Efficient Continuous Speech Recognition Approaches for Dravidian Languages

J. Sangeetha
Department of CSE
Annamalai University
Chidambaram 608 002.
sangita.sudhakar@gmail.com

S. Jothilakshmi
Department of CSE
Annamalai University
Chidambaram 608 002.
jothi.sekar@gmail.com

R. N. Devendrakumar
Department of CSE
Sri Ramakrishna Institute of Technology
Coimbatore 641010.
devendrakumar.cse@srit.org

***Abstract*— Nowadays, Continuous Speech Recognition (CSR) deals with developing a healthy approach, which can handle changeability in environment, utterance, speaker and language. CSR has been created for numerous languages because each language has its own specific features. This research work mainly focuses on developing a recognition system for Dravidian languages such as Tamil, Malayalam, Kannada and Telugu. This system can also be used to provide social security wherein the voices of the native speakers can be utilized to create authentication identification such as used in ATM, home automation etc. The proposed CSR system comprises of three steps namely pre-processing, feature extraction and classification. In the preprocessing step, the input signal is preprocessed in the course of pre emphasis filter, framing, windowing and filtering has been done in order to remove the background noise and to enhance the signal. The output of the pre-processing step is taken as the input for the further process of speech segmentation system. The speech has been segmented automatically using Zero Crossing Rate (ZCR) and Energy profile. Then features of the segmented speech signals are analyzed and extracted via Mel Frequency Cepstral Coefficients (MFCC) and Linear Predictive Cepstral Coefficients (LPCC). These feature vectors are given as the input to the classifiers such as an Auto Associative Neural Network (AANN), Hidden Markov Model (HMM) and Support Vector Machine (SVM) for classifying and recognizing Dravidian languages. Experiments are carried out with real time Dravidian languages speech signal. It is observed from the results that the proposed system gives significant results in which HMM classifier gives 92% accuracy in Malayalam language where as SVM and AANN classifiers provide 90.8% and 89.9% accuracy for Tamil and Kannada languages respectively.**

Keywords- Auto Associative Neural Networks, Automatic Speech Recognition, Mel Frequency Cepstral Coefficients (MFCC), Dravidian languages, Linear Predictive Cepstral Coefficients, Hidden Markov Model, Support Vector machine.

I. INTRODUCTION

Continuous Speech Recognition (CSR) facilitates the system to identify the spoken words and convert them into written text [1]. The main benefit of CSR is its ability to achieve hands-free computing. CSR also offers huge social benefits for people with disabilities who find difficulties in using a keyboard and mouse. Thus, it has become an attractive alternate choice for many users to manage applications through speech rather than a mouse or keyboard. The main applications of the CSR system include voice dialing, call routing, automatic transcriptions, information searching, data entry, speech-to-text processing etc.

The important aspect in CSR is its performance level in noisy environment. Most of the recognition systems achieve reliable performance in noise free environments but under performs in noisy conditions. Developing a highly effective speech recognition system, which achieves greater accuracy in noisy conditions, is a challenging task [2]. There are several techniques available in the literature for improving the efficiency of speech recognition systems. The modeling accuracy is to relax the HMM conditional-independence assumption, and condition the distribution of each examination of the previous studies in addition to the state that generates it [3][4]. This technique is known as conditional Gaussian HMMs or autoregressive HMMs. However, it has been observed that the conditional Gaussian HMMs do not offer significant advantages if the dynamic features are used [5].

There are other forms of approaches, which investigate the utilization of more complex HMM structures, such as multiple-path modeling [6][7]. This formulation consists of multiple parallel paths, each of which may account for the acoustic variability from a particular source. The multiple-path model may over-correct the trajectory-folding problem connected with the GMM-HMM, as the acceptable mixture paths are exponentially minimized. Most of these systems have only been validated on certain simple recognition actions using a small number of parallel paths. However, developing a model that is essentially robust to speaker and environmental alterations is still a challenging issue.

There have been certain prominent advances in discriminative training as studied in [8] such as Maximum Mutual Information (MMI) estimation [9], Minimum Classification Error (MCE) training [10], and Minimum Phone Error (MPE) training. A subspace GMM method [28] has similarities to Joint Factor Analysis as used in speaker recognition and to Eigen voices and Cluster Adaptive Training (CAT) proposed for speech recognition.

Large-margin approaches [11], large-margin MCE approaches [12] and boosted MMI [13] ,as well as novel acoustic models such as Conditional Random Fields (CRFs) [14], hidden CRFs [12], and segmental CRFs [16] has been proposed for speech recognition. Among these approaches, the hidden markov model technique is often considered as

being inaccurate to formulate heterogeneous data sources. The mixture segments that are attained in different acoustic conditions for one sound can be concatenated to match at a high probability with the speech observations from another sound, a problem referred to as trajectory folding [17].

Hence, a number of research works have been developed in the last decade, which mainly focuses on developing speech recognizers for their native languages. Even after decades of research and many successfully deployed approaches, the performances of Speech Recognition (SR) systems in real usage scenarios for Indian languages is still in its earlier stage. South Indian languages such as Tamil, Malayalam, Kannada and Telugu are among the most popular spoken languages worldwide with 77 million speakers. Hence, there is an urgent need for the recognition system to interact with south Indian or Dravidian languages. This research work is mainly applicable for native speakers where the people do not know any other languages other than native languages and it is applied in many real time environments such as railway station, banking & ATM service, weather forecasting, etc.

