# A Hierarchical Automatic Language Identification System for Indian Languages Using Acoustic Features

*S. Jothilakshmi[1] V. Ramalingam[2] S. Palanivel[3]*

**Abstract**

Automatic spoken language identification (LID) is the task of identifying the language from a short utterance of the speech signal uttered by an unknown speaker. This paper describes a novel two level identification system for Indian languages using acoustic features. In the first level, the system identifies the family of the spoken language, and then it is fed to the second level which aims at identifying the particular language in the corresponding family. The proposed system has been modelled using Hidden Markov Model (HMM) and utilizes the acoustic features namely Mel frequency cepstral coefficients (MFCC) and Shifted delta cepstrum (SDC). A new database has been created for 11 Indian languages. The proposed system achieves a high accuracy of 62.36% for MFCC features and 71.2% for SDC features.

**Keywords:** Language identification, Indian languages, Hidden Markov model, Mel frequency cepstral coefficients (MFCC), Shifted delta cepstrum (SDC).

## 1. Introduction

The automatic language identification (LID) [1], [2] is a process by which the language spoken in a particular speech utterance is identified. It is an important technology in many applications, such as spoken language translation [3], multi lingual speech recognition [4], and spoken document retrieval [5].

[1, 2 & 3]Department of Computer Science and Engineering, Annamalai University, Annamalainagar-608 002, India. Email : jothi.sekar@gmail.com, aucsevr@yahoo.com, spal_yughu@yahoo.com

Humans are the best LID systems in the world today. Just by hearing one or two seconds of speech of a familiar language, they can easily identify the language. The performance of any LID system depends on the amount of information and the reliability of information extracted from the speech signal and how efficiently it is incorporated into the system [1].

Existing spoken language identification systems can be broadly classified into two groups namely, explicit and implicit systems, The LID systems that require speech recognizers of one or several language, in other words, the systems that require a segmented and labelled speech corpus are termed as explicit LID systems. The language identification systems which do not require phone recognizers (or rather segmented and labelled speech data) are termed here as implicit LID systems. In other words, these systems require only the raw speech data along with the true identity of the language spoken [1], [6].

In general, LID features fall into five groups according to their level of knowledge abstraction [7]. Lower level features, such as spectral feature, are easier to obtain but volatile because speech variations such as speaker to channel variations are present. Higher level features, such as lexical /syntactic features, rely on large vocabulary speech recognizer, which is language and domain dependant. They are therefore difficult to generalize across languages and domains. Phonotactic features become a trade-off between computational complexity and performance. It is generally agreed that phonotactics i.e. the rules governing the sequences of admissible

phones/phonemes, carry more language discriminative information than the phonemes themselves. This work focuses on acoustic only LID system for which Hidden Markov Modelling is the state of the art classifier.

In this paper, a hierarchical language identification system has been proposed for Indian languages. The languages of India belong to four major linguistic families namely Indo Aryan, Dravidian, Austro-Asiatic and Tibeto-Burman [8]. The largest of these in terms of speakers is Indo Aryan which is spoken by 75.278% of the people. The second largest is the Dravidian family which is spoken by 22.5% of the people whereas Austro Asiatic is spoken by 1.132% of the people and Tibeto- Burman by 0.965%. So, nearly 98% of the people in India are speaking languages from Aryan family and Dravidian family. Hence the proposed two level systems are designed to identify the languages of Aryan and Dravidian family. In the first level, the system identifies the family of the spoken language, and then it is fed to the second level which aims at identifying the particular language in the corresponding family.

The aim of this work is to design a less complex system for Indian language identification, so only the acoustic features are utilized in the system. The most widely used features for LID are Mel frequency cepstral coefficients (MFCC). Traditionally, language and speaker identification tasks use feature vectors containing cepstra and delta and acceleration cepstra. Recently, however, the shifted delta cepstrum (SDC) has been found to exhibit superior performance to the delta and acceleration cepstra in a number of language identification studies [9] due to its ability to incorporate additional temporal information, spanning multiple frames, into the feature vector. The performances of MFCC and SDC features are compared in this paper.

The paper is organized as follows: section 2 briefly reviews the feature extraction used for LID. In section 3, we briefly review Hidden Markov model. Our method is presented in detail in section 4. Experimental settings and results are described in section 5. Finally the conclusion is given in section 6.

