Pak. J. Biotechnol. Vol 14 (1) 77-81 (2017)
www.pjbt.org
ISSN Print: 1812-1837
ISSN Online: 2312-7791

A HIGH-PERFORMANCE HTK BASED LANGUAGE IDENTIFICATION SYSTEM FOR VARIOUS INDIAN CLASSICAL LANGUAGES

[a]Karthick, M., T. Muruganandam[b], C.Jeyalakshmi[c] and A.Revathi[d]

[a,b,c]Department of ECE, K.Ramakrishnan college of Engineering, Tamilnadu, India, [d]School of EEE, Sastra university, Tamilnadu, India.

Article received 2.2.2017; Revised 28.2.2017; Accepted 6.3.2017

ABSTRACT

Language identification is one of the major research areas in the field of speech processing and tremendous works has been done on that. One of the great bottleneck of language identification system is, for languages which are having closely related pronunciation, it is very difficult to classify them. In our experiments, we have considered seven Indian classical languages and Mel frequency cepstrum with their delta cepstral feature are utilized as features and HMM is used as a classifier. Performance of the system is analysed using HTK based continuous density HMM with MFCC features. The state of the art HTK with MFCC features produced 98.1% with 3 Gaussian mixtures and 100% accuracy with 10 Gaussian mixtures even with small amount of training samples. The same method can be utilized for other languages also.

Key words: Mel frequency cepstral coefficients (MFCC), Language identification (LID), Indian languages, Hidden Markov Model tool kit (HTK), Recognition accuracy, Continuous density HMM(CDHMM)

INTRODUCTION

Automatic language identification is the method of identifying the spoken language from the set of speech utterances. It finds application in call routing systems, multilingual dialog systems and identifying proper signal routing path in multilingual communication systems etc. It has also importance in areas of intelligence and security. A native Language Identification method is much useful in situations where we have only the simple written text paper. From this we can identify the person who writes that letter for investigation. Frequently people find convenient to converse in their native language with the system. This makes necessity to introduce a system to recognize the spoken utterances in any language so that, effective communication can be made between people and computer. Since the operators are not good verse in all the languages, they need an interpreter which takes a longer time for call routing.

The proposed work is on developing language identification system for Indian languages. By considering the Indian languages, since they are from the same origin, mostly all the languages have the same set of phonemes. In the present work, the aim is to suggest proper value of mixture values when considering MFCC features and HMM models for Indian languages to achieve good accuracy. In General, two levels are used i.e in the first level, the main origin of the spoken language is identified, then from that the language is found. But in our system, we have not utilized that method instead, directly created the models for the input languages and matching any input language with this models for identification. The recognition method in the literature for LID are hidden Markov model (HMM), Gaussian mixture models (GMM), artificial neural networks (ANN) and support vector machines. The modeling used in the proposed system uses HMM for Indian languages.

Thus, a LID can quickly find the language from the speech data base which reduces the time required for an interpreter. To assess the performance of LID system four different methods has been adopted like GMM and various phone models (Mark, 1995). Review about explicit and implicit LID system has been presented (Rao and Nandi, 2015). For six Indic languages, text independent LID system was developed (Sadanandam, et al., 2012a). LID system using parallel phone recognition language modeling is discussed (Suo, et al., 2008). Language modeling PRLM and GMM is utilized for classification (Santhi and Rajasekar, 2013). Language models using MFCC features with Discrete HMM are developed (Sadanandam, et al., 2012b). Using the same method but for six Indic languages has been discussed Sadanandam, et al., 2012c). The authors made use of ASM frame work for text independent LID systems (Li, et al., 2007). The cluster based computation is used to get new feature from MFCC and GMM is used for classification (Sadanandam, et al., 2014). The development of four approaches for LID system has been proposed (Mark, 1996). The use of SVM for identifying languages has been described (Jang, et al., 2006; Lee, et al., 2008; Yan and Liu, 2008; Ziaei, et al., 2008 and Verma and Khanna, 2013). Nicolai et al., 2013 discusses the development of probabilistic graphical models for identifying languages and five languages are considered. Vyas and Dutta, 2014) describes the LID system for two languages English and French. Most LID systems comprise of two main phases namely training and recognition. During training, feature vectors depicting the characteristics of the speeches corresponding to the languages are extracted and training models are created. During recognition phase, feature vectors are extracted for test speeches and applied to the models and based on the specific parameter corresponding to the modeling technique, classification is done.