The present research work mainly focuses on providing the banking facility to the native speakers of Dravidian languages who are not illiterate enough to use the banking services. In such case, CSR systems would be a great help in which without typing or writing anything, the Dravidian language speaker can get their work done easily. This would elevate their standard of living as they could make use of many banking facilitates available.

The following section clearly describes the proposed CSR model, which could be used for the above said applications.

II. METHODOLOGY

There is variety of speech recognition as stated in Hasnain and Azam (2008) approaches, available such as Neural Networks, Hidden Markov Models, Bayesian networks and Dynamic Time Warping etc. Among these approaches Neural Networks (NNs) have proven to be a powerful tool for solving problems of prediction, classification and pattern recognition. The speech recognition system can effectively handle low quality, noisy data and speaker independence applications along with general-purpose speech recognition applications.

The proposed Continuous Speech Recognition (CSR) system comprises of three stages namely pre processing, Segmentation and Classification.

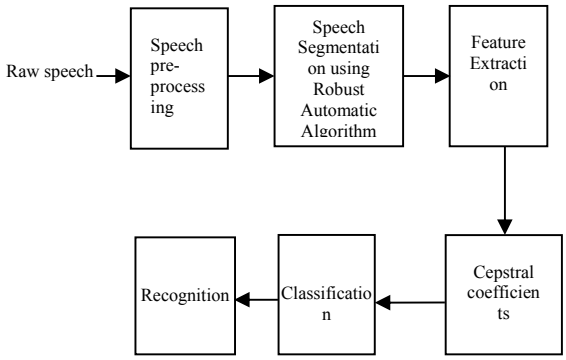

Fig. 1.System Overview of Speech Recognition

A. Preprocessing

- *Digital Speech Acquisition*

Digital speech acquisition is obtaining the analog speech signal through the microphone and getting a digital representation of speech signals. Speech capturing or speech recording is the initial step of implementation. In the proposed algorithm, the sampling frequency is 8 KHz; sample size is 8 bits, and mono channel is used.

- *Signal Pre-processing*

It is very essential to pre-process the speech signal in the applications where silence or background noise is completely objectionable.

- *Band Stop Filter*

A Band-Stop filter discards frequencies that are within a particular range, giving easy passage only to frequencies outside of that range. It is also called as band-elimination, band reject, or notch filters. A low-pass filter placed in parallel with a high-pass filter forms a band-stop filter. The frequencies that a band-stop filter blocks are bounded by a lower cutoff frequency and a higher cutoff frequency. The frequency of maximum attenuation in it is called the notch frequency.

- *Framing*

A speech signal cannot be taken as such for evaluations in most of the processing tools. A speech signal is often partitioned into a number of segments called frames. A continuous speech signal has been blocked into *N*models, with adjacent frames being partitioned by M ($M<N$). In this work, after the Band Stop filtering, the filtered samples have been converted into frames, having a frame size of 25 msec. Each frame overlaps by10 msec.

- *Windowing*

The window w(n), identifies the portion of the speech signal that is to be processed by zeroing out the signal outside the region of interest. Windowing is carried out to minimize the edge effect of each frame segment. Rectangular window has been used in this work.

B. Speech Segmentation

Automatic speech segmentation is an essential process that is used in speech recognition and synthesis systems. Speech segmentation is breaking continuous streams of sound into certain basic units like words, phonemes or syllables that can be identified. The main concept of segmentation is to partition something continuous into discrete, non-overlapping entities. Segmentation can be also utilized to differentiate different types of audio signals from huge amounts of audio data, often considered as audio classification as stated in [18].

Automatic speech segmentation approaches have several categories, but one very common classification is the division to blind and aided segmentation algorithms. The main difference between aided and blind approaches lies in the fact the amount of previously obtained data or external knowledge utilized by the segmentation algorithm for processing the expected speech.

C. Proposed Speech Segmentation Approach

The short time energy and zero crossing rates are used to process speech models to achieve the proposed segmentation approach. This proposed approach utilizes the following process for automatically marking the boundaries in the sound file.

- Short-term energy and zero crossing rates are computed for the preprocessed frames.
- A certain threshold value for short time energy, which is dynamically formed, has been taken and signals having a value less than this threshold value has been altered to zero as signal having syllable will have a data value more than the threshold value.
- Then signal has been checked for value not equal to zero and greater than certain particular value and that point will be marked as starting location of the boundary.
- After getting the starting location, the zero values of signal have been checked and if there are suitable numbers of continuous zeros then it has been defined as the end of the boundary. Once an endpoint has been detected, it is possible to analyze signal from the endpoint of the first one looking for the starting position of next one.

