Journal of Engineering Science and Technology
Vol. 10, No. 5 (2015) 617 - 632
© School of Engineering, Taylor's University

ISOLATED SPEECH RECOGNITION SYSTEM FOR TAMIL LANGUAGE USING STATISTICAL PATTERN MATCHING AND MACHINE LEARNING TECHNIQUES

VIMALA C.*, RADHA V.

Department of Computer Science, Avinashilingam Institute for
Home Science and Higher Education for Women, Coimbatore, India
*Corresponding Author: vimalac.au@gmail.com

Abstract

In recent years, speech technology has become a vital part of our daily lives. Various techniques have been proposed for developing Automatic Speech Recognition (ASR) system and have achieved great success in many applications. Among them, Template Matching techniques like Dynamic Time Warping (DTW), Statistical Pattern Matching techniques such as Hidden Markov Model (HMM) and Gaussian Mixture Models (GMM), Machine Learning techniques such as Neural Networks (NN), Support Vector Machine (SVM), and Decision Trees (DT) are most popular. The main objective of this paper is to design and develop a speaker-independent isolated speech recognition system for Tamil language using the above speech recognition techniques. The background of ASR system, the steps involved in ASR, merits and demerits of the conventional and machine learning algorithms and the observations made based on the experiments are presented in this paper. For the above developed system, highest word recognition accuracy is achieved with HMM technique. It offered 100% accuracy during training process and 97.92% for testing process.

Keywords: Tamil speech recognition, Dynamic time warping, Hidden Markov model, Gaussian mixture models, Support vector machine (SVM), Neural networks, Decision trees.

1. Introduction

Automatic Speech Recognition is used to convert the spoken words into written text. It usually involves the extraction of patterns from digitized speech samples and representing them using an appropriate data model. These speech patterns are subsequently compared to each other using mathematical operations to determine their contents [1]. ASR has been applied in many different sectors namely medical, military, education, telephony and commercial applications. Apart from these benefits, speech technology mainly offers its major contribution to help the differently abled people to access their computer and internet. These speech recognition systems can be categorized into various types based on the type of utterance, speaker model, channel, vocabulary size and environment for which it is developed. Developing speech recognition systems are becoming more complex and it is a challenging task because of this variability [2]. The types of speech recognition systems are explained in Fig. 1.