Utility of Speech Database: Performance of any speech recognition system depends on the speech database. While Hindi is the official language in India, there are 17 Indian languages other than Hindi such as Tamil, Malayalam, Guajarati and Telugu. In the proposed work, IIIT-H Indic speech databases were utilized. It consists of text and speech data in seven Indian languages such as, Bengali, Hindi, Kannada, Malayalam, Marathi, Tamil and Telugu. Database is obtained from the native speakers of the language with minimum background environment. If there are any mistakes during recording it is re-recorded. To cover all possible dialectic variations of the language, speakers from different states and regions are considered. From each language 1000 sentences were considered which covers 5000 most frequently used words in the corresponding language.

Characteristics of input speech in Indian Languages: To show the variations in the characteristics between various languages spectrogram is shown for all the seven Indian languages. Using the spectral representation also, we can characterize a signal to show the information associated with it. For this purpose, sound spectrogram in three dimensional views is considered by clearly indicating all frequency bands. Figure 1 shows the spectrogram of speech signal for the four Indian languages Bengali, Hindi, Malayalam and Kannada. Figure 2 shows the spectrogram of speech signal for the three Indian languages Marathi, Tamil, Telungu. The spectral intensity at each point in time is indicated by the intensity (darkness) of the plot at a frequency.

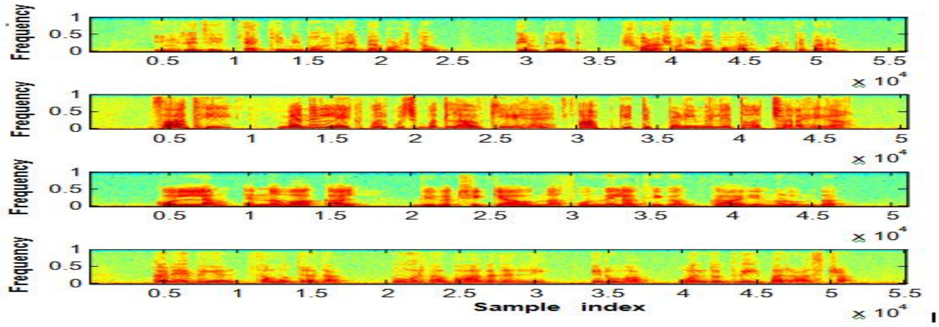

Figure 1: Spectrogram of speech in four Indian languages Bengali, Hindi, Malayalam and Kannada

In general, the pitch of the voiced regions in the input speech signal is illustrated by the dark regions of the spectrogram with horizontal lines. During periods of unvoiced speech, primarily high-frequency energy in the spectrograms can be seen and during periods of silence, there is no spectral activity. Similar to time domain representation of the speech signal spectrogram plots are also completely different for the different languages due to variation in the speaking rate, Pronunciation, abstraction, directness etc.

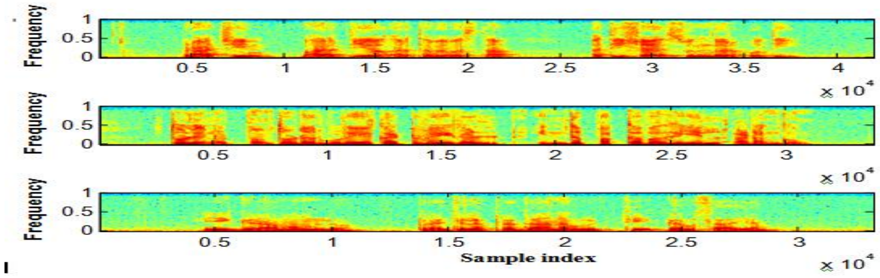

Figure 2: Spectrogram of speech in four Indian languages Marathi, Tamil and Telungu

LID system with MFCC features and HMM models: The most frequently used speech features are MFCC, for speech recognition, speaker recognition and language identification (Murty and Yegnanarayana, 2006). LID system is created using MFCC acoustic feature and HTK to get the training models.

Feature extraction: The front end spectral analysis of any speech recognition system is the feature extr-action and they should exhibit statistics which are highly invariant across speakers and speaking environment. Since from most of the research it is understood that, MFCC is the widely-used features for speech recognition, it has been utilized in this present work also. Filters are spaced linearly at low frequencies and logarithmically at high frequencies when extracting MFCC to capture the phonetically important characteristics of speech. The figure 1 shows the process of computing MFCC.

