International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:8, No:12, 2014

Hybrid Modeling Algorithm for Continuous Tamil Speech Recognition

M. Kalamani, S. Valarmathy, M. Krishnamoorthi

Abstract**—**In this paper, Fuzzy C-Means clustering with Expectation Maximization-Gaussian Mixture Model based hybrid modeling algorithm is proposed for Continuous Tamil Speech Recognition. The speech sentences from various speakers are used for training and testing phase and objective measures are between the proposed and existing Continuous Speech Recognition algorithms. From the simulated results, it is observed that the proposed algorithm improves the recognition accuracy and F-measure up to 3% as compared to that of the existing algorithms for the speech signal from various speakers. In addition, it reduces the Word Error Rate, Error Rate and Error up to 4% as compared to that of the existing algorithms. In all aspects, the proposed hybrid modeling for Tamil speech recognition provides the significant improvements for speech-to-text conversion in various applications.

Keywords**—**Speech Segmentation, Feature Extraction, Clustering, HMM, EM-GMM, CSR.

I. Introduction

THE Continuous Speech Recognition (CSR) for human-machine interface still remains a challenging problem today. It requires the more sophisticated recognizers in order to handle the word boundary issues. Speech segmentation and Feature extraction are used as a preprocessing step in the CSR system. Speech segmentation algorithm is used to segment the speech sentences into words. A feature extraction algorithm is used for data reduction by converting the segmented speech signal into a compact set of parameters. These features are used for training and testing phase of modeling in speech recognition.

Rabiner et al. proposed the short time energy and zero crossing rate of the speech signal for speech segmentation in which it is used to detect the word boundaries at low noise environments [19]. Rahman et al. proposed the short time energy with spectral centroid based hybrid speech segmentation algorithm in which it provides the better segmentation accuracy as compared to other algorithms [20], [25], [8], [9].

Deller proposed the mel scale conversion for Cepstral analysis and the resulting Mel Frequency Cepstral Coefficients (MFCC) are used as a feature vector in the modeling of CSR [5], [7], [1], [10], [11]. Davis et al. introduced the feature extraction with MFCC and their dynamic derivatives, which improves the performance of the recognizer [6], [15], [22], [26].

Deller proposed the unsupervised K-means clustering algorithm for labeling the feature vectors before modeling. In this speech features are clustered around some centroid locations. The resulting cluster centers constitute a codebook which is used for training phase [7]. Linde, Buzo and Gray et al. proposed the K-means algorithm for vector quantization with a large class of distortion measures [14]. Bezdek proposed the Fuzzy C-means (FCM) clustering, which is used for clustering the data vectors and reduces the word error rate in CSR system [4], [3], [12].

Rabiner describes the theoretical aspects of HMM and describes their basic problems in HMM [18]. Rabiner provides the first implementation of HMM in speech processing applications. In this method, the speech signal is characterized as a random stochastic process and its parameters are estimated using well defined model structure. It is tedious to implement for continuous speech recognition [17], [2], [13], [16]. Thangarajan et al. proposed the HMM based Tamil speech recognition method and it is used for limited Tamil words [23]. Manan Vyas proposed the Gaussian Mixture Models for isolated word recognition. In GMM, the maximization will leads problem and it is less efficient [24], [21].

In order to improve the recognition accuracy for continuous Tamil speech sentences, the FCM clustering with EM-GMM based hybrid modeling algorithm for CSR is proposed in this paper. This method reduces the Word Error Rate at a significant level for Tamil speech-to-text conversion as compared to that of various algorithms is discussed in previous.

This paper is organized as follows: Section II describes about the Speech preprocessing techniques and Section III provides some of the existing modeling techniques for CSR. The proposed FCM clustering with EM-GMM based hybrid modeling algorithm for CSR is described in Section IV. Section V illustrates the performance evaluation of the existing and proposed algorithms and Section VI concludes this paper.

II. Speech Preprocessing

The essential step in Speech Recognition and Synthesis systems is speech segmentation and Feature extraction. Speech segmentation is used to break the continuous speech into basic units like words, sub words and syllables and that can be recognized in Automatic Speech recognition (ASR) system. Feature Extraction is used to compute the raw, static

M.Kalamani is with the ECE Department, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu 638401 India (phone: 91-9442354033; e-mail: kalamani.mece@gmail.com).

S. Valarmathy and M. Krishnamoorthi are with the ECE Department, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu 638401 India (e-mail: atrmathy@rediffmail.com, krishna_bit@yahoo.co.in).

and dynamic feature vector from the segmented speech signal. This process is used to reduce the dimensions of the input data for modeling in the speech recognition system.