## 2. Feature Extraction For Language Identification

Speech signals need to be parameterized prior to identification process. Parameterization consists of the extraction of a set of features from the speech waveform, which may present two main characteristics: they must provide a reasonable and compact representation of the speech signal and they must have adequate discrimination capabilities for discriminating between sounds.

### 2.1 Mel Frequency Cepstral Coefficients (MFCC)

Mel frequency cepstral coefficients (MFCC) have proved to be one of the most successful feature representations in speech related recognition tasks [10]. The mel-cepstrum exploits auditory principles, as well as the decorrelating property of the cepstrum. The computation of MFCC is shown in Fig.2.1 and described as follows.

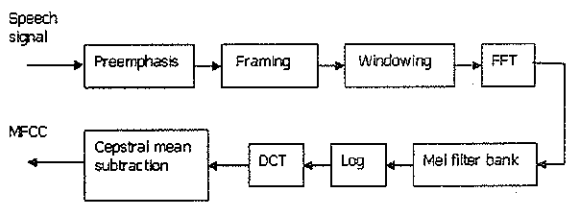


**Figure 2.1 : Extraction of MFCC from Speech Signal**

**Preemphasis**

The digitized speech signal *s(n)* is put through a low order digital system (typically a first–order FIR filter), to spectrally flatten the signal and to make it less susceptible to finite precision effects later in the signal processing. The output of the preemphasis network, $\hat{s}(n)$ is related to the input *s(n)*, by the difference equation

$$\hat{s}(n) = s(n) - \alpha s(n-1)$$

The most common value for $\alpha$ is around 0.95,

**Frame Blocking**

Speech analysis usually assumes that the signal properties change relatively slowly with time. This allows examination of a short time window of speech to extract parameters presumed to remain fixed for the duration of the window. Thus to model dynamic parameters, the signal must be divided into successive windows or analysis frames, so that the parameters can be calculated often enough to follow the relevant changes. In this step the preemphasized speech signal, $\hat{s}(n)$ is blocked into frames of $N$ samples, with adjacent frames being separated by $M$ samples. If we denote the $l^{th}$ frame speech by $x_l(n)$, and there are $L$ frames within the entire speech signal, then

$$x_l(n) = \hat{s}(Ml + n), n = 0,..., N-1, l = 0,..., L-1$$

**Windowing**

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of the frame. The window must be selected to taper the signal to zero at the beginning and end of each frame. If we define the window as $w(n)$, $0 \le n \le N-1$ then the result of windowing the signal is

$$\overline{x}_l(n) = x_l(n)\, w(n), 0 \le n \le N-1$$

The Hamming window is used for this work, which has the form

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1$$

**Computing Spectral Coefficients**

The spectral coefficients of the windowed frames are computed using Fast Fourier Transform, as follows:

$$X(k) = \sum_{n=0}^{N-1} \overline{x}_l(n) \exp^{-jk(2\pi/N)n}, \quad 0 \le n \le N-1$$

**Computing mel Spectral Coefficients**

The spectral coefficients of each frame are then weighted by a series of filter frequency response whose center frequencies and bandwidths roughly match those of the auditory critical band filters. These filters follow the mel scale whereby band edges and center frequencies of the filters are linear for low frequency and logarithmically increase with increasing frequency as shown in Fig. 2.2. These are called as mel-scale filters and collectively a mel-scale filter bank [11]. As can be seen, the filters used are triangular and they are equally spaced along the mel scale which is defined by

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

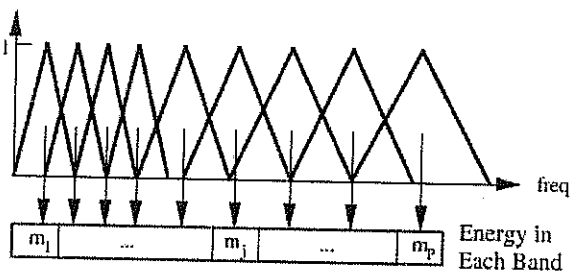


**Figure 2.2 : Mel- Scale Filters**

Each short term Fourier transform (STFT) magnitude coefficient is multiplied by the corresponding filter gain and the results are accumulated.

