SLID: Hybrid Learning Model and Acoustic Approach to Spoken Language Identification using Machine Learning

R. MADANA MOHANA[1], DR. A. RAMA MOHAN REDDY[2]

[1] Research Scholar, Department of Computer Science & Engineering, Sri Venkateswara University College of Engineering, Sri Venkateswara University, Tirupathi - 517 502, Andhra Pradesh, India.
rmmnaidu@gmail.com
[2] Professor, Department of Computer Science & Engineering, Sri Venkateswara University College of Engineering, Sri Venkateswara University, Tirupathi - 517 502, Andhra Pradesh, India.
ramamohansvu@yahoo.com

Abstract: - Spoken Language Identification (SLId) is the process of identifying the language of an utterance from an anonymous speaker, irrespective of gender, pronunciation and accent. In this paper we present acoustics based learning model for spoken language identification. An acoustic feature representing the short term power spectrum of sound called Mel Frequency Cepstral Coefficients (MFCC) is used as a part of the investigation in this paper. The proposed system uses a combination of Gaussian Mixture Model (GMM) and the Support Vector Machines (SVM) to handle the problem of multi class classification. The model aims at detecting English, Japanese, French, Hindi, and Telugu. A speech corpus was built using speech samples obtained from a plethora of online podcasts and audio books. This corpus comprised of utterances spanning over a uniform duration of 10 seconds. Preliminary results indicate an overall accuracy of 96%. A more comprehensive and rigorous test indicates an overall accuracy of 80%. The acoustic model combined with learning techniques hence proposed proves to be a viable approach for Language Identification.

Key-Words: - MFCC, Language Identification, SVM, GMM, LongRun technique.

1 INTRODUCTION

Language document identification is the process of identifying the language uttered in a given audio excerpt. The advent of artificial Intelligence gave rise to Computational Linguistics, a new branch of NLP, to devise algorithms for intelligently processing language data taking human-machine interaction to a new level. Extensive work was done to model language from a computational perspective. The initial years were dedicated to research in speech and speaker recognition systems, making speech a seamless input to machines.

The first question to consider before processing speech is what characteristics of speech could be used in computations. Speech is nothing but an audio signal which is characterized by many parameters. There are three approaches to analyze these parameters for linguistic computation, they are 1) Prosodic 2) Phonotactic 3) Acoustic.

Prosodic approach: Prosody is the rhythm, stress, and intonation of speech. The prosodic of oral languages involve variation in syllable length, pitch, loudness and the formant frequencies of speech sounds. This includes phoneme length and pitch contour. Some of the prosodic cues which have been proposed for Language Identification System are pitch contour shape of the pitch contour on the syllable, rhythm and phrase location initial/mid/final in breadth.

Phonotactic approach: Phonotactics are rules that govern permissible sequence of phonemes in speech signals. Phonotactics defines acceptable syllable structure, vowel sequences and consonant clusters by means of phonotactical constraints. This approach becomes more meaningful when the linguistics of the language is thoroughly known. The phonemes used while identifying language is a daunting task, because many phonemes overlap across languages. Hence the model should have good set of phonemes which can help identify languages accurately.

Acoustic approach: 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. Acoustic features are independent of speaker's intrinsic characteristics and hence their performance is unprejudiced.

The power spectrum of a signal is indicative of acoustic information in speech. The Cepstral analysis of the power spectrum of the speech signal is the most common acoustic feature. A cepstrum is the result of taking the Inverse Fourier transform of the logarithm of the spectrum of a signal. This data can be used to model the language feature space [23] & [24].

Some of the Cepstral coefficients which can be used are Mel Frequency Cepstral Coefficient (MFCC), Perceptual Linear Predictive Cepstrum Coefficient (PLPCC) and Linear Predictive Coding (LPC). The Mel-frequency Cepstrum is a illustration of the short term power spectrum of a noise based on a linear cosine transform of a log power spectrum on a non-linear Mel scale of frequency.

Mel-frequency Cepstral coefficients (MFCCs): MFCCs are coefficients that together build up an MFC. The MFC frequency bands are uniformly spaced on the Mel scale, which approximates the human acoustic system's reply more strongly than the linearly-spaced frequency bands.

Linear Predictive Cepstrum Coefficients (LPCC): It is a tool used for audio signal processing and is based on short term spectrum of the speech. The basic idea behind LPCC is to approximate a current sample to a series of past samples. The predictive and actual samples are used to obtain the coefficients.

Perceptual Linear Predictive Cepstrum Coefficient (PLPCC): This is also a short term spectrum of the speech. It modifies the short-term spectrum of the speech by several psychophysically based transformations.

This paper proposes an acoustic model to develop a language identification system. We will study the nature of other two approaches here in order to justify the selection of the acoustic model. The prosody of oral languages involves variation in syllable length, pitch, loudness and the formant frequencies of speech sounds. This approach may seem convincing but when considered for a diverse dataset of language its accuracy will get in jeopardy due to the infiltration of speaker innate features like irony, sarcasm, focus and other emotion.