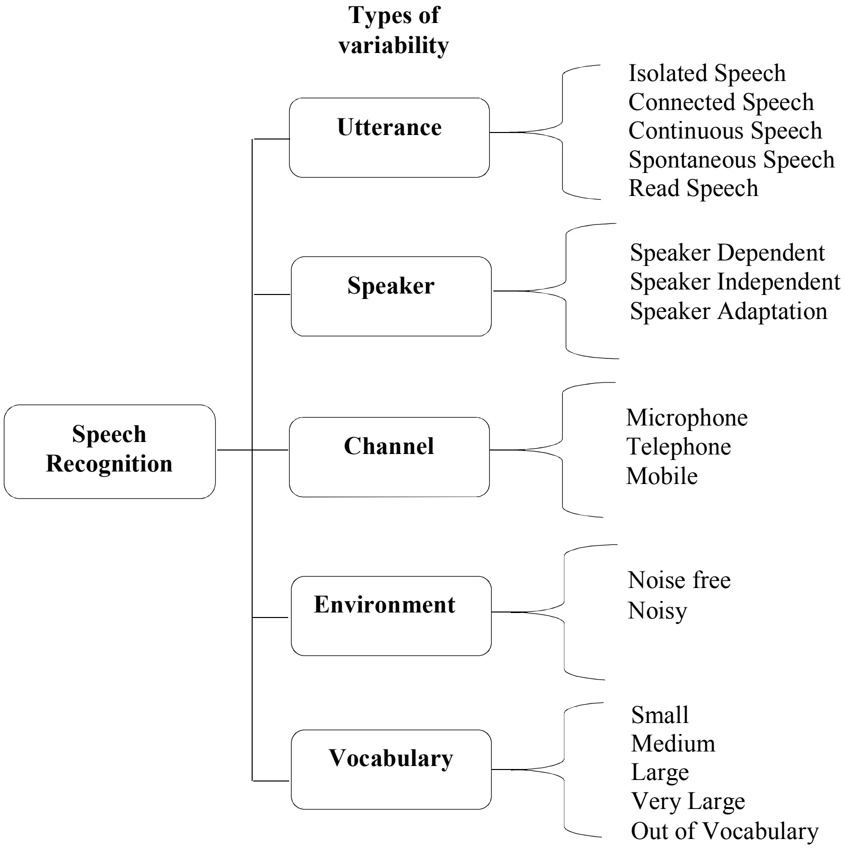

Fig. 1. Types of Speech Recognition.

In general, small or medium vocabulary speaker dependent isolated speech recognition systems are comparatively simple and easy to implement. They can achieve more accuracy but lacks in offering flexibility when compared with speaker adaptive or speaker independent systems. In contrast, large vocabulary speaker independent continuous speech recognition systems recognize continuous speech patterns from a large vocabulary and from a large group of people. This kind of system is most difficult to implement and achieves quite less accuracy compared to speaker dependent isolated speech recognition systems [3]. These systems offer more flexibility, but when speaker and vocabulary size grows larger, the confusability between different speech patterns also grows larger. Apart from the above characteristics, the environmental variability, channel variability, speaking style, gender, age, rate of speech also makes the ASR system more complex [2]. For several years, lots of research efforts have been put into practice to build an efficient ASR system that works irrespective of this variability. The main aim of this research work is to develop a system for recognizing Tamil Spoken words. ASR for Tamil language is still at the early stage. Among the few attempts on Tamil ASR most of the experiments are based on HMM and NN. Hence the particular contribution of this paper is involving GMM and other machine learning techniques like SVM and Decision tree algorithms. The following section discusses some related works on Tamil Speech recognition.

2. Related Work

B. Bharathi et al. have proposed a neural network approach to build a speaker independent isolated word recognition system for Tamil language [4]. Sequence of steps were implemented in order to improve the performance of a system namely de-noising using wavelet transform, silence removal using energy values and number of zero crossings. Next, Mel Frequency Cepstral Coefficients (MFCC) feature vectors were extracted and they are normalized by Cepstral Mean Normalization (CMN) to reduce specific variation between different people. The author have used Self Organizing Map (SOM) neural network to make each variable length MFCC trajectory of an input into a fixed length MFCC trajectory. Finally feed forward neural network is used to recognize the spoken words. Their system has achieved 90% accuracy for recognizing 10 spoken digits from 10 speakers.

Sigappi et al. have proposed an approach for recognizing the preferred vocabulary of words uttered in Tamil language [5]. They have chosen HMM and Auto Associative Neural Networks (AANN) along with the MFCC features. The HMM has been employed to model the temporal nature of speech and the AANNs to capture the distribution of feature vectors in the feature space. The developed model has provided a way to examine an uncharted speech recognition arena for the Tamil language. The dataset consists of 100 railway station names in TamilNadu. Each station names are uttered five times by 20 speakers (total=100*20*5=10000). Out of 10000 samples, 8000 samples are used for training purpose and the remaining 2000 samples were used for testing. Their proposed system has offered 95% for HMM and 90% for AANN.

Rojathai et al., have proposed a Novel speech word recognition system for Tamil language [6]. In their work, FFBNN (Feed Forward Back Propagation Neural Network) is trained with the MFCC features and this trained FFBNN is further used for testing new feature vectors extracted from new input signals. The dataset consists of ten Tamil words from ten people out of which eight person’s data are utilized for the training process and the remaining two person’s data are utilized for the testing process. The authors have compared their proposed work with HMM speech word recognition system proposed by Sigappi et al. They strongly state that the recognition rate of the proposed system has achieved high accuracy compared to the existing technique. The accuracy, sensitivity, and specificity values of proposed technique are 90%, 90% and 90.5% respectively where the existing system results has achieved 61.11%,79.5%,82.7% for their work. The next section describes the basic steps involved in developing an ASR system.

