I J C T A, 9(7), 2016, pp. 3391-3400

Syllable Based Tamil Language Continuous Robust Speech Recognition Using MGDF-GWCC with DNN-HMM

Sundarapandiyan S.*, Shanthi N. and Mohamed Yoonus M.*****

ABSTRACT

Speech recognition is the important concept in the field of computational linguistics which is used to recognize the translated spoken language. Even though various speech recognition systems are available in the computer-based computational filed, this paper deals with the Tamil based automatic speech recognition system. During the speech recognition process, feature extraction and classification of the particular speech are the complex tasks. So, the proposed system uses the three different models such as syllable segmentation, Acoustic model and Language model for extracting the meaningful Modified Group Delay Function with Gammatone Cepstral Coefficients related features from the audio signal in an efficient manner. Finally, the extracted features are efficiently classified by applying the Deep Neural Network with the Hidden Markov Model process. The performance of the system is implemented with the help of the IIIT-H Indic Speech database, and the efficiency is analyzed regarding the sensitivity, specificity, and recognition accuracy.

Keywords: Speech recognition, syllable segmentation, Acoustic model, Language Model, Modified Group Delay Function with Gammatone Cepstral Coefficients, Deep Neural Network with the Hidden Markov Model.

1. INTRODUCTION

Speech recognition is also called the speech to text or automatic speech recognition which is the subfield of the computational linguistic that is used in various research areas such as computer science, automatic translation, electrical engineering, multi-model interaction and so on. The speech recognition system analyzes the speaker's voice and tune because the speech is the most important communication tool between the people [1]. The speech recognition system uses the different types of language tone like Hindi, English, Sanskrit, Telugu, Tamil and so on which is used to make the speech recognition software fill the communication gap between the deaf and dumb people [2]. Even though various language based speech recognition systems are present, Tamil language based speech recognition system is playing a crucial role because it is the important Dravidian language which existence in several thousand years ago.

This language consists of various consonants and multiple rhotic which are categorized into the consonants, secondary character, vowels and rhythm [3]. The sample Tamil language based text and the related syllable is shown as follows,

அம்மா கடைக்கு சென்று இருக்கிறார்

அம்+மா கடைக்+கு சென்+று இ+ருக்+கி+றார்

[*] Assistant Professor,Department of Computer Science and Engineering, PeriyarManiammai University, Vallam.

[**] Professor, Department of Computer Science & Engineering, Nandha Engineering College, Erode.

[***] Senior Lecturer/Junior Research Officer, Linguistic Data Consortium for Indian Languages (LDC-IL), Central Institute of Indian Languages, Mysore.

Due to the enormous amount of the constants and rhotic, the spoken language is difficult to identify sometimes. So, models such as Acoustic model and Language model are used to recognize the spoken language efficiently. In which the acoustic model has analyzed the relationship between the audio signal and the phonemes in the spoken language and the language model is used to determine the probability distribution of the words present in the spoken language. Based on the relationship between the words, phoneme, and phrase the features are extracted which is used to recognize the spoken language efficient way. There are several methods such as Linear Frequency Cepstral Coefficients (LFCC) [4],Mel Frequency Cepstral Coefficients(MFCC) [5], Neural Networks, and Hidden Markov Model [6] are used to recognize the speech in the existing research work. These methods are sometimes difficult to extract the exact features from the spoken Tamil language which reduces the efficiency of the entire recognition system. So, the paper uses the Tamil spoken language to recognize the speech by utilizing the three different models like syllable segmentation, Acoustic model, and Language model. The model extracts the Modified Group Delay Function with Gammatone Cepstral Coefficients related features and those features are classified by applying the Deep Neural Network with the Hidden Markov Model process which matches the testing features with trained features.

Then the remaining of this paper organized as follows, Section 2 summarizes the related works for speech recognition system, Section 3 deals with the detailed proposed methodology and Section 4 discusses the results and Section 5 is the conclusion.

2. RELATED WORKS

This section discusses various surveys about the speech recognition processes. Lakshmi et al.,[7] implementing the speech recognition system using the syllable based continuous approach. The system segments the images into different syllable by applying the group delay segmentation method in both training and testing process. In the training process, the rules are listed based on the rules the language is automatically segmented and the information is used to recognize the testing phase. In the testing phase, the syllable boundary information is collected and mapped with the trained features. The implemented group delay based syllable segmentation approach reduces the error rate up to 20% which leads to increase the recognition accuracy effectively. Thangarajan et al., [8] implementing the acoustic syllable modeling based speech recognition system to reduce the contextual variation, agglutination and morphophonology problems. The system uses the small vocabulary model to analyze the speech by using the two different ways. In the first stage context independent syllables are analyzed and the analyzed information is transmitted to the second stage. The second stage the analyzed syllable information is integrated with the triphone modeling which replaces the mono-phones. The implemented two stage speech recognition systems reduce the error rate up to 22.76% and increase the recognition rate in an efficient way.