The next approach is based on the phonemes of a language. Phonemes are the building blocks of the speech. Phonotactics define permissible syllable structure, consonant clusters, and vowel sequences. A model built upon these structural components i:e phonemes looks more appropriate when compared to a prosodic approach. This approach is resource and computation intensive as it should maintain intense phoneme dictionary consisting of syllables, vowels, consonants, and phonotactic syntax for all languages.

The acoustic feature that we have adopted is Mel Frequency Cepstral Coefficient (MFCC). MFCC by far has proved to be an efficient tool in various speech processing systems. The selection of MFCC is justified by the fact that MFCC is modelled to align the human auditory system.

As it is clearly evident that language identification is classification problem. Machine Learning will help in making the final discretion. We have used Support Vector Machines (SVM) as the learning engine for the LiD system. The SVM basically defines vectors, and uses them to draw boundaries between languages. These boundaries are a result of the training phase of the SVM. Once the boundaries are defined the system is subjected to test cases in the testing phase. The system uses the model developed during training to make a decision about the language of the test sample.

2 BACKGROUND

Research in the field of Spoken Language Identification (SLId) started in the 1970s. During the four decades of research, many methods in different aspects were studied to achieve high performance language recognition. Of many approaches the phonotactic approach deals with modeling 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. Different phoneme based approaches are proposed by Berkling et al [1].

Hieronymous and Kadambe proposed a task independent spoken language identification which uses a Large Vocabulary Automatic Speech Recognition (LVASR) [2]. The LVASR system has many differences in the language model. Different languages have different number of phonemes, word length, and word. 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 the following four approaches [6] for automatic language recognition of speech utterances, they are single language phone recognition followed by language dependent, interpose n-gram language modelling (PRLM); Parallel PRLM, which uses several single language phone recognizers, each one trained in a unusual language; language dependent Parallel Phone Recognition (PPR) and Gaussian Mixture Model (GMM) classification.

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 recognition using Gaussian Mixture Model Tokenization [9] in which a Gaussian Mixture Model rather than a phone recognizer was used. To accomplish SLId a variety of methods using Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) are proposed. Phonetic, acoustic and discriminative approaches to automatic language identification [10] describes and evaluates the three techniques that have been helpful to the language identification problem: phone recognition, support vector machine classification and Gaussian mixture modeling.

The next approach to SLId 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, and PLPCC. The advantage of applying the Mel scale is that, it approximates the non-linear frequency resolution of the human ear. 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 is further described in detail by Hasan et al [13] where they have applied it to speaker verification.

Mathematically, Language Identification is nothing but a maximum likelihood classification problem. A 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 a learning technique yields a good model for Language Identification.

3 THE CORPORA

One of the challenges faced during the development of a speech based intelligent system is the requirement of accurate and adequate data for training and testing. In our experimentation we handle this problem in a more realistic manner rather than the more conventional counterpart. Instead of building a system and testing for a set of standard input sample, we have used a speech corpus which consists of real-time/non-standard speech input from different users with different origin and background over the selected set of languages. The speech dataset is derived from podcasts and online audio books. This corpus comprises of utterances each of which span over a uniform duration of 10 seconds. All samples used are recorded in studio environment with reduced noise and glitches. It is semi spontaneous and colloquial in nature which resembles the real world closely. This approach has its own limitations and one of it is reduced system accuracy. It is mainly due to the fact that there is no standardization of the training data and is diverse in terms of distribution of speech instance and in terms of speech and sound characteristics.

4 LANGUAGE IDENTIFICATION SYSTEM

4.1 Stages in Language Identification System

All Language Identification systems irrespective of the type of the model should follow some basic steps for processing utterances. These steps can be visualized in Figure1:

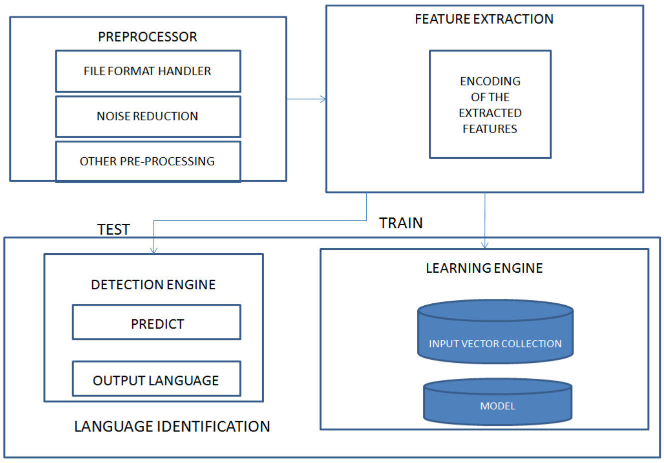


Fig. 1 Stages in Language Identification System.

Pre-processing: Pre-processing is the tuning stage of the system. The basic pre-processing involves background noise reduction and trimming the audio samples to durations which are suitable to extract sufficient features. This step should not be overlooked as the accuracy may change with the type of modelling and duration.

