SPOKEN LANGUAGE IDENTIFICATION USING MACHINE LEARNING

A project report submitted to

M S Ramaiah Institute of Technology

An Autonomous Institute, Affiliated to

Visvesvaraya Technological University, Belgaum

in partial fulfillment of the requirements for the degree of

Bachelor of Engineering in Computer Science & Engineering

Submitted by

Adarsh.D.Patil	1MS08CS004
Akshay Vishwas Joshi	1MS08CS008
Harsha.K.C	1MS08CS034
Pramod.N	1MS08CS073

Under the guidance of
Dr.K.G.Srinivasa
Professor
Department of Computer Science and Engineering
M.S.Ramaiah Institute of Technology

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

M.S.RAMAIAH INSTITUTE OF TECHNOLOGY

(Autonomous Institute, Affiliated to VTU)

BANGALORE-560054

www.msrit.edu

May 2012

SPOKEN LANGUAGE IDENTIFICATION USING MACHINE LEARNING

A project report submitted to

M S Ramaiah Institute of Technology

An Autonomous Institute, Affiliated to

Visvesvaraya Technological University, Belgaum

in partial fulfillment of the requirements for the degree of

Bachelor of Engineering in Computer Science & Engineering

Submitted by

Adarsh.D.Patil	1MS08CS004
Akshay Vishwas Joshi	1MS08CS008
Harsha.K.C	1MS08CS034
Pramod.N	1MS08CS073

Under the guidance of
Dr.K.G.Srinivasa
Professor
Department of Computer Science and Engineering
M.S.Ramaiah Institute of Technology

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

M.S.RAMAIAH INSTITUTE OF TECHNOLOGY

(Autonomous Institute, Affiliated to VTU)

BANGALORE-560054

www.msrit.edu

May 2012

Department of Computer Science & Engineering

M.S. Ramaiah Institute of Technology

(Autonomous Institute, Affiliated to VTU)

BANGALORE-560054

CERTIFICATE

This is to certify that the project work titled **Spoken Language Identification using Machine Learning** is carried out by *Adarsh.D.Patil-1MS08CS004, Akshay Vishwas Joshi-1MS08CS008, Harsha.K.C - 1MS08CS034 and Pramod.N-1MS08CS073* in partial fulfillment for the award of degree of Bachelor of Engineering in Computer Science and Engineering during the year 2012. The Project report has been approved as it satisfies the academic requirements with respect to the project work prescribed for Bachelor of Engineering Degree. To the best of our understanding the work submitted in this report has not been submitted, in part or full, for the award of any diploma or degree of this or any other University.

Dr. K. G. Srinivasa

Dr. R. Selvarani
Head, Dept of CSE

(External Examiner)

DECLARATION

We hereby declare that the entire work embodied in this report has been carried out by us at M S Ramaiah Institute of Technology under the supervision of Dr. K. G. Srinivasa. This report has not been submitted in part or full for the award of any diploma or degree of this or any other University.

Adarsh.D.Patil
Akshay Vishwas Joshi
Harsha.K.C
Pramod.N

ABSTRACT

Spoken Language Identification is the process of detecting the language of an utterance by an anonymous speaker, irrespective of gender, accent and pronunciations. Implementation of an acoustic model for Spoken Language Identification is to be carried out in this project. The major task is to identify those features or parameters which could be used to clearly distinguish between languages. This acoustic model makes use of mean values of Mel Frequency Cepstral Coefficients (MFCC). The system uses Support Vector Machine (SVM) to the handle the problem of multi class classification. The project aims at detecting English, Japanese, French, Hindi, and Kannada.

Experiments were conducted by forming a speech corpus using speech samples obtained from online podcasts and audio books. This corpus comprises of utterances, each of them spanning over a uniform duration of 10 seconds. The entire corpus is split into two sets, larger unit as the training dataset and a smaller set as the test set. Preliminary results indicate an overall accuracy of 96%. A more comprehensive and rigorous test indicates an overall accuracy of 80%. Thus, the acoustic model employing mean values of MFCC proves to be a viable approach for Language Identification.

ACKNOWLEDGEMENTS

We are thankful to MSRIT CSE department for having provided us with an opportunity to implement this project. We would like to thank our principal Dr. N. V. R. Naidu and Dr. R. Selvarani, HOD, Department of CSE for this opportunity.

We would like to thank our internal project guide Dr. K. G. Srinivasa, Professor, Department of CSE, MSRIT for his time and support throughout the course of the project which set the foundation for the successful completion of the project.

Contents

List of Figures

CHAPTER 1

INTRODUCTION

1.1 GENERAL INTRODUCTION

The phenomenon of globalization has brought together people from around the world. However one barrier to this increase in global communication is that many people speak different languages and effectively lack a common communication medium. That is, in order to communicate effectively a language which is mutually understandable by both parties is required. Language Identification offers a means for providing this medium.

Language as a communication system is thought to be fundamentally different from and of much higher complexity than those of other species as it is based on a complex system of rules relating symbols to their meanings, resulting in an indefinite number of possible innovative utterances from a finite number of elements. Among the various factors that define different cultures and communities, an important factor is language. The importance of speech and language for human to human communication can be over emphasized. Speech would thus be the most natural medium of interaction between humans and machines too. Language can be in the spoken or textual form.

Spoken Language Identification (LiD) is the process of identification of the language spoken in an utterance. Automatic language identification is the problem of identifying the language being spoken from a sample of speech by a speaker. As with speech recognition, humans are the most accurate language identification systems in the world today. Within seconds of hearing speech, people are able to determine whether it is a language they know. If it is a language with which they are not familiar, they often can make subjective judgments as to its similarity to a language they know.

Any utterance is nothing but a speech or audio signal. Speech processing is the study of speech signals and the processing methods of these signals. The signals are usually processed in a digital representation, so speech processing can be regarded as a special case of digital signal processing.

There are different aspects incorporated in speech which can be employed to represent the characteristics of a language. The raw speech signal is complex and may not be suitable

for feeding as input to the language identification system, hence the need for a good front-end arises. The task of this front-end is to extract all relevant acoustic information in a compact form. In other words, the pre-processing should remove all non-relevant information such as background noise, and encode the remaining (relevant) information in a compact set of features that can be given as input to the classifier.

The major task is to identify what features have to be extracted in order to discriminate between languages. Feature is a broad term with respect to speech signals. They could be acoustic features, prosodic features or phonotactic features. Prosody is the rhythm, stress, and intonation of speech. The prosodic of oral languages involve variation in syllable length, loudness, pitch, and the formant frequencies of speech sounds. This includes phoneme length and pitch contour. These prosodic units are the actual phonetic "spurts", or chunks of speech. Phonotactics: are rules that govern permissible sequence of phonemes in speech signals. Phonotactics defines permissible syllable structure, consonant clusters, and vowel sequences by means of phonotactical constraints.

The acoustic features are the low level features from which the prosodic and phonotactic features are derived. The acoustic features deal with modelling those parameters which are obtained from digital signal processing techniques. The power spectrum of a signal is indicative of acoustic information in speech. We make use of the cepstral analysis of the power spectrum of the speech signal. A cepstrum is the result of taking the Inverse Fourier transform of the logarithm of the spectrum of a signal. This data is used to model the language feature space.