A. Speech Segmentation Algorithms

In this paper, the blind speech segmentation procedure is used. This is carried out using end point detection techniques. This technique is used to detect the segment boundaries in which the spectral changes exceed a minimum threshold level. Time domain and frequency domain features of the signal are used to segment the speech signal.

Due to simple implementation and efficient calculation, the time domain feature is widely used to extract the speech segment from its continuous speech signal. For this speech segmentation, Short-time signal energy and Short-term average zero-crossing rate is frequently used. The speech signal is mostly concentrated in 330Hz to 3300 KHz. DFT is used to extract the frequency domain features which give the spectral composition of the signal. Spectral centroid and Spectral flux are widely used as frequency domain features. In this paper, the combination of time domain (Short-time signal energy) and frequency domain (Spectral Centroid) features of the signal is used to extract the speech segments.

1.Short-Time Signal Energy

This is the most natural feature which is used to measure the signal densities at each sample index. This is used to discover the voiced speech which has higher energy than the unvoiced speech signal. Energy is calculated by squaring the samples and taking the average for each windowed speech sequence. Short time signal energy in the N length speech frame is defined as,

$$E_n = \frac{1}{N}\sum_{m=1}^{N}[x(m)w(n-m)]^2 \quad (1)$$

where, x(n) is the speech signal and w(n) is the window function which is described as,

$$w(n) = \begin{cases} 1, & for\ 0 \le n \le (N-1) \\ 0, & otherwise \end{cases} \quad (2)$$

2.Spectral Centroid

This is used to characterize and indicates the center of gravity of the spectrum. Also, it is a measure of spectral position. For voiced speech frame, this measure has a high value. Spectral Centroid for ith frame is defined as,

$$SC_i = \frac{\sum_{m=0}^{N-1} f(m)X_i(m)}{\sum_{m=0}^{N-1} X_i(m)} \quad (3)$$

where, f(m) is the center frequency of ith bin with N length speech frame and $X_i(m)$ is the i^{th} frequency bin DFT of the of the input speech frame.

B. Feature Extraction Algorithms

The performance of speech recognition mainly depends on this feature extraction. Feature Extraction is used to compute the raw, static and dynamic feature vector which provides a compact representation of the given signal while preserving spectral and/or temporal characteristics of the speech signal information. It is used to reduce the dimensions of the input data for modeling in the speech recognition system. Commonly used feature extraction methods are Ceptral analysis, Linear Predictive Coding (LPC) coefficients, Perceptually Linear Predictive coefficients (PLP) and Mel-frequency Cepstral coefficients (MFCC) for various speech processing applications.

LPC is an autoregressive (all pole) model which is used to estimate a short-term spectral envelope for the speech spectrum. In this spectrum, the voiced sounds are characterized by peak spectral values. Linear prediction model can cause poor modeling in noisy environments. Cepstral analysis has been extensively used for feature extraction in speech recognition. Perhaps, the most popular derivation of Cepstral analysis combines the cepstrum with a nonlinear frequency-warping, known as Mel-scale conversion.

1.Mel Frequency Cepstral Coefficients

This is the most dominant method used to extract the spectral features such as Mel Frequency Cepstral Coefficients (MFCC) in the frequency domain using mel scale and it is more accurate than the time domain features. For human auditory system, the spectral resolution is nonlinear along the frequency axis. In order to mimic the spectral characteristics of the human ear, the mel-scale conversion is expressed as linear frequency spacing below 1 kHz and a logarithmic spacing above 1 kHz. An acoustic frequency is mapped to a perceptual frequency scale as follows,

$$F_{Mel} = 2595 * log_{10}\left(1 + \frac{f(Hz)}{700}\right) \quad (4)$$

Cepstral coefficients are estimated based on short time analysis and the MFCC vectors are computed in each frame.

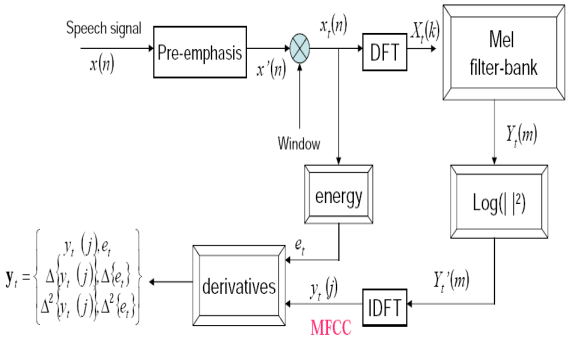

Fig. 1 Block Diagram of Feature Extraction using MFCC

Fig. 1 shows the block diagram of feature extraction using MFCC. Generally, speech signal has a high dynamic range and it suffers from additive noise. In order to overcome this, the pre-emphasis filter (FIR HPF) is used and its transfer function is described as,

$$H(z) = 1 - \alpha z^{-1} \quad (5)$$

where, α is the pre-emphasis parameter which has a range of $0.9 \leq \alpha \leq 1.0$.

