An iterative approach is followed for computing the GMM model parameters using EM algorithm [1]. The aim of training is to obtain the mean, variance, and weighting of each Gaussian distribution (λ).

Steps for training

1. Begin with an initial model λ then calculate the new mean, variance weighting for the new model $\bar{\lambda}$.
2. Check if the newly calculated parameters are more suitable to model the language by using the following formula.

$$\rho(i|\bar{x}_t, \lambda) = \frac{\rho_i b_i(\bar{x}_t)}{\sum_{k=1}^{M} \rho_k b_k(\bar{x}_i)} \quad [5]$$

3. If the $\rho(X|\bar{\lambda})$ is larger than the $\rho(X|\lambda)$, then the new model $\bar{\lambda}$ is used to do the training again. i.e.

$$\rho(X|\bar{\lambda}) \geq \rho(X|\lambda) \quad [6]$$

4. Continue to do the training by repeating step (2) and step (3).

Where λ_i is model for $i = 1, 2, 3 \ldots N$ and N is the number of languages, $\bar{x}$ is a D dimensional observed data, $b_i(\bar{x})$ is the component density, ρ_i is the mixture weight for $i = 1, \ldots, M$ and M is the number of component densities and $\rho(x|\lambda)$ is the conditional probability and vector $X = \{x_1, x_2, \ldots, x_t\}$

When procedure is repeated to train the new model $\bar{\lambda}$, the new parameters are more close to the actual parameter for modeling the language. The error between the actual parameter for the model and λ become smaller and smaller through training. This procedure is repeated until the error is reached to certain threshold or stopping criterion is met [1].

In [15] it is stated that the iteration of the EM-algorithm was tested in 500, 250, 100, 50 and 10. The research shows that the EM algorithm converges quickly to a good state due to its local search nature even in 10 iterations. Due to this in this research an iteration of 100 was chosen for the EM algorithm of theresearch work. However, the GMM mixture order is dependent with the processing power of the device used for the training and the research given in [5] shows with the increased number of GMM mixture order, the classification accuracy increases. In this research since the device used for the training was having limited computational capability (Intel core i3 laptop with 4GB memory), 16 GMM mixture order is used.

LID IMPLEMENTATION AND TEST RESULT

LID Implementation and the Test Metrics

To train and test the LID model an auditory tool box was used [14]. The paper [2] [5] also uses the same auditory MATLAB toolbox for training and testing the LID model. Additional MATLAB codes were used in addition to the auditory toolbox. The paper uses two modules of which the first module combines theselected features extraction algorithm, MFCC, and GMM training algorithm implementations. The

second module also contains MFCC along with the testing code for GMM classifier which has been trained and its parameters being tuned by the first module.
With the training parameters given in section 3.5.1, the time duration it takes to train the system is about 56 min for the selected four languages. However, the decision time of the LID system for all conditions and tests is so fast. For any test either for 1min or (3-5) tests it gives decision within a second.

The LID system accuracy is tested for an utterance dependent, an utterance independent and speaker independent systems.
Here, Single Language Accuracy (%) and System Accuracy (%) is defined:

$$\text{Single Language Accuracy}(\%) = \left(\frac{\text{correct}}{\text{total}}\right) * 100 \quad [7]$$

$$\text{System Accuracy}(\%) = \frac{\text{Accuracy}(\%)\text{ of L1} + \text{Accuracy}(\%)\text{ of L2} + \cdots + \text{Accuracy}(\%)\text{of Ln}}{\text{n}} \quad [8]$$

Where correct is number of samples correctly classified, total is total number of samples given for testing, $\text{Accuracy}(\%)\text{ of L1}, \ldots, \text{Ln}$ are Accuracy of individual language accuracies and n is number of language

Results of the Tests on LID System

It is clear that it is somehow difficult to train GMM and get a better LID system performance with an utterance independent system with such a small recorded database. But even under such condition, the test has shown an excellent result for utterance dependent LID and a promising result for the utterance independent LID system. The experimental results of each test for the LID performance for two, three and four language classification was done and the result shows that the accuracy of classification decreases as the number of languages in the LID increases as shown in Table 2 to 5 and Figures 6 to 9 below.