**Computing MFCC**

The discrete cosine transform (DCT) is applied to the log of the mel spectral coefficients to obtain the MFCC as follows:

$$x(m) = \sqrt{\frac{2}{M}} \sum_{i=0}^{M-1} E(i) \cos\left(\frac{(2x+1)m\pi}{2N}\right), m = 1,..., M$$

Where $M$ is the number of filters in the filter bank, finally, cepstral mean subtraction is performed to reduce the channel effects.

### 2.2 Shifted Delta Cepstrum (SDC)

The shifted delta cepstral features have been introduced to improve the LID performance with respect to the classical cepstral and delta cepstral features [12].

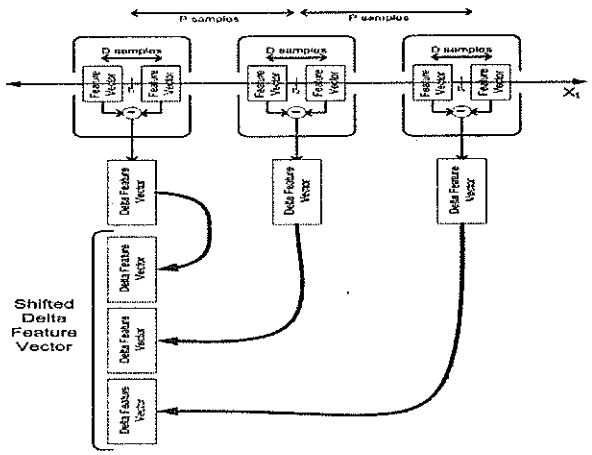


**Figure 2.3 : Calculation of the Shifted Delta Feature Vectors**

The SDC coefficients are computed, for a cepstral frame at time $t$, according to:

$$\Delta c_n(t,i) = c_n(t+iP+D) - c_n(t+ip-D),$$
$$n = 0,\ldots N-1, i = 0,\ldots, k-1$$

Where $n$ is the $n^{th}$ cepstral coefficients, $D$ is the lag of the deltas, $P$ is the distance between successive delta computations, and $i$ is the SDC block number. The final feature vector is obtained by concatenation of $k$ blocks of $N$ parameters.

The computation of the Shifted Delta feature vectors is a relatively simple procedure. The process is as follows: The MFCC feature vectors are first computed as described above. Then, the acoustic feature vectors spaced $D$ sample frames apart are first differenced. Then $k$ differenced feature vector frames, spaced $P$ frames apart, are then stacked to form a new feature vector. Fig.2.3 gives a graphical depiction of this process.

### 3. Hidden Markov Model

Hidden Markov model [11], [13] is used in the problem of making a sequence of decisions on temporal basis. It is a statistical model and a variant of finite state machine. In Markov model the states are directly accessible to the observer. But in HMM the states are not directly accessible to the observer only the variables influenced by the states are accessible to the observer.

### 3.1 Notations used in HMM

$w \rightarrow$ Hidden State

$v \rightarrow$ Visible state

$a_{ij} \rightarrow$ Transition probability to make transition from $i^{th}$ state

at $t$ to $j^{th}$ state at $(t+1)$

$b_{jk} \rightarrow$ Emission probability to emit $k^{th}$ visible state at $j^{th}$ hidden state.

$N \rightarrow$ Number of hidden states (Guess this number)

$M \rightarrow$ Number of visible states (obtained from the training set)

### 3.2 Design Issues

The HMM will be useful in real world applications, if three basic problems of HMM are solved. These problems are the following.

1. Learning problem.
2. Evaluation problem.
3. Decoding problem

**Learning Problem**

Given the values of $N$, $M$

- The goal of learning is to determine model parameters-$\{a_{ij}, b_{jk}\}$ from the training samples.
- Forward – backward algorithm, also known as Baum Welch algorithm is used for learning problem.

- Forward algorithm will generate $\alpha$ values. By using backward algorithm we should find $\beta$ value.

