Automatic Language Identification for Languages of Pakistan

Madiha Itrat, *Syed Abbas Ali, Raheela Asif, Kamran Khanzada, Mukesh Kumar Rathi

Department of Computer & Information Systems Engineering, N.E.D. University, Pakistan.

Summary
Language identification and research in its related areas is gaining more and more importance and becoming the focus of research these days. People from different backgrounds talk in different languages which creates a language barrier for communication among individuals but this problem can be resolved using emerging and latest techniques of speech technology. This paper presents an automatic language identification system that differentiates between two different spoken utterances in Urdu and Sindhi which are national and one of the provincial languages of Pakistan respectively. The proposed approach in this paper is based on audio feature extraction, vector quantization for phoneme codebook generation and multi class like support vector machine for classification and identification of the respective languages. The experimental result is encouraging and indicates that the proposed approach is effective by identifying the spoken utterances for two languages of Pakistan in real environment.

Key words:
Language Identification, MFCCs (Mel Frequency Cepstral Coefficients), Vector Quantization (VQ), Support Vector Machine (SVM), Languages of Pakistan

I. Introduction

Language Identification is the process of recognizing the spoken language in a brief speech signal of digitally recorded audio uttered by an anonymous speaker. It is the first stepping stone for the language based technology, especially where people are hoping to implement automatic language translators in the near future. Many speech technology systems today, are restricted to working with only a single language which makes them impractical and non-useful in today's world of huge multicultural exchange where the people speak different languages which creates a barrier in communication. Hence, language identification is a very desirable feature in many practical applications and the automation of this task is highly useful. It is generally assumed that language identification systems perform the best when they work with phono tactic rules i.e.: the modelling of each phoneme and its location in the signal. However, recent researches show that good results may also be derived when working with acoustic features. The front-end conversion of raw speech to parametric form can be done by several methods such as Bank-of-Filters Processor [1], Linear Predictive Coding model analysis or MFCCs [2]. The former two though proved to be effective but proved to be too complex to use in practical systems. A lot of machine learning algorithms also were suggested to work well with language identification systems such as Neural Networks and Gaussian Mixture Model (GMM)[3] , Hidden Markov Model[4], Deep Learning [5] etc. A number of options were also available for the selection of classifier such as probability matrix, bag of sounds, linear discriminant analysis etc. that could also have been used in combination with vector quantization as were used in various previous researches. Due to a wide variety of techniques it is problematic to compare each other directly because each processes their own merits and demerits so the decision should be made according to the requirements of the work. The automatic language identification System aims to contribute to the dream of creating an artificial system that would possess the ability to store the models of all known languages in the world. This research proposes a real time identification system in national and one of the provincial languages of Pakistan. The proposed system could easily be applied for the automatic redirection of calls in call centers or customer services where provision of quick and efficient service is required, in combination with an automatic real-time speech translator for real time speech-to-text conversion, on airports, for the analysis of spoken natural language, study of language dialects and linguistic research where the spoken language identification is the shortcoming.

Rest of the paper is organized as follows, The idea of the language identification systems including the techniques used for its design and implementation are supported by work of the various researchers in the literature review are presented in section II. Section III, describes the corpus collection and system's modules including the description of sub-techniques used in developing the proposed system. Experimental results are presented in section IV. Finally the conclusion is drawn in section V.

II. Related Work

In the beginning of 1970, research was started in the field of language identification system. But since, the pace of research in this area was very slow for almost two decades. Afterwards, public domain multi -Lingual Corporation of speech was come. Then, many people started to show their interests in this area of research. As a result of this interest, lots of progress had been made [6]. Dodington and