D. Feature Extraction For Speech Recognition

- Mel Frequency Cepstral Coefficients (MFCC)

MFCC has demonstrated to be one of the most important feature denotations in speech related recognition assignments. The mel-cepstrum makes use of auditory principles, as well as the decor relating property of the cepstrum [20]. Fig. 2 illustrates the computation of MFCC features for a segment of speech signal, which is, described as follows :

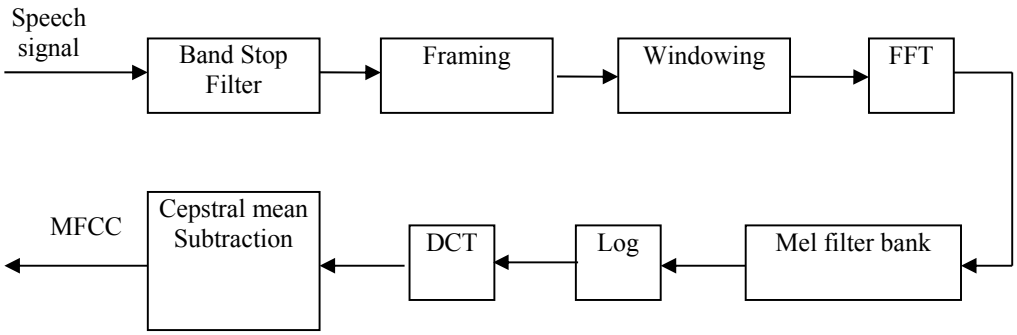

Fig. 2.Extraction of MFCC from speech signal

- The speech waveform is first windowed with analysis window and the discrete short time Fourier transform (STFT) is computed.
- The magnitude is then weighted through a series of filter frequency responses whose centre frequencies and bandwidths approximately match those of the auditory critical band filters. These filters follow the mel scale where by band edges and centre frequencies of the filters are linear for low frequency and logarithmically increase with increasing frequency. These filters are collectively called as Mel-scale filter bank. This filter bank, with 24 triangularly shaped frequency responses, is a rough approximation to actual auditory critical band filters covering a 4000 Hz range as discussed in [22].
- The log energy in the STFT weighted by each mel-scale filter frequency response is computed.

- Finally, Discrete Cosine Transform (DCT) is applied to the filter bank output to produce the Cepstral coefficients.

- Linear Prediction Cepstral Coefficients(LPCC)

A given speech sample at time n, s(n), can be approximated as a linear combination of the past p speech samples, such that

$$s(n) \approx a_1 s(n-1) + a_2 s(n-2) + a_3 s(n-3) + \cdots + a_p s(n-p)$$

where the coefficients are the assumed constants over the analysis frame. The steps for computing LPC is illustrated in Fig. 4. After obtaining the autocorrelation of a windowed frame, the linear prediction coefficients are obtained using Levinson-Durbin recursive algorithm. The cepstrum is a common transform used to gain information from a person's speech signal. It can be used to separate the excitation signal (which contains the words and the pitch) and the transfer function (which contains the voice quality). The cepstrum can be seen as information about the rate of change in the different spectrum bands. The cepstral coefficients are the coefficients of the Fourier transform representation of the logarithm magnitude spectrum. Cepstral coefficients of a sequence x are the coefficients of the inverse discrete Fourier transform (IDFT) of the log magnitude short-time spectrum

$$IDFT\ (log\ (|DFT(x)|))$$

If x is LPC, the cepstral coefficients are known as Linear Prediction Cepstral Coefficients (LPCC). The LPC parameter conversion block in Fig. 3 will convert LPC to LPCC.

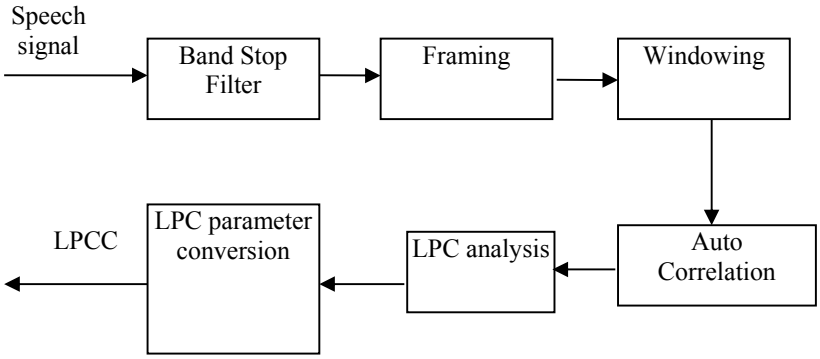

Fig.3. Extraction of LPCC from speech signal

- Classification

In this research work, three classifiers are analyzed for obtaining better classification accuracy. The output of feature extraction is given as an input to the classifiers and the obtained results are compared with each other.

- Hidden Markov Model (HMM)

Hidden Markov model [21] is used in the problem of making a sequence of decisions on a temporal basis. It is a statistical model and a variant of finite state machine. In Markov model the states are directly available to the observer. But in HMM the states are not directly attainable to the observer only the variables influenced by the states are accessible to the observer.

Notations used in HMM

$w \rightarrow$ Hidden State

$v \rightarrow$ Visible state

$a_{ij} \rightarrow$ Transition probability to make the 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

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

HMM will perform better in real world applications, if three basic design issues such as learning problem, evaluation problem and decoding problem are taken into consideration.

- *Auto Associative Neural Network (AANN) Model*

AANN models are belongs to feed forward neural networks accomplishing an identity mapping of the input space and are used to confine the distribution of the input data [19]. Limitation of PCA is to represent an input space using a linear subspace motivated the researchers to investigate a method of projecting the input data onto a nonlinear subspace using AANN models [23]. AANN consists of three types of layers that are input layer, hidden layer and output layer. An AANN is with the preferred output being the same as the input vector. The number of hidden layers and the number of units in each hidden layer depend on the problem. A three-layered AANN representation clusters the input data in the linear

subspace, while a five-layer AANN model captures the nonlinear subspace momentary through the distribution of the input data [24].