It is advisable to have raw speech data, which is without any lossy compression, given as input to the system. Hence the file formats, bit rate, frequency and number of channels of recording have to be taken care of, to avoid losing meaningful information in the audio signal.

In general the pre-processing is required to handle dissimilarities in the input and converging them on common the grounds.

Feature Extraction: Transforming the key data into various set of features is called feature extraction. Feature extraction is a very pivotal stage in language identification system. Whenever the features extracted are circumspectly preferred and it is estimated that the features set will extract the proper information from the input data. In order to perform the desired task this condensed representation used as an alternative of the full size input. Feature extraction is a common phrase for methods of constructing combinations of the variables to find around these problems whereas at a standstill relating the data with adequate 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 feature extraction is not a one-step extraction process but involves many sequential phases. Generally followed steps for feature extraction are described as follows. The first stage is to apply a window function to the input signal. In signal processing, window function is a statistical function that is zero valued outer surface of a number of chosen intervals. Hamming and Hanning are the prominently used window functions. Transformations are applied to shift domains. The frequency domain analysis holds grater relevance for audio processing; hence Fourier transforms (DFT, FFT) are usually used. Further filters are used to analyse the audio signal in different frequency bands.

Language Identification: The identification of language is nothing but a classification problem. Hence various approaches for classification be used majority of them being machine learning techniques like Hidden Markov Model (HMM), Gaussian Mixture Model (GMM).

Hidden Markov Model (HMM): The Hidden Markov Model (HMM) is a statistical model which follows the Markov property. A model which abides by the Markov property is called a Markov model. The Markov property states that one should be able make predictions for the future of a process based solely on its present state just as accurately as one could do so by knowing the process's full history. For any model there are three basic aspects, an input, states and the outputs. In case of a Hidden Markov model the states are hidden where as the outputs are known. But each state is mapped with certain probability over all the possible outputs. Thus a HMM can be used to decipher the states or parameters behind an output.

In case of a Language Identification system the HMM is used to learn the states or parameters behind a specific language sample. The system when trained with a large corpus of languages generates a probabilistic model of states for each language. In other words the highly probable state route by the language is marked. Any test sample is then featured against the available language equations. The model it matches with highest probability can be taken as the resulting language. The HMM is defined as follows [16] & [21]:

A HMM is a directed graph $<V, A>$ with vertices representing states $V = \{v_1, v_2, ..., v_n\}$ and arcs $A = \{<i, j> | v_i, v_j \in V\}$, showing transitions between states. Each HMM has the following additional components:

i) Initial state distribution used to determine the starting state at time 0, v_0.

ii) Each arc <i, j> is labeled with a probability p_{ij} of transitioning from v_i to v_j. This value is fixed.

iii) Given a set of possible observations, O{o_1, o_2,....,o_k}, each state, v_i contains a set of probabilities for each observations, {p_{i1}, p_{i2},...., p_{ik}}.

Gaussian Mixture Model (GMM): The Gaussian mixture model estimates probability density functions for each class (here each language will be a class), and then performs classification based on Bayes' rule. The Baye's theorem is used to calculate the probability of an event (say event 1) given another event (one or more, event 2) has already occurred. Extrapolating this to the Language Identification system scenario the parameters of languages sample are event 2 and the language itself is event 1. Hence GMM can be used as a classifier for a Language Identification system. The Bayes' theorem is given in equation (1).

$$P(H / X) = P(X / H)P(H) / P(X) \tag{1}$$

Here, P(H/X) is Posterior Probability, P(H) is Prior Probability associated with Hypothesis H, P(X) is Probability of the occurrence of data value X and P(X/H)

IS THE Conditional Probability that given a Hypothesis H, the tuple X satisfies it.

The Gaussian function is defined as [16]: The Gaussian function is a bell-shaped curve without values in the range [0, 1]. A typical Gaussian function is shown in equation (2)

$$f_i\ (S)=e^{-\frac{s^2}{v}} \qquad (2)$$

Here s is the mean and v is the predefined positive variance of the function. A typical Gaussian function is shown in Figure2.

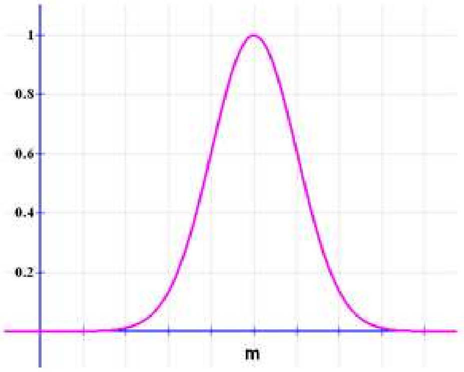


Fig. 2 A typical Gaussian Function.

4.2 Applications

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

The Language Identification 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 Language Identification system can act as a switchboard routing for incoming calls to operators fluent in the language. Internationally operating companies maintain customer care centres to assist their clients. Hence such centres handle 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.