Manuscript received February 5, 2017
Manuscript revised February 20, 2017

Leonard [7] have calculated frequency of appearances of certain related sounds in different languages. They founded the result as the average accuracy of language identification system was achieved 64% when there were five languages considered and 80% in case of seven languages. House and Neuberg in [8] studied on phonetic transcribed data and extracted the useful information of languages instead of acoustic features using Hidden Markov Model (HMM) for training of system in eight languages with 80% accuracy. Foil used two different approaches to study language identification in noisy background. First approach processed pitch and energy contours and then language features are captured for language identification. In second approach, computation of formant vectors (K-means clustering algorithm) is done for each language to distinguish between same phonemes of different languages [9]. Similarly, Goodman et al. had modified the training algorithm proposed by Foil He splitted the training two vectors namely "clean" and "noisy" .K-means clustering algorithm was used to check whether pitch information is useful in noisy background or not. Syllabic rate was used for distinguishing features of one language from other languages [10]. In early nineties, Zissman and Singer have performed comparative analysis on four approaches: (i) GMM based classification, (ii) Phoneme identification by using Phone Recognition Language Modelling (iii) Parallel PRLM and (iv) PPR [11]. In 1997, Hazen and Zue have developed the idea to identify language by doing comparative analysis (calculating probabilities) on phonotactic, acoustic-phonetic and prosodic information [12]. In the era of early twenties, Gleason and Zissman have explored two methods to improve the accuracy of PPRLM approach. They used a modelling technique i.e. (Composite background) which gave opportunity to identify targeted language in an environment where training data is limited or even unavailable [13]. Rouas had used an approach that was based on modelling rhythm of sound (phonetics and phonotactics). He proposed an algorithm to extract the rhythm for language identification [14]. In continuation his work in 2007 Rouas again used another model (n-gram models) that was based on prosodic variations that method was used to model two sequences of labels based on language dependent (short-term and long-term labels) [15]. Sangwan had done an analysis on identification system based on production of speech knowledge by extracting the important language distinguishable feature from speech in five closely related languages and founded 65% result [16]. In 2012, Martnez used an approach of i-vector whose basis was on prosodic information (rhythm, stress, and intonation) to made decision about the language by using classifier that was based on i-Vectors [17]. There are other techniques as well for language classification including but not limited to Hidden Markov Model (HMM) and neural networks [18]. Basically, HMMs work in a similar manner to Finite State Machines (FSMs). The transitions from one state to another are based upon the training set's probabilistic data, whereas neural networks learn and acquire knowledge in a manner not dissimilar to humans. Sugiyama has done classification of acoustic features such (like LPC, autocorrelation and delta cepstral coefficients) with the help of Vector Quantization and found the differences between using different code books for different languages and one common code book for all languages [19]. Itahashi and Liang performed language identification whose basis was on fundamental frequency and energy contours. They modelled the LID system by using a linear function (piece-wise) [20, 21]. Li gave an idea to extract features on the basis of syllables (vowels) by computing the feature vectors having spectral information in [22]. Chung-Hsien segmented speech utterance (input) into language-dependent segments using delta Bayesian information criterion (delta-BIC) and then a VQ-based bi-gram model characterized the acoustic features of two consecutive code words in a language [23]. Nagarajan made use of syllable like units and parallel syllable like unit to generate code words for recognition [24]. Mary had explored models to extract spectral features with neural network for language identification. Test speech samples were collected with varying duration [25-27]. In the era of 2012, Botha and Barnard worked on n-gram statistics for getting features of language identification by approaching different classifiers like support vector machines (SVMs), naive Bayesian and difference-in-frequency classifiers [28]. In the same era, Barroso had used hybrid approaches for language identification that were based on the selection of system elements by several classifiers (Support Vector Machines (SVMs), Multilayer Perceptron classifiers and Discriminant analysis for improves the system performance [29]. Siniscalchi worked on a novel acoustic characterization approach for language identification process. Fundamental units for universal set had been founded, which can be used for all the languages that have to be identified by a specific system [30]. In 2013, Bhaskar performed a research study on gender independent, gender dependent and hierarchical grouping approaches on different languages. Features of vocal tract were identified for capturing specific information of a particular language [31]. Polasi worked on the improvement of the ability of the machines to distinguish between languages. In this paper author made use of the MFCCs (Mel-Frequency Cepstral Coefficients) and GMM (Gaussian Mixture Models) to perform the language identification studies on the databases of 27 different Indian languages, which addressed the capability of Automatic Language Identification Systems comparing them in both clean and noisy environments. [32]. Malmasi along with his fellows performed a shared task of language dialect identification in speech transcripts on Arabic language and found highorder character n-grams as the most successful feature while classification approaches such as supervised learning methods including SVM (Support Vector Machine), logistic regression in [33]. In this paper, automatic language identification system is proposed for national and one of the provincial languages of Pakistan. The proposed language identification module is comprises on MFCC technique for the parametric representation of spoken audio utterances, vector quantization is used for unforeseen audio samples and SVM classifier for classifying parametric data samples.

III. Corpus Collection and system Modules

The recording for training set audio was done with the built-in sound recorder of Microsoft Windows 8 in standard environmental conditions, having SNR>=45dB to record complete speech utterances using MATLAB environment and a microphone connected to the desktop PC. These audio recordings were read and saved in an audio file using MATLAB platform functions. Mono recording format is selected, with 16 bit PCM and a default sampling rate of 8 KHz due to microphone properties and sensitivity of 2.2W and 54dB±2dB respectively, a pulp stereo-type of 3.5mm and the length of the cable is 1.8m. To develop the training set corpus for language identification system, input is collected from male and female speakers aged between 20 to 25 years. Table 3.1 is comprised on the stated some sentences, phrases and words that have been spoken during the recording of the training set audio along with the times they have been spoken (x times it is spoken) for each of the two languages. Note that the sentence and phrase selection for each language was based on the fact to incorporate maximum common sounds used within that language and the frequency of occurrence of a sentence/phrase was based on the fact that how much chances are there that it will be spoken during the testing phase.