Table 2: Test result for utterance dependent/ independent of Amharic and Oromiffa language

	Utterance dependent		Utterance independent	
Test Utterance	Amharic (1min Long)	Oromiffa (1min Long)	Amharic ((3-5)sec Long)	Oromiffa ((3-5)sec Long)
Accuracy (%)	100	100	91.43	77.14

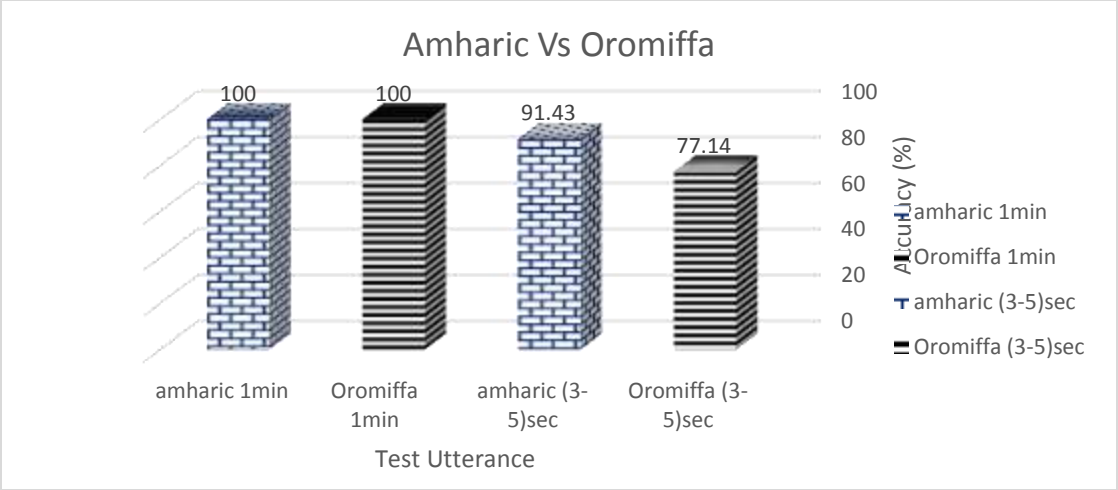

Figure 4: Test result for utterance dependent and independent tests for two languages

As shown in the Table 2 and Figure 6, the minimum accuracy for two language classification is 77.14% for utterance independent LID of Oromiffa and the maximum is 100% for utterance dependent classification of both languages. Two language classification for all other pairs of the four languages have been used.

Table 3: Test result for utterance dependent/ independent of Amharic/Guragegna/Tigregna language.

	Utterance dependent			Utterance independent		
Test Utterance	Amharic (1min Long)	Guragegna (1min Long)	Tigregna (1min Long)	Amharic ((3-5)sec Long)	Guragegna ((3-5)sec Long)	Tigregna ((3-5)sec Long)
Accuracy (%)	100.00	100.00	100.00	97.14	62.86	74.29