Figure 5 shows the five-layer AANN model, which has three hidden layers. The first and third hidden layers are nonlinear, and the second hidden layer can be linear or nonlinear. Since the error within the actual and the preferred output vectors is minimized, the cluster points in the input space concludes the shape of the hyper surface, attain by the projection onto the lower dimensional space. The second and fourth layers of the network have more units than the input layer. The third layer has fewer units than the first or fifth. The activation functions in the second, third and fourth layers are nonlinear. The structure of the AANN model used in our study is 19L 38N 5N 38N 19L, where L denotes a linear unit and N denotes a nonlinear unit. The nonlinear output function for each unit is tanh(s), where s is the activation value of the unit. The standard back propagation learning algorithm[32] is used to regulate the weights of the network to reduce the mean square error for each feature vector. The AANN captures the distribution of the input data depending on the constraints imposed by the structure of the network; just as the number of mixtures and Gaussian, functions do in the case of GMM. The choice of parameters such as feature vectors, initial weights and structure of AANN is not very critical, as variation of these parameters does not affect the performance of the system abruptly [26].

- *Support Vector Machine (SVM)*

The support vector machine is a valued machine learning technique that has been effectively applied in the pattern recognition tasks [27]. If the data are linearly indivisible but nonlinearly separable, the nonlinear support vector classifier will be applied. The fundamental idea is to make over input vectors into a high dimensional feature space by means of nonlinear transformation and then to do a linear separation in feature. Nonlinear support vector classifier employed the optimal separating hyper plane in the feature space with a kernel function K (xi, x_{new}) is given by

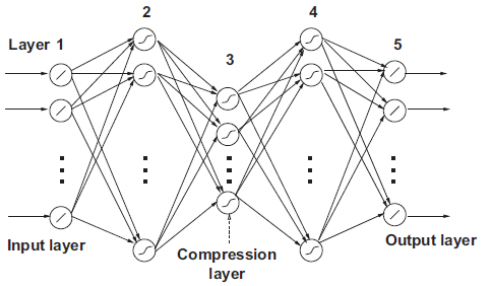

Fig. 4. Five layer AANN model

$$f(X_{new}) = sgn\left(\sum_{i=1}^{SV} \alpha_i y_i K(X_i, X_{new}) + b\right)$$

The SVM consist of two layers, the first layer selects the fundamental K (x_i, x_{new}), i= 1, 2,. . . N, from the given set of bases defined by the kernel; the second layer constructs a linear function in this space during learning phase. This is totally equal to constructing the optimal hyperplane in the corresponding feature space. The SVM algorithm can build a variety of learning machines by use of different kernel functions,

Four kinds of kernel functions are usually used. They are

(1) Linear kernel
$K(x_1, x_2) = (x_1, x_2)$
(2) Polynomial kernel of degree
$K(x_1, x_2) = (\gamma(x_1, x_2) + C_o)^d$
(3) Gaussian radial basis function (RBF)
$K(x_1, x_2) = \exp(-\gamma\|x_1 - x_2\|^2)$
(4) Sigmoidal kernel
$K(x_1, x_2) = \tan h(\gamma(x_1, x_2) + C_0)$
where kernel parameters

- γ: width of RBF coefficient in polynomial.
- d : degree of polynomial.
- c0 : additive constant in polynomial.

III. Ease of Use Experimental Results And Discussion

A. *Database*

In this research work, real time speech database is considered for the banking sector and in ATM centres. For the experimental simulation, around 100 people's voice samples based on the banking sectors were collected and evaluated. The speech samples mainly based on the transaction, account type, branch name, account number, cheque, deposit, credit, date, etc. was collected from the native speaker and stored in the database.

For the ATM centre applications, the voice samples for making transaction in the ATM, balance enquiry, pin change, account type, etc were collected from various native speakers. These samples were collected over a period of time and are used in this experimental simulation.

The dataset is divided into the development corpus and evaluation corpus. The development corpus is used

for training the system and tuning the parameters, which is composed of six 20-minute speech of each language. The evaluation corpus is composed of three 10-minute speech of each language, which is for validation. Different types of speech from native speakers have been used to test the performance of the algorithm.

B. *Parameter tuning phase*

In order to adjust the parameters of this proposed recognition task several experiments have been accompanied on the Dravidian language database, whose results are reported in this subsection. Three classifiers are used in this proposed work. Based on the classifiers, different parameters have to be tuned. For the AANN classifier, the parameters to be tuned are the number of epochs (one epoch of training is a single presentation of all the training vectors to the network). For each epochthe distribution of the MFCC feature vectors and LPCC feature vectors are captured by means of the AANN model. The feature vectors of are given as input to the AANN model, the average confidence score is calculated for different values of epochs such as 100, 200, 500 and 1000 for both MFCC and LPCC features.

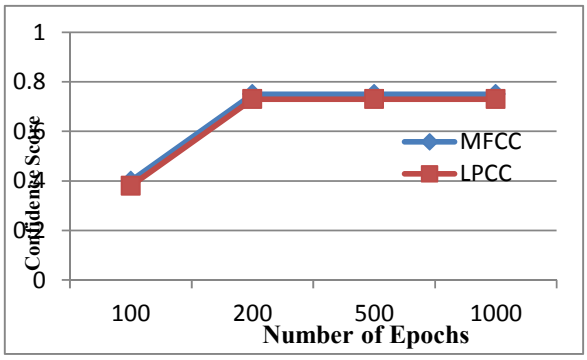

Fig. 8. Confidence Score based on Epochs

Figure 8 clearly shows that the evaluation of confidence score based on epochs when increased from 100, 200, 500 and 1000. After 100 epochs, the confidence score gets saturated. Hence, the AANN models are trained for 200 epochs.