In order to improve the time and frequency resolution, the speech signal is divided into frames. Windows used to select a portion of the signal that can reasonably be assumed stationary over the frame. For short time analysis, each frame is multiplied by a window function with overlapping which is allowed up to 50%. It tapers down at the frame end, which minimize the spectral distortion. For speech analysis, the Hamming window is widely used.

After windowing, FFT is calculated for each frame in order to extract the frequency components of the signal. For this transformed frame, the mel scaled filter bank is applied. At last, the DCT is computed for the outputs of the filter to get a central coefficient. Generally, 12 to 14 Mel Frequency Cepstral Coefficients is computed for the each frame. In addition, the energy in each frame is considered as a one feature. In the frame-by-frame analysis causes similarity losses in time domain. To recover this, delta (1st derivative) and delta-delta (2nd derivative) features are used. Totally 39 features such as 12 original, 12 delta, 12 delta-delta and 3 energy features as considered as MFCC acoustic feature vectors for each frame. It is very useful for further analysis and processing in speech recognition.

III. Existing Modeling Techniques for CSR

A. Clustering Techniques

Clustering is used to convert continuous-valued observation vectors into quantized discrete observation symbols before the classification/modeling process. It will generate the code book for training sequence which contains the acoustic information of the signal. For this purpose, K-means clustering is used in speech recognition. The K-means algorithm is an algorithm to cluster 'n' objects based on attributes into 'K' partitions (where K < n). Given K, the number of clusters and training data is clustered iteratively to minimize the objective function,

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2 \quad (6)$$

where, $x_i^{(j)}$ is the ith data point in cluster j and c_j is the centroid of cluster j.

Then the feature vector in the training matrix is clustered based on the squared Euclidean distance between a training vector and each of the code vectors in the codebook. The total distortion is computed once when all the training vectors have been labeled with the corresponding code vectors. Total quantization error is the summation of all these distances.

B. Modeling of Speech Signal

Several types of models are used to characterize the properties of the signal. Broadly, it is classified as deterministic and statistical models. In deterministic models, the signal model specifications are straightforward and it requires the determine values of the parameters of the signal. Statistical models are used to characterize only the statistical properties of the signal. In statistical model, the signal is well characterized into a parametric random stochastic process and it is estimated in a precise, well defined manner. Statistical models include Gaussian processes, Passion processes, Markov processes and Hidden Markov processes. For a nonstationary speech signal, HMM can employ as a piecewise stationary model and implemented as discrete observations. Two main reasons for accepting HMM for ASR systems as:

- Its capability to model the non-linear dependencies of each speech unit on the adjacent units.
- A powerful set of analytical approaches provided for estimating model parameters.

Once these fundamental problems are solved, then apply HMM to speech recognition. The three fundamental problems for HMM design are:

- Evaluation of the probability of a sequence of observations given a specific HMM
- Determination of a better sequence of model states
- Adjustment of model parameters so as to best account for the observed signal

The HMM model parameters are,

- Number of hidden states in the model, S
- Number of distinct observation symbols, N
- State transition probability distribution, $A = a(i|j)$, where

$$a(i|j) = P(q_t = i|q_{t-1} = j) \text{ and } 1 \leq i, j \leq S \quad (7)$$

- Observation symbol probability distribution, $B = b(k|i)$, where

$$b(k|i) = P(o_t = k|q_t = i), 1 \leq i \leq S, 1 \leq k \leq K \quad (8)$$

- Initial state distribution, $\pi = \{P(q_1 = i)\}$, where $1 \leq i \leq S$
- The compact notation, $\lambda = (A, B, \pi)$ is usually used to define the model parameters of an HMM for convenience.

It is difficult to optimize the HMM training parameters and inefficient analytical method for maximizing the joint probability of the trained data. Baum-Welch algorithm is used to optimize the parameters $\lambda(A, B, \pi)$ such that the likelihood of occurrence for the training data $P(O|\lambda)$ is locally maximized. It is an iterative method, also known as the Forward-Backward algorithm. In this method, initially started with the arbitrary model λ and searches for a new model $\hat{\lambda}$. At each iteration, this improves the maximum probability that the given observation sequence which is generated by the new model.