Radha et al.,[9] proposes the automatic Tamil speech recognition system by utilizing the multilayer feed forward neural network to increase the recognition rate. The author introduced system eliminates the noise present in the input audio signal by applying the pre-emphasis, median, average and butter worth filter and the Linear Predictive Cepstral Coefficients features are extracted from the preprocessed signal. The extracted features are classified by applying the multi-layer feed forward network which classifies the Tamil language efficiently. The performance of the system is analyzed with the help of the experimental results which reduces the error rate and increase the recognition accuracy. Alex Graves et al.,[10] recognize the speech features by applying the deep neural network because it works well for sequential data. The network works are based on the long and short term memory process to analyze the interconnection between the speech features. The extracted features are classified in terms of the connectionist temporal classification process. The implemented system reduces the error rate up to 17.7% which is analyzed using the TIMIT phoneme recognition database.

AniruddhaAdiga et al., [10] recognizes the speech from the human audio signal by utilizing the Gammatone wavelet function. The method analyzes the audio signal spatial frequency with different modulation using the Gammatone filterbank. The method extracts the audio features same as the Mel filter bank which is fed into the efficient classifier for identifying the human speech with effectively. The efficiency of the system is analyzed using the two AURORA database and the obtained result is compared with the GWCC and GCC method which achieves the low SNR and increases the recognition accuracy. Mark Galeset al., [11] implement the large vocabulary continuous speech recognition system for improving the recognition rate. The author reduces the assumptions about the particular speech features which is classified by applying the hidden Markov model. This model uses the various process like feature projection, discriminative parameter estimation, covariance modeling, adaption, normalization, multipass and noise compensation process while detecting the speech feature. The author proposed system reduces the error rate also increases the recognition rate in an effective manner. So, the proposed system uses the Tamil language while recognizes system using the hybrid feature extraction process and the extracted features are classified by applying the DNN with HMM process. The following section explains the proposed methodology in detail.

3. PROPOSED METHODOLOGY

In the modern world, the speech recognition system is used for the purposes such as controlling the user data access, biometric systems, and document classification and so on. So, in the paper has been implemented the Tamil language speech recognition system by utilizing theIIIT-H Indic Speech databases. It consists of three different stages such as syllable segmentation, feature extraction, and recognition stage. During the recognition step, the system uses both the acoustic and language model for improving the recognition rate. The proposed system block diagram is shown in the following figure 1.

The above figure 1 explains that the proposed system architecture which is explained that how the proposed system recognizes the input speech signal. It consists of two stages namely, training and testing.