For the SVM classifier, kernels of the SVM classifier have to be tuned. A key feature of the support vector machines is the capability to restore the input data by a non-linear function φ(x) operating on the input data. This can be viewed since mapping the input data to a higher dimensional space, to enable classification of data that is not linearly divisible in the original input space. An equivalent elucidation is that the kernel function is a defined dot product < > replacing in the Hilbert space defined by the mapping φ.In this way, avoid signifying the mapping φ explicitly. In either case, the use of kernel function permits the SVM representation to be self-governing of the dimensionality of the input space. There are different kernel functions such as that the provided SVM as the ability to model complex separating hyperplanes. In this research, RBF kernel is used.

Then for the HMM classifier, mixtures and states are tuned to optimize value for attaining the desired results. The number of states for HMM classifier considered are 5, 7, 9 and 11 and in this, 7 states provide good results when compared with other states.

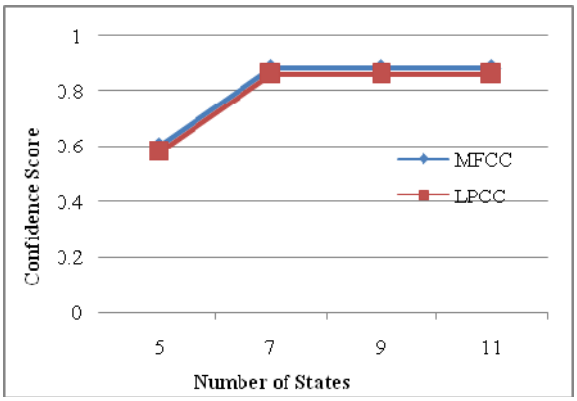

Fig. 9.Confidence Score based on Number of States

Therefore, 7 states have been changed for the mixtures such as 2, 4,6 and 8 for LPCC and MFCC.

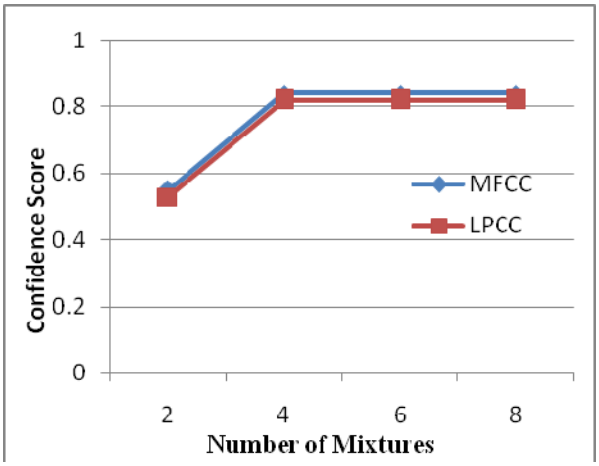

Fig.10.Confidence Score based on Number of Mixtures

Fig 10 shows the confidence score based on number of mixtures. When the number of mixtures is varied as 2, 4, 6 and 8, the significant result is attained for mixture 4.

C. *Performance Evaluation*

The performance of the proposed CSR system is specified in terms of accuracy and error rate. Accuracy may be measured in terms of performance accuracy,

which is usually based on the recognition accuracy of classifiers and based on the hit rate. The error rate is rated with word error rate (WER), false alarm and miss detection rate. The other parameters used to evaluate the performance of the proposed CSR are Mean Square Error Rate (MSE) and Peak Signal to Noise Ratio (PSNR).

D. Mean Square Error Rate (MSE)

Mean Square Error Rate (MSE) and Peak Signal to Noise Ratio (PSNR) values are used as the parameters for performance measures to find out the better filtering signal for feature extraction. The MSE value is calculated by the following equation.

$$MSE = \frac{\Sigma(y_i - \hat{y}_i)}{n - p}$$

– Vector of n predictions,
- Vector of the true values,
- Number of independent variable.