1.2 PROBLEM STATEMENT

Given an utterance that belongs to the vocabulary of a language, the language identification systems should identify the language irrespective of gender and accents or pronunciations using an acoustic model, which extracts suitable features from speech samples across different languages and different speakers in each language. The language set considered in our project includes English, French, Japaneese, Hindi and Kannada. The generic LiD should be adopted to include regional languages like Hindi and Kannada. The system should evolve over a time with better accuracy and use continuous learning mechanism by incorporating machine learning techniques.

1.3 OBJECTIVES OF THE PROJECT

- To devise a system: confined to identify language of utterances from English, French, Hindi, Kannada and Japanese.
- The system should not restricted to a limited vocabulary, in other words the detection is independent of the content of speech.
- The system should not depend on any prosodic features like rhythm, stress and punctuation. The system performance should not be affected by the nature and gender of the speaker.

1.4 CURRENT SCOPE

The LiD system can be used in any Contact Centre deployment to pre-sort the callers based on the language they speak so that the required service or IVR can be provided in the language appropriate to the caller. Global call centres would benefit from this as callers from any part of the world may be redirected to the centres of their local native language without human intervention. The LiD system can act as a switchboard routing for incoming calls to operators fluent in the language.

Other services which make use of LiD are tourist information retrieval system, which lets the tourists retrieve information by querying the system in their native language.

LiD can be used for any indexed search engines and could be coupled with multilingual speech recognition systems to switch between recognizers. Spoken language interpretation and dialogue systems are other services which use LiD. Currently AT&T and Language Line Services partner to provide Customer Service Assistance in more than 170 Languages.

1.5 FUTURE SCOPE

LiD systems are still in the early stages of their deployments in real time systems. But there is a lot of potential for taking speech as input to communicate with a machine. The majority of the present real time deployments support few languages and English being the prominent. Hence multi lingual support can come up if systems are devised which could be accurate across languages.

There are various dimensions for improvements in the field of LiD. The systems should be made more robust, that is to improve the accuracy of identifying languages. Apart from this add support to a wider set of languages than the ones existing.

The major scope of LiD systems lies in making the LiD as the pre-processing stage for Language translation. This could be the biggest contribution to the field of speech translation and many applications can make use of speech as a more dominant input to machine.

CHAPTER 2

LITERATURE SURVEY

Research in the field of Spoken Language Identification started in the 1970s. During nearly 4 decades of research, many methods in different aspects were studied to achieve high performance language recognition. Spoken language identification is one of several processes in which information is extracted from a speech signal and this information is used to detect the language.

This information extracted could be in the following forms: phonotactic, prosodic or acoustic. Phonotactic approach deals with modelling speech at the phoneme or syllable level. A phoneme is a sound or a group of sounds that is the smallest unit which can be used to differentiate between utterances. Many approaches based on the Phoneme based features have been proposed in the field LiD [1].

Hieronymous and Kadambe proposed a task independent spoken language identification which uses a Large Vocabulary Automatic Speech Recognition (LVASR) [2]. The LVASR LiD system has many differences in the language model. Different languages have different number of phonemes, different word length, different word sequences which may contain high frequency words.

A Broad Phoneme [3] approach for Language identification was proposed by Berkling and Barnard. Their system claims 90% accuracy to discriminate between Japanese and English. The duo also proposed a theoretical error prediction for language identification system [4].

A segmental approach to Automatic Language Identification is based on the assumption that the acoustic structure of language can be estimated by segmenting the speech into phonetic categories [5]. Zissman has compared the performance of four approaches [6] for automatic language identification of speech utterances: Gaussian mixture model (GMM) classification; single-language phone recognition followed by language dependent, interpolated n-gram language modelling (PRLM); parallel PRLM, which uses multiple single-language phone recognizers, each trained in a different language; and language dependent parallel phone recognition (PPR).

Prosodic approach is another path taken for spoken language identification. Prosodic features encompass a large number of vocal tract dependent features like rhythm, pitch and stress. An approach to automatic language identification using pitch contour information is proposed by Lin and Wang [7]. A segment of pitch contour is approximated by a set of Legendre polynomials so that coefficients of polynomials form a feature vector to represent this pitch contour. Biadsy and Hirschberg [8] examined the role of intonation and rhythm across four Arabic dialects: Gulf, Iraqi, Levantine and Egyptian for the purpose of automatic dialect identification. This method gave good results with the duration of utterances being two minutes.

A novel phonotactic approach to LiD was described in Language Identification using Gaussian Mixture Model Tokenization [9] in which a Gaussian Mixture Model rather than a phone recognizer was used. To accomplish LiD a variety of methods using Gaussian Mixture Model and Hidden Markov Model are proposed. Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification [10] describes and evaluates three techniques that have been applied to the language identification problem: phone recognition, Gaussian mixture modelling, and support vector machine classification.

The next approach to LiD is the acoustic model. This aims at obtaining cepstral data from speech samples. The cepstral data which are used in majority of LiD systems are MFCC, LPC, PLPCC. Most of today's automatic speech recognition (ASR) systems are based on some type of Mel-frequency cepstral coefficients, which have proven to be effective and robust under various conditions. The advantage of applying the mel scale is that, it approximates the non-linear frequency resolution of the human ear.

The research with respect to speech recognition, speaker verification and language identification are derived from speech signal analysis and has the acoustic model as their bottom line. MFCC and other cepstral data are used to model them. The work [11] provides an insight to compute Mel frequency cepstral coefficients on the power spectrum. An adaptive algorithm for Mel cepstral analysis of speech was proposed by Fukada et al [12]. The process of generation of MFCC can be understood well in the work of Hasan et al [13] where they have applied it to speaker verification.

Mathematically, Language Identification is nothing but a maximum likelihood classification problem. Thus machine learning techniques can be applied to determine the

language. Many such approaches have been taken. One such system for Speaker and Language Recognition using Support vector machine was proposed by Campbell et al [14]. Artificial neural network based LiD system was also proposed [15]. This work makes use of two different statistical parameters namely prosodic and segmental features extracted from fundamental frequency contour (F0) and frequency spectrum were used for language classification.

From the detailed examination of the literature, it can be observed that acoustic model analysis coupled with machine learning techniques yields a good model for Language Identification.

CHAPTER 3

SOFTWARE REQUIREMENTS SPECIFICATION

3.1 INTRODUCTION

3.1.1 PURPOSE

The sole purpose of this project is to correctly identify the utterance in any language available in the training set. In this project we make use of the acoustic model for Language Identification.

The acoustic model means that only those features which are independent of prosodic or phonotactic information are used to model languages. One such feature is Mel Frequency Cepstral Coefficients (MFCC). The Mel-frequency Cepstrum is a representation of the short term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a non-linear Mel scale of frequency. Mel-frequency cepstral coefficients (MFCCs) are coefficients that collectively make up an MFC. They are derived from a type of cepstral representation of the audio sample which is a nonlinear "spectrum-of-a-spectrum". The MFC frequency bands are equally spaced on the mel scale, which approximates the human auditory system's response more closely than the linearly-spaced frequency bands.

The LiD system which we have proposed takes the mean of the MFCC at every nth order cepstrum. Similar means are computed till the 20^{th} order resulting in 20 mean MFCC values per speech sample. These are indicative of the acoustic data which is specific to each language and are independent of prosodic features.