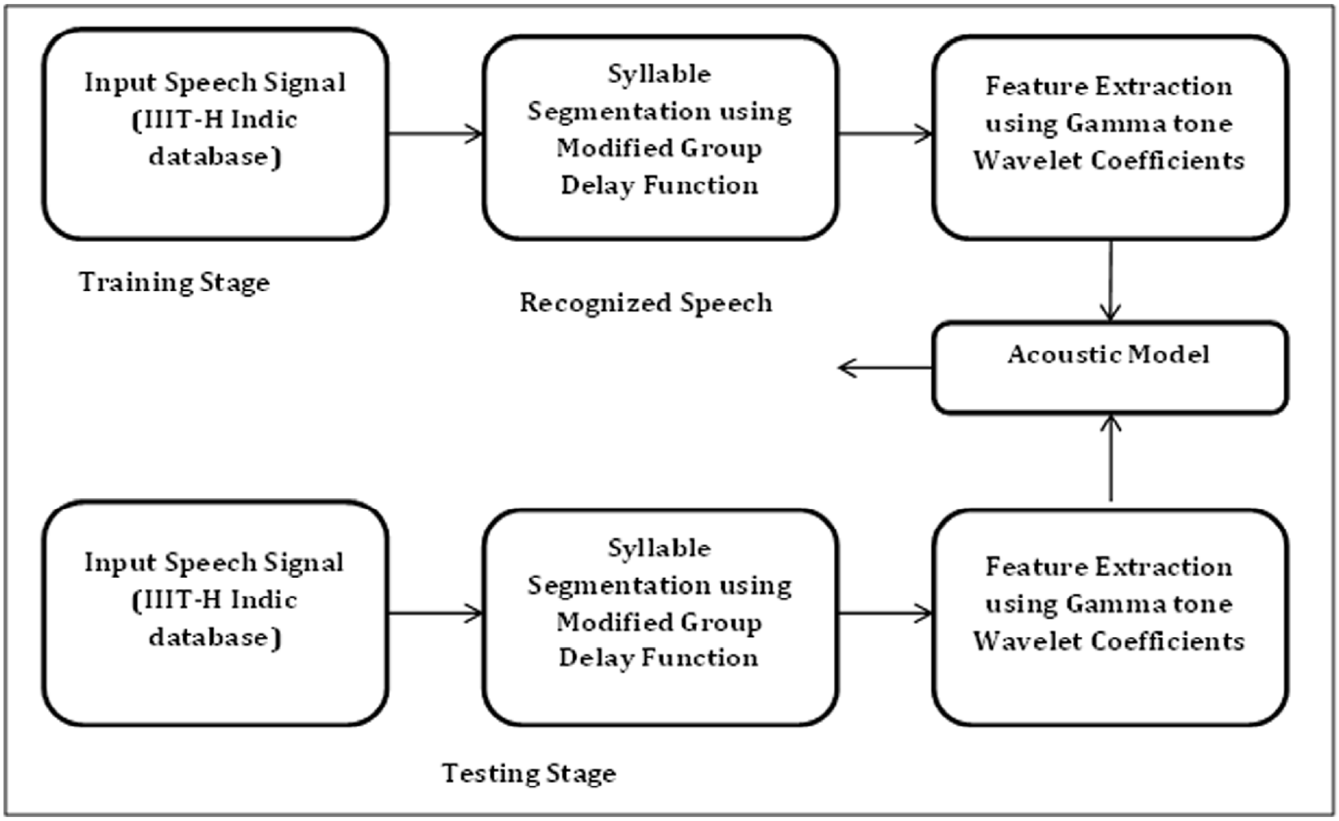
Input Speech Signal (IIIT-H Indic database)
Syllable Segmentation using Modified Group Delay Function
Feature Extraction using Gamma tone Wavelet Coefficients
Training Stage
Recognized Speech
Acoustic Model
Input Speech Signal (IIIT-H Indic database)
Syllable Segmentation using Modified Group Delay Function
Feature Extraction using Gamma tone Wavelet Coefficients
Testing Stage

Figure 1: Proposed System Architecture

In the training stage, the input is collected from IIIT-H Indic Speech databases and those signals are segmented by applying the modified group delay function. From the segmented syllable, the features are extracted using the Gammatone wavelet coefficients. The extracted features are fed into the acoustic model which constructs the model with the help of the HMM with DNN method. In the testing stage, the extracted features are fed into the acoustic model which uses the language model for matching the extracted features into the trained features performed by the Deep Neural Network with efficiently. The rest of the section discusses that the proposed speech recognition process and related methodology.

3.1. Syllable Segmentation

The first step of the speech recognition is syllable segmentation. It is the process of dividing the given speech text into the n-grams. In this paper, the syllable segmentation is performed by applying the modified group delay function. The group delay function analyzes the speech signal in the negative derivative of the phase that is used to analyze the various speech spectrum parameters. Then the group delay function [13] is defined as follows,

$$\tau(\omega) = -\frac{d(\vartheta(\omega))}{d\omega} \tag{1}$$

Where, $\vartheta(\omega)$ is represented as the unwrapped phase function of the signal. In addition the group delay function is also calculated from the speech signal as follows,

$$\tau_x(\omega) = \frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{|X(\omega)|^2} \tag{2}$$

In the above Eqn (2) R and I is represented as the real and imaginary part of the Fourier transform that is used to minimize the phase of the speech signal pole. In addition the $|X(\omega)|^2$ value becomes zero which is located to the unit circle. The group delay function spike nature has been overcome by applying the two important parameters such as γ, α. Based on these new parameters, the new modified group delay function is defined as follows,

$$\tau_m = \left(\frac{\tau(\omega)}{\tau(\omega)}\right)\left(|\tau(\omega)|\right)^{\alpha} \tag{3}$$

$$\tau(\omega) = \left(\frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^{2\gamma}}\right) \tag{4}$$

Where, $S(\omega)$ *represented as the smoothed version of* $|X(\omega)|$, and the parameters γ, α is varied from 0 to 1. Then the syllable segmentation steps have been explained as follows,

Steps for Syllable Segmentation

Step 1: Consider the x(n) is the given speech sequence

Step 2: Compute the Fourier transform of $x(n)$ and $nx(n)$ and choose $X(k)$ and $Y(k)$.