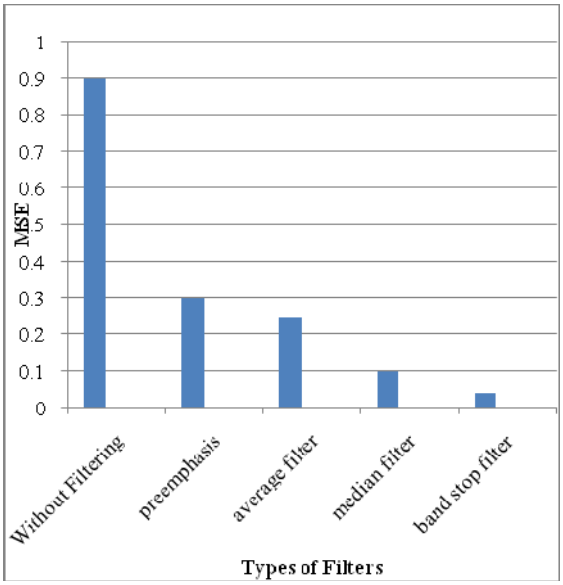

Fig. 11. Performance Evaluation of Filtering based on MSE

E. Peak Signal to Noise Ratio (PSNR)

PSNR value is calculated by the equation

$$PSNR = 10 \log_{10}\left[\frac{R^2}{MSE}\right]$$

R - Maximum fluctuation in the input

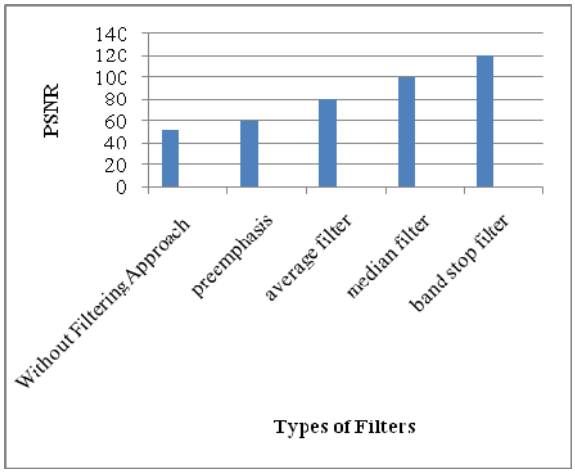

Fig. 12. Performance Evaluation of Filters based on PSNR

The performance measurements PSNR and MSE show the importance of the filtering technique in the preprocessing stage. In this approach, the band stop filtering approach is used. The performance of the band stop filtering approach is compared with other filtering approaches such as pre emphasis, average filter and median filter. It is observed that, the band stop filtering approach outperforms the other filtering approaches. Moreover, it is observed that, the CSR with filtering approach provides better performance when compared with the CSR without filtering approach.

F. Hit rate and false alarm rate

Percentage of correctness regarding extraction of voice samples from a speech signal is defined as follows:

$$Hitrate = \frac{Numberafcorrectlyidentifiedwords}{Nummerafwordboundariestestutterance}$$

$$falsealarmrate = \frac{Numberaferraneauswordboundariestestdentified}{Numberafwordboundariestestutterance}$$

$$Misseddetectionrate = \frac{Nummerafmisseddetection}{Numberafwordboundariestestutterance}$$

Table I summarizes the results showing the percentage of correctness in detection of word boundaries for four Dravidian languages.

TABLE I: PERFORMANCE MEASURES

Sl.No	Language	Total no of words present	Number of correctly identified words	Hit rate (%)	Missed detection rate	False alarm rate
1.	Tamil	100	92	92	5	3
2.	Telugu	100	92	92	4	4
3.	Malayalam	100	90	90	6	4
4.	Kannada	100	91	91	4	5

G. *Word Error Rate (WER)*

Word error rate is a parameter used to find the performance of a speech recognition or machine translation system. The common complexity of measuring performance lies in the fact that the familiar word series can have a different length from the reference word sequence. The WER is resulting from the Levenshtein distance, functioning at the word level as an alternative of the phoneme level. This problem is solved by first aligning the recognized word sequence with the reference (spoken) word sequence using dynamic string alignment. Word error rate can then be computed as

$$WER = \frac{S + D + I}{N}$$

Where

- S is the number of substitutions,
- D is the number of the deletions,
- I is the number of the insertions,
- N is the number of words in the reference

When reporting the performance of a speech recognition system, sometimes the word recognition rate (WRR) is used instead.

$$WRR = 1 - WER$$

Figure 15 shows the WRR comparison for the Dravidian languages for both MFCC and LPCC features. It is observed from the figure that the WRR by MPCC and LPCC features are almost similar. Best results are attained for the Tamil language among other languages. MPCC obtained 89% WRR whereas LPPC attained 88% WRR.

C. *HMM*

The recognition accuracy for training and testing set with the feature LPCC and MFCC for proposed speech recognition system with HMM classifier is presented in figure 16. The performance is evaluated for all the four Dravidian languages taken into consideration. It is observed that the recognition accuracy of HMM with MFCC and LPCC feature is significant for Malayalam language with 89%.

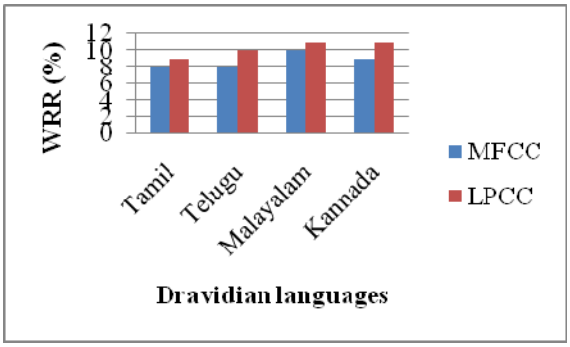

Fig.15. WRR Comparison for Dravidian languages

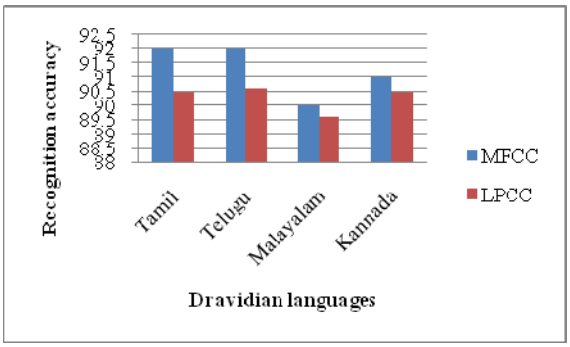

Fig. 16. Recognition accuracy for proposed speech recognition system with HMM classifier

D. *SVM*

The recognition accuracy for SVM classifier with LPCC and MFCC features for proposed speech recognition system is shown in figure 17. It is observed from the figure that the proposed CSR system with SVM classifier provides significant results for Tamil language with 89.2% recognition accuracy of MFCC and 89% accuracy for LPCC. The recognition accuracy of other Dravidian languages is slightly lesser than Tamil language.