IV. Proposed Modeling Techniques for CSR

A. FCM Based Clustering Algorithm

Fuzzy C-means (FCM) clustering is the most powerful fuzzy clustering techniques. This is an unsupervised method for the analysis of data and construction of models. It is converged to the local minima and this process is more natural than hard clustering. FCM clustering employs fuzzy partitioning such that a data point can belong to all groups with different membership grades between 0 and 1. In this, membership values are assigned based on the distance between the cluster center and the data points. In this

approach, the membership values and cluster centers are updated iteratively. The FCM clustering algorithm has the following steps:

- Let us suppose that N data points represented by $x_i, i = 1,2,\dots,N$ are to be clustered.
- Assume the number of clusters to be made, that is, C, where $2 \le C \le N$.
- Choose an appropriate level of cluster fuzziness, $f > 1$.
- Initialize the $N \times C$ sized membership matrix U, at random, such that $U_{ij} \in [0,1]$ and $\sum_{j=1}^{C} U_{ij} = 1$ for each i.
- Determine the cluster centers CC_j, for jth cluster by using:

$$CC_j = \frac{\sum_{i=1}^{N} U_{ij}^{f} x_i}{\sum_{i=1}^{N} U_{ij}^{f}} \quad (9)$$

- Calculate the Euclidean distance between ith data point and jth cluster center as follows,

$$D_{ij} = \|x_{im} - CC_j\| \quad (10)$$

- Update fuzzy membership matrix U according to D_{ij}. If $D_{ij} > 0$, then

$$U_{ij} = \frac{1}{\sum_{c=1}^{C} \left(\frac{D_{ij}}{D_{ic}}\right)^{\frac{2}{f-1}}} \quad (11)$$

- If $D_{ij} = 0$, then the data point coincides with the corresponding data point of jth cluster center CC_j and it has the full membership value, that is, $U_{ij} = 1$.
- Repeat from Step 5 to Step 7 until the changes in $U < \varepsilon$, where ε is a pre-specified termination criterion.

B. Speech Modeling by EM-GMM

Gaussian Mixture Model (GMM) is a parametric model which is represented as a weighted sum of Gaussian Mixture component densities under continuous measurements of labeled features from FCM clustering techniques. GMM parameters are estimated using well trained prior model by iterative Expectation-Maximization (EM) or Maximum A Posteriori (MAP) estimation. This model is parameterized by mean vectors, covariance vectors and mixture weights from all component densities. Several techniques are available for estimating these parameters and in which the well established one is Maximum Likelihood (ML). This is a nonlinear function of model parameters and direct maximization is not possible. Hence, the iterative EM algorithm is used for training as well as matching. The idea behind this algorithm is to start with initial model λ and to estimate a new model $\bar{\lambda}$, such that,

$$p(X|\bar{\lambda}) \ge p(X|\lambda) \quad (12)$$

Then this new model becomes the initial model for the next iteration and it is repeated till convergence is reached. The re-estimation formulas are used in order to increase the model likelihood as follows:

Mixture Weights:

$$\bar{p_i} = \frac{1}{T}\sum_{t=1}^{T} p(i|\vec{x_t}, \lambda) \quad (13)$$

Means:

$$\vec{\bar{\mu}} = \frac{\sum_{t=1}^{T} p(i|\vec{x_t},\lambda)\vec{x_t}}{\sum_{t=1}^{T} p(i|\vec{x_t},\lambda)} \quad (14)$$

Variances:

$$\bar{\sigma_i}^2 = \frac{\sum_{t=1}^{T} p(i|\vec{x_t},\lambda){x_t}^2}{\sum_{t=1}^{T} p(i|\vec{x_t},\lambda)} - \bar{\mu_i}^2 \quad (15)$$

where, σ_i^2, x_t and μ_i refer to aribitary elements of the vectors $\vec{\sigma}_i^2$, $\vec{x}_t$ and $\vec{\mu}_i$ respectively.

A posteriori probability for acoustic class i is given by,

$$p(i|\vec{x_t}, \lambda) = \frac{p_i b_i(\vec{x_t})}{\sum_{k=1}^{M} p_k b_k(\vec{x_t})} \quad (16)$$

where, M is the order of the mixture. The selection of mixture order and initializing the model parameters are the critical factors in GMM.

V. Performance Evaluation

This section wholly describes the comparative results between the proposed and the existing Continuous Tamil Speech Recognition algorithms. Throughout the evaluation process, different input speech signals for different speakers are obtained from the database of IIIT consisting of various Tamil speech sentences. Around 500 different sentences from 9 different speakers are used for training and testing. In which, 75% of the words from each speaker used for training and 25% words used in testing phase.

Table I shows the performance comparison of Segmentation accuracy for various speakers of the proposed and existing speech segmentation algorithms. From the tabulated results, it is observed that the hybrid speech segmentation algorithm improves the segmentation accuracy as 0.9% to 7.55% as compared to that of the existing algorithms.