Step 3: Then estimate the cepstrally smoothed spectra of $| X(k) |$ and $S(\omega)^2$

Step 4: Apply the group delay function as follows,

$$\tau(\omega) = \left(\frac{X_R(\omega)Y_R(\omega) + Y_I(\omega)X_I(\omega)}{S(\omega)^{2\gamma}}\right)$$

Step 5: Then tune the parameters such as γ, α according to the speech environment.

Based on the above steps the syllables are extracted from the speech sequence and the features are extracted by applying the Gammatone wavelet coefficients which are discussed as follows.

3.2. Feature Extraction

The next important step is feature extraction which is done with the help of the Gammatone wavelet coefficients approach [14]. The feature extraction phase includes the various processes such as windowing, FFT, Gammatone (GT) bank filter, log and DCT approach. Initially, the segmented syllable related audio signal needs to be windowed into short frames like 10 to 50ms. During the windowing process, the non-stationary signal is considered as the stationary one for some short period of time to increase the feature extraction accuracy. Then the windowing process is done with the help of the hamming windowing process which is defined as follows,

$$Y(n) = X(n)*W(n) \tag{5}$$

Where $Y(n)$ is the output of the windowing signal, $X(n)$ is the input of the audio signal, n is the number of samples present in each frame and W is the hamming window. After windowing the signal which is fed into the FFT section. The FFT method converts each frame of N samples into frequency domain as follows,

$$Y(w) = FFT\left[h(t)*X(t)\right] = H(w)*X(w) \tag{6}$$

Then the FFT resultant value is passed to the Gammatone filterbank to extracting the meaningful sound frequencies. The GT bank filter consists of the variety of filters which is described as the Equivalent rectangular bandwidth (ERB). Then the ERB frequency of each filter is calculated as follows,

$$ERB(f) = \frac{f}{9.26} + 24.7 \tag{7}$$

After that the center frequency of the channel is computed as follows,

$$f_c(k) = -C + e\frac{k\log\left(\frac{f\min+C}{f\max+c}\right)}{K}.(f\max+C) \tag{8}$$

Where, fmin and fmax is the lowest and highest cutoff frequency of the filter bank. The estimated center frequency value is applied to Gammatone wavelet for obtaining various features from the audio signal. The computed filter output is applied to the log energy function and discrete cosine transform to obtain the human voice based features effectively. From the extracted feature the acoustic model has been created for recognizing the speech by matching the testing and training features.

3.3. Acoustic Model Creation

The next step is acoustic model creation which is done by using the deep neural network with the hidden Markov model process [15]. This method developing the acoustic model in terms of two forms such as without speaker adoption and with the speaker I vector process. The method trains the model according to the IF based and ONC based model. The extracted features are fed into the acoustic model which uses the lexicon that converts the word-level transcriptions into the mapped phones and sequence. The mapped sequence consists of independent states retrieve the mixture probability of the particular feature sequence. Based on the states the features are trained by applying the deep neural network [16]. The neural network uses the multiple hidden layers while training the features. The network analyzes the non-linear relationship between the each feature state and mapping the input feature according to the relationship because it is efficiently mapping the complex speech data patterns. During the training process, the network uses the learning rate for efficiently matching the features, so the system uses the 0.001 as the learning rate at the first epoch which is continuously updated for next 15 training epoch. For each iteration, the feature HMM hidden states have frequently been changed is used to train the feature for making the efficient acoustic model. After that, the Tamil speech signal related syllable and features are stored in the acoustic is used during the testing process.

3.4. Language Model based Recognition

The last phase is language model [17] based speech recognition which is used to identify and distinguish between the words and sounds between the similar Tamil word. The language model uses the deep neural network to recognize the speech character by using the multiple hidden layers. Initially, the sequence of words is represented as the N-dimensional sparse vector which has the index value as 1 and remaining words are 0. Each word is the layer which is mapped with the help of the linear projections. Based on the projections, the look-up table has been created the table includes the list of words, vocabularies that are used to the mapping process. When the new features are entered into the system, the neural network mapping the new entries with the look table in terms of the ith row in the feature space.The sequence feature vector is stored in the history which is concatenated with the projection layers that is used to retrieve the speech with minimum time. The collected information is summed up in the output layer for achieving the target effectively. During the target speech recognition process, the error rate has been reduced by updating the weights and bias value which is defined as follows,

$$d = \tan\ h\left(\sum_{l=1}^{(n-1)*P} M_{jlcl} + b_j\right)\forall_j = 1...H \tag{9}$$

$$o_i = \sum_{j=1}^{H} V_{ij} d_j + k_i \forall_i = 1,\ ...\ N \tag{10}$$

$$p_i = \frac{\exp(o_i)}{\sum_{r=1}^{N} \exp(o_r)} = P\left(w_j = \left\langle i \middle| h_j \right\rangle\right) \tag{11}$$

Where,

M_{jlcl} represented as the weight of the projection and hidden layer in the network,

V_{ij} represented as the weight of the hidden layers and output layer in the network

b_j is the bias of the network,

$P\left(w_j = \left\langle i \middle| h_j \right\rangle\right)$ *represented as the output layer posterior probability*

Based on the mapping process, the testing features are mapped with the trained acoustic features which efficiently retrieve the Tamil speech text with minimum computation time also it eliminates the errors present in the recognition process. Then the performance of the proposed system is analyzed using the experimental results and discussions the performance is explained as follows.