Table 3.1: Sentences/Phrases/Words spoken in Urdu & Sindhi Languages

Sentences/Phrases/Words	Urdu	Sindhi
What is your name?	X2	X2
My name is ----------	X2	X2
How are you?	X3	X2
Where do you study?	X2	X1
We are working very hard.	X1	X1
She is a very good girl.	X1	X2
Today is our exam.	X1	X1
Do this right now at the moment.	X2	X2
How much work have you completed?	X2	X1
How long will it take?	X1	X1
Tell me the reason.	X1	X1
He did not tell me.	X1	X2
Make it fast.	X1	X1
Something is missing here.	X1	X1
Kindly help me.	X2	X2
What, where, why, how , which	X3	X3
No, yes, ok, fine	X3	X3
Is, was, not, have, give, take	X3	X3
Writing, reading, listening	X2	X2
My, mine, day	X2	X2
Come, go, now, never, do	X3	X3
Nice, good, bad, this, that	**X3**	**X2**

To begin with proposed language identification system, first of all we had to train the system through samples of spoken utterances which were pre-recorded for both languages. The microphone is connected to system which is used for audio input (and also for pre-recording of training audios), which will be processed, synthesized and signal processing algorithms will be applied on them, to generate reasonably accurate results, hence indicating that which of the two languages was spoken in the real-time input (Urdu or Sindhi). The proposed system is capable enough to deal with two different languages and provides a very easy way to extend this work for the identification among a desired number of languages. MATLAB platform is used for training and such algorithms are implemented which can recognize the spoken speech utterance. The very first step in any speech recognition system is extracting features and identifying the audio components of the signal that are very good in identifying the linguistic content and omitting all the redundant things which causes problems like background noise, music etc. The main point to understand regarding speech is that the sounds are filtered by the shape of the vocal tract (teeth tongue etc.) that are generated by a human which helps us in determining what sound came out. If we are able to determine the shape this will further helps us to obtain the representation of phoneme being produced. Speech recognition In the language identification systems have two phases; the training phase i.e. to train the System from audio samples of multiple languages and the testing phase i.e. to test the System on real time input to determine the language being spoken. Table 3.2 states the differences between these two phases in this work.

Table 3.2: Difference between Training & Testing Sets

Training Set	Testing Set
□ Recording training Audio Sample.	□ Taking input through Microphone.
□ MFCC calculation (Feature Extraction)	□ MFCC calculation (Feature Extraction)
□ Generation of *Vector Quantized Codebook* for training.	□ Generation of *Vector Quantized Codebook for testing.*
□ Language Model formation by *Support Vector Machine* Technique.	□ Test Sample Model formation by *Support Vector Machine* Technique.
□ Train the System.	□ Decision Making and display on Hardware.

The work flow of the proposed systems is provided in Fig.3.1 to show the functions involved in developing automatic language identification system for two languages of Pakistan:

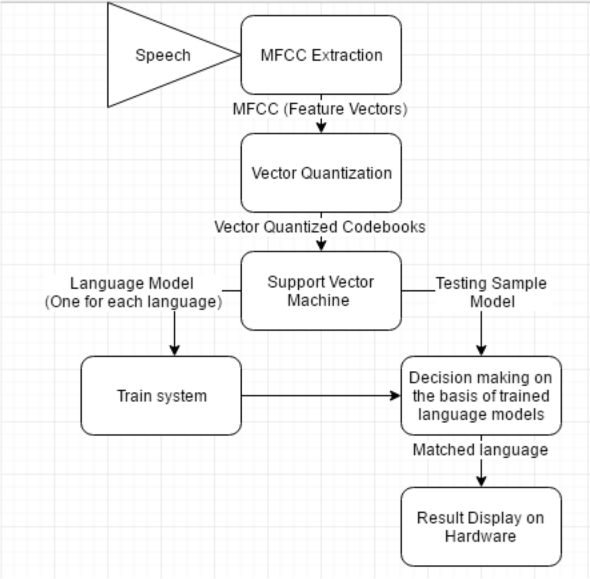
Speech
MFCC Extraction
MFCC (Feature Vectors)
Vector Quantization
Vector Quantized Codebooks
Support Vector Machine
Language Model (One for each language)
Testing Sample Model
Train system
Decision making on the basis of trained language models
Matched language
Result Display on Hardware

Fig 3.1. System Functionality Flow Chart

The proposed automatic language identification systems module is divided into three different phases:

1. Feature Extraction (parameter representation of spoken utterances)
2. Vector Quantization (analysis for unforeseen data samples)
3. Classification

A) Mel Frequency Cepstral Coefficient (MFCC) Extraction:

To take an input speech sample and then converting it to a set of feature vectors is a difficult task but not very difficult to understand. First, we have to convert input speech in to its raw form. Then in order to make the signal spectrally flatten and to make it less susceptible to finite-precision effects, we used pre-emphasis. First-order FIR digital filter is used to handle pre-emphasis. Afterwards, the pre-emphasized signal is divided up into frames, where each frame is just a segment of the speech data over a very small amount of time. In order to minimize signal discontinuities at the beginning and end of each frame so the frames are windowed using a Hamming window. After windowing, the frames get Fourier transformed to get their Fourier spectrum. Then the resulting spectrum is processed through the Mel-scaling filter bank. Once the Mel frequency spectra are generated, they are put through cepstral analysis (taking the logarithm of the spectra) to finally produce the MFCCs. We used MATLAB platform for extracting MFCCs and the entire coding of the MFCC and its sub-functions. Next we take the discrete cosine transform which results in Mel spectrum. And at last the final Delta Energy Spectrum is generated. The entire process is concisely illustrated in Fig 3.2.