Deployment of a Language Identification system 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 identify their language.

Language Identification system 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 other's 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. Language Identification systems at airports can assist the airport authorities to gratify the needs of foreign tourist. Hence it voids the effect of language barrier on the service of the airport to the customers. Various speech queries from across the globe, and these may not be in the same language. Presence of an automatic speech activated systems which can understand limited range of languages can be expanded to cater to a larger language space.

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 are 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 prevailing speech recognition systems like siri and iris are proficient in doing so for only English. This language limitation can be surpassed by efficient Language Identification systems. Language Identification system 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.

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.

Language Identification system 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 Language Identification system. Currently AT&T and Language Line Services partner to provide Customer Service Assistance in more than 170 Languages.

4.3 Challenges

The major constraint in the field of Language Identification system is the lack of suitable resources. The initial problem in the formative years of Language Identification system 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 Language Identification system research. Most of the speech databases available for research were telephonic speech corpus.

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.

Apart from the limitations of the dataset, another obstacle is the variation in the same language. 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 Language Identification system.

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.

Another pivotal constraint is feature selection, as there is no unique feature which can be used to discriminate between languages accurately. Hence selecting an optimum set of relevant features is an important decision to make while implementing a language identification system.

5 IMPLEMENTATION

The following section describes the architecture of SLId. SLId follows an acoustic model and this type of modelling makes use of lesser resources for training and testing when compared to a large speech vocabulary method and similar techniques(phonemes). In contrast, the phonotactic model maintains a database of all possible phonemes (a sound or a group of sounds that is the smallest unit that can be used to differentiate between utterances) occurring in a particular language. And in case of a prosodic approach the system would rely on features which may be morphed due to extraneous factors like emotions. The design includes various phases based on the flow of data and the action performed on this data.

The following Figure3 represents the overall system architecture:

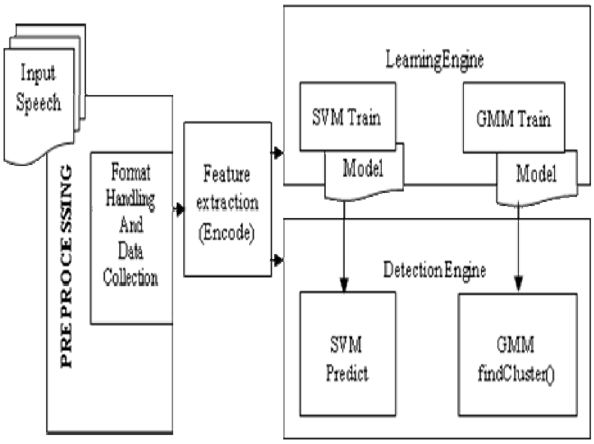


Fig. 3 SLId Architecture

The process of language identification is carried out progressively in three stages 1) Pre-processing 2) Feature Extraction 3) Machine Learning phase

Pre-processing: In the pre-processing stage the emphasis is on collecting data which would help in achieving the goal of identifying language and also to make further stages in the system easier by maintaining technical consistencies. Methods are adopted to bring all the input data at the same configuration of the concerned attributes. This involves background noise reduction, re-sampling and file format handling.

- The first step in pre-processing is file format handling and it takes care of format of the speech data sample. This is ensured to avoid loss in energy or power information which is required to measure the acoustic behaviour and hence the language behaviour. The input samples are all in WAV format. This is particularly useful because no data is removed in as part of compression unlike the compressed counterparts.

- Sampling rate is in direct effect with the amount of information contained by a speech sample. The sampling rate defines the number of samples per unit of time taken from a continuous signal to make a discrete signal. This process has a conditional execution in the system. Re-sampling processes the input audio and tunes the sampling rate of every audio file to 44.1 KHz.

Feature Extraction: Feature extraction is a pivotal stage in language identification system. If the features extracted are appropriately chosen it is likely that the feature set will extract the related information from the source input data in order to complete the required assignment using this reduced depiction instead of the full size input. The chosen acoustic feature is MFCC [18].

MFCC extraction is carried out in the following steps 1) windowing 2) Discrete Fourier Transformation 3) Mel filter bank 4) Discrete Cosine Trans-form 5) Mean MFCC. The block diagram for MFCC extraction is given in figure4:

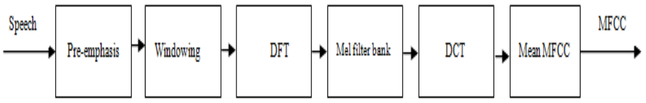


Fig. 4 Extraction of MFCC.