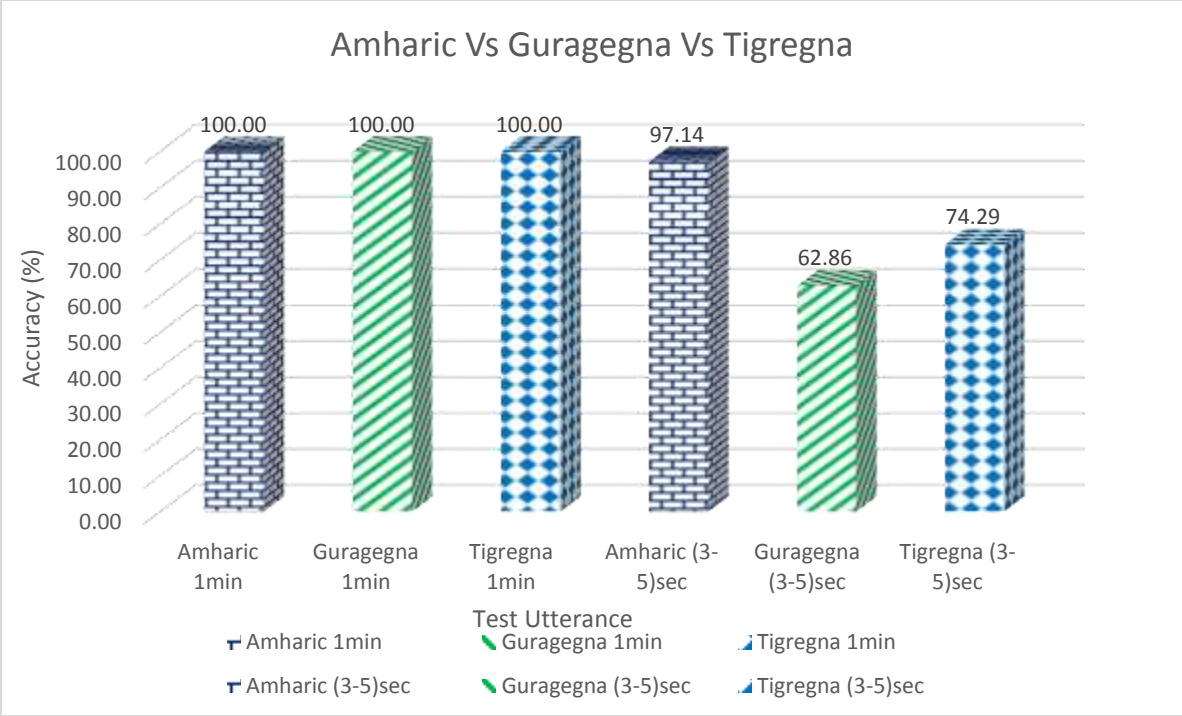

Figure 5: Test result for utterance dependent and independent tests for three languages

For three language classification, the minimum has decreased to 62.86% for utterance independent LID system while the utterance dependent one is 100% accurate for both languages. All possible triple languages classifications have been also tested.

Table 4: Utterance dependent LID System Accuracy taking four languages at a time

	Utterance dependent			
Test Utterance	Amharic (1min Long)	Guragegna (1min Long)	Oromiffa (1min Long)	Tigregna (1min Long)
Accuracy (%)	91.43	97.14	97.14	85.71

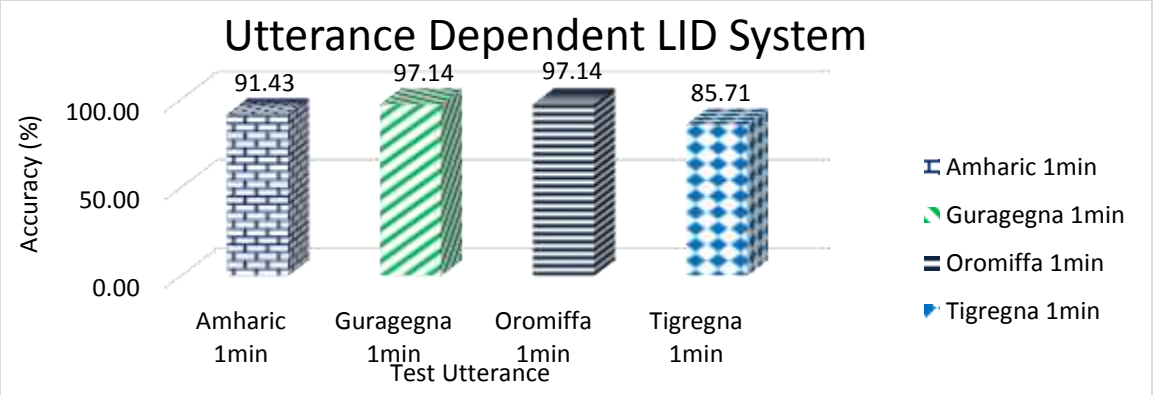

Figure 6: Utterance dependent LID system accuracy taking four language classifications

The above Table 4 and Figure 8 shows the LID system accuracy for the four language task of an utterance dependent system. It is clearly shown that the accuracy has decreased from average accuracy of 98.10% observed for two and 96.43% observed for three language classifications to an average of 92.85%.

Table 5: Utterance independent LID system accuracy taking four languages at a time

	Utterance independent			
Test Utterance	Amharic ((3-5)sec long)	Guragegna ((3-5)sec long)	Oromiffa ((3-5)sec long)	Tigregna ((3-5)sec long)
Accuracy (%)	91.43	62.86	60.00	65.71