This data constitutes the feature space for different languages which form the knowledge base for the classifier. This project makes use of Support Vector Machine (SVM) as the classifier. The process has two phases: training the SVM with cepstral data (mean MFCC) and testing it with speech samples.

3.1.2 SCOPE OF THE PROJECT

This project aims at identifying languages spoken by anonymous speaker. The current LiD system is capable of identifying five languages, which are English, Japanese, French, Hindi, and Kannada.

Also, the system is independent of the input file format and properties like sampling rate. Currently .WAV and .MP3 files are supported

The system is unaffected by the background noise in the speech samples.

This system has a client side and a server side. The client provides a system uploading speech samples to the server. The server extracts the acoustics features and uses SVM to detect the language.

The LiD system is capable to detecting languages across gender and accent.

3.1.3 ACRONYMS AND ABBREVATIONS

ASR – Automatic Speech Recognition

DCT – Discrete Cosine Transform

DSP – Digital Signal Processing

DFT – Discrete Fourier Transform

LiD - Language Identification

LPCC – Linear Predictive Cepstral Coefficient

LRE – Language Recognition Evaluation

MFCC - Mel-Frequency Cepstral Coefficient

PLPCC – Perceptual Linear Predictive Cepstral Coefficient

SVM - Support Vector Machine

3.1.4 OVERVIEW OF THE DOCUMENT

As part of general description the project perspective gives the general outlook of the project with regard to the features and functionality. Product functions details the description of the functionality of the end product. The subsequent subdivision talks about whom the product is designed for and who will be users of the devised system followed by the general constraints and common restraints of the system is subject to.

This is ensued by the assumptions and dependencies of the system and expectations from its end users

The next subsection elucidates the general operative necessities required by the system i.e. the functional, software and hardware requirements of the system on client and server side. The ensuing subsection describes the end user interface with the system which will serve as the foreface of the project and a portal for usage by the users.

This is followed by the comprehensive performance analysis of the devised system under various datasets and constraints. Graphs and plots cover show the efficiency and robustness of the system. Limitations of the system are covered in the next section which is concluded from the performance analysis.

This section illustrates the architectural design and components of the system. Hierarchy of components involved in processing of speech utterances to detection of language is shown. The detailed explanation of the functioning of individual modules of the system is taken up in detail designed. Each module is explained in terms of its functioning, processing and its input and output. Implementation of the above described modules and related algorithms for operation of each are detailed in following subsection. The implemented system is then tested for various input samples by subjecting it to diverse inputs to check the range of the system and robustness.

Following this is the project prospective vision and improvements are projected. Enhancements to make the system rigid and tolerable to all types of inputs and enhancements to make the system more practical for usage are stated. Next the references to material of research and articles for compiling this report and the project itself are stated. Lastly some screenshots of the implemented system and its working are shown.

3.2 GENERAL DESCRIPTION

3.2.1 PROJECT PERSPECTIVE

The project aims to devise a spoken language identification system which can detect languages in an unbiased manner, not giving regard to who is producing those utterances and how it is pronounced.

The major function of the Language Identification system is to recognize the language of the speaker accurately. This task has nothing to do with understanding what the speaker is

trying to communicate, which is the task of a Speech recognition system. LiD can be used as a pre-processing stage for multilingual speech recognition, that is, the language of the speaker is identified prior to the recognition, with an LiD which can switch between recognizers. This reduces the overhead of multilingual recognizers.

There are three approaches to LiD: Acoustic, Prosodic and Phonotactic. The approach taken here to model speech is the acoustic approach. This type of modelling comparatively makes use of lesser resources for training and testing when compared to a large speech vocabulary method (phonemes). The phonotactic model should maintain a database of all possible phonemes in a particular language which is a daunting task. Phonemes are a set of recurring, distinctive speech sounds. A phoneme is a sound or a group of sounds that is the smallest unit which can be used to differentiate between utterances. One phoneme may be more frequent in one language than another. An example of a phoneme is the /k/ sound in the words kit and skill.

Preparing this phoneme set for each language can be cumbersome and requires sufficient knowledge about the syllables, vowels and consonants of each language. Thus training the classifier requires more resources and training time is sufficiently large.

The syllables which are stored may be common to many languages. Thus the semantic rules of the language have to be understood in order to define syntax which describes syllable patterns for a language. The sentence patterns are different. Even when two languages share a word, e.g., the word “bin” in English and German, the sets of words that may precede and follow the word will be different.

The prosodic model requires modelling the vocal tract information and details regarding the rhythm, intonation and other vocal tract parameters like stress and pauses. If speech is perceived as a sequence of events in time, then word rhythm is used to refer to the way these events are distributed in time. Obvious examples of vocal rhythms are chanting as part of games (for example, children calling words while skipping, or football crowds calling their team's name) or in connection with work (e.g. sailor's chants used to synchronise the pulling on an anchor rope). In conversational speech the rhythms are vastly more complicated, but it is clear that the timing of speech is not random.

Conventional definition of speech rhythm as defined by Gibbon & Gut (2001) as “The recurrence of perceivable temporal patterning of strongly marked and weakly marked values of some parameter of a constant temporal domain”

Languages have been classified as either stress-timed, which refer to regularly occurring beats or stresses such as in British English and German, or syllable-timed, which refers to regularly timed syllables as in French. Off late a third classification has been introduced which is the 'mora', a sub-syllabalic timing unit that occurs mostly in Japanese and languages of that family.

Thus the prosodic and phonotactic approach have a higher overhead of understanding the linguistics of each language before extracting valuable speech data. The acoustic model has a universal approach across languages, which are standard digital signal processing techniques.

3.2.2 PRODUCT FUNCTIONS

The essential task of the LiD system is to identify spoken language. The system identifies languages for the ones it is trained to detect. The decision making is enabled by the SVM which is the machine learning technique tool used.

The system performs irrespective of the human innate features in their speech. This is accomplished by modelling the system about acoustic characteristics. Acoustic features are independent of speaker's intrinsic characteristics and hence their performance is unprejudiced.

The system functions by taking in speech sample from the user and processing it to detect the language. The system performs feature extraction on the input data. These extracted features enable the SVM to enroute the detection to a particular language.

This LiD is a mobile application as it rests on a client-server infrastructure. The client side can be used on a wide variety of systems and act as a link for user to upload the speech sample. The server in turn handles the incoming data by processing it and finally identifying the language.

3.2.3 END USERS

The end users may range from casual to industrial users each of them seeking a common response of language identification.

Rapid language identification can even save lives . There are many reported cases of 911 operators being unable to understand the language of the destressed caller. The current response service uses trained human interpreters who can handle about 140 languages.

The drawback with this system is that it has an innate delay because of human interpreters. An automated system can thus provide a more reliable and give faster responses.

Deployment of a LiD in a hotel lobby could cater to the queries of international customers. They can pose questions in their native languages and get help accordingly. The customers can make reservations, set menus, set cleaning schedules if they have a system that can recognize their language.

LiD finds extensive use in the tourism industry, as tourists may or may not know language used in the visited place. Hence such systems can act a link, enabling people from diverse community to be able to identify and by further introspection understand each others languages. This helps in propagation of correct information to the tourists which otherwise may get distorted due to limited understanding of languages.