3. Basic Steps Involved in Developing an ASR System

Making a speech signal understandable by the computer and converting them into a written data consists of several steps. Figure 2 shows the steps involved in the development of ASR system.

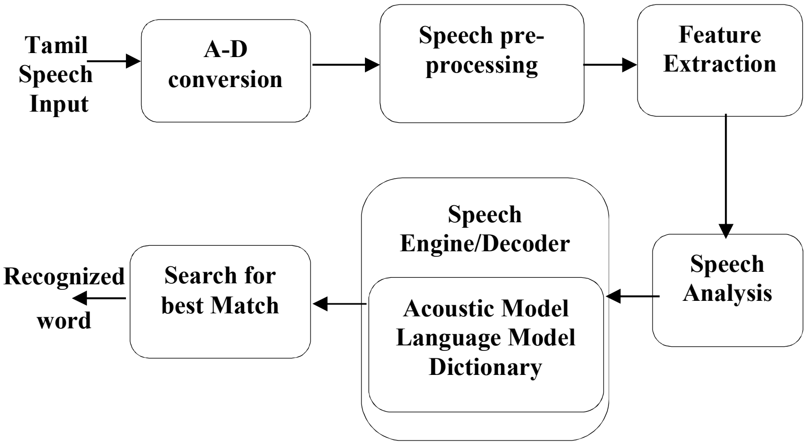

Fig. 2. Steps Involved in the Development of an ASR System.

The first step is digitization where the analog to digital converter translates analog signals into digital signals for the computer to process further. Next, preprocessing step is carried out to check for better frequency levels and to produce the optimal input speech signal for the recognition system. Then the speech signal is divided into a smaller number of frames in order to characterize the information reside in the signal in compact form. In every 10 ms the useful feature vectors are extracted from these smaller windows for feature matching and classification. Since the feature extraction is an important task to produce a better recognition performance it must be chosen carefully. Hence, in this research work, the most popular MFCC technique is adopted for the best parametric representation of acoustic signals. After that, the speech engine will develop a search/decoder for comparing the phonemes/words with the model developed and trained with these feature vectors.

The speech engine consists of three important components to perform the recognition task namely, acoustic model, language model and dictionary. Finally, the most probable word sequence can be identified as the recognized text based on the knowledge acquired from these components. As mentioned in the previous section, number of techniques was proposed for developing an efficient speech recognition system and they are explained below.

4. Speech Recognition Techniques

The ambitious speech research for more than 50 years is having a machine to fluently understand the spoken speech. For this purpose, various techniques were proposed and successfully applied in many significant areas. Initially, the dynamic programming techniques have been proposed for spoken word recognition based on template matching approach. Subsequent researches were focused on statistical pattern matching approach such as HMM and GMM. These statistical methods use probability distribution density and build a model with the entire data. So, it will have the knowledge and complete description of the actual problems [7]. Next to these approaches, the machine learning techniques like Artificial Neural Networks(ANNs), SVM, Deep Belief Networks (DBNs) are proposed to replace the conventional HMM/GMM systems [8]. Recent researches are focusing on hybrid techniques to combine the potent metrics of these methods for increased accuracy. The techniques adopted for this work are briefly explained below.