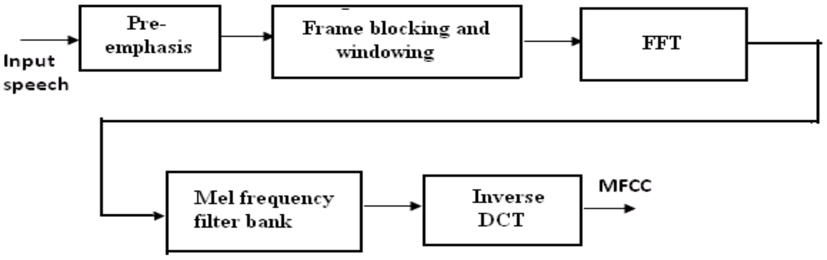

Figure -3: Block diagram of MFCC feature extraction method

The first order FIR filter is used as a pre-emphasis filter for the input speech. Then it is converted into several frames. The 20 msec. is taken as a frame length and 10 msec. overlap between the adjacent frames. This process corresponds to 320 samples per frame (N) with adjacent frame being separated by 160 samples (M). If M≤ N, then adjacent frames overlap and the resulting spectral estimates will be correlated from frame to frame; If M is very much less than N, then the spectral estimates from frame to frame will be quite smooth (Rabiner and Juang, 1993). Then Hamming window is used to reduce the signal discontinuities and FFT is applied for each frame. Forty band pass filters are used to filter the spectrum of each frame and then power of each band is calculated. For the given frequency ‘F’, mel frequency is calculated using equation (1) for conversion of linear frequency into mel frequency (Umesh, et al., 1997).

$$mel(f) = 2595 * \log(1 + f(Hz)/700) \quad (1)$$

Mel-Frequency cepstrum is calculated from the output power of the filter bank using Equation (2).

$$C_n = \sum_{k=1}^{K} (\log S_k) \cos[n(k-0.5)\pi / K] \quad (2) \quad n=1, 2 \ldots L$$

where S_k is the output power of the kth filter of the filter bank and L is the desired length of the cepstrum.

4.2 Language Model creation based on HMM: Many researchers for so many years utilized the very powerful algorithm Hidden Markov Models for speech recognition for calculating the model parameters and it is one of the statistical models (Rabiner, 1989). There are two types of HMM, model one is ergodic model and another is left to right or Bakis model. In ergodic model, every state of the model could be reached from every other state of the model. It has the property that every state can be reached from every other state in a finite but aperiodic number of steps. Similarly, Bakis model has the property that, as time increases, the state index increases or stays the same – that is, the system states proceed from left to right. It has the desirable property that it can readily model signals whose properties change over time in a successive manner e.g. speech. In the present work, left to right HMM is utilized.

The issue in implementing HMMs is the choice of model type, choice of model size and choice of observation symbols. According to the choice of observation symbols, HMM can be discrete or continuous type and these choices must be made depending on the signal being modeled. Speech signal is first converted into discrete sequence of feature vectors which is assumed to contain information about given utterance that is important for its correct recognition. Feature extraction is performed to reduce dimensionality of original speech signal and to preprocess that signal into a form suitable for classification stage.

During training phase, training speeches for each language are considered separately to develop language models for all the languages. Initially, the input speeches are preprocessed and MFCC features are extracted. These features are used to develop the language models and after training, model parameters are stored for testing. During testing phase, MFCC features are extracted from the test speech and test feature vectors applied to all HMMs and log likelihood values are compared and the model with maximum log likelihood value is taken as the recognized word.

Language models using Continuous density HMM by HTK: In HTK HCompV, HInit and HRest are used for estimating the system parameters. From the input speech training data (Steve, et al., 2001), overall mean and variance is calculated and this is set equal to the mean and variance of every Gaussian component in a HMM definition by HComp V. Since HTK uses conti-nuous density models, mixture Gaussian density is the observation probability distribution of the input data. Hence, to define a HMM, number of states, type of observation vector, number and width of the data stream and mixture weights for each emitting state are to be defined. Using the model definition, proto type HMM models are developed.

In this work, prototype models are initialized with number of mixtures and number of states and models are trained. The model parameters such as mean vector and covariance matrix are re-estimated using Baum Welch re-estimation algorithm. Now the training models are developed for all the languages. During the recognition phase, when unknown sentence in a language is given, MFCC features are extracted from them and they are compared with the HMM models developed for all the language by considering the dictionary, word network and task grammar, the language is identified.