International airports are common hosts to foreign travellers, as they might be present on a direct visit or hop journey. Such systems at airports can assist the airport authorities to gratify the needs of foreign tourist. Hence is voids the effect of language barrier on the service of the airport to the customers. Various speech activated systems which can understand limited range of languages can be expanded to cater to a larger language space.

Internationally operating companies maintain customer care centres to assist their clients. Hence such centres handle queries from across the globe, and these may not be in the same language. Presence of an automatic language identification module in such centres can help routing the customer's call to the language specific location. For instance a call from Germany can be automatically switched to a German proficient operator. This increases the efficiency of the organization's in understanding the customer's problems.

Dialogue systems are becoming common in places like parliament. These systems can identify the language being spoken and simultaneously broadcast it in multiple languages. One such implementation is found in the parliament. At present, in Lok Sabha, there is a facility for simultaneous interpretation in the following languages namely: Assamese, Bengali, Kannada, Malayalam, Manipuri, Maithili, Marathi, Nepali, Oriya, Punjabi, Sanskrit, Tamil, Telugu and Urdu is available. Hence in parliaments and such conventions like United Nations Organizations where representatives from across the globe gather a language identification system can be very useful.

Another implementation can be found in the audio and video media. The majority of communication is through television and radio, provisioning of a language identification system followed by a speech interpreter can bridge the language barrier for the viewers.

The next set of end users could be the people who query the search engine for supermarkets through speech. It is commonly observed that people tend to identify certain commodities with names in their native languages. In a supermarket one might find it difficult to spot the required commodity. A language identification system can identify the language being spoken and the further speech interpretation system can help getting the common name for the commodity.

The prevailing speech recognition systems like siri and iris are proficient in doing so for only English. This language limitation can be surpassed by efficient LiD systems. LiD can be the initial module where in the recognizer can first detect the language and then accordingly interpret them. Imagine using siri in your native language. The primary perspective of enabling diverse language in such systems is to wrap a larger community into the user space.

3.2.4 GENERAL CONSTRAINTS

The major constraint in the field of LiD is the lack of suitable resources. The initial problem in the formative years of LiD research was the lack of speech data across multiple languages. Over the years more speech data became available including multi-lingual speech database suitable for LiD research.

However recording across multiple languages is a start and obtaining accurate phonetic transcriptions of the speech data are mandatory. The utilization of word level information therefore, becomes a more serious problem.

The standard dataset provided by NIST for the Language Recognition Evaluation (LRE) is the largest corpus available only prior to the evaluation and the registrations for the upcoming evaluation in 2012-13 required registration in 2011.

Most of the languages have many dialects and sub categories. The speakers of the same language may sound different or have different accents in different parts of the world. For example English spoken in the United States and in India have a significant difference in the accent. Apart from this, within India the accents further change based on the location. Identifying a language irrespective of these constraints is not an easy task. Thus the

datasets must include a large variety of speakers, both male and female, having different accents to make the system more robust. Collecting such speech samples is a serious constraint in the field of LiD.

Determining the best duration of speech samples required for training is another task which cannot be overlooked. The feature space varies with the duration of the speech samples used. So fixing on an optimum duration of the utterances is important.

Speech signal processing involves the task of pre-processing and resampling data. The data provided for training should have a uniform sample rate and frequency. Apart from this, noise is an inherent property in any signal and may hamper performance if not handled appropriately. Noise could be the background noise or the noise from the recording devices.

3.2.5 ASSUMPTIONS AND DEPENDENCIES

The system developed assumes that the input test samples contain segments of the same language. The input speech sample is hence monolingual in nature.

We have considered all the dialects or variants of a language to be contributing to the same language. Hence English in any of its forms is still recognised as English. The speech input is assumed not to contain numbers as a significant part of the utterance.

The features are derived by applying transformations like discrete cosine transformation and Fourier transformations. Thus the feature computation is largely dependent on the digital signal processing parameters being imposed in the transformations.

3.3 SPECIFIC REQUIREMENTS

3.3.1 FUNCTIONAL REQUIREMENTS

- The service expected from the system is to identify the language in which the speaker is talking. The system recognises the language irrespective of the type of speaker, gender and accent of the speaker.
- Given a speech sample input to the system it should extract the acoustic feature MFCC and use this information to detect the language
- The input speech sample should not contain abrupt utterances, noise or background music

3.3.2 SOFTWARE REQUIREMENTS

Client side

The user should have a browser to access the language identification service.

The users system should be internet enabled.

Server side

The server should have the support vector machine libraries.

libsndfile library to enable reading WAV files format.

libmpg123 library to enable reading MP3 audio files.

liblapack library to enable general audio features like linear algebra routines..

fftw3 library to use FFTW for fast fourier transform computations.

3.3.3 HARDWARE REQUIREMENTS

Server Side

Intel core 2 duo processor

Linux Server environment with python support.

2 gigabytes of random access memory (RAM)

3.4 INTERFACE REQUIREMENTS

3.4.1 USER INTERFACE

The client side interface is an intuitively crafted web page, which is makes it easy to use. The web interface guides the user through the steps to be followed to upload an existing audio file. The audio file is then uploaded to the server where the language detection takes place.

3.5 GENERAL CONSTRAINTS

The utterances should be spelled at a uniform and slow rate. The speech sample should be clear and not contain noise in any form like laughter, music.

CHAPTER 4

SYSTEM DESIGN

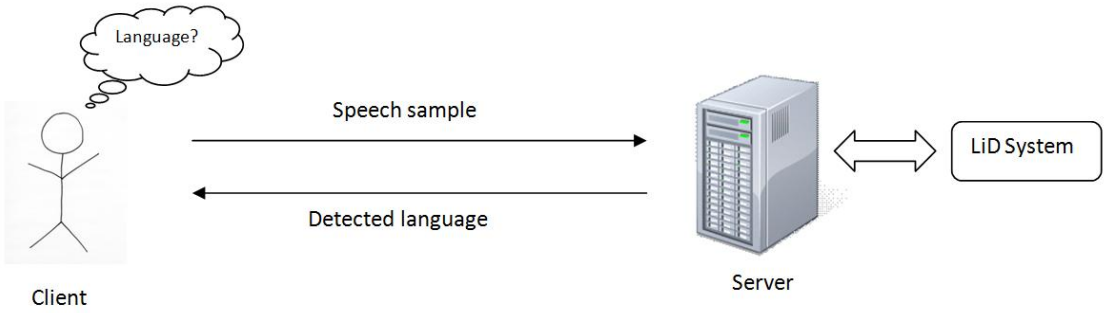

Figure 1: System Overview

Figure 1 depicts a high level design of the system and it follows client server architecture. The client side comprises of a portal which allows the user to upload an audio speech sample using a browser. The sample is sent over the network to the remote server running an LiD system. The LiD system processes this request and detects the language of the speech sample. The server returns a response indicating the language identified.