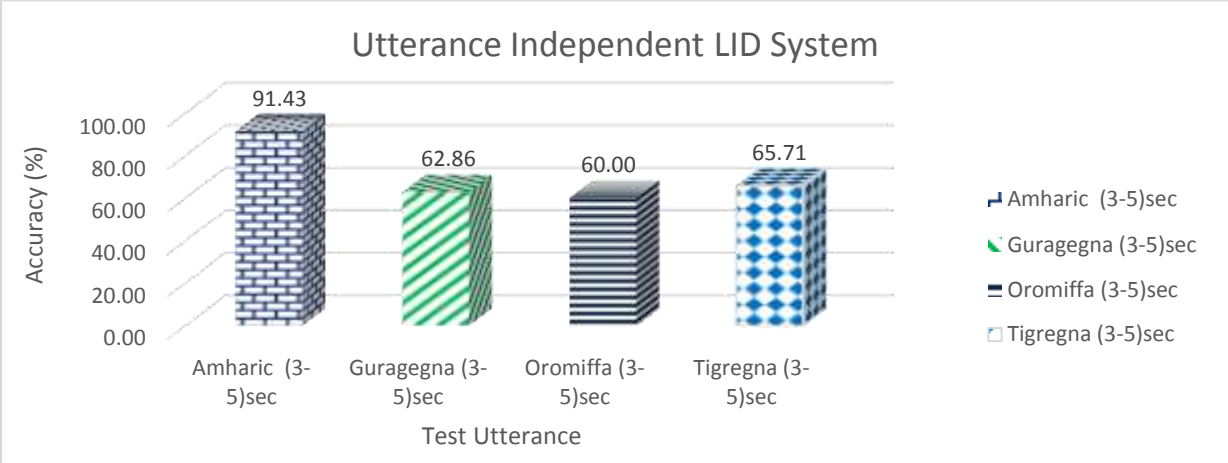

Figure 7: Utterance independent LID system accuracy taking four languages classifications

From Table 5 and Figure 9, it can be seen that when the number of languages increases, the utterance independent classification accuracy also decreases from 85.24% observed for two language classifications to 76.47% observed for three language classifications and then finally to 70% for the four languages classification LID system.

Table 6: Summary of the performance of the LID system for increasing number of Languages

Test Languages	Accuracy (%)	
	Utterance Dependent	Utterance Independent
LID by taking Two languages task	98.10	85.24
LID by taking Three languages task	96.43	76.47
LID by taking Four languages task	92.85	70.00

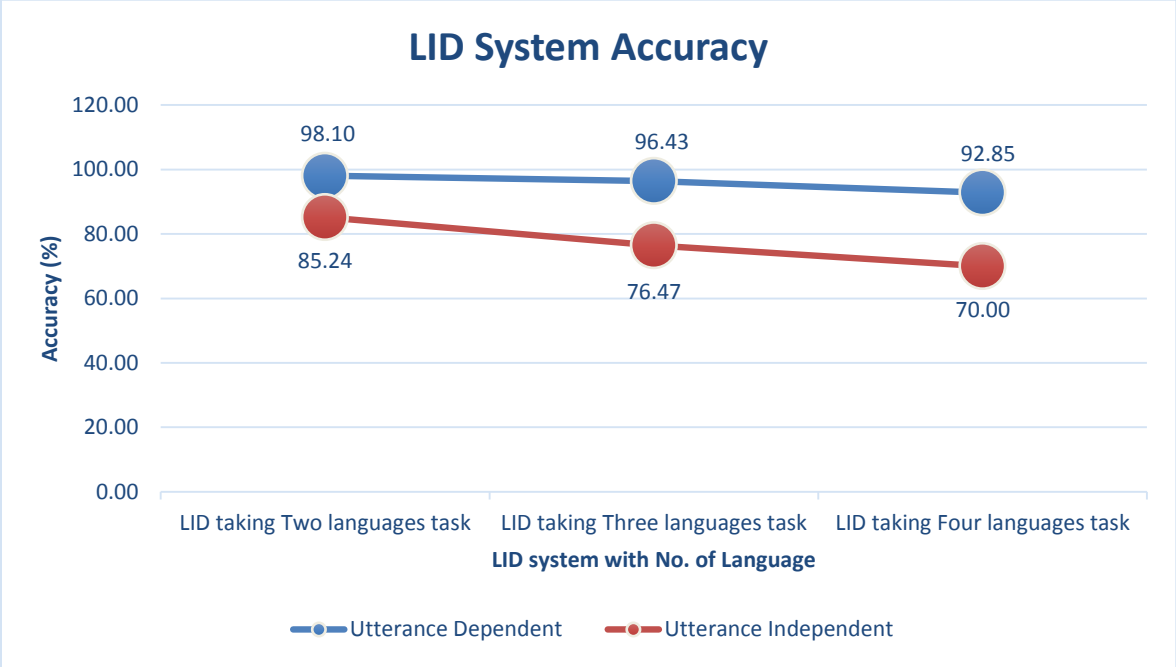

Figure 8: Summary of the performance of the LID system as the number of languages increases

But this gap due to number of languages can be reduced by collecting more utterance dataset samples for more phrases and words and re-training the model. The training of the LID model with rich spectral diversity dataset not only closes the gap between the accuracy for different number of languages but also improves the accuracy of the system in all respects. The overall comparison of the test results made for two, three and four languages classification is given in Table 6 and Figure 10 above.