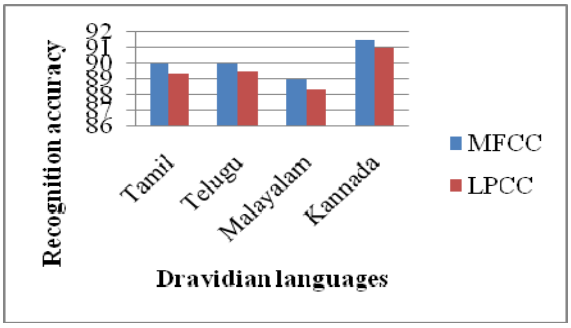

Fig 17. Recognition accuracy for proposed speech recognition system with SVM classifier

E. AANN

The recognition accuracy for AANN classifier with LPCC and MFCC features for proposed speech recognition system is shown in figure 18. It is observed from the figure that the proposed CSR system with AANN classifier provides significant results for Tamil language with 89% recognition accuracy of MFCC and 89% accuracy for LPCC.

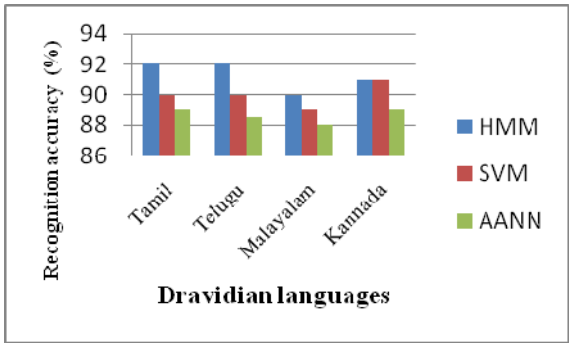

Fig.18. Recognition accuracy for proposed speech recognition system with AANN classifier

F. Classification Accuracy

Figure 19 shows the accuracy comparison of the classification techniques such as AANN, HMM and SVM. It is clearly observed from the results that the HMM classification approach provides better classification accuracy when compared with AANN and SVM techniques. HMM classifier provides significant results to Malayalam language with 92% accuracy. When SVM and AANN classifiers are considered, they provide good results with Tamil language with 90.8% and 89.9% respectively.

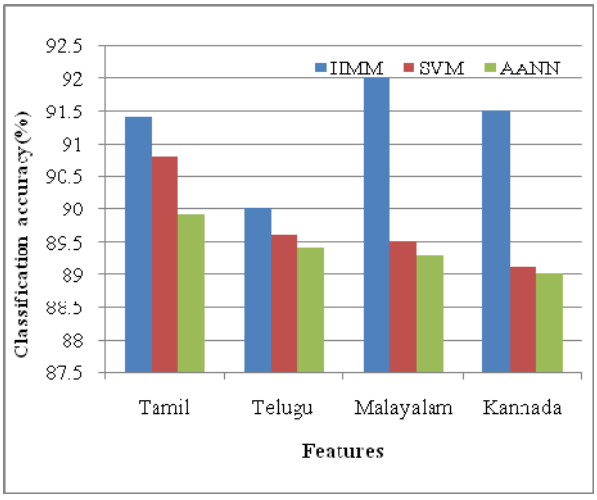

Fig. 19.Classification Accuracy Comparison

IV. CONCLUSION

In recent years, there is a momentary need for the CSR system to be developed in Dravidian languages such as Tamil, Malayalam, Kannada and Telugu. In this paper, such an important effort is carried out for recognizing Dravidian Language since it would help the native speakers of these languages to a great extent in various applications such as hands free computing, interactive voice response, multimodal interaction etc. To accomplish this task, feature extraction is done after employing required preprocessing techniques. The most widely used MFCC, LPCC method is used to extract the significant feature vectors from the enhanced speech signal and they are given as the input to the multi classifiers. The adopted classifiers are trained with these input and target feature vectors. The classifiers are compared with the each other for better classification accuracy. The results with the specified parameters were found to be satisfactory considering the less number of training data. The main application of this research work is to assist the native speakers of Dravidian languages in applications such as authentication in ATM centers, home automation etc. This would also be used as the security system where the speech could be taken as the identity. This research work provides significant performance where in HMM classifier gives 92% accuracy in Malayalam language where as SVM and AANN classifiers provide 90.8% and 89.9% accuracy for Tami and Kannada languages respectively. The future scope of this research work would be to use more training samples. This preliminary experiment will helps to develop CSR system for Dravidian language using different hybrid approaches.

As a future work, CSR would be used to make Information Technology (IT) relevant to rural people. The present research work could be applied for speech interfaces in Dravidian languages. Application specific Dravidian CSR could be proposed to make computer aided teaching, a reality in rural schools.