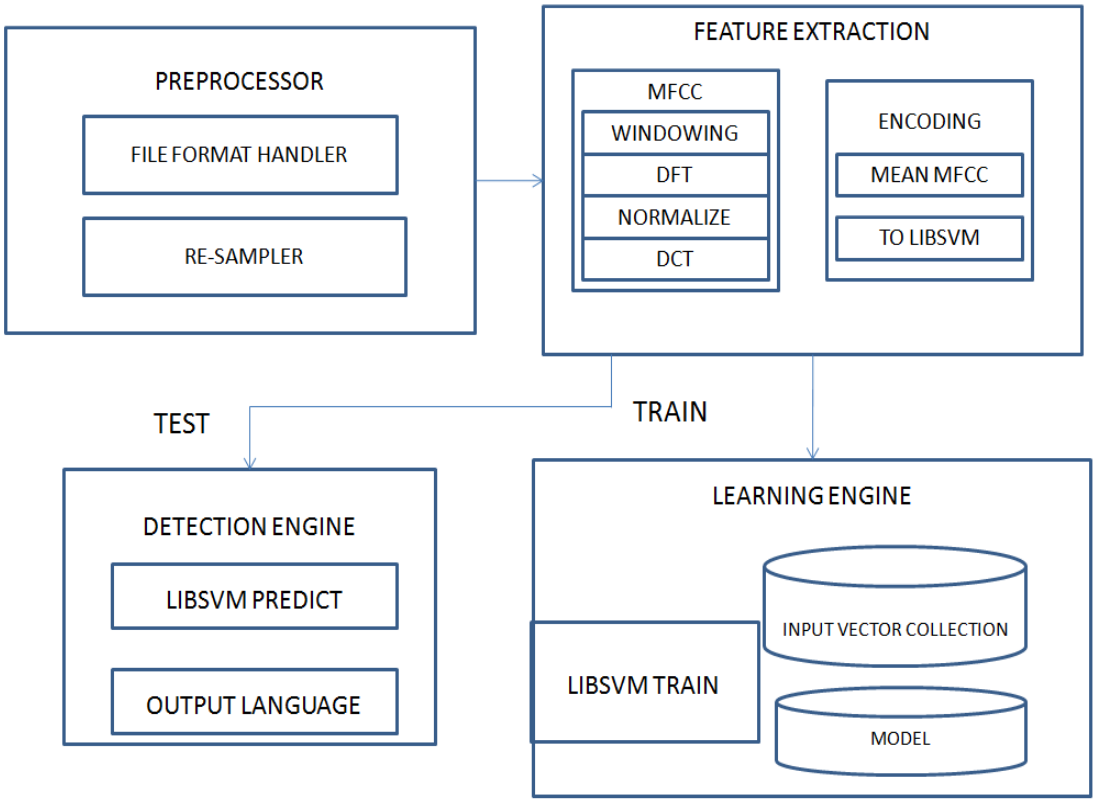

Figure 2: System Architecture

System Architecture: The proposed system architecture is shown in Figure 2. The LiD system takes speech samples as the input. There are three processing blocks in the architecture; they are, Pre-Processing Block, Feature Extraction Block and Machine Learning Block.

The Pre-Processing block is concerned with making the speech sample suitable for acoustic feature extraction. At this stage the audio file format is checked and converted to .wav if the input file format is .mp3. The audio file is resampled at 44.1 kHz before further processing. The final output of this block is a Pre-Processed Speech sample.

This sample is the input to the feature extraction block. The feature extraction block initially incorporates a windowing function, in our case a Hamming window aimed at making the signal zero-valued outside the chosen interval. The main task of the feature extraction block is to extract MFCC for the audio sample. It then computes the mean MFCC.

This cepstral feature vector serves as the input to the next block which is the Machine Learning Block. The SVM in the training phase creates a model based on the input feature vectors for different languages. This model file is used by the classifier in the testing phase to predict language. The final result is the language identified for the given test samples.

CHAPTER 5

DETAILED DESIGN

A detailed explanation of the LiD system is explained in this section. The process of language identification is carried out progressively in three stages 1) pre-processing 2) feature extraction 3) machine learning. These three phases together contribute to the language identification.

The pre-processing phase is concerned with changing properties of the input data to a same required value. Feature extraction phase deals with the processing of the input speech sample in order to extract required features. Features extraction generally involves performing various transformations and hence mathematical operations on the data. Machine learning phase operates in two phases which are training and testing. During the training phase the knowledge base is built using the vector space provided by the feature extraction phase. Whereas in the testing input speech sample are tested against the knowledge base developed in the training phase.

And use of machine learning inherently calls for two phases of operation, 1) training and 2) testing. The system is first trained with the available dataset and then tested with samples.

Pre-processing:

Pre-processing is the tuning stage of the system. In the pre-processing stage various methods are adopted to bring all the input data at the same configuration of the concerned fields or attributes. The basic pre-processing involves background noise reduction and re-sampling. The pre-processing performed has two steps 1) re-sampling and 2) file format handling.

- The first pre-processing is re sampling of the input audio data. The sampling rate defines the number of samples per unit of time (usually seconds) taken from a continuous signal to make a discrete signal. Re-sampling processes the input audio and tunes the sampling rate of every audio file to 44.1 KHz.
- File format handling takes care of format of the speech data sample. The file format handlers check the format of the speech sample for its format. If the format

is anything other than WAV then it converts it to WAV format. Hence the file format handler makes sure that all the input data is in the same format.

- Overall the pre-processor is concerned with providing the feature extraction stage with input having same file attributes.

Feature extraction:

Transforming the input data into the set of features is called feature extraction. Feature extraction is a very pivotal stage in language identification system. If the features extracted are carefully chosen it is expected that the features set will extract the relevant information from the input data in order to perform the desired task using this reduced representation instead of the full size input. When performing analysis of complex data one of the major problems stems from the number of variables involved. Analysis with a large number of variables generally requires a large amount of memory and computation power or a classification algorithm which over fits the training sample and generalizes poorly to new samples. Feature extraction is a general term for methods of constructing combinations of the variables to get around these problems while still describing the data with sufficient accuracy.

With respect to language identification the first task is to identify features which may provide us with information relevant to the task at hand. The audio features can be classified as high level and low level. Low level features are the ones which can be directly obtained from the speech samples without any additional operations. High level features are the ones that are derived from the audio by performing mathematical operations like transformations on them.

The feature extraction is not a one-step extraction process but involves many sequential phases. The feature extraction stage of the LiD is illustrated here. The data coming in after pre-processing undergoes the following steps 1) windowing 2) Discrete Fourier Transformation 3) Mel filter bank 4) Discrete Cosine Transform 5) Mean MFCC.

- **Windowing:** In signal processing, a window function is a mathematical function that is zero-valued outside of some chosen interval. The window is optimized to minimize the maximum (nearest) side lobe, giving it a height of about one-fifth that of the Hann window. We apply a hamming window to the speech utterance.

$$w(n) = 0.54 - 0.64 \cos\left(\frac{2\pi n}{N-1}\right)$$

- **Discrete Fourier Transform:** The DFT is defined mathematically as:

$$X_k = \sum_{n=0}^{N-1} x_n . e^{-i2\pi n \frac{k}{N}}$$

Given the sequence of N complex numbers $x_0 \ldots x_{N-1}$ is transformed into another sequence of N complex numbers according to the DFT formula shown above. The input signal which is in the time domain is converted to frequency domain by applying discrete Fourier transform. Data in frequency domain is easier to handle than in time domain hence the conversion.