The speaker independent system is tested by creating biometrical disjoint in the training and testing dataset (i.e. out of seven speakers of each language, four speakers utterances is used to train the system and the other 3 speakers utterances is used to test the accuracy of the system). The speaker independence test has been carried out for utterance dependent LID for the four languages classification. The result of the test is shown below in Figure 11.

The average accuracy of the test for all four languages is 91.67% a little bit lower than the result that have been gotten (about 92.85%) when there is nobiometrical disjoint sets between the training and testing datasets.

This is because the GMM classifier also learns the biometric identity of the speakers and takes that learning into classification. But the biometrical identity influence is almost very

small which only created a 1.15% in this small dataset. Such gaps can be made nil if the spectral diversity of the dataset is enriched by collecting more audio samples.

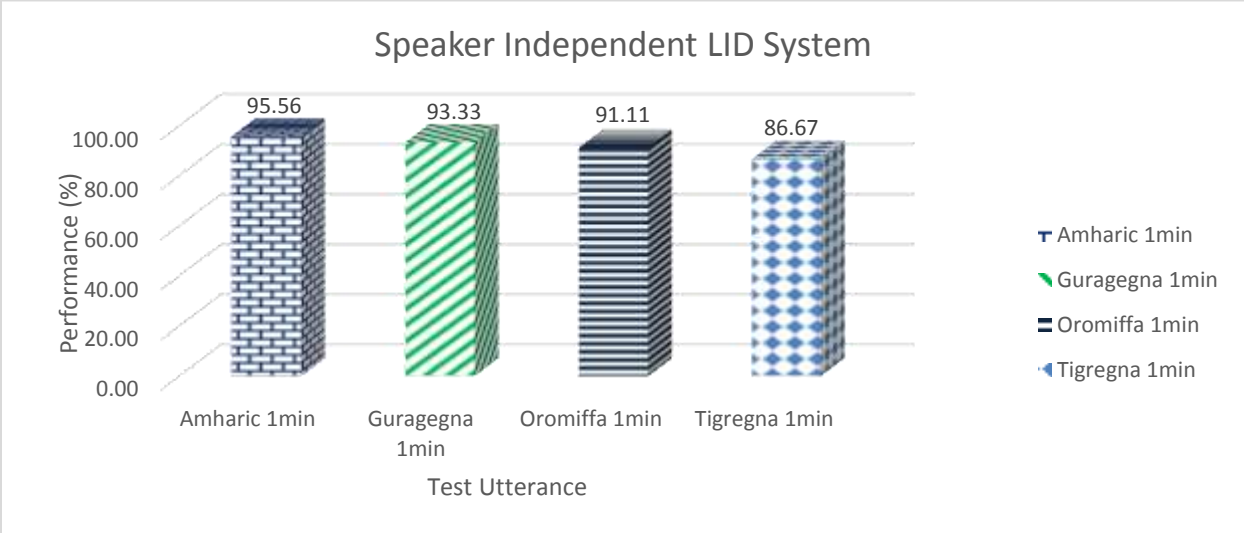

Figure 9: Speaker independent LID system for utterance dependent scenario.

CONCLUSIONS

It has been developed and tested an LID system for the four Ethiopian languages (Amharic, Oromiffa, Guragegna and Tigregna). To develop the LID system, train it and test the system, dataset was prepared by recoding 7 speakers for each language taking 15 samples of a single sentence and one paragrapgh from each volunteer speaker.

After preparation of the database pre-processing is done using software called waveserfer. We have used MFCC as feature extraction and GMM as language classifier and we have trained the system. The training time was about 56 minutes in Intel core i3 laptop with 4GB memory.

The classification task after being trained was within a second for both single sentence and paragraph tests we made. To test the performance of the LID system, experimental scenarios are designed and carried out by taking two, three and four languages task at a time. The LID system accuracy by taking two language classifications for the utterance dependent LID system was excellent and it was 98.10% accurate on average. For the utterance independent system even though the performance was decreasing compared to the utterance dependent system, but it shows a good performance 85.24% accurate on average. The result for taking three languages classifications for the utterance dependent LID system was also 96.43% accurate on average, whereas the LID system accuracy for the utterance independent system was 76.47% accurate on average.