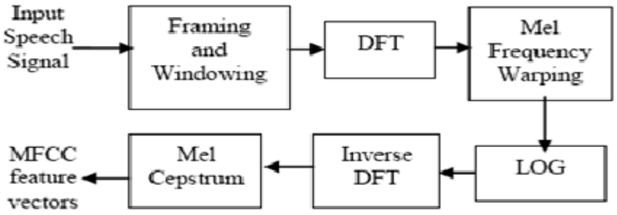

Fig 3.2: MFCC Feature Extraction [34]

2) Vector Quantization:

Vector quantization (VQ) is used for codebook generation comprising of code vectors. It is a lossy data compression method which is based on the principle of block coding. The LBG VQ algorithm is an iterative algorithm used in this technique. An initial codebook is required which is obtained by splitting method. An initial code vector is set as the average of the entire training sequence and then split into two. Initially the algorithm runs with these two vectors which are further splitted into four and the process is repeated until the desired number of code vectors is obtained. The codebook created at the end of this algorithm is the codebook of phonemes. Phonemes are distinct units of sound within a given language that clearly identify one word of the language from its another word. For Example: p, b, d, and t in the English words pad, pat, bad, and bat. These distinct sounds are here used as a basic classification and distinction criteria among the different languages used in the system. So, the automatic language identification system distinguishes between the sounds of the different languages based on the language's distinct sound units (phonemes). Fig 3.3 shows the flowchart of the LBG algorithm:

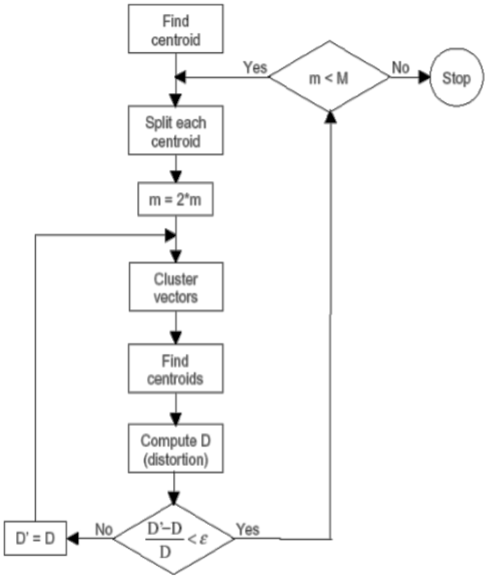

Fig 3.3. LBG Algorithm Flowchart

3) Support Vector Machine:

Support vector machine is used when there are exactly two classes (sets of data) to distinguish among, but since here more than two different languages can also be used to train the system so we used support vector machine classifier in an iterative way once for each language (a multi-class like support vector machine to make the idea feasible for implementation in the desired number of languages even more than two). The model for each language is formed in this step by which the system is trained. Support Vector Machine (SVM) is a discriminative type of classifier. It works on the concept of supervised learning. When an input training data is given, it provides an optimal hyperplane or a separating "classifier" that clearly classifies the training data. Unlike many other classifiers SVM holds an advantage that along with linear default classification it can also provide non –linear classification through its very popular kernel function. A kernel function is basically an equation that can pull data points apart into a 3-Dimensional space using a classifier called the hyper plane. It is a special kind of classifier that can take non-linear form providing a more generic and sophisticated plus better and in depth classification. Here in this project, the most appropriate results were derived by using the Quadratic Non-linear Kernel function in the model selection phase. The geometrical margin is proved to be equal to the inverse of the norm of the gradient of the decision function. After scaling the vector quantized codebooks (for the sake of simplicity in SVM model formation), the kernel type is selected. This is followed by a cross-validation of the training model at the end resulting in the creation of a final concrete model of the input sample in training phase as well as in testing phase. Fig 3.4 shows how a training model is generated using SVM classifier.

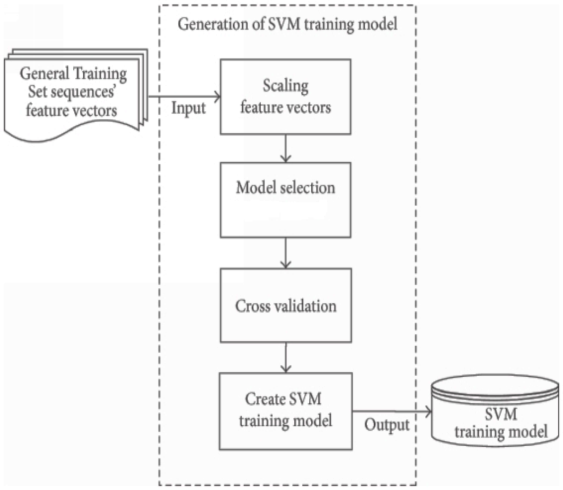
Generation of SVM training model
General Training Set sequences' feature vectors
Input
Scaling feature vectors
Model selection
Cross validation
Create SVM training model
Output
SVM training model

Fig 3.4. Training Model generation using SVM

Integration is the final and most important part of proposed automatic language identification module. After taking input through microphone, it will be saved and pre-processing will be applied on the audio which includes MFCCs extraction. Afterwards codebook will be generated from the MFCCs obtained through famous technique vector quantization. This codebook will be given as an input to support vector machine for classification and decision making which will then show the results in the final output to indicate the language being spoken, using the pre-built language models.