4. RESULT AND DISCUSSIONS

The speech recognition system efficiently retrieves the Tamil speech characters by segmenting the syllable present in the audio. The proposed system uses the IIIT-H Indic Speechdataset[18] for recognizing the user audio. The dataset consists of various languages such as Bengali, Hindi, Kannada,Malayalam, Marathi, Tamil, and Telugu but the proposed system only uses the Tamil language for speech recognition. The sample Tamil language related statistics is shown in Table 1.

Table 1
IIIT-H Indic based Tamil Speech text Corpus

S. No	Text Corpus	No of Sentence	No of words		No of syllables		No of phones	
			Total	Unique	Total	Unique	Total	Unique
1.	Wikipedia	99650	1888462	857850	3193292	10525	5688710	35
2	Optimal	1000	7045	2182	232284	930	42134	35

These text corpus information are collected for the duration of 1 hr and 28 minutes in an average of 5.27sec. The collected information is segmented into the different syllable and the different features such as phone type, vowel length, lip rounding, vowel height, vowel frontness, consonant voicing, cluster, aspirations and place of articulation are extracted by applying the modified global delay function with Gammatone wavelet coefficients. Those features are fed into the next DNN with HMM process for developing the efficient acoustic model which is used in the time of testing. Finally, the Tamil speech has been recognized by utilizing the deep neural network which is done with the help of the mapping process. Then the performance of the system is evaluated with the help of Error Rate, recognition time, recognition accuracy, sensitivity and specificity measures.

4.1. Error Rate

The error rate is the important measure while analyzing the performance of the proposed system. The proposed system analyzes the speech features with the minimum error because of using the efficient trained acoustic model based features. The error rate during the performance is graphically represented is figure 2.

The above figure 2 clearly shows that the proposed system has the minimum error rate while matching the testing features (language model) with the trained features (acoustic model). The minimum error rate shows that the proposed system eliminates the unwanted speech features and also it increases the recognition accuracy which is analyzed by using the sensitivity, specificity and accuracy metrics. The sensitivity and specificity value are calculated as follows,