4.1. Dynamic time warping for speech recognition

One of the oldest and most important algorithms in speech recognition is Dynamic Time Warping (DTW) technique. DTW approach was originally developed to compare different speech patterns which are applied in speech signal processing related tasks. It uses a very simple concept to recognize a word sample by comparing it against a number of stored word templates and determines the best match. For this purpose, initially normalization will be carried out for speech template since different samples of a given word will have somewhat different durations. Since the voice signal consists of different temporal rate, the alignment is important to produce the better performance. DTW is an efficient method for finding this optimal nonlinear alignment so that it computes distance between the given feature vectors effectively. Then it finds an optimal alignment between two given time-dependent sequences which are nonlinear. The algorithm travels through a matrix of frame scores once and computes the locally optimized segments of the global alignment path. Generally, the process consists of two phases and the algorithm is explained below [9]:

Training Phase:

Generate the speech signal S_w for each word *W* in the vocabulary

Extract feature vectors for each word *W* for creating templates (*A*'w)

Testing Phase:

Generate feature vectors for given test speech signal $S_{test} \rightarrow S'_{test}$

Compute the distance between each test word *W* and template feature vectors using DTW using Eq. (1).

$$W^{*} = \underset{w \in vocab}{argmindistance}(A'_{test}, A'_{w}) \tag{1}$$

Finally, find the minimum distance for a given test word *W*.

For finding the distance between the signals, the Euclidean distance measurement has been used. It is simply the sum of the squared distances from each nth point in one time series to the nth point in the other [10]. This technique has achieved a reasonable success in digit recognition tasks. The major advantages and disadvantages of DTW are concluded below [8, 11].

Advantage:

- Easy to implement and it can model arbitrary time warping

Disadvantage:

- Distance measures and warping paths are heuristic,
- No guarantees of optimality or convergence,
- Doesn’t scale well for large vocabulary,
- Doesn’t support mix and match between templates, and
- Works poor if environment changes.

To overcome these issues, the statistical approaches were proposed and they are briefly discussed in the subsequent sections.

4.2. Statistical pattern matching approach for speech recognition

The solution to the issues mentioned in the template matching techniques is, to make the algorithm to learn as much as possible from data. It can be achieved by using probabilistic modeling since it applies well described theories and models from probability, statistics, and computer science [8, 11]. The most popular and widely used statistical probabilistic modeling techniques for speech recognition are explained below.

4.2.1. Hidden Markov model (HMM)

The most dominant technique for speech recognition based on statistical acoustic and language model is HMM. It uses the automatic learning procedures to model the variations of speech. HMMs are simple networks that can generate speech (sequences of cepstral vectors) using a number of states for each model [12]. Modeling the short-term spectra associated with each state usually mixtures of multivariate Gaussian distributions (the state output distributions).The parameters of the model are, the state transition probabilities, means, variances and mixture weights that characterize the state output distributions [13]. Each word, or each phoneme, will have a different output distribution. A HMM for a sequence of words or phonemes is made by concatenating the individual trained HMM for the separate words and phonemes [13].

It works by starting at upper left corner of trellis and generate observations according to permissible transitions and output probabilities. The output and transition probabilities define a HMM. This algorithm not only can compute likelihood of single path and can compute overall likelihood of observation string as sum of overall paths in trellis. The HMM addresses three problems to perform ASR and the solutions are given by the following three algorithms.

Three key problems

i. Computing the overall likelihood generating strings of observations from HMM by using the *Forward algorithm.*

ii. Decoding the most likely state sequence/best path from HMM by using the *Viterbi algorithm.*

iii. Learning parameters (output and transition probabilities) of HMM from data by using Baum-Welch also called *Forward-Backward algorithm*.

The advantages and disadvantages of HMM are clearly explained below [12].

Advantages:

- Easily extendable, because each HMM uses only positive data so they scale well.
- Extremely reduce the time and complexity of recognition process for training large vocabulary.

Disadvantages:

- Must make a large priori modeling assumptions about the data,
- The number of parameters that need to be set in an HMM is huge,
- Amount of data that is required to train an HMM is very large, and
- It does not minimize the probability of observation of instances from other classes.