Let the model is in state $w_i(t)$ by generating part of the given visible sequence, $\alpha$ is nothing but theprobability taken so far to come to the current state from the initial state. We express $\alpha_j(t)$ as

$$\alpha_j(t) = \begin{cases} 0 & ; t = 0 \text{ and } j \neq \text{initial state} \\ 1 & ; t = 0 \text{ and } j = \text{initial state} \\ \left[\sum_i \alpha_i(t-1)a_{ij}\right] b_{jk}\, v(t); \text{Otherwise} \end{cases}$$

- $\beta$ is the probability of the model to generate the remainder of the target sequence. We express $\beta_i(t)$ as

$$\beta_i(t) = \begin{cases} 0 & ; w_i(t) \neq 1 \ \textit{for final state and } t = T \\ 1 & ; w_i(t) = 1 \ \textit{for final state and } t = T \\ \sum_j \beta_j(t+1)\, a_{ij}\, b_{jk}\, v(t+1) & ; \textit{Otherwise} \end{cases}$$

Evaluation process started initially by randomly selecting the value of $a_{ij}$ and $b_{jk}$ (such that the summation of each row of $a_{ij}$ and $b_{jk}$ is equal to 1). Then re- estimation of $a_{ij}$ and $b_{jk}$ will be done to achieve the true values of $a_{ij}$ and $b_{jk}$. For the same training data again $P\left(\frac{V^T}{\theta}\right)$ is calculated by using re estimated $\{a_{ij}\}$, $\{b_{jk}\}$. This re-estimation for the same training data will be done repeatedly until the value of $\{a_{ij}\}$ and $\{b_{jk}\}$ is constant for subsequent iterations or negligible change in the estimated values of the parameters on subsequent iterations. Now the values of $\{a_{ij}\}$, $\{b_{jk}\}$ are the true values. So it can be applied to test data.

**Evaluation Problem**

The goal is to find the probability to generate a particular sequence of visible state $V^T$ by the model when the HMM parameters $\theta$ is given. $\theta = \{a_{ij}, b_{jk}\}$. The probability of each possible sequence of hidden states to produce $V^T$ is calculated and then the probabilities are added up. So

$$P\left(\frac{V^T}{\theta}\right) = \sum_{r=1}^{N^T} P\left(\frac{V^T}{W_r^T}\right) P\left(W_r^T\right)$$

But this type of calculation is much complex. It will take $O(N^T T)$ calculation. A computationally simpler recursive algorithm for the same goal is the forward algorithm.

**Decoding Problem**

The decoding problem is to find the most probable sequence of hidden states for the given sequence of visible states $V^T$. For decoding Viterbi algorithm is used. The decoding algorithm finds at each time step $t$, the state that has the highest probability $(\alpha_j(t))$. The full path is the sequence of hidden states to generate the given visible state sequence optimally.

## 4. Language Identification System

The acoustic systems are an interesting compromise between complexity and performance. We have implemented a simple acoustic system for Indian languages using MFCC, SDC coefficients and Hidden Markov model. An acoustic language identification system based on Hidden Markov Model (HMM) works in two phases, a learning procedure to create the models, and a testing procedure.

To identify N number of languages, $N$ numbers of HMMs are to be modeled because each language is to be modeled by a distinct HMM. For each language, a training set of $K$ speech segments spoken by many talkers. For each language many observation sequences $(V^T)$ will be there. Each HMM will be trained by using the observation sequences of the corresponding language by doing

learning process. The training will be stopped after obtaining optimal values for $\{a_{ij}\}$ and $\{b_{jk}\}$. Likewise all $N$ HMMs will be trained.

During testing phase, each unknown speech segment, the language of which is to be identified is applied to the system. The observation sequences $(V^T)$ are obtained and applied to all HMMs (from $HMM_0$ to $HMM_N$). Each HMM will compute the probability $P\left(\frac{V^T}{\theta}\right)$ for the particular observation sequence $(V^T)$ by using evaluation process. From all $P\left(\frac{V^T}{\theta}\right)$ values the maximum value will be selected by using Viterbi algorithm. This is the unknown language, i.e. for the particular language $P\left(\frac{V^T}{\theta}\right)$ generated by the corresponding HMM will be greater than other HMMs.