The next experiment was done by taking four languages at a time for both the utterance dependent and independent system. The utterance dependent system performs with 92.85% accuracy and on the other side the utterance independent system shows a performance of 70.00% accuracy. The last experiment was done by taking four languages classification at a time for speaker independent LID system in utterance dependent setting and the system performance was 91.67% accuracy on average.

Future Works

Even though it has been seen that our development of LID system has very good results for utterance dependent scenarios, there are still some issues which are not addressed. These limitations of the current research project can be interesting future research directions. Some of these limitations are: As the GMM LID system provides an efficient means to identify spoken languages automatically, it is worth the effort to develop techniques to further improve the accuracy of the system since most LID applications require faster and accurate LID systems by increasing the spectral diversity of the dataset with more audio samples. The system can be tested with a higher hardware resource with higher GMM mixture order to improve the accuracy.

The LID system can also be tested using other classification algorithms and its performance can be compared with this research work.

REFERENCES

[1]. Nagesh A. , "Automatic Text Independent Language Identification," International Journal of Emerging Technology & Research, April 2013.

[2]. Nkadimeng Calvin, "Language Identification Using Gaussian Mixture Models," Stellenbosch: University of Stellenbosch, March 2010.

[3]. Manas A. Pathak and Bhiksha Raj, "Privacy-Preserving Speaker Verification and Identification Using Gaussian Mixture Models," IEEE Transactions on Audio, Speech, and Language Processing, vol. 21(2), February 2013.

[4]. Prasad Bhanu and Mahadeva S.R. Prasanna (Eds.), Speech, Audio, Image and Biomedical Signal Processing using Neural Networks, vol. 83, 2008.

[5]. Pinki Roy, Pradip K. Das, "Language Identification of Indian Languages Based on Gaussian," International Journal of Wisdom based Computing, vol. 1(3), pp. 54-59, December 2011.

[6]. Pedro A. Torres-Carrasquillo , Douglas A. Reynolds and J.R. Deller, "Language Identification using Gaussian Mixture Model Tokenization," IEEE Conference on ICASSP, May 2002.

[7]. David Martinez, Lukas Burget, Luciana Ferrer and Nicolas Scheffer, "Ivector-Based Prosodic System for Language Identification," IEEE Conference on ICASSP, 2012.

[8]. Lamel L.F. and Gauvain J.L., "Language Identification Using Phone-based Acoustic Likelihoods," ICASSP-94, 1994.

[9]. Koena R. Mabokela, Madimetja J. D. Manamela and Nalson Gasela, "Automatic Language Identification Using Word Segments on Mixed-Language Speech," Telkom Centre of Excellence for Speech Technology, 2011.

[10]. Heuser Julian, "Speech Recognition Wiki," 28- 12- 2014. [Online]. Available: http://recognize-speech.com/preprocessing. [Accessed 15- 04- 2015].

[11]. Bourlard H, Hermansky H., N. Morgan, "Towards increasing speech recognition error rates," Speech Communications, vol. 18, pp. 205-231, 1995.

[12]. Liu Leo, "Acoustic Models for Speech Recognition Using Deep Neural," Massachusetts Institute of Technology, 2015.

[13]. "https://github.com/lucyd/independent_study/tree/master/GMM,"[Online]. [Accessed 26- 11- 2014].

[14]. Stromhaug Tommy, "Discriminating Music,Speech and other," Norwegian University of Science and Technology, August 2008.

[15]. Pratik K. Kurzekar, Ratnadeep R. Deshmukh, Vishal B. Waghmare, Pukhraj P. Shrishrimal, "A Comparative Study of Feature Extraction," International Journal of Innovative Research in Science, vol. 3, no. 12, 2014.

[16]. Eslam Mansour mohammed and Mohammed Sharaf Sayed,, "LPC and MFCC Performance Evaluation with Artificial Neural Network for Spoken Language Identification," International Journal of Signal Processing, Image Processing and Pattern Recognition, vol. 6, no. 3, 20

[17]. Poonam Sharma and Anjali Garg, "Feature Extraction and Recognition 2013.

of Hindi Spoken Words using Neural Networks," International Journal of Computer Applications, vol. 142, no. 7, 2016.

[18]. Poonam Sharma and Anjali Garg, "Feature Extraction and Recognition of Hindi Spoken Words using Neural Networks," International Journal of Computer Applications, vol. 142, no. 7, 2016.

[19]. Varsha Singh, Vinay Kumar Jain and Dr. Neeta Tripathi, "A Comparative Study on Feature Extraction Techniques for Language," International Journal of Engineering Research and General Science, vol. 2, no. 3, 2014.