Windowing: A window function in signal processing is a mathematical task that is zero-valued outside of some preferred interval. Window is optimized to de-crease the maximum (nearby) side lobe, giving it a stature of about one fifth that of the Hann window. We pertain a hamming window to the speech utterance. Hamming Windowing is given by equation (3).

$$W(n) = 0.54 - 0.46 * \cos(2\Pi \frac{n}{N}), 0 \le n \le N \qquad (3)$$

The window length is L = N + 1

Discrete Fourier Transform (DFT): The DFT is defined mathematically shown in equation (4).

$$F[n] = \sum_{k=0}^{N-1} f[k] e^{-\frac{j2\Pi}{N}nk} \qquad (4)$$

Where n is 0 to N-1, F[n] is the DFT of the sequence f[k], Given the sequence of N instants or samples denoted f[0], f[1], f[2],….,f[N-1] and f[k] be the continuous signal which is the source of the data. Complex numbers x0,... ,xN-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 DFT.

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 Mel mf is given by equation (5).

$$m_f = 2595 \log_{10}(1 + \frac{f}{700}) \qquad (5)$$

Thus, with the help of Filter bank with proper spacing done by Mel scaling it becomes easy to get the estimation about the energies at each spot and once this energies are estimated then the log of these energies also identified as Mel spectrum can be used for calculating first thirteen coefficients using DCT. Since, the increasing numbers of coefficients represent faster change in the estimated energies and thus have less information to be used for classifying the given images. Hence, first thirteen coefficients are calculated using DCT and higher are unused. The following two famous experiments generated the Bark and Mel scales, given below figure5 describe the experiments. So, we make use of the Mel scale to manage the filter bank used in MFCC computation. Using function melbankm. This function returns a sparse matrix, so using command full to convert it to a regular matrix [17].

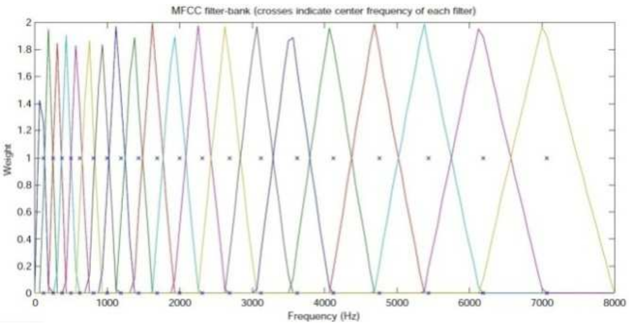


Fig. 5 Mel scale to organize the filter bank used in MFCC calculation

Discrete Cosine Transformation (DCT): A DCT expresses a series of finitely various data points in terms of a sum of cosine functions oscillating at different frequencies. DCT make similar functions and both decompose a finite length discrete time vector into a sum of scaled-and-shifted basis functions. DCT property makes it relatively suitable for density is its high degree of spectral compaction at a qualitative level, a signals Discrete Cosine Transformation representation tends to have more of its energy determined in a small number of coefficients when compared to other transforms like the DFT. The output of the band pass filter is used for MFCC extraction by application of discrete cosine transforms. DCT of Log of the Spectrum Energies given in equation (6).

$$y(k) = w(k)\sum_{n=1}^{N} x(n)\cos(\frac{\Pi(2n-1)(k-1)}{2N}) \qquad (6)$$

Where k = 1, 2, 3, …., N and w(k) is given by equation (7).

$$w(k) = \begin{cases} \frac{1}{\sqrt{N}}, k = 1 \wedge \sqrt{2/N}, 2 \le k \le N \end{cases} \qquad (7)$$

The output after applying DCT is known as MFCC (Mel Frequency Cepstrum Coefficient) given by equation (8).

$$c_n = \sum_{k=1}^{m} (\log D_k)\cos[m(k-\frac{1}{2})\frac{\Pi}{k}] \qquad (8)$$

Where m = 0, 1… k- 1 and cn represents the MFCC and m is the number of the coefficients.

Mean MFCC: A mean of all the MFCC is taken at every cepstrum. The mean is calculated given by equation (9):

$$Mean = \frac{1}{m}\sum_{n=2}^{k}(ABS(MFCC(n) - MFCC(1)) \qquad (9)$$

Where m is input parameters, n is the designed cases varies from 2 to k.

Reconstructing the spectrum based on MFCC. Some examples below figur6 and figure7: one segment of voice speech and another unvoiced [17], [19] & [20].

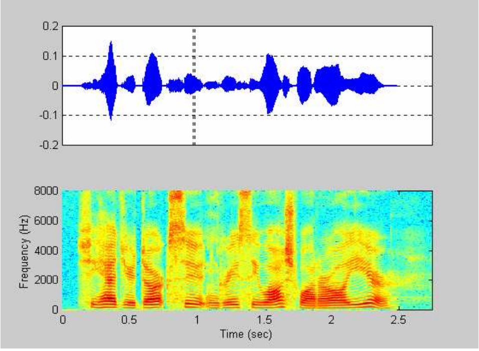


Fig. 6 Voiced Speech

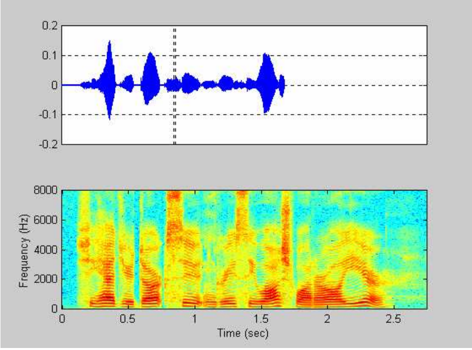


Fig. 7 Unvoiced Speech

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

We apply machine learning techniques for the language classification problem.We make use of support vector machine, as the machine learning block. Support vector machines are a set of related supervised learning methods used for classification and regression. As it is required to identify the language of the speech sample from the set of languages which it is trained with, we make use of multi class Support Vector Machine (SVM).