- **Mel filter bank:** MFCCs are one of the most popular filter bank based parameterization used in speech technology. As with any filter bank based analysis technique an array of band pass filters are utilized to analyse the speech in different frequency bandwidths. A popular formula to convert *f* hertz into *m* mel is

$$m = 2595\ log_{10}\left(1 + \frac{f}{700}\right) = 1127\ log_e\left(1 + \frac{f}{700}\right)$$

- **Discrete cosine transformation:** A discrete cosine transform (DCT) expresses a sequence of finitely many data points in terms of a sum of cosine functions oscillating at different frequencies. DCT perform similar functions: they both decompose a finite-length discrete-time vector into a sum of scaled-and-shifted basis functions. The property of the DCT that makes it quite suitable for compression is its high degree of "spectral compaction;" at a qualitative level, a signal's DCT representation tends to have more of its energy concentrated in a small number of coefficients when compared to other transforms like the DFT. The output of the bandpass filter is used for MFCC extraction by application of discrete cosine transforms.

$$X_k = \sum_{n=0}^{N-1} x_n \cos\left[\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right] \qquad k = 0,1 \ldots N$$

- **Mean MFCC:** A mean of all the MFCC is taken at every cepstrum. We derive 20 such coefficients, which make up the feature space.

$$\text{Mean MFCC} = \frac{\sum X_i}{N}, \text{ where } X_i = \text{MFCC values}$$

Support Vector Machines:

Machine learning, a branch of artificial intelligence, is a scientific discipline concerned with the design and development of algorithms that allow computers to evolve behaviours based on empirical data.

The system uses machine learning techniques for the language classification problem. It makes use of support vector machine, as the machine learning block. Support vector machine is a supervised learning method used for classification and regression. LiD involves identification of languages from a set of languages hence it is a multi class classification problem. Therefore the system employs multi class support vector machine in its machine learning block.

Basic SVM algorithm is an efficient binary classifier. The idea behind SVM approach to language detection is that we map our data to a feature space. This feature space is the basis for the SVM algorithm which determines a linear decision surface (hyperplane) using the set of labelled data within it. This surface is then used to classify future instances of data. Data is classified based upon which side of the decision surface it falls. SVM is applicable to both linearly separable and non-linearly separable patterns. Patterns not linearly separable are transformed using kernel functions- a mapping function, into linearly separable ones. It can be formulated as follows. The optimal hyper plane separating the two classes can be represented as:

$$\omega.X + \beta = 0$$

where, X – sample input vectors defined as

$$\{(x_1,y_1),(x_2,y_2).....(x_k ,y_k)\}\ x_k \in R_n \quad y_i \in \{1,-1\}$$

ω,β - non zero constants ω indicating the weight component and β indicating the bias component

The ordered pair <x, y> is the representation of each input used to form hyper plane which are N dimensional vectors labelled with corresponding y.

$$\omega.X + \beta \geq 1 \quad \text{if } y_i = 1$$

$$\omega.X + \beta \leq -1 \text{ if } y_i = -1$$

These can be combined into one set of inequalities:

$$y_i(x_i.\ \omega + \beta) \geq 1 \quad \forall$$

The above inequalities hold for all input samples (linearly separable and suffice the optimal hyper plane equation). The optimal hyper plane is the unique one which separates the training data with a maximal margin. The figure 3 depicts the above mathematical representation. Figure 4 sketches the idea of hyper planes and classification in case of multi class data. One of the highlighting difference between the binary and multi class SVM is the set y={1,2,3,..k} and operations which are dependent on this set.

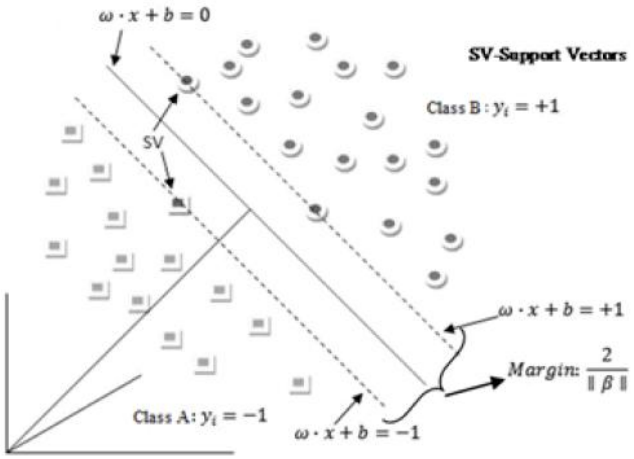

Figure 3: 2D representation of Support Vector Machine

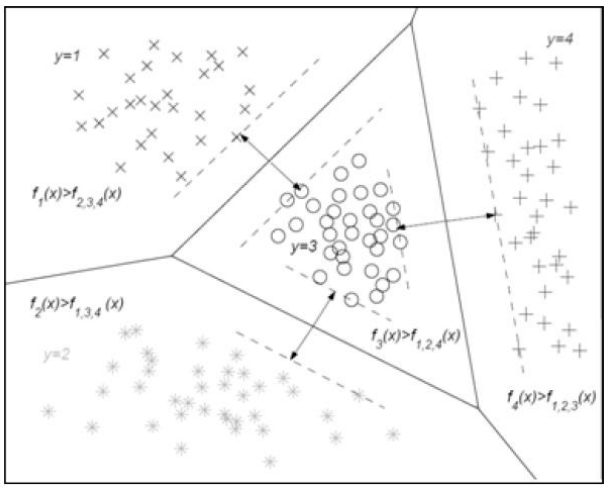

Figure 4: Classification in multi-class data

CHAPTER 6

IMPLEMENTATION DETAILS

The proposed system implements the LiD which makes use of python bindings for audio feature extraction. The libraries mentioned in the Software requirements section provide capabilities to extract mean MFCC values for the given sample.