RESULTS AND DISCUSSION

For each language, a total of 60 utterances are considered for training and 440 utterances are considered for testing. In the testing phase, the test speeches are converted into series of acoustical vectors, i.e. feature vectors. These vectors are compared with the models already developed for each language using **HVite** tool. It compares the speech file against the HMM networks and produces a transcription for it.

Experimental analysis based on HTK: The present LID system is evaluated using the data containing utterances in six Indic languages. Training models are developed using 60 utterances spoken in particular language from all the speakers. For the resultant signal Hamming window is applied on frames of 20msecs duration and 10msecs of overlapping. Then, MFCC features are extracted for each speech frame in a particular language. Initially, proto type HMM models are created for different states and mixtures. The number of states roughly corresponds to number of phonemes in a word. In our work, number of mixture is initially taken as 3 and number of states is taken as 5. The proto type models are generated for 5s-3m (5 states and 3 mixture). This is clearly explained with an example. For example, the Tamil word *ondru* has 4 phonemes (sounds oh, in, ir, uw) in which the emitting states are 4 with 3 component mixture Gaussians. Hence, the size of the transition matrix is chosen as 4 x 4. In our work, it is not possible to initialize the no. of states corresponds to no. of phonemes, since lengthy sentences are considered. It is initially chosen as 5 and the mixture value is taken as 3. Then, for the input utterances in all the languages HMM parameters are re estimated using Baum Welch algorithm. During testing phase, 440 utterances of each language is considered and MFCC features are extracted for the test speech. Out of the 3080 utterances, 3022 are correctly recognized and the overall recognition accuracy obtained is 98.1%.

We can express the recognition result using the confusion matrix. It is a table form of matrix for comparing the wrongly recognized words with the original words while the row shows the actual word and the column shows the recognized word. Table 1 shows the confusion matrix for the highest accuracy 98.1 for MFCC with 3 mixtures. In this matrix, each row represents actual spoken language and each column represents, no. of times the languages recognised by the system for 440 test utterances of each language.

The total no. of misidentified is 58 out of 3080 and of these 58 utterances, six are due to confusion by the language Malayalam with Bengali and one is due to the confusion between Malayalam and Marathi and 51 are due to confusion between Marathi and Bengali. By carefully viewing the above matrix, it is concluded that no language has been misidentified as Hindi, kannada, Tamil and Telungu. Maximum Marathi is misidentified as Bengali because their features are slightly similar. Even though the system produces higher recognition rate, in order to maximize the recognition accuracy, the no. of mixtures are increased from 3 to 10 and the states are randomly chosen between 2 to 13. Now the prototype models are developed using these new parameters and then it is re-estimated using Baum Welch algorithm. To validate the recognition accuracy, the same 440 utterances of each language are considered.

Table 1 Confusion matrix corresponding to HMM of 3 mixtures with 5 states model

	Bengali	Hindi	Kannada	Malayalam	Marathi	Tamil	Telungu
Bengali	440						
Hindi		440					
Kannada			440				
Malayalam	6			433	1		
Marathi	51				389		
Tamil						440	
Telungu							440

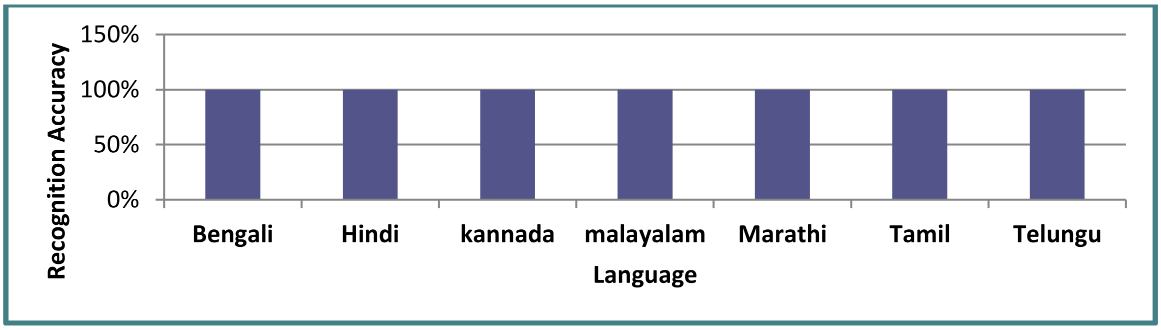

Fig.-4: Performance of the system for 10 mixtures with maximum of 13 states

These 3080 utterances are correctly recognized and overall recognition accuracy obtained is 100% for mixture value 10. It is illustrated in figure 4. Since the no. of mixtures corresponds to the variations in the speech input, when it is increased from 3 to 10 the maximum recognition accuracy is achieved.