Support Vector Machine (SVM): The system uses support vector machines, as the learning technique for the language classification problem. SLId involves identification of languages from a set of languages therefore it employs multi class support vector machines.

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 given in equation (10):

$$w.X+\beta=0 \tag{10}$$

Where, X - sample input vectors defined as $\{(x_1,y_1),(x_2,y_2)............(x_k\ ,y_k)\}$ $x_k \in Rn$, $y_i \in \{1,-1\}$ and 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 hyperplane which are N dimensional vectors labelled with corresponding y are given in equations (11) and (12).

$$w.X+\beta \geq 1 \qquad \text{if } y_i = 1 \tag{11}$$

$$w.X+\beta \leq -1 \qquad \text{if } y_i = -1 \tag{12}$$

These can be combined into one set of inequalities is given in equation (13):

$$y_i\,(x_i\,.w+\beta) \geq 1 \tag{13}$$

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. 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.

A SVM operation consists of the two phases. In the training phase SVM plots the vectors on an N-dimensional space. Hence in the case of LiD the mean MFCC form the vector space for the SVM [22].

In The testing phase, the speech sample is subjected to feature extraction and similar data, that is, 20 orders of mean values of MFCC. Using the model built in the training phase, the SVM predicts the language of the test sample.

The proposed system implements the SLId which makes use of python bindings for audio feature extraction. The libraries used in our proposed system provide capabilities to extract mean MFCC values for the given sample. In our SLId system, the server should have libraries shown in Table1:

TABLE 1
SLID SYSTEM LIBRARIES

SLId System Libraries
The server should have the *SVM libraries*.
libsndfile library to enable reading WAV file formats.
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.

The various steps in our proposed SLId system testing for the given speech sample inputs are shown in Figure8:

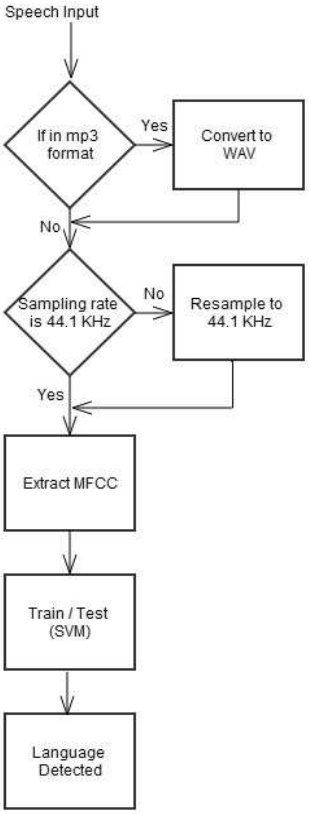


Fig. 8 Steps in testing SLId system

The pseudo codes for our implementation are shown in the following tables 2 to 5 which consisting of algorithms for SLId speech, generate_MFCC (resampled_speech), SVM_Train (Vector) and SVM_Predict (Model,TestSample):

TABLE 2
ALGORITHM FOR SLID (SPEECH)

Algorithm SLiD (speech)

Input: The speech sample whose language has to be identified

Output: Mean values of MFCC which represent the language information

```
If( File_Type  =  Mp3)
        Convert_to_wav( speech )
If( Sampling_Rate  !=  44.1kHz )
        resampled_speech = resample( speech )
Vector = generate_MFCC( resampled_speech )
```

TABLE 3
ALGORITHM FOR GENERATE_MFCC(RESAMPLED_SPEECH)

Algorithm generate_MFCC(resampled_speech)

Input: The pre-processed speech sample

Output: Mean MFCC values upto 20thcepstrum

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

TABLE 4
ALGORITHM FOR SVM_TRAIN (VECTOR)

Algorithm SVM_Train(Vector)

Input: The Support Vectors which have the cepstral data

Output: Model file which represents the knowledgebase

```
Model = SVM( type = C-SVM, kernel = RBF )
```

TABLE 5
ALGORITHM FOR SVM_PREDICT (MODEL, TESTSAMPLE)

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

```
Language = SVM ( LiD (TestSample) , Model)
```

6 EXPERIMENTAL RESULTS

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 input speech samples are given in Table6.The experiments are conducted to analyse the response of the proposed SLId against the considered languages (English, Hindi, French, Japanese and Telugu). The result is depicted in the form of a confusion matrix (Table7) and graphically represented in figure9. This test case is the trivial case of testing the SLId with the entire training set. Confusion matrix is defined as follows:

Confusion matrix illustrates the accuracy of the solution to a classification problem. Given m classes, a confusion matrix is mxn matrix where c_{ij} indicates the number of tuples from D that were assigned to class c_j but where the correct class is c_i

TABLE 6
INPUT SPEECH SAMPLES

Number of Speech samples	
English	1093
French	1069
Hindi	853
Japanese	539
Telugu	868

TABLE 7
CONFUSION MATRIX

	Eng	Fr	Hin	Tel	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
Tel	2.9935	0. 4608	0.1152	96.42	0
Jap	0	1.29	0	0.371	98.3302

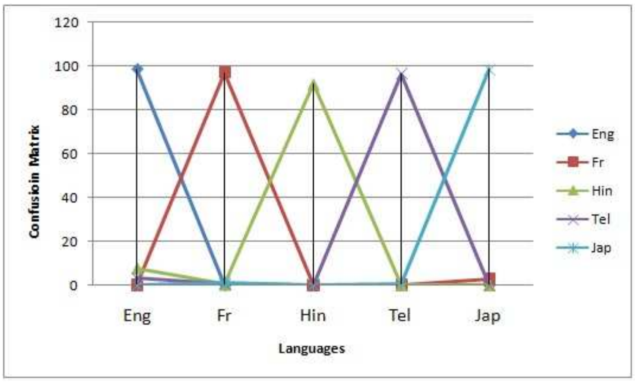
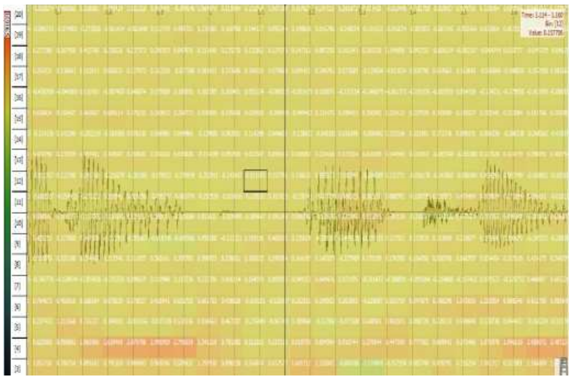


Fig. 9 Performance analysis of SLId

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. 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.

Further, experiments are conducted to demonstrate the system accuracy for a chosen language. Around 105 English speech samples were used to test the system and the LiD demonstrated around 80% classification accuracy. The graph of classification accuracy of the system against English is shown in Figure10, it is evident that the system 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%.



The

Fig. 10 Classification Accuracy of English language

The

The TIMIT database, that consists of labe segments of speech. As an example, a phrase is shown Figure7. The TIMIT database gives the first and l samples of each word, and the first and last sample each phoneme too. In Figure11, the vertical line in b indicates the beginning of a word and the red line indicates its end [17].

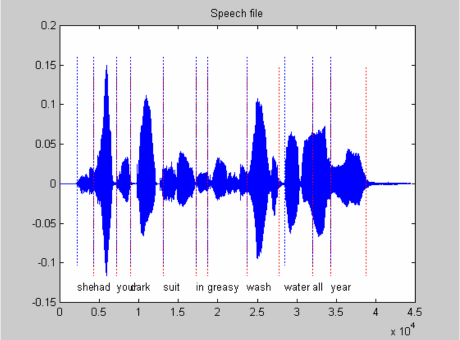


Fig. 11 Example of a phrase in the TIMIT database: "she had your dark suit in greasy wash water all year".

A snapshot is the state of a system at a particular point in time. It can refer to an actual copy of the state of a system or to a capability provided by certain systems. These snapshots output a set of Classification Accuracy of languages used in our SLId system. The following Figure12 shows the GUI of SLId system.

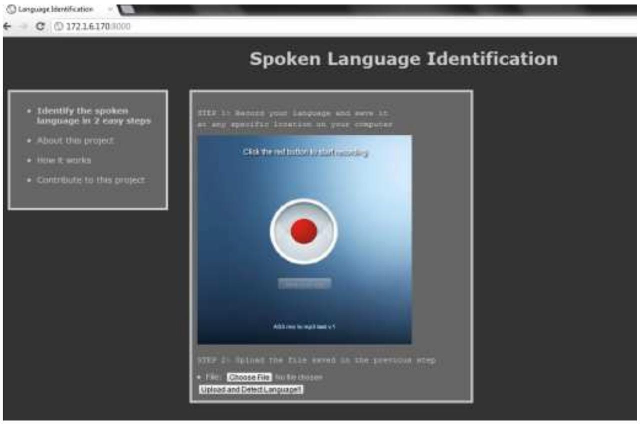

Language Identification
172.1.6.170:8000
Spoken Language Identification
- Identify the spoken language in 2 easy steps
- About this project
- How it works
- Contribute to this project
STEP 1: Record your language and save it at any specific location on your computer
Click the red button to start recording
STEP 2: Upload the file saved in the previous step
- File: Choose File No file chosen
Upload and Detect Language!!