IV. Experimental Results and Discussion

This section presents demonstrative experiments to design different modules of proposed automatic identification system under following conditions: 1) noiseless and noisy environment with good and bad training samples 2) considering accent of speaker in the training Sample for both languages separately 3) comparing the different kernel functions of SVM , and 4) changing length of training sample in term of time. The experimental framework is divided into three phases, initially, Feature extraction technique MFCCs were used to convert raw audio samples into parametric form. In the second phase, vector quantization technique was used for the classification and clustering of phonemes into fixed dimensional phoneme (vector) codebooks which consider to be work effectively well in real-time systems and in case of unforeseen testing samples. Lastly, Support Vector Machine classifier was used to classify the data present in the codebooks and create relative language models for each language as it provides multiple kernel functions and regularization parameter making it versatile and consistently efficient when used with different types of training audio samples.The classification of the real-time testing input will be done on the basis of these pre-built language models. The actual training audio is of about 2 minutes to 2 minutes and 35 seconds for each language. For simplicity, Fig 4.1 and Fig 4.2 show the audio samples plots of only about the first 15 seconds of Urdu and Sindhi language respectively. Fig 4.1 and Fig 4.2 shows the plots of raw audio recordings (training sets) of Urdu and Sindhi languages before processing, where x-axis represents the time of the audio recording or in other words the audio length, while the y-axis here specifies the range of values of the audio samples. So, it is a time versus samples plot for both languages.

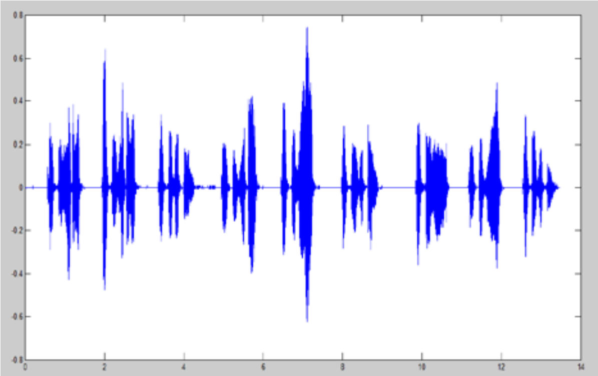

Fig 4.1. Plot of raw Urdu audio recording

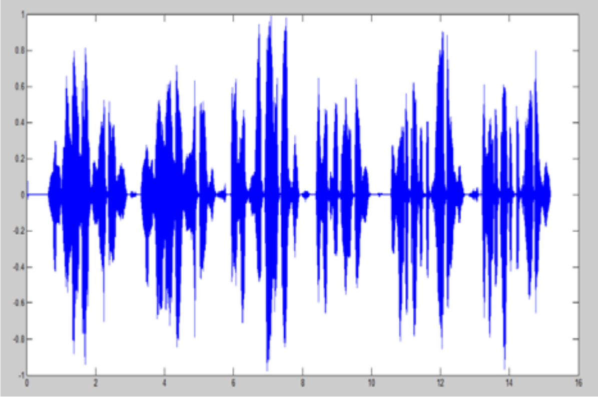

Fig 4.2. Plot of raw Sindhi audio recording

Fig 4.3 and Fig 4.4, present the plots of the pre-emphasized training sets of Urdu and Sindhi languages respectively. In other words, MFCCs have been extracted after filtering, framing, windowing, Fourier transformation, and Mel-scale warping and cosine transformation. Number of MFCC coefficients per frame may differ in different applications. However, they have certain fixed values, for example 4,12,13,20 etc., and only a value out of these fixed values must be chosen to get a reliable spectral estimate otherwise, the signal content can be lost or even some undefined values can be processed along with the useful values. To obtain a medium amount of features per frame for simplicity we have considered 20 MFCC coefficients per frame and avoid small values like 4 and very large values like 41. Hence the dimensions of resultant MFCC matrix becomes 20 times x, where x depends on the length of the audio sample. In Fig 4.3 and Fig 4.4, x-axis represents the number of MFCC coefficients per frame, while the y-axis specifies the range of values within the MFCC matrix or simply the different MFCCs of the samples.