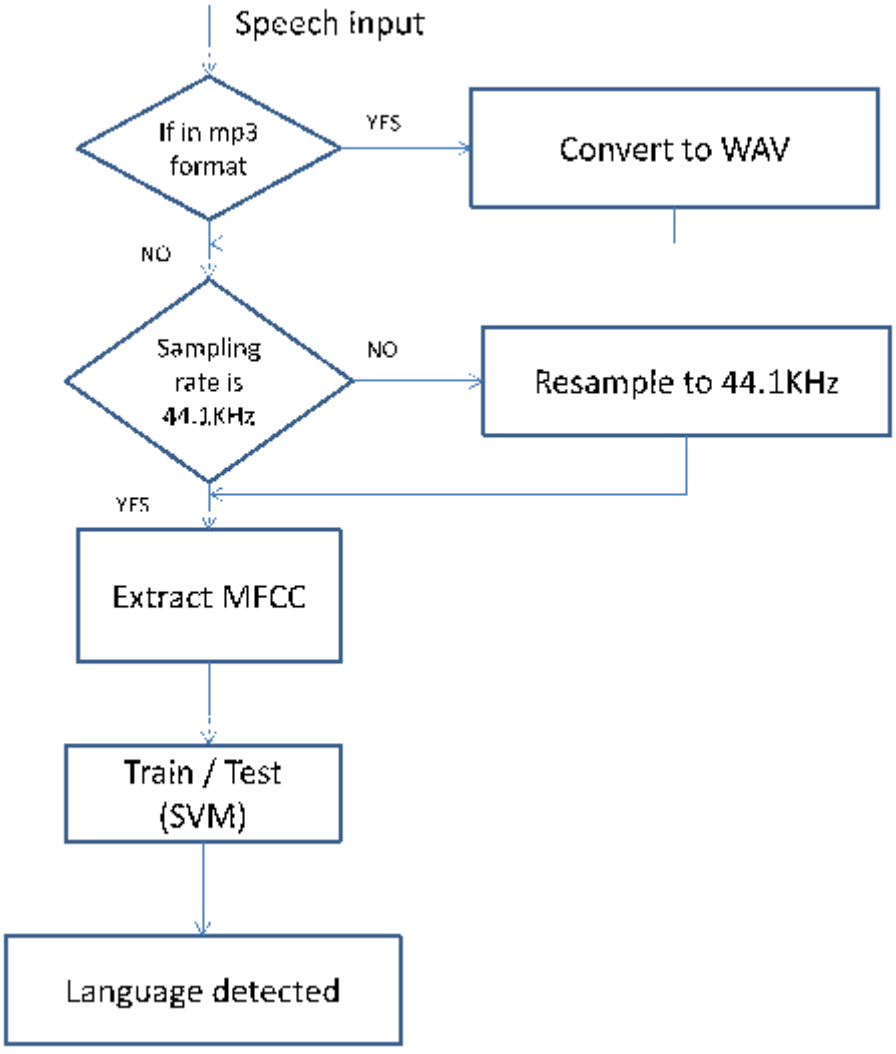

Figure 5 – Flow Diagram of the proposed LiD

The pseudo code of the system is as follows:

ALGORITHM LiD (speech_sample)

Input: The speech sample whose language has to be identified

Output: Mean values of MFCC which represent the language information

//Convert_to_wav(speech) : function to convert mp3 to wav

// resample(speech): re-samples input sample to 44.1KHz

// generate_MFCC(window, blocksize, stepSize, CepsNbCoeffs, computeMean): extracts mean MFCC considering

If(File_Type = Mp3) **then** Convert_to_wav(speech)

If(Sampling_Rate != 44.1kHz) **then** resampled_speech = resample(speech)

Vector = generate_MFCC (MFCC:Window = Hamming, blockSize=1024, stepSize=2048, CepsNbCoeffs=20, computeMean = True)

Return Vector;

ALGORITHM SVM_Train(Vector)

Input: The Support Vectors which have the cepstral data

Output: Model file which represents the knowledgebase

//svm-train(type, kernel, vector) : return a model file based on the parameters set for vectors

Model = svm-train(type = C-SVM, kernel = RBF, vector)

ALGORITHM SVM_Predict (Model,TestSample)

Input: The Model file built in the training phase and the test speech sample

Output: The language of the test sample

//svm-predict(vectors, model) : returns language identified for the given model and input vectors

Language = svm-predict (LiD (TestSample) , Model)

CHAPTER 7

PERFORMANCE ANALYSIS

The datasets for all our experiments are randomly taken from different parts of Web like podcasts and online audio books. The datasets are divided into two parts: Training Data and Testing Data. *N*-fold cross validation is adopted for training the machine for different languages. The system is trained over a large corpus of data and a small subset is used for testing to achieve better accuracy. The experiments are conducted to analyse the response of the proposed LiD against the considered languages (English, Hindi, French, Japanese and Kannada). The result is depicted in the form of a confusion matrix (Table 1).

INPUT : **OUTPUT OBTAINED :**

Number of Speech samples

English: 1093
French : 1069
Hindi : 853
Japanese : 539
Kannada : 868

Table 1: Confusion matrix

	Eng	Fr	Hin	Kan	Jap
Eng	98.558	0.0108	0.0027	0	0
Fr	0.0935	97.0065	0	0.0935	2.8
Hin	7.735	0.351	91.79	0.1172	0
Kan	2.9935	0.4608	0.1152	96.42	0
Jap	0	1.29	0	0.371	98.3302

From the confusion matrix it is evident that the system is reliable as the diagonal elements of the matrix holds highest values when compared to its row contemporaries. The accuracy for individual languages is as follows:
English : 98.558%
French : 97.0065%
Hindi: 91.79%
Kannada: 96.42%
Japanese: 98.3302%

The overall accuracy of the system for this sample data is **96.42%**.

Further, experiments are conducted to demonstrate the system accuracy for a chosen language. Around 105 English speech samples are fed to the system and the LiD demonstrated around 80% classification accuracy. The graph of classification accuracy of the system against English is shown in Figure 6 and it is evident that the systems perform well as it comes across more evidences against each language. The correctly classified instances of English language from a subset of the open source speech corpus, Vox Forge reveals that 85 out of the 125 samples were classified correctly as English The accuracy is found to be **80.95%**.

INPUT :

Number of English
Speech samples : 105

OUTPUT:

Samples classified as :
English : 85
French : 11
Hindi : 2
Kannada : 3
Japanese : 4

Accuracy : 80.95 %

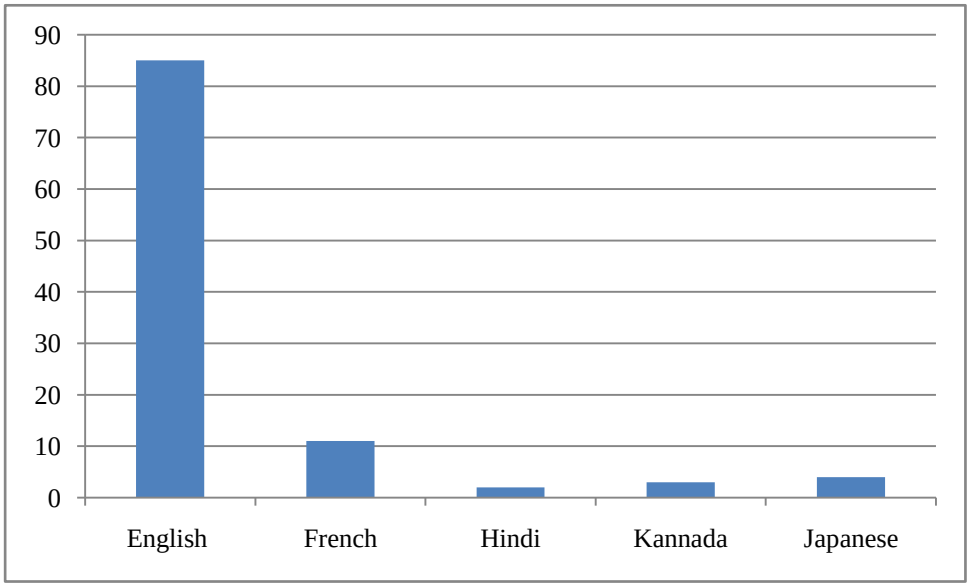

Figure 6 : Classification Accuracy of English dataset

CHAPTER 8

CONCLUSION AND FUTURE ENHANCEMENTS

8.1 CONCLUSION

- The current system is capable of identifying English, Kannada, Hindi, French and Japanese with an appreciable accuracy.
- An attempt to customize the LiD for regional languages like Kannada, Hindi is made.
- The major barrier with any LiD research is the availability of standard multi lingual speech corpus for training. This project has not made use of any standard dataset, but still competes for a good accuracy.