REFE'RENCES

[1] Vimala C, Radha V, " Efficient Speaker Independent Isolated Speech Recognition for Tamil Language Using Wavelet Denoising and Hidden Markov Model", Proceedings of the Fourth International Conference on Signal and Image Processing, 2012 (ICSIP 2012).

[2] Wellekens CJ, "Explicit time correlation in hidden Markov models for speech recognition", In *Proc. ICASSP*, 1987, pp. 384–386.

[3] Bilmes JA, "Graphical models and automatic speech recognition. In Mathematical foundations of speech and language processing", Springer- Verlag, New York, 2003.

[4] Kenny P, Lennig M, Mermelstein P, "A linear predictive HMM for vector-valued observations with applications to speech recognition", *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 38, no. 2, Feb. 1990, pp. 220– 225.

[5] Su J, Li H, Haton JP, Ng KT, "Speaker time-drifting adaptation using trajectory mixture hidden Markov models," In *Proc. ICASSP*, 1996, pp. 709–712.

[6] Korkmazskiy F, Juang BH, Soong FK, "Generalized mixture of HMMs for continuous speech recognition", In *Proc. ICASSP*, 1997, pp. 1443–1446.

[7] George E Dahl, Dong Yu, Li Deng, Alex Acero, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition", IEEE Transactions On Audio, Speech, And Language Processing, Vol. 20, NO. 1, January 2012.

[8] He X, Deng L, Chou W, "Discriminative learning in sequential pattern recognition—A unifying review for optimization-oriented speech recognition", *IEEE Signal Process, Mag.*, vol. 25, no. 5, Sep. 2008, pp. 14–36.

[9] Kapadia S, Valtchev V, Young SJ, "MMI training for continuous phoneme recognition on the TIMIT database", In *Proc. ICASSP*, 1993, vol. 2, pp. 491–494.

[10] Juang BH, Chou W, Lee CH, "Minimum classification errorrate methods for speech recognition". *IEEE Trans. Speech AudioProcess.*, vol. 5, no. 3, May 1997, pp. 257–265.

[11] Jiang H, Li X, "Incorporating training errors for large margin HMMs under semi-definite programming framework" In *Proc. ICASSP*, 2007, vol. 4, pp. 629–632.

[12] Yu D, Deng L, He X, Acero A, "Use of incrementally regulated discriminative margins in MCE training for speech recognition," in *Proc. ICSLP*, 2006, pp. 2418–2421.

[13] Hifny Y, Renals S, "Speech recognition using augmented conditional random fields. *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 17, no. 2, Feb. 2009, pp. 354–365.

[14] Morris J, Fosler-Lussier E, "Combining phonetic attributes using conditional random fields" In *Proc. Interspeech*, 2006, pp. 597–600.

[15] Yu D, Deng L, "Deep-structured hidden conditional random fields for phonetic recognition," in *Proc. Interspeech*, 2010, pp. 2986–2989.

[16] Gong Y, "Stochastic trajectory modeling and sentence searching for continuous speech recognition", *IEEE Trans. Speech Audio Process.*, vol. 5, no. 1, 1997, pp. 33–44.

[17] Amin AshouriSaheli, Gholam Ali Abdali, Amir Abolfazlsuratgar, "Speech Recognition from PSD using Neural Network", *Proceedings of the International Multi Conference of Engineers andComputer Scientists 2009,* Vol I,IMECS 2009, March 18 – 20.

[18] Kvale K, "Segmentation and Labeling of Speech. PhD Dissertation", The Norwegian Institute of Technology, 1993.

[19] Yegnanarayana B, Kishore S, "AANN: an alternative to GMM for pattern recognition", Neural Networks, 2002.

[20] Naresh P Jawarkar, Vasif Ahmed, "Micro-controller based Remote Monitoring using Mobile through Spoken Commands", Journal of Networks, Vol. 3, No. 2, 2008.

[21] Young et al, " The HTK Book" (for HTK version 3.2.1), Cambridge University Engineering Dept, 2002.

[22] Alexandros Georgogiannis, Vassilis Digalakis, "Speech Emotion Recognition Using Non-Linear Teager Energy Based Features in Noisy Environments", 20th European Signal Processing Conference (EUSIPCO 2012).

[23] Kramer MA, "Nonlinear principal component analysis using auto associative neural networks" AIChE 7, 1991, pp. 233–243.

[24] Bianchini M, Frasconi P, Gori M, (1995) Learning in multilayered networks used as autoassociators. IEEE Trans. Neural Networks 6, 512 - 515.

[25] Haykin S (1999) Neural Networks: A Comprehensive Foundation. Prentice Hall, Englewood Cliffs, NJ.

[26] Kishore SP (2000) Speaker verification using autoassociative neural networks model. M.S. Thesis, Department of Computer Science and Engineering, Indian Institute of Technology Madras.

[27] Vapnik V, "The Nature of Statistical Learning Theory" Springer Verlag, 1995.

[28] Daniel Poveya,, LukásBurget b, MohitAgarwalc, Pinar Akyazid, Feng Kai, Arnab Ghoshalf, Ondrej Glembekb, Nagendra Goelg, Martin Karafiátb, AriyaRastrowh, Richard C, Rosei Petr Schwarzb, Samuel Thomas, "The subspace Gaussian mixture model—A structured model for speech recognition", Computer Speech and Language, Vol. 25, 2011,pp. 404–439.