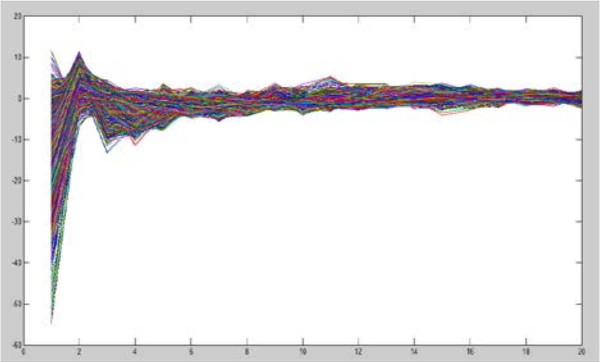

Fig 4.3. Pre-Emphasized Training Set of Urdu Language

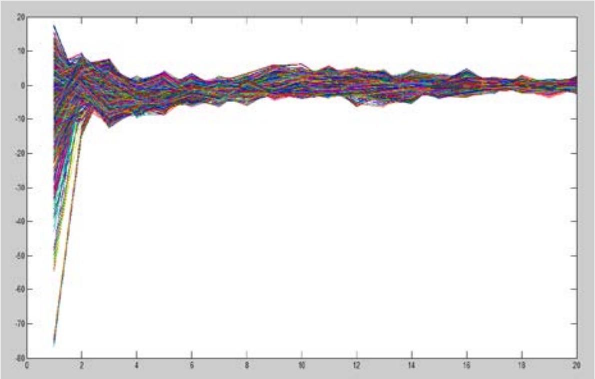

Fig 4.4. Pre-Emphasized Training Set of Sindhi Language

Similarly, Fig 4.5 and Fig 4.6, display the plots of the vector quantized codebooks of phonemes after applying the LBG vector quantization algorithm on the obtained MFCC feature vectors for Urdu and Sindhi languages respectively. Among the MFCC values, some NaN (no value) and Infinity values were obtained. We made assumptions and designed an algorithm to map the NaN values to 0 (as it represents no value) and Infinity values to 100 (because it represents a very large value) because the LBG algorithm does not function properly in the presence of NaN and Infinity values. The dimension of the codebooks is 20 times 128. 20 because of 20 MFCC coefficients per frame and 128 is a value that specifies that into how many phonemes (set of features) require for codebook. It can be selected however it must be in a power of 2 for the efficient working of the LBG algorithm. Again, to work with a medium amount of phonemes the value of 128 phonemes is selected here. In Fig 4.5 and Fig 4.6, x-axis represents the number of MFCC coefficients per frame, while the y-axis specifies the range of values within the codebook which are now normalized and clustered into 128 clusters.

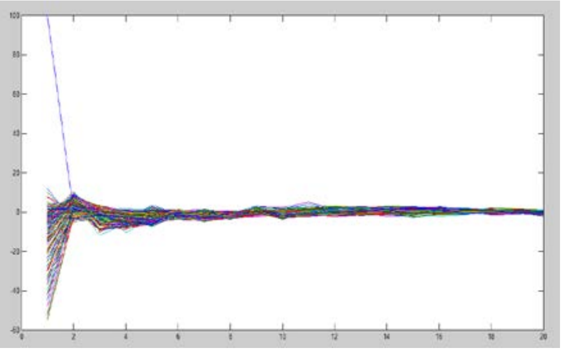

Fig 4.5. Plot of Vector Quantized Codebook for Urdu Language

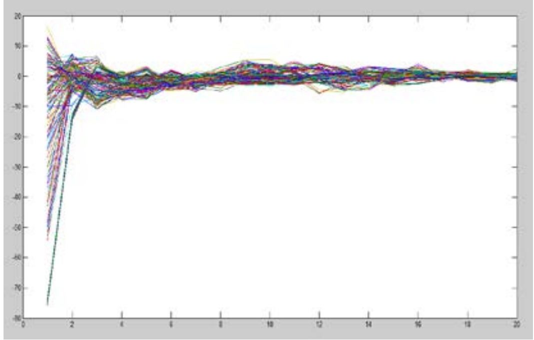

Fig 4.6. Plot of Vector Quantized Codebook for Sindhi Language

In the end, a classifier or language model is built for each language using these codeboks and the support vector machine technique. Consider building a classifier for Urdu language. This technique will consider all the sample values of urdu language as positive samples (+1) and all the samples belonging to every other language being observed in the system will be considered as negative samples (-1) for Urdu language model. We have used already available functions svmtrain and svmclassify available in the bio-informatics toolbox of the MATLAB 2015 version. Hence models will be built to train the system and decision making will be performed for the testing inputs. About 50 testing samples altogether were used to test the system. Out of these 25 testing samples were of Sindhi language and 25 testing samples were of Urdu language. 10 testing samples from each language were picked from the training sample randomly (Any 10 phrases or sentences used during the system training were spoken again for testing purpose). The rest of the 15 testing samples for each language were unforeseen samples for the system (New phrases or sentences or words were spoken that were not spoken during the training phase). Table 4.1 shows some observed results and their accuracies in different Environmental conditions and different qualities of training audio corpus. These observations are recorded in ideal conditions i.e.: keeping all the other factors affecting the system's efficiency ideal.

Table 4.1: Effect of Noise and Training Set Quality on results

Environment		**Testing Language**	**Correct Results %**
Noiseless + good Training Set	□	Urdu	90 %
Noiseless + good Training Set	□	Sindhi	80%
Noisy + good Training Set	□	Urdu	75%
Noisy + good Training Set	□	Sindhi	55%
Noiseless + Bad Training Set e	□	Urdu	50%
Noiseless + Bad Training Set	□	Sindhi	40%
Noisy + Bad Training Set	□	Urdu	25%
Noisy + Bad Training Set	□	Sindhi	10%

Note that a good training set here refers to a training audio which is closest to the testing audio expected as an input to the system in the future. For example, consider that our proposed system is put into operation in a call center where the majority of the calling customers are expected to have an English accent; they speak fast and are expected to speak a few common phrases like hello, hi etc. In such conditions a good training set will contain a recording of a person having English accent, speaking relatively fast and probably speaking the commonly expected phrases more than once for making the system able to identify specially those phrases very easily. In other words a good training set is the one that is closest to the testing input (a quite reverse engineering type phenomenon). Consider Table 4.1, here one thing is observable that in most of the cases the Urdu language correct identification rate is more than Sindhi language. This is primarily because the accent is also but minimally affecting the results. In this case the training audios were recorded of a person having Urdu as their mother tongue so the best results produced are of Urdu language. Similarly this factor may be adjusted by adjusting the training sets according to the requirements. It is also noted that the system's overall efficiency also increases when we use the training audios of each language having that language spoken by people of at least two to three different accents (preferably also having an accent of other languages considered in the system design). Also, that if each language is spoken by only one person in the training set having it as their mother tongue so the efficiency of the system detecting that language increases but it greatly decreases the efficiency of the system to detect the other languages. This observation is shown in the Table 4.2. These observations are recorded in ideal conditions i.e.: keeping all the other factors affecting the system's efficiency ideal.