The Recognition Accuracy (%) and F-Measure for the existing and proposed speech recognition algorithms under various MFCC (16, 26, 39) features for each frame are shown in Figs. 2 and 3. From these results, it is observed that the recognition accuracy and F-Measure are improved at the MFCCs is 26. Hence, that the optimal number of Mel Frequency Cepstral coefficients is found at 26 and this optimal MFCC feature for each frame are extracted and it is used for speech recognition system.

TABLE I
COMPARISON OF SEGMENTATION ACCURACY FOR VARIOUS SPEAKERS BY THE PROPOSED AND EXISTING SPEECH SEGMENTATION ALGORITHMS

No. of Speakers	Segmentation Accuracy (%)				
	Spectral Flux	ZCR	Spectral Energy	Spectral Centroid	Hybrid (Energy+Centroid)
1	91.84	92.86	94.90	96.22	97.96
2	91.24	92.78	94.23	95.88	96.91
3	91.03	92.41	93.79	95.52	96.55
4	90.70	92.19	93.22	95.50	96.48
5	90.15	91.42	92.94	95.42	96.32
6	89.89	91.23	92.56	95.24	96.24
7	89.39	91.30	92.33	95.18	96.18
8	88.88	91.19	92.29	95.10	96.00
9	88.32	91.04	91.88	94.85	95.87

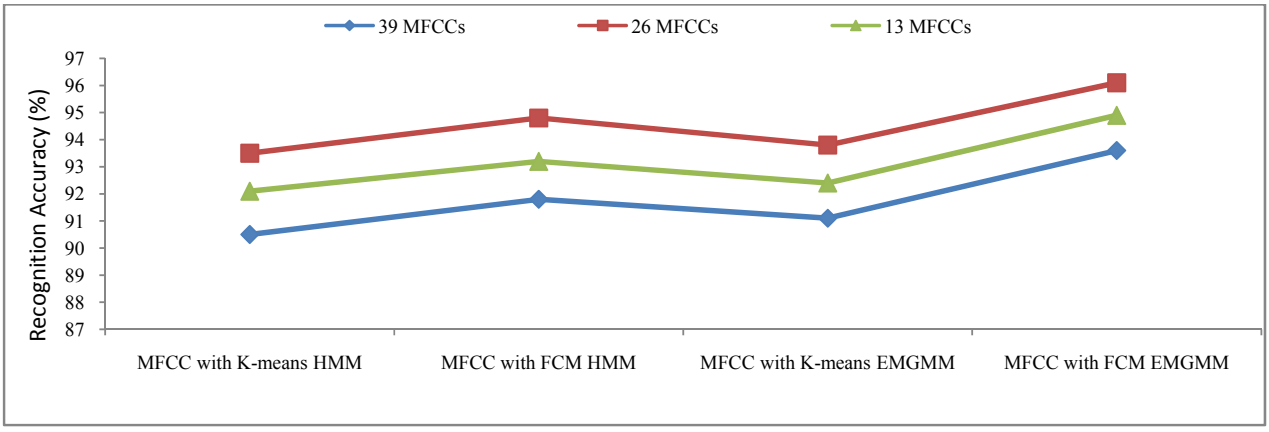

Fig. 2 Recognition Accuracy (%) for the existing and proposed speech recognition algorithms under various MFCCs for each frame

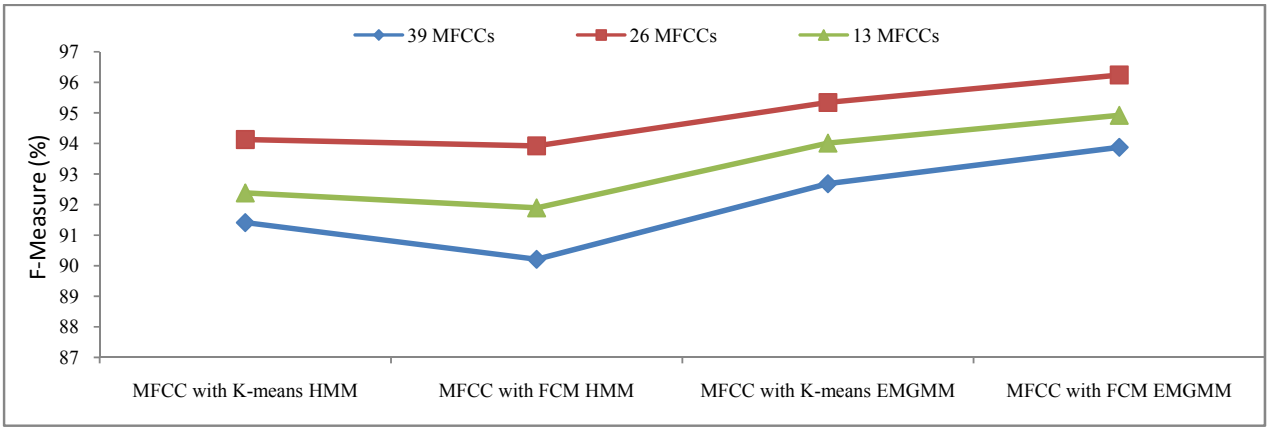

Fig. 3 F-Measure (%) for the existing and proposed speech recognition algorithms under various MFCCs for each frame