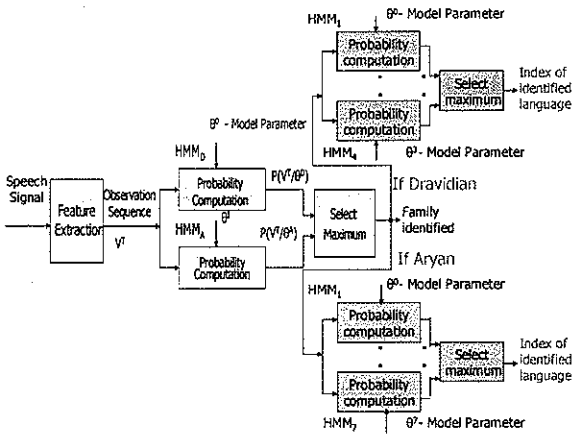


**Figure 4.1 : Overview Of The Proposed Acoustic Language Identification System**

The proposed system for identifying Indian languages is a two level system as shown in Fig. 4.1. In the first level, it will identify whether the language belongs to Dravidian family or Aryan family. Then in the second level it will identify the corresponding language.

In Dravidian family, all languages namely Tamil, Telugu, Kannada and Malayalam are considered in this system. Even though many languages are in Aryan family, the languages spoken by large number of peoples are considered in this system. The languages spoken by less than 2% of the people of the country are not included.

## 5. Experiments And Results

### 5.1 The Database

All the experiments described in this paper were conducted on our own database. It comprises broadcast news shows in 11 languages. In Dravidian family, all four languages namely Tamil (Ta), Telugu (Te), Kannada (Ka) and Malayalam (Ma) languages are used. In Aryan family the major languages namely Hindi (Hi), Bengali (Be), Marathi (Mar), Gujarati (Gu), Oriya (Or), Kashmiri (Kas) and Punjabi (Pu) are selected. This database contains a total of 10h of broadcasts from Doordharsan television network because the network is available in all regional languages of India.

Train and test sets have been defined for each language. For each language, 30 speakers are selected as the training set, and the duration of each speaker is about 60 seconds. The testing set consists of 10 speakers and the duration of each speaker is 10 seconds.

### 5.2. Feature Extraction

The selected properties for the speech signals are a sampling rate of 8 kHz, 16 bit monophonic PCM format. We used a frame rate of 125frames/s, where each frame is 16ms in duration with an overlap of 50% between adjacent frames. All the training and test data are pre-processed to remove silence from the speech signals.

The feature vectors used consist of 13 Mel frequency cepstral coefficients (MFCC). Finally, the delta and acceleration coefficients are appended to the features. So for each frame, a 39 dimensional feature vector is calculated.

The configuration 7-1-3-7 for N-D-P-k has been used to extract SDC feature vector. For each frame, with 7 direct MFCC coefficients 49 SDC coefficients are appended, so totally 56 coefficients are used.

### 5.3 Hidden Markov Model Classifier

The first level of the system uses two HMMs and the second level uses four HMMs for Dravidian family and seven HMMs for Aryan family. All HMMs are initialized with five states and two Gaussian mixtures/state.

### 5.4 Results

Investigations were conducted to compare the performance of the system with MFCC with delta and acceleration coefficients and SDC individually. The performance of the system is given in Table 1. The results of experiments indicate that the proposed system is able to help in distinguishing between languages with greater accuracy. The average performance is affected by the poor performance for the languages Kannada and Oriya.

### 5.5 Discussion

The major challenge in Indian languages is the similar characteristics of the languages. So it is very difficult to distinguish one from the other and it is a challenging task to design a language identification system for these languages. In this system, we used continuous speech for both training and testing. The purpose of hierarchical system is to reduce the complexity of the system. Once the family is identified then it is enough to compare the test utterance within the languages of the family. But the drawback in multi stage system is each stage of the multi stage model exploits results from the previous stage, errors introduced by a stage certainly affects the accuracy of next stage. In this system also the second level results are affected by the first stage. Here we selected the features and all the parameters based on the best features and best parameter values used in the existing LID systems.

## 6. Conclusion

In this work a novel two level language identification system is proposed for Indian languages using acoustic features. The acoustic systems are an interesting compromise between complexity and performance. Investigations were conducted to compare the performance of the system with MFCC with delta and acceleration coefficients and SDC individually. The proposed system has been designed to identify 11 major Indian languages. We created a new database to investigate the performance of this system. The system with SDC performs better than the system with MFCC features.