Like HMM, the other technique which is efficiently used for many pattern matching tasks is GMM. The basic concepts of GMM and its merits and demerits are discussed here.

4.2.2. Gaussian mixture model (GMM)

GMM is commonly used in pattern matching problems since it involves an efficient mathematical straightforward analysis with a series of good computational properties [7]. GMM is a mixture of several Gaussian distributions and can therefore represent different subclasses inside one class [14, 15]. It is a weighted sum of Gaussian probability density functions which are referred to as Gaussian components of the mixture model describing a class. The probability density function is defined as a weighted sum of Gaussians using Eq. (2).

$$p(x;\theta) = \sum_{c=1}^{c} \alpha_c N(x;\mu_c,\Sigma_c) \tag{2}$$

where α_c is the weight of the component c, $0 < \alpha_c < 1$ for all components, and $\sum_{c=1}^{c} \alpha_c = 1$ .The parameter list $\theta = \{\alpha_1, \mu_1, \Sigma_1, ...\alpha_C, \mu_C, \Sigma_C\}$ defines a particular Gaussian mixture probability density function [12]. For speech recognition, the GMM makes a probabilistic model of feature vectors associated with a speech sound. It works based on the principled distance between test frame and a set of template frames. The GMM algorithm for speech recognition is explained below [12]:

1. P(X) defines a GMM distribution over individual feature vectors X

2. To find out P ($A'=X_1, X_2 \ldots\ldots\ldots X_T$) make distribution over sequences of feature vectors

3. Make probabilistic model of training samples using Eq. (3).

$$distance\,T_X T_Y(X,Y) = \sum_{t=1}^{T} framedist(X_{TX(t)}, Y_{TY(t)}) \qquad (3)$$

4. Build a separate model P (A'|W) for each word W

5. Find the maximum probability of test signal from the model using Eq. (4)

$$W^* = \underset{w \in vocab}{argmax}\, P(A'_{test} \mid W) \qquad (4)$$

The model can use Expectation Maximization (EM) algorithm to do Maximum Likelihood (ML) estimation. EM algorithm calculates the maximum likelihood distribution parameter from the training data based on some iterative process [8].

Advantage:

- It can estimate probability perfectly and it can perform classification optimally.

Disadvantage:

- Time complexity

In order to overcome the issues mentioned for the above techniques, the machine learning approaches are developed which are described below.

4.3. Machine learning approach for speech recognition

The most popular approach which is recently applied in speech technology is Machine Learning. The approach combines the study of pattern recognition with the machine's ability to analyze, learn and make a decision accordingly [16]. Several methods exist for this task such as Artificial Neural Networks, SVM, Decision Trees and combination methods. The performance of these algorithms can vary according to the task for which it is applied.

4.3.1. Neural Networks for speech recognition

During the last two decades some alternative approaches to HMMs and GMMs have been proposed which are mostly based on ANNs. Generally, ANNs are represented as an important class of discriminative techniques, which are very well suited for classification problems [17]. In this paper, MLP is adopted and successfully applied for Tamil speech recognition.

Multi-layer perceptron (MLP)

MLP is a modification of the standard linear perceptron and it is a feed forward artificial neural network model which maps set of input data onto a set of suitable output data. The MLP contains multiple layers of nodes in a directed graph, with each layer fully connected to the subsequent layer where learning will be done based on these patterns. The data travels in one direction and the output of one layer will be given as an input to the next layer [18]. Here, each node is represented as a neuron or processing element with a nonlinear activation function other than an input node [18]. These interconnected groups of artificial neurons are used for computation. It utilizes a supervised learning technique called back propagation for training the network and the algorithm is given below [4].

```
1. Initialize the input layer: $y_0 = x$
2. Propagate activity forward: for l = 1, 2... L, $yl = f_l (w_l y_{l-1} + b_l)$,
   Where $b_l$ is the vector of bias weights.
3. Calculate the error in the output layer: $\delta_L = t - y_L$
4. Back propagate the error: for l = L-1, L-2,.. 1, $\delta_L = (w_{l+1}^T \delta_{l+1}). f_l^1$ (net l)
   Where $T$ is the matrix transposition operator.
5. Update the weights and biases: $\Delta W_l = \delta_l y_{l-}$ ; $\Delta bl = \delta l$
```