TABLE II
COMPARISON OF RECOGNITION ACCURACY FOR VARIOUS SPEAKERS BY THE PROPOSED AND EXISTING SPEECH RECOGNITION ALGORITHMS

No. of Speakers	Total No. of words for Test (After Segmentation)	Recognition Accuracy			
		K-Means HMM	FCM HMM	K-Means EM-GMM	Proposed (FCM EM-GMM)
1	240	95.8	96.7	96.3	97.9
2	470	95.1	96.0	95.7	97.7
3	700	94.4	95.7	95.1	97.4
4	960	93.9	95.2	94.4	96.8
5	1210	93.5	94.8	93.8	96.1
6	1440	93.1	94.2	93.4	95.8
7	1700	92.8	93.9	93.1	95.2
8	1920	92.6	93.5	92.9	94.9
9	2205	92.3	93.2	92.6	94.5

Table II shows the performance comparison of recognition accuracy for the proposed and existing speech recognition algorithms. From these results, it is observed that the proposed algorithm improves the recognition accuracy from 1.2 to 3% as compared to that of the existing algorithms for various speakers.

TABLE III
COMPARISON OF WER FOR VARIOUS SPEAKERS BY THE PROPOSED AND EXISTING SPEECH RECOGNITION ALGORITHMS

No. of Speakers	Total No. of words for Test (After Segmentation)	WER			
		K-Means HMM	FCM HMM	K-MeansEM-GMM	Proposed (FCM EM-GMM)
1	240	5.83	4.58	3.75	2.50
2	470	5.96	5.11	4.68	2.98
3	700	6.71	5.86	5.14	3.14
4	960	7.29	6.67	5.73	3.75
5	1210	7.69	7.36	6.20	4.38
6	1440	8.13	7.71	6.81	4.72
7	1700	8.35	8.00	7.06	5.41
8	1920	8.65	8.18	7.97	5.78
9	2205	8.98	8.53	7.98	6.21

TABLE IV
COMPARISON OF ERROR RATE FOR VARIOUS SPEAKERS BY THE PROPOSED AND EXISTING SPEECH RECOGNITION ALGORITHMS

No. of Speakers	Total No. of words for Test (After Segmentation)	Error Rate			
		K-Means HMM	FCM HMM	K-Means EM-GMM	Proposed (FCM EM-GMM)
1	240	5.74	4.55	3.73	2.49
2	470	5.89	5.06	4.65	2.96
3	700	6.64	5.80	5.10	3.13
4	960	7.21	6.60	5.68	3.73
5	1210	7.60	7.27	6.14	4.36
6	1440	8.02	7.62	6.74	4.70
7	1700	8.26	7.91	6.99	5.38
8	1920	8.54	8.09	7.89	5.74
9	2205	8.87	8.43	7.89	6.17

TABLE V
COMPARISON OF F-MEASURE FOR VARIOUS SPEAKERS BY THE PROPOSED AND EXISTING SPEECH RECOGNITION ALGORITHMS

No. of Speakers	Total No. of words for Test (After Segmentation)	F-measure, F			
		K-Means HMM	FCM HMM	K-Means EM-GMM	Proposed (FCM EM-GMM)
1	240	95.63	96.45	96.87	97.71
2	470	95.51	96.05	96.37	97.66
3	700	94.77	95.28	96.13	97.57
4	960	94.49	94.47	95.66	96.92
5	1210	94.13	93.92	95.34	96.24
6	1440	93.94	93.63	94.86	96.00
7	1700	93.76	93.36	94.63	95.29
8	1920	93.65	93.23	94.03	95.04
9	2205	93.39	92.94	94.09	94.66

TABLE VI
COMPARISON OF ERROR FOR VARIOUS SPEAKERS BY THE PROPOSED AND EXISTING SPEECH RECOGNITION ALGORITHMS

No. of Speakers	Total No. of words for Test (After Segmentation)	Error, E			
		K-Means HMM	FCM HMM	K-Means EM-GMM	Proposed (FCM EM-GMM)
1	240	4.37	3.55	3.13	2.29
2	470	4.49	3.95	3.63	2.34
3	700	5.23	4.72	3.87	2.43
4	960	5.51	5.53	4.34	3.08
5	1210	5.87	6.08	4.66	3.76
6	1440	6.06	6.37	5.14	4.00
7	1700	6.24	6.64	5.37	4.71
8	1920	6.35	6.77	5.97	4.96
9	2205	6.61	7.06	5.91	5.34

Performance comparison of Word Error Rate (WER) and Error Rate (ER) for the proposed and existing speech recognition algorithms are shown in Tables III and IV. From the tabulated results, it is observed that the proposed algorithm reduces the WER from 1.25 to 3.57% and ER from 1.24 to 3.51% as compared to that of the existing algorithms for various speakers.