Fig. 12 GUI of SLId system

7 CONCLUSION

SLId system should be capable of accurately identifying the language of a speech sample for which it is trained. The current system is capable of identifying English, Telugu, Hindi, French and Japanese with an appreciable accuracy.

There are very few SLId systems which provide support for regional languages and adding Telugu to the set of languages is a minimal contribution.

The major barrier with any SLId research is the availability of standard multi lingual speech corpus for training. Our proposed system has not made use of any standard dataset, and still competes for a decent accuracy.

A novel approach of creating the dataset was tried. As SLId systems doesn't require phoneme level description or syllable database but needs noise free speech at a constant level. Thus the online audio books without background sound and few podcasts were used to build the corpus.

The SLId 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 (1994). "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 (1996). "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 (1994). "Language Identification of Six Languages Based on a Common Set of Broad Phonemes", *in Proceedings of the 1994 International Conference on Spoken Language Processing (ICSLP94),* Yokohama, Japan, September 1994.

[4] K. M. Berkling and E. Barnard (1995). "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 (1993). "A Segmental Approach to Automatic Language Identification", *Ph.D thesis, Oregon Graduate Institute of Science & Technology,* July 1993.

[6] M. A. Zissman (1996). "Comparison of Four Approaches to Automatic Language Identification of Telephone Speech", *in IEEE Transaction Speech and Audio Processing*, SAP-4(1), January 1996.

[7] Chi-Yueh Lin, Hsiao-Chuan Wang (2005). "Language identification using pitch contour information", from *IEEE ICASSP-2005,* Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan.

[8] Fadi Biadsy, Julia Hirschberg (2009). "Using Prosody and Phonotactics in Arabic Dialect Identification", *In Proceedings of Interspeech 2009*, Brighton, UK.

[9] Pedro A. Torres-Carrasquillo et al (2006). "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] E.Singer et al (2003). "Acoustic, Phonetic, and Discriminative Approaches to Automatic Language Identification", *In Proc. Eurospeech, 2003*.

[11] Sirko Molau et al (2001). "Computing mel-frequency cepstral coefficients on the power spectrum", Proceedings. (ICASSP '01). *2001 IEEE International Conference*, Salt Lake City, UT, USA.

[12] Fukada et al (1992). "An adaptive algorithm for mel-cepstral analysis of speech", *IEEE conference on Acoustic, Speech and Signal Processing (ICASSP-92),* 1992, Information Systems Research Center, Canon, Japan.

[13] Hasan et al (2004). "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 (2006). "Support Vector Machines for Speaker and Language Recognition", *Computer Speech and Language*, 2006, Elsevier, MIT Lincoln Laboratory.

[15] Javad Shiekzadagen and Mahamood Reza Roohani (2000). "Automatic spoken language identification based on ANN using fundamental frequency and relative changes in spectrum", *International Conference on Speech Science and Technology (SST-2000),* 2000, Research centre of intelligent signal processing, Iran.

[16] Margaret H. Dunham (2008). "Data Mining Introductors and advanced topics", *Pearson Education*, 2008.

[17] Davis, S.; Mermelstein, P. (1980). "Comparison of Parametric Representations for Monosyllabic Word

Recognition in Continuously Spoken Sentences”, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 4 (1980).

[18] Wei Han, Cheong-Fat Chan, Chiu-Sing Choy and Kong-Pang Pun (2006). “ An Efficient MFCC Extraction Method in Speech Recognition”, Department of Electronic Engineering, The Chinese University of Hong Kong, Hong, *IEEE – ISCAS*, 2006.

[19] Lindasalwa Muda, Mumtaj Begam and I. Elamvazuthi (2010). “Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques”, *Journal of Computing*, Volume 2, Issue 3, March 2010, Malaysia.

[20] Shikha Gupta, Jafreezal Jaafar, Wan Fatimah wan Ahmad and Arpit Bansal (2013). “Feature Extraction using MFCC”, *Signal & Image Processing : An International Journal (SIPIJ)* Vol.4, No.4, August 2013.

[21] Mark Gales and Steve Young (2007). “The Application of Hidden Markov Models in Speech Recognition”, *Foundations and Trends in Signal Processing*, Vol. 1, No. 3 (2007), UK.

[22] Shi-Huang Chen and Yu-Ren Luo (2009). “Speaker Verification Using MFCC and Support Vector Machine”, *Proceedings of the International MultiConference of Engineers and Computer Scientists 2009*, Vol.1, IMECS 2009, March 18 - 20, 2009, Hong Kong.

[23] Katrin Kirchhoff, Gernot A. Fink, Gerhard Sagerer (2002). “Combining acoustic and articulatory feature information for robust speech recognition”, *Speech Communication* 37 (2002), Elsevier, USA.

[24] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and Brian Kingsbury (2012). “Deep Neural Networks for Acoustic Modeling in Speech Recognition”, *IEEE Signal processing Magazine*, November, 2012.