CONCLUSIONS: The performance of the LID system is analyzed to recognize seven Indian languages using Text and speaker independent input speech utterances. The performances of LID system using the basic features MFCC and HMM for the development of training models are evaluated for different mixtures and states. These features and modeling techniques are evaluated by testing 440 test utterances for each language. For a mixture value of 3 and fixed state value of 5, MFCC with its delta, acceleration coefficients and HMM models using HTK has provided 98.1%. Whereas when the mixture value is increased from 3 to 10 with a maximum state value of 13, 100% RA has been achieved. The present system utilized the database created by IIIT which produced good accuracy. We can also develop the same system for other Indian languages using different data base.

REFERENCES

Jang, W., B. Li, D. Qu and B. Wang, Automatic language identification using support vector machines, Proceedings of 8th International Conference on signal processing (2006)

Lee, K.-A., C. You and H. Li, Spoken language identification using support vector machine with generative front end, Acoustics, IEEE International Conference on Acoustics, Speech and Signal Processing Pp. 4153 – 4156 (2008).

Li, H., B. Ma and C. Lee, A vector space modelling approach to spoken language identification. IEEE Transactions on Audio, Speech and Language Processing 15(8): 271-284 (2007).

Mark,A.Z., Automatic identification of telephone speech. The Lincolns laboratory Journal 8(2): 115-143 (1995).

Mark, A.Z., Comparison of four approaches to automatic language identification of telephone speech. IEEE Transactions on Speech and Audio Processing 4(1): 31-44 (1996).

Murty, K.S.R. and B. Yegnanarayana, combining evidence from residual phase and MFCC features for speaker recognition. IEEE Signal Processing Letters 13(1): 52–55 (2006).

Nicolai,G., M.A.Islam and R.Greiner, Native language identification using probabilistic graphical models, International Conference on Electrical Information and Communication Technology (EICT) Pp.1-6 (2013).

Rabiner, L.R. and B.H. Juang, Fundamentals of Speech Recognition, Prentice Hall, New Jersey (1993).

Rabiner, L.R., A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, Proce-edings of the IEEE 77(2): 257-286 (1989).

Rao, K.S. and D. Nandi, Language identification – a brief review. Springer briefs in Speech Technology Pp. 11-30 (2015)

Sadanandam,M., V.Kamakshiprasad and V.Janaki, Automatic language identification using new features and their weightage. Int. Journal of Advanced Computing 35(7): 380-385 (2012a).

Sadanandam, M., V.Kamakshiprasad and V.Janaki, DHMM based automatic language identification system. Int. Journal of Information Technology and Knowledge Management 6(1): 93-97 (2012b).

Sadanandam, M., V. Kamakshiprasad and V. Janaki, Text independent language identification using DHMM. IJCA 48(7): 42-45 (2012c).

Sadanandam, M., V. Kamakshiprasad and V. Janaki, GMM based language identification system using robust features. Int. Journal of Speech Technology 17: 99-105 (2014).

Santhi and Rajasekar, An automatic language identification using audio features. IJETAE, Special issues, January (2013).

Steve, Y., G. Evermann, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev and P. Woodland, The HTK Book, Cambridge University Engineering department (2001).

Suo, H., M. Li, P. Lu and Y. Yan, Automatic language identification with discriminative language characterization based on SVM. IEICE Trans. On Information and Systems E 91-D(3): 567-575 (2008).

Umesh, S., L. Cohen and D. Nelson, Fitting the mel scale, Proc. of ICASSP, IEEE 1: 217-220 (1999)

Vyas, G. and M.K. Dutta, An integrated spoken language recognition system using support vector machine, Seventh International Conference on Contemporary Computing (IC3) Pp.105-108 (2014).

Verma, V.K. and N. Khanna, Indian language identifycation using k means clustering and support vector machine, Students Conference on Engineering and Systems (SCES) Pp.1-5 (2013).

Yan, D. and J. Liu, Automatic language identification using support vector machine and phonetic N gram, International Conference on Audio, Language and Image Processing ICALIP Pp.71-74 (2008).

Ziaei, A., S.M. Ahadi, S.M. Mirrezaie and H. Yeganeh, Spoken language identification using a new sequence kernel- based SVM back-end classifier, IEEE International Symposium on Signal Processing and Information Technology, ISSPIT, Pp.324-329 (2008)