Table V and VI shows the performance comparison of F-Measure (F) and Error (E) for the proposed and existing speech recognition algorithms. From these results, it is observed that the proposed algorithm increases the F-Measure

from 0.57 to 2.8% and reduces Error from 0.57 to 2.8% as compared to that of the existing algorithms for various speakers.

VI. Conclusions

In this paper, FCM clustering with EM-GMM based hybrid modeling algorithm is proposed for Continuous Tamil Speech Recognition. Speech sentences from nine speakers are used for training and testing phase. Performance evaluation is made between the existing and proposed algorithms. From the tabulated results, it is observed that the hybrid speech segmentation algorithm improves the segmentation accuracy up to 8% as compared to that of the existing algorithms. From these segment words, the raw feature vectors (26 feature values/frame) are extracted using Feature extraction by delta MFCC. The extracted raw features are labeled by using various clustering techniques, and then modeled by the proposed and existing algorithms for speech-to-text conversion.

From the simulated results, it is observed that the proposed modeling algorithm for CSR improves the recognition accuracy as 1.2 to 3% and F-measure as 0.57 to 2.8% when compared to that of the existing algorithms under speech signal from nine speakers. In addition, it reduces the Word Error Rate (WER) as 1.25 to 3.57%, Error Rate as 1.24 to 3.51% and Error as 0.57 to 2.8% when compared to that of the existing algorithms. In all aspects, the FCM clustering with EM-GMM based hybrid modeling algorithm for Tamil speech recognition provides the significant improvements for speech-to-text conversion in various applications. In future, this may be extended to unlimited speakers with large vocabularies.

Acknowledgment

The authors would like to thank the anonymous reviewers for all their valuable comments and suggestions.

References

[1] T.B.Adam, M.Salam, "Spoken English Alphabet Recognition with Mel Frequency Cepstral Coefficients and Back Propagation Neural Networks", *International Journal of Computer Applications*, vol. 42, no.12, pp. 21-27, March 2012.

[2] M.A.Al-Alaoui, L.Al-Kanj, J.Azar and E.Yaacoub, "Speech Recognition using Artificial Neural Networks and Hidden Markov Models", *IEEE Multidisciplinary Engineering Education Magazine*, vol. 3, no. 3, pp. 77-86, September 2008.

[3] J. C.Bezdek , Robert Ehrlich and William Full, "FCM: The Fuzzy c-means clustering algorithm", *Computers & Geosciences*, vol. 10, no. 2-3, pp. 191-203, 1984.

[4] S.Chattopadhyay, "A Comparative study of Fuzzy C-Means Algorithm and Entropy-based Fuzzy Clustering Algorithms", *Computing and Informatics*, vol. 30, pp.701–720, 2011.

[5] D.Chazan, R.Hoory, G.Cohen and M.Zibulski, "Speech reconstruction from mel frequency cepstral coefficients and pitch frequency," *Proc. ICASSP*, vol. 3, pp. 1299–1302, 2000.

[6] S.Davis and P.Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 4, pp. 357-366, 1980.

[7] J.R.Deller, J.H.L.Hansen and J.G.Proakis, *Discrete-Time Processing of Speech Signals*, IEEE Press, New York, 2000.

[8] S.Furui, "Speaker-independent isolated word recognition using dynamic features of speech spectrum," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 34, no. 1, pp. 52-59, 1986.

[9] G.Hemakumar and P.Punitha, " Automatic Segmentation of Kannada Speech Signal into Syllables and Sub-words: Noised and Noiseless Signals", *International Journal of Scientific & Engineering Research*, vol.5, no.1, pp. 1707- 1711, January 2014.

[10] H.Hermansky and N.Morgan,"RASTA processing of speech", *IEEE Transactions on Speech and Audio Processing*, vol. 2, no.4, pp. 578–589, 1994.

[11] H.Hermansky, "Perceptual linear predictive (PLP) analysis for speech", *Journal of Acoustic Society of America*, pp. 1738–1752,1990.