Table 4.2: Effect of Speaker Accent on results

Accent of Speaker in the Training Set audio	Language	Correct Results %
Urdu only	Urdu	90%
	Sindhi	72%
Sindhi only	Urdu	75%
	Sindhi	89%
Half audio of Urdu speaker + Half audio of Sindhi speaker	Urdu	85%
	Sindhi	82%

While experimenting with several kernel functions of the support vector machine classifier, certain observations and different levels of result accuracies with each kernel function can be seen in Table 4.3. These observations are recorded in ideal conditions i.e.: keeping all the other factors affecting the system's efficiency ideal.

Table 4.3: Comparison between kernel functions of SVM w.r.t. results

Kernel Function	Language	Correct Results %
Linear	Urdu	60%
	Sindhi	10%
Quadratic	Urdu	90%
	Sindhi	84%
Polynomial	Urdu	70%
	Sindhi	60%
RBF	Urdu	55%
	Sindhi	55%
MLP	Urdu	45%
	Sindhi	55%

Table 4.3 clearly shows that performing experiments with different kernel functions the best results were obtained by using quadratic kernel function of the support vector machine classifier. However, this observation is expected to change with the changing nature of the training audio samples. For recorded audio training set, the quadratic kernel function proved to be comparatively the most accurate one. During the study, it was also observed that as we varied the length of the training audio sample, the output result accuracy also varied proportionally and hence the testing sample result accuracy was noted by varying training sample length which is recorded in Table 4.4. These observations are recorded in ideal conditions i.e.: keeping all the other factors affecting the system's efficiency ideal.

Table 4.4: Effect of varying Training Set Audio Length on results

Length of Training Set	Language	Correct Results %
> 20 mins	Urdu	5%
	Sindhi	5%
10 mins- 20 mins	Urdu	40%
	Sindhi	30%
5 mins- 10 mins	Urdu	65%
	Sindhi	55%
2 mins- 5mins	Urdu	90%
	Sindhi	84%
< 2 mins	Urdu	72%
	Sindhi	68%

Table 4.4 indicates that the best results were obtained when the training samples were of a length between 2 to 5 minutes. This can be understood as the testing sample given in our case was between 10 and 30 seconds. Most of the expected words and sounds during these 10-30 seconds can be incorporated (once or more than one time) in at least duration of 2-5 mins training set. This shows that a training set whose duration is closest to the testing set duration but has the maximum possible sounds of a language in it (it is robust) can be of about 2-5 minutes which is the comparatively most ideal training set audio length.

V. Conclusion

This paper presented an automatic language identification system in national and one of the provincial languages of Pakistan. The proposed methodology for developing identification system is tested on Urdu and Sindhi spoken utterances to provide reasonably accurate results on real-time input testing sample. The proposed identification modules were comprised on three phases including; feature extraction, vector quantization for analysis the unforeseen data samples and classification. The Experiments were performed on spoken sentences of recorded audios in Urdu and Sindhi languages using MATLAB Toolkit platform in addition with a microphone. Premilarly results of demonstrative experiments under different conditions and time frames evident that proposed automatic language identification system provide expected results with less noise interference (correct spoken language was identified in about every 82 out of 100 testing inputs).Authors are focusing two major aspects of these proposed systems 1) Including VAD (Voice Activity Detection) module and 2) considering other provincial languages of Pakistan to improve and enhance their design and identification capability of the system.

References

[1] J. Kovacevic and M. Vetterli. "Perfect reconstruction filter banks with rational sampling factors." IEEE Transactions on Signal Processing, vol. 41, no. 6, pp. 2047 – 2066, Jun. 1993.

[2] E.M. Mohammed et al. "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, pp. 55-66, 2013.

[3] T. Hazen. "Automatic Language Identification Using a Segment-based Approach." PhD thesis, MIT, USA, Aug. 1993.

[4] S. Nakagawa and Hashimoto. "A method of continuous speech segmentation using HMM," in Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing, 1988, pp. 960-962.

[5] G. Montavon. "Deep learning for spoken language identification,"in NIPS workshop on Deep Learning for Speech Recognition and Related Applications, 2009.

[6] LDC. (1996). Philadelphia. Available: www.ldc.upenn.edu/Catalog.LDC96S46–LDC96S62.

[7] R. Leonard and G. Dodington, "Automatic language identification," Air Force Rome Air Development Center, New York, Tech. Rep. RADC-TR- 74-200, Aug. 1974.

[8] A.S. House and E.P. Neuberg. "Toward Automatic Identification of Language of an Utterance. I. Preliminary Methodological Considerations." Journal of the Acoustical Society of America, vol. 62, no. 3, pp. 708-713, 1977.