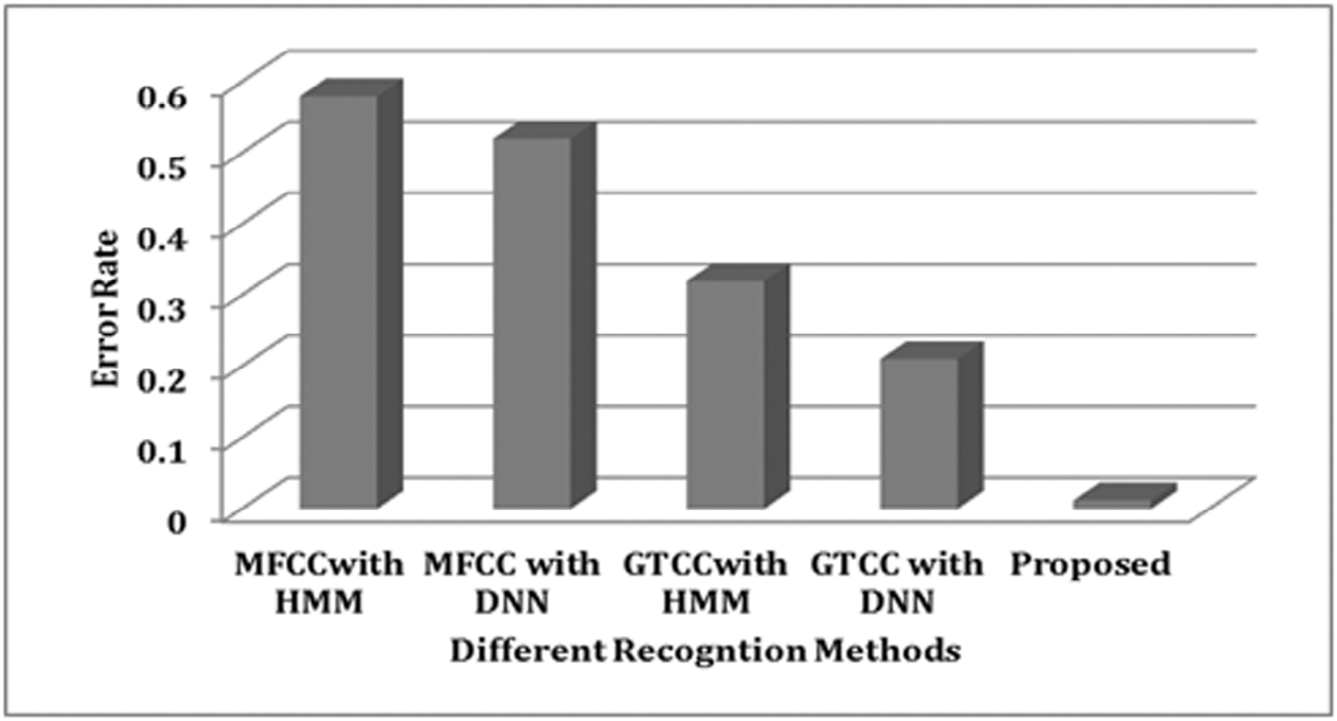

Figure 2: Graph representation of Error Rate

$$\text{Sensitivity} = TP/((TP + FN)) \tag{12}$$

$$\text{Specificity} = TN/((TN + FP)) \tag{13}$$

Where, TP = True Positive, TN = True Negative

FP = False Positive, FN = False Negative.

The following figure 3 shows that the sensitivity and the specificity value of different speech recognition methods such as MFCC with HMM [19], MFCC with DNN [20], GTCC with HMM [21]and GTCC with DNN [22].

From the above figure 3, it is easy to identify that the proposed system efficiently recognizes the Tamil speech features which are shown in by using the sensitivity and specificity. The increase performance metrics leads to improve the recognition accuracy which is shown in the table 2.

The above table 2, clearly shows that the proposed system efficiently recognizes the Tamil speech up to 98.3% compared to the other existing methods. Also, the proposed method recognizes the speech with minimum time because of using the language model and the performance is analyzed using the following figure 4.