8.2 FUTURE ENHANCEMENTS

- The LiD system can be made more robust by increasing the number of samples for each language. Adding more speech samples from different speakers and incorporating different accents of the same language can improve the accuracy.
- The immediate improvement could be to add more languages to the existing dataset to enhance the boundary of identification of languages.
- The feature space can be enhanced by considering more acoustic parameters apart from MFCC and could incorporate a hybrid model comprising of many parameters.
- The biggest improvement to the system could be to incorporate incremental machine learning technique, that is, to learn from the utterances which the system had wrongly classified via a user feedback mechanism.

REFERENCES

1. K. M. Berkling, T. Arai and E. Barnard, "Analysis of phoneme-based features for language identification", in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing 94, Adelaide, Australia, April 1994
2. J. Hieronymous and S. Kadambe, "Spoken Language Identification Using Large Vocabulary Speech Recognition", in Proceedings of the 1996 International Conference on Spoken Language Processing (ICSLP 96), Philadelphia, USA, 1996.
3. K. M. Berkling and E. Barnard, "Language Identification of Six Languages Based on a Common Set of Broad Phonemes", in Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP 94), Yokohama, Japan, September 1994.
4. K. M. Berkling and E. Barnard, "Theoretical Error Prediction for a Language Identification System using Optimal Phoneme Clustering", in Proceedings 4rd European Conference on Speech Communication and Technology (Eurospeech 95), Madrid, Spain, September 1995.
5. Y. K. Muthusamy, "A Segmental Approach to Automatic Language Identification", Ph.D thesis, Oregon Graduate Institute of Science & Technology, July 1993.
6. M. A. Zissman, "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", in IEEE Trans. Speech and Audio Proc., SAP-4(1), January 1996.
7. Chi-Yueh Lin, Hsiao-Chuan Wang, “Language identification using pitch contour information”, from Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan
8. Fadi Biadsy, Julia Hirschberg, “Using Prosody and Phonotactics in Arabic Dialect Identification”, Department of Computer Science, Columbia University, New York, NY, 10027
9. Pedro A. Torres-Carrasquillo et al, “Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features”, 2002 International Conference on Spoken Language Processing (ICSLP 2006), Denver, USA, 2006.

10. Singer et al," Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification", MIT Lincoln Laboratory Lexington, MA, USA
11. Sirko Molau et al, "Computing mel-frequency cepstral coefficients on the power spectrum", from Computer Science Department, RWTH Aachen – University of Technology, 52056 Aachen, Germany
12. Fukada et al, An adaptive algorithm for mel-cepstral analysis of speech", Information Systems Research Center, Canon, Japan
13. Hasan et al, "Speaker identification using mel frequency cepstral coefficients, 3rd International Conference on Electrical & Computer Engineering ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh
14. Campbell et al, "Support Vector Machines for Speaker and Language Recognition", MIT Lincoln Laboratory.
15. Javad Shiekzadagen and Mahamood Reza Roohani, "Autoamtic spoken language identification based on ANN using fundamental frequency and relative changes in spectrum", Research centre of intelligent signal processing,Iran.

SCREENSHOTS

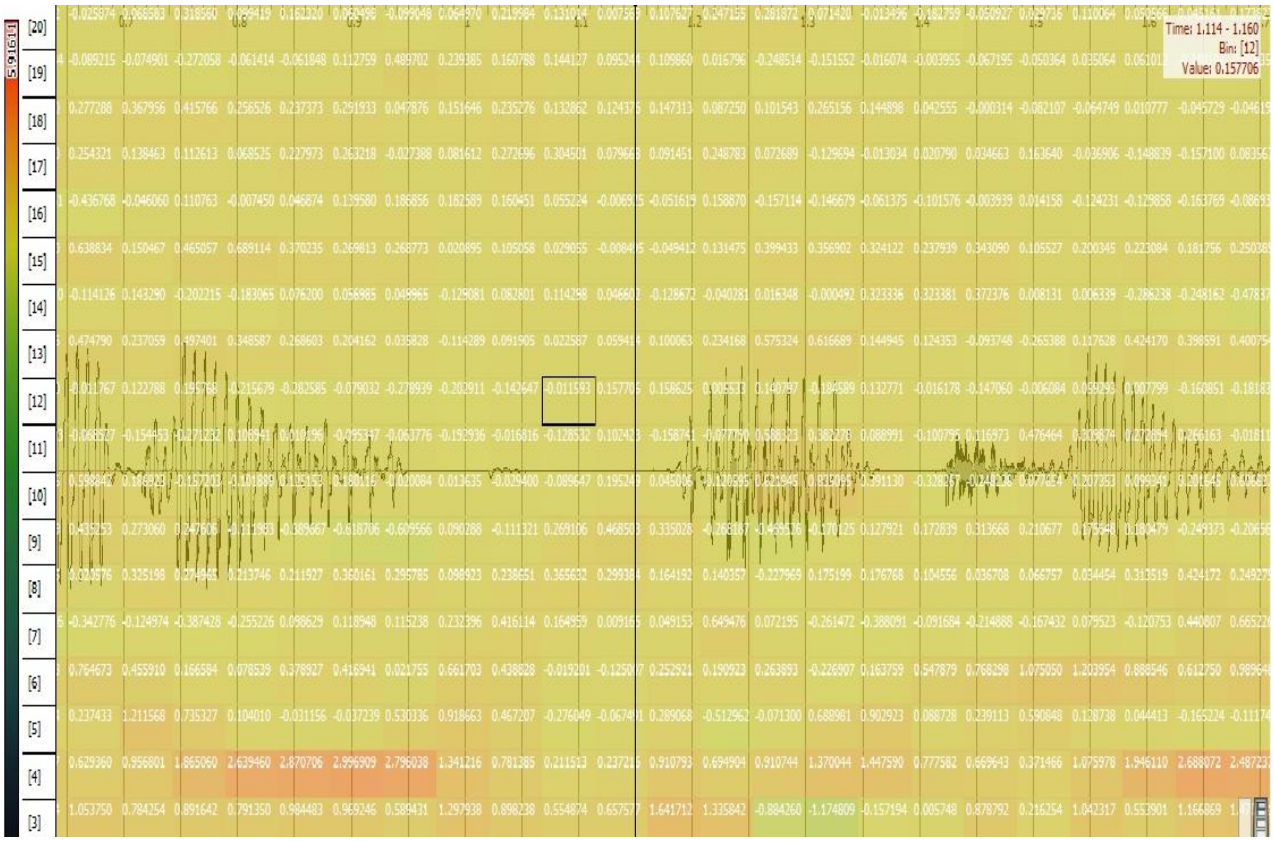

Figure 7: Visualization of MFCC Values

Figure 8: Web Interface for file upload

Table 2: Table depicting the MFCC values for different speech samples

INPUT :	**OUTPUT :**
Speech samples of ten seconds duration for each language considered.	1:2.83852212608 2:0.536402824333 3:0.8337923467 4:0.0633935242364 5:0.0832426107773 6:-0.156224855311 7:-0.132890900028 8:0.0401785314832 9:0.00606324821814 10:0.0293816095602 11:0.304801685214 12:-0.425143449145 13:0.151562740019 14:0.104476270797 15:-0.0376416936493 16:0.00582682680278 17:0.0531998953335 18:-0.0678337803949 19:0.0203557533966 20:-0.0177320342163

The feature extractor unit successfully provided 20 cepstrums of MFCC for the given speech sample.