[9] J.T. Foil. "Language identification using noisy speech," in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 1986, pp. 861–864.

[10] F. Goodman, A. Martin and R. Wohlford. "Improved automatic language identification in noisy speech," in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, May 1989, pp. 528–531.

[11]]M.A. Zissman and E. Singer. "Automatic language identification of telephone speech messages using phoneme recognition and N-gram modeling," in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, 1994, pp. I/305–I/308.

[12] T.J. Hazen and V.W. Zue. "Segment-based automatic language identification." Journal of the Acoustical Society of America, vol. 101, pp. 2323–2331, 1997.

[13] T. Gleason and M. Zissman. "Composite background models and score standardization for language identification systems," in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, 2001, pp. 529–532.

[14] J.L. Rouas, J. Farinas, F. Pellegrino and R. Andr-Obrecht. "Rhythmic unit extraction and modelling for automatic language identification." Speech Communication, vol. 47, pp. 436–456, 2005.

[15] J.L. Rouas. "Automatic prosodic variations modeling for language and dialect discrimination." IEEE Transactions on Audio, Speech and Language Processing, vol. 15, no. 6, pp. 1904–1911, 2007.

[16] A. Sangwan, M. Mehrabani and J. Hansen. "Automatic language analysis and identification based on speech production knowledge," in International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Mar. 2010, pp. 5006–5009.

[17] D. Martinez, L. Burget, L. Ferrer and N. Scheffer. "i-vector based prosodic system for language identification," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Mar. 2012, pp. 4861–4864.

[18] J. Braun and H. Levkowitz. "Automatic Language Identification with Recurrent Neural Networks," in Neural Networks Proceedings- IEEE World Congress on Computational Intelligence, 1998.

[19] M. Sugiyama. "Automatic language recognition using acoustic features," in Proceedings of IEEE international conference on acoustics, speech, and signal processing, May 1991, pp. 813–816.

[20] S. Itahashi, J. Zhou and K. Tanaka. "Spoken language discrimination using speech fundamental frequency," in Proceedings of international conference on spoken language processing (ICSLP), 1994, pp. 1899–1902.

[21] I. Shuichi and D. Liang. "Language identification based on speech fundamental frequency," in Proceedings of EUROSPEECH, 1995, pp. 1359–1362.

[22] K. Li. "Automatic language identification using syllabic features," in Proceedings of IEEE international conference on acoustics, speech, and signal processing, 1994, pp. 297–300.

[23] 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, vol. 14, no. 1, pp. 266–276, Jan. 2006.

[24] T. Nagarajan and H.A. Murthy. "Language identification using spectral vector distribution across languages," in Proceedings of international conference on natural language processing, 2002, pp. 327–335.

[25] L. Mary and B. Yegnanarayana. "Auto associative neural network models for language identification," in Proceedings of international conference on intelligent sensing and information processing (Chennai, India), 2004, pp. 317–320.

[26] L. Mary, K.S. Rao and B. Yegnanarayana. "Neural network classifiers for language identification using syntactic and prosodic features," in Proceedings of IEEE international conference on intelligent sensing and information processing (Chennai, India), Jan. 2005, pp. 404–408.

[27] L. Mary and B. Yegnanarayana. "Extraction and representation of prosodic features for language and speaker recognition." Speech Communication, vol. 50, pp. 782–796, 2008.

[28] G.R. Botha and E. Barnard. "Factors that affect the accuracy of text-based language identification." Computer Speech and Language, vol. 26, no. 5, pp. 307–320, 2012.

[29] N. Barroso, K. Lopez de Ipina, C. Hernandez, A. Ezeiza and M. Grana. "Semantic speech recognition in the Basque context, Part II: language identification for under-resourced languages." International Journal of Speech Technology, vol. 15, no. 1, pp. 41–47, 2012.

[30] S.M. Siniscalchi, J. Reed, T. Svendsen and C.H. Lee. "Universal attribute characterization of spoken languages for automatic spoken language recognition." Computer Speech and Language, vol. 27, no.1, pp. 209–227, 2013.

[31] B. Bhaskar, D. Nandi and K.S. Rao. "Analysis of language identification performance based on gender and hierarchial grouping approaches," in International Conference on Natural Language Processing, Dec. 2013.

[32] P.K. Polasi and K.S.R. Krishna. "Performance of Speaker Independent Language Identification System under various Noise Environments," in Proceedings of Third International Conference INDIA, vol. 1, 2016, pp. 315-320.

[33] S. Malmasi, M. Zampieri , N. Ljubesiˇ ,P. Nakov , A. Ali and J. Tiedemann. "Discriminating Between Similar Languages and Arabic Dialect Identification: A Report on the Third DSL Shared Task, in Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 3)," presented at the 25th Int. Conf. on Computational Linguistics (COLING), Osaka, Japan, Dec. 2016.

[34] S. Gaikwad, B. Gawali and P. Yannawar. "A Review on Speech Recognition Technique." International Journal of Computer Applications, vol. 10, no.3, 2010.

[35] Moler,C The Origins of MATLAB(http://www.mathworks.com/company/newsletters/news_notes/clevescorner/dec04.html,2011