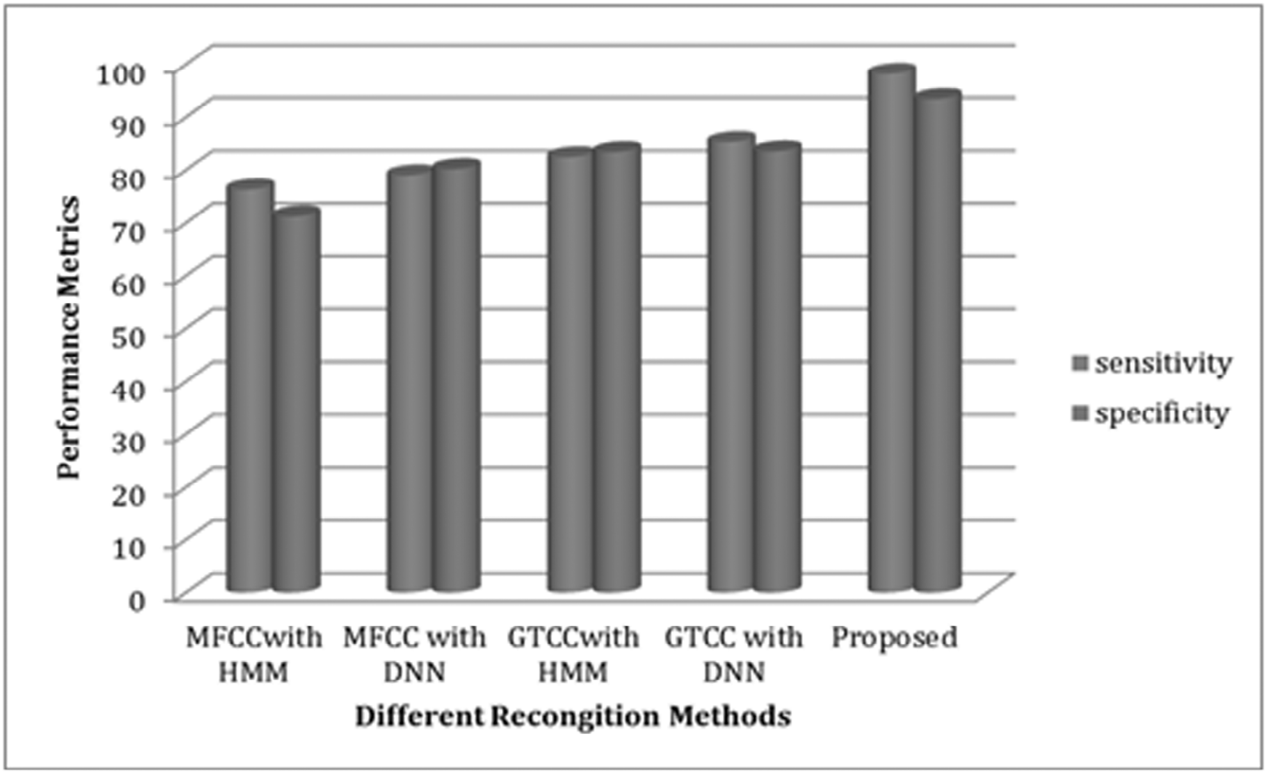

Figure 3: Sensitivity and Specificity

Table 2
Speech Recognition Accuracy

S. No	*Recognition Techniques*	*Recognition Accuracy (%)*
1	MFCC-HMM	85
2	MFCC-DNN	82.2
3	GTCC- HMM	85.6
4	GTCC-DNN	88.32
5.	MGDF-GWCC WITH DNN-HMM	98.3

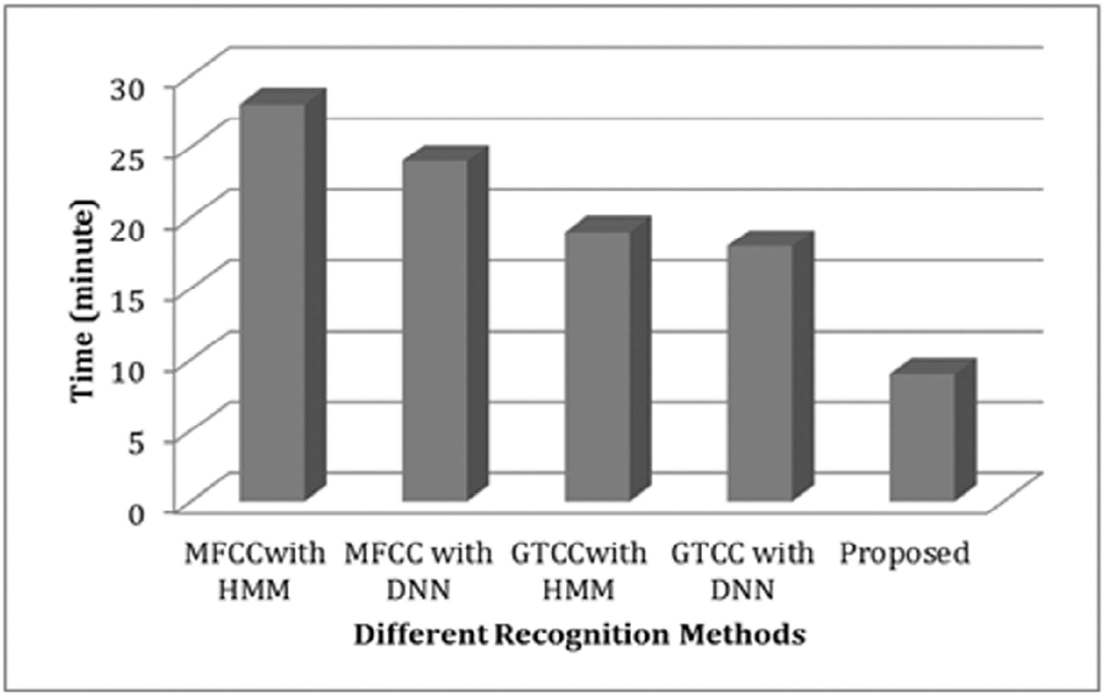

Figure 4: Graph Representation of Recognition Time

The above discussions clearly show that the proposed system efficiently recognizes the Tamil speech with minimum time. Thus the proposed modified global delay function with Gammatone wavelet coefficient based features are effectively trained and matched with language model which has improved the overall recognition accuracy.

5. CONCLUSION

Thus the paper implemented the Tamil speech recognition system by utilizing the three different models such as syllable segmentation, acoustic model creation and language model on the IIIT-H Indic speech database. The collected Tamil speech text or audio is segmented by applying the modified global delay function which segments the words into the syllable and the efficient Gammatone based features are extracted from the syllable. Those extracted features are formed as the acoustic model by using the DNN-HMM which considered each feature as the independent state. Based on the state the features are trained and stored as the acoustic model. During the testing process, the non-linear relationship between the features is estimated and formed the look-up table. From the look-up table, the testing and training features are mapped and the related Tamil speech is recognized. Finally, the performance of the system is analyzed using the experimental results and the proposed system recognizes the character with minimum complexity, time and high accuracy.