**Table 1 : Language wise Performance in %**

| Performance for 10s test utterances | | |
|---|---|---|
| Language | MFCC | SDC |
| Ta | 70 | 75 |
| Te | 55 | 65 |
| Ma | 60 | 75 |
| Ka | 35 | 60 |
| Gu | 60 | 70 |
| Mar | 60 | 65 |
| Pu | 85 | 80 |
| Hi | 60 | 65 |
| Be | 76 | 88 |
| Kas | 85 | 80 |
| Or | 40 | 60 |
| Average | 62.36 | 71.2 |

In future the research will be conducted in the direction to improve the performance of the system. This can be achieved by combining the prosodic features with the acoustic features, by using other modeling techniques and by improving the training and testing data sets. As the Indian languages are similar in characteristics, designing a less complex system with the best performance is a challenging task. This work is the first step in this direction.

## References


[1] T. Nagarajan and H. A. Murthy, *"Language identification using acoustic log-likelihoods of syllable-like units"*, Speech communication, Vol. 48, 913-926, 2001.

[2] Muthusamy et al., *"Reviewing automatic language identification"*, IEEE Signal process, Mag., 33-41, 1994.

[3] A. Waibel et al., *"Multilinguality in speech and spoken language systems"*, Proc. IEEE, Vol. 88(8), 1181-1190, 2000.

[4] B. Ma et al., *"Multilingual speech recognition with language identification"*, Proc. ICSLP, 505-508, 2002.

[5] P. Dai et al., *"A novel feature combination approach for spoken document classification with support vector machine"*, Proc. Multimedia information retrieval workshop, 1-5, 2003.

[6] Jean-Luc Rouas et al., *"Language and variety verification on broadcast news for Portuguese"*, Speech communication, Vol. 50, 965-979, 2008.

[7] Rong Tong et al., *"Integrating acoustic, prosodic and phonotactic features for spoken language identification"*, Proc. ICASSP, 205-208, 2006.

[8] www.ciil-spokencorpus.net.

[9] F. Allen et al., *"Language identification using warping and the shifted delta cepstrum"*, Presented at IEEE MMSP'05, 2005.

[10] Davis. S.B and Mermelstein. P , *"Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences"*, IEEE Trans. Acoust. Speech Signal Process, Vol. 28, 357–366, 1980.

[11] S. Young et al., *"The HTK Book (for HTK version 3.2.1)"*, Cambridge University Engineering Department, 2002.

[12] Bo Yin et al., *"Combining Cepstral and Prosodic features in language identification"*, IEEE, ICPR'06, 2006.

[13] L. R. Rabiner, *"A tutorial on hidden Markov models and selected applications in speech recognition"*, Proc. IEEE, Vol. 77(2), 257-286, 1989.


### *Author's Biography*

S. Jothilakshmi received the B.E degree in Electronics and Communication Engineering from Govt. College of Engineering, Salem in 1994. She received the M.E. degree in Computer Science and Engineering from Annamalai University in the year 2005. She has been with Annamalai University, since 1999. She is pursuing her Ph.D degree in Computer Science and Engineering at Annamalai University. She published 5 papers in international conferences and journals. Her research interest includes speech processing, image and video processing, and pattern classification.

Dr. V. Ramalingam received the M.Sc degree in Statistics from Annamalai University in 1980. He received the M.Tech degree in Computer Applications from Indian Institute of Technology Delhi in the year 1995. He has been with Annamalai University, since 1982. He completed his Ph.D degree in Computer Science and Engineering at Annamalai University in 2006. He published 27 papers in international conferences and journals. His research interest includes image and video processing, natural language processing and neural networks.

Dr. S. Palanivel received the B.E (Hons) degree in Computer Science and Engineering from Mookambigai College of Engineering in 1989. He received the M.E degree in Computer Science and Engineering from Government College of Technology in the year 1994. He has been with Annamalai University, since 1994. He completed his Ph.D degree in Computer Science and Engineering at Indian Institute of Technology Madras under quality improvement programme (QIP) sponsored by Annamalai University, in the year 2005. He published 27 papers in international conferences and journals. His research interest includes speech processing, image and video processing, pattern classification and neural networks