[12] T.Kinnunen, T.Kilpeläinen and P.Fränti "Comparison of Clustering Algorithms in Speaker Identification," *Proceedings of International Conference on Signal Processing and Communications (SPC 2000)*, Spain, pp. 222-227, September 2000.

[13] R.S.Kurcan, "*Isolated word recognition from in-ear microphone data using Hidden Markov Models (HMM)*", Ph.D. Thesis, March 2006.

[14] Y.Linde, A.Buzo and R.M.Gray, "An algorithm for vector quantizer design," *IEEE Transactions on Communications*, vol. 28, pp. 84-95, 1980.

[15] B.Milner and X.Shao, "Prediction of fundamental frequency and voicing from mel-frequency cepstral coefficients for unconstrained speech reconstruction," *IEEE Trans. Audio, Speech, Lang. Process*., vol. 15, no. 1, pp. 24–33, January 2007.

[16] S.Moon and J.Hwang, "Robust speech recognition based on joint model and feature space optimization of hidden Markov models," *IEEE Trans. Neural Networks*, vol. 8, pp. 194–204, March 1997.

[17] L.R.Rabiner and B-H. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[18] L.R.Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proceedings of the IEEE*, vol.77, no.2, pp. 257- 286, February 1989.

[19] L.R.Rabiner and M.R.Sambur, "An algorithm for determining the endpoints of isolated utterances," *The Bell System Technical Journal*, February 1975.

[20] M.M.Rahman and M.A.Bhuiyan, "Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches", *International Journal of Advanced Computer Science and Applications*, vol. 3, no. 11, pp. 131-138, 2012.

[21] D.A.Reynolds and R.C.Rose, "Robust Text-Independent Speaker Identification using Gaussian Mixture Speaker Models", *IEEE Transactions on Speech and Audio Processing*, vol.3, no.1, pp. 72-83, January 1995.

[22] X.Shao and B.Milner, "Clean speech reconstruction from noisy mel-frequency cepstral coefficients using a sinusoidal model," *in Proc. ICASSP*, 2003, vol. 1, pp. 704–707.

[23] R.Thangarajan , A.M.Natarajan and M.Selvam, "Syllable modeling in continuous speech recognition for Tamil language", *International Journal of Speech Technology*, vol.2, pp. 47–57, 2009.

[24] M.Vyas, "A Gaussian Mixture Model based Speech Recognition system using MATLAB", *Signal & Image Processing: An International Journal (SIPIJ)*, vol.4, no.4, pp.109-118, August 2013.

[25] G.S.Ying, C.D.Mitchell and L.H.Jamieson, "Endpoint detection of isolated utterances based on a modified Teager energy measurement," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93)*, vol. 2, pp. 732-735, April 1993.

[26] Q.Zhu and A.Alwan, "Non-linear feature extraction for robust speech recognition in stationary and non-stationary noise", *Speech Communication*, March 2003.

Mrs.M.Kalamani received her B.E. (Electronics and Communication Engineering) from Bharathiar University, Coimbatore and M.E. (Applied Electronics) from Anna University, Chennai in April 2004 and April 2009 respectively. She is currently pursuing her research in the area of Speech signal processing under Anna University, Chennai. She is presently working as Assistant Professor (Senior Grade) in the department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam. She is having 10 years of teaching experience in various engineering colleges. She has published 8 papers in International Journals, 12 papers in International and National Conferences.

Dr.S.Valarmathy received her B.E. (Electronics and Communication Engineering) and M.E. (Applied Electronics) degrees from Bharathiar University, Coimbatore in April 1989 and January 2000 respectively. She received her Ph.D. degree at Anna University, Chennai in the area of Biometrics in 2009. She is presently working as a Professor and Head of the Department of Electronics and Communication Engineering, Bannari Amman Institute of Technology, Sathyamangalam. She is having 21 years of teaching experience in various engineering colleges. Her research interest includes Biometrics, Image Processing, Soft Computing, Pattern Recognition and Neural Networks. She is the life member in ISTE and Member in Institution of Engineers. She has published 38 papers in International and National Journals, 68 papers in International and National Conferences.

Mr.M.Krishnamoorthi received his B.E.(Electrical and Electronics Engineering) from Bharathiar University, Coimbatore and M.E. (Computer Science and Engineering) from Annamalai University, Chidambaram in April 2002 and April 2004 respectively. He is currently pursuing his research in the area of Data Mining and Optimization algorithms under Anna University, Chennai. He is presently working as Assistant Professor (Senior Grade) in the department of Computer Science and Engineering, Bannari Amman Institute of Technology, Sathyamangalam. He is having 10 years of teaching experience. He has published 4 papers in International Journals, 7 papers in International and National Conferences.