REFERENCE

[1] Sivaranjani, B. Bharathi, “Syllable Based Continuous Speech RecognitionFor Tamil Language”, International Journal of Advanced Engineering Technology, 2016.

[2] Sigappi and S. Palanivel, “Spoken Word Recognition Strategy for Tamil Language”, International Journal of Computer Science Issues, Vol. 9, Issue 1, No 3, January 2012.

[3] Hanitha Gnanathesigar, “Tamil Speech Recognition using Semi-Continuous Models”, International Journal of Scientific and Research Publications, Volume 2, Issue 6, June 2012.

[4] TomyslavSledevic, Ž Artu¯rasSerackis, GintautasTamulevici Ž us, DaliusNavakauskas, “Evaluation of Features Extraction Algorithms for a Real-Time Isolated Word Recognition System”, World Academy of Science, Engineering and TechnologyInternational Journal of Electrical, Computer, Energetic, Electronic and Communication Engineering Vol: 7, No: 12, 2013.

[5] Rajesh M.Hegde, HemaA.Murthy and VenkataRamanaRaoGadde, " Application of the Modified GroupDelay Function to Speaker Identification and Discrimination,"in Proceedings of the ICASSP, SP-P6.4, May2004.

[6] Tomas Mikolov, Stefan Kombrink, Lukas Burget, JanCernocky, and SanjeevKhudanpur.2011b.Extensions of recurrent neural network language model.InProc. ICASSP 2011, pages 5528–5531.

[7] Lakshmi A, Hema A Murthy "A New Approach to Continuous Speech Recognition in Indian Languages", available at., www.ncc.org.in/download.php?f=NCC2008/2008_D3_5.pdf.

[8] Thangarajan , A. M. Natarajan, M. Selvam, "Syllable modeling in continuous speech recognition for Tamil language", International Journal of Speech TechnologyMarch 2009.

[9] Radha, Vimala, M.Krishnaveni, "Isolated Word Recognition System ForTamil Spoken Language Using BackPropagation Neural Network Based OnLpcc Features", An International Journal (CSEIJ), Vol.1, No.4, October 2011.

[10] Alex Graves, Abdel-rahman Mohamed and Geoffrey Hinton, "Speech Recognition With Deep Recurrent Neural Networks", http://www.cs.toronto.edu/~fritz/absps/RNN13.pdf.

[11] AniruddhaAdiga, Mathew Magimai ; Chandra SekharSeelamantula, "Gammatone wavelet Cepstral Coefficients for robust speech recognition", TENCON 2013-2013 IEEE Region 10 Conference (31194).

[12] Mark Gales and Steve Young, "The Application of Hidden Markov Models in Speech Recognition", Foundations and TrendsinSignal Processing, Vol. 1, No. 3, 2007.

[13] Prasad, V. K., Nagarajan, T., and Murthy, H. A., "Automatic segmentation of continuous speech using minimum phase group delay functions,in Speech Communication, vol. 42, Apr. 2004, pp. 1883–1886.

[14] Lakshmi, A, "A Syllable-based Continuous Speech Recognizer for Indian Languages," MS Thesis, Indian Institute of Technology, Department of Computer Science and Engg., Madras, India, August 2007.

[15] Chunsheng Fang, From Dynamic time warping (DTW) to Hidden Markov Model (HMM), University of Cincinnati,2009.

[16] Suma Swamy and K.V Ramakrishna, " An Efficient Speech Recognition System ", An International Journal (CSEIJ), Vol. 3, No. 4, pp 21-27, August2013.

[17] Saraswathi,Geetha, "Design of language models at various phases of Tamil speech recognition system",International Journal of Engineering, Science and Technology Vol. 2, No. 5, 2010.

[18] Kishore Prahallad, E.Naresh Kumar, Venkatesh Keri, S.Rajendran, Alan W Black, "The IIIT-H Indic Speech Databases".

[19] Dalmiya C, Dr. Dharun V, Rajesh K, "An Efficient Method for Tamil Speech Recognition using MFCC and DTW", IEEE Conference on Information and Communication Technologies (ICT), pp 1263-1268,2013.

[20] V.Kamakshi Prasad, T. Nagarajan and HemaA.Murthy," Continuous speech recognition using automatically segmented data at syllabic units," in Proceedings of the Sixth International Conference on Signal Processing,ICSP, pp. 235-238, August 2002

[21] Q. Li and Y. Huang, "An auditory-based feature extraction algorithm for robust speaker identification under mismatched conditions," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19,no. 6, pp. 1791–1801, 2011.

[22] R. Schluter, L. Bezrukov, H. Wagner, and H. Ney, "Gammatone features and feature combination for large vocabulary speech recognition," in IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007., vol. 4, April 2007, pp. IV–649 –IV–652.