MLP can distinguish data that are not linearly separable. It can be defined as follows: The MLP contains *c* classes $w_1,\ldots.w_c$, and a supervised training sample $T = \{(x_i, w(x_i)) \mid i = 1, . . . ,N\}$, where $w(x_i)$ denotes the class which pattern x_i belongs to. In this work, different network parameters are tested in order to improve the accuracy. Here, the learning rate is set to 0.3 and the momentum rate is set to 0.2 and the number of epochs is set to 500. It is found that, better accuracy is obtained when the neurons in hidden layer are used according to the attributes and the number of class involved in the experiments. In this work, MLP technique has achieved satisfactory results. The major advantages and disadvantages of NN are discussed here.

Advantages:

- It can learn according to discriminative criteria,
- It can approximate any continuous function with a simple structure,
- Do not require strong assumptions about the input data, and
- It produces reasonable outputs for inputs which have not been taught before how to deal with.

Disadvantages:

- Does not perform well for more complex tasks as continuous speech recognition, and
- Inability to build speech model even though the recurrent structures are defined.

4.3.2. Support vector machine (SVM)

SVMs are theoretically well motivated algorithm developed from Statistical Learning theory. The main goal of SVM is to produce a model based on the training data which predicts the target values of the test data using kernel Adatron algorithm [19]. The data analysis and pattern recognitions are performed using a collection of learning methods. The important merit of SVM is it constructs a set of hyper planes in a high dimensional space for classification tasks. Here, a good separation is achieved by the hyper plane which has the largest distance to the nearest training data points of any class. In general, larger the margin, lower the generalization error of the classifier [20]. In accordance with the above mentioned merits, SVMs are implemented in this paper. The advantage of using SVM is, it globally replaces all missing values and transforms nominal attributes into binary ones and also applies normalization to all values. Since ASR needs involving multiclass SVM, one Vs one type is chosen to test against all the other classes separately. The size of the training set needed for each SVM in one-Vs-one solution leads to a smaller computational effort with comparable accuracy rates. The significant factor in using SVM is choosing a good kernel function. In this work, the polynomial kernel of first order is used. The cost parameter value is set as 1. For this work, SVM has offered considerable results like HMM and MLP. SVMs have several advantages which are listed below:

Advantages:

- Minimize the structural risk which results in better generalization ability,
- Increase the robustness of the system,
- Training is relatively easy,
- Local optimality is not needed, and
- It scales relatively well for high dimensional data.

Disadvantage:

- Good kernel function is needed.

4.3.3. Decision trees

The Decision Tree is one of the most popular classification algorithms which are currently used in Data Mining and Machine Learning tasks. A decision tree works based on the hierarchical manner where the classification procedure determined by the sequence of questions. The first question will be asked initially and the subsequent sections operate depends on the previous answer. This can be represented by a directed graph known as a tree. Here, “Memory” is a tree of rules about the features for which it need an algorithm to learn the rules. It begins by applying the rules to an unforeseen feature vector from the root node. It travels through the root node and the leaf nodes, then it predicts the new object belongs to the class represented by the leaf node [21]. Recently, some attempts have been made to apply the decision trees for speech recognition tasks; hence it is applied in this paper. This technique has offered less accuracy when compared with other machine learning algorithms involved in this work. The merits and demerits of decision trees are given below [22].

Advantages:

- Easy to understand and interpret, and
- Can be combined with other decision techniques.

Disadvantage:

- Complexity increases when data size is increased.

All the above techniques were successfully implemented for Tamil speech recognition. The experimental results are briefly given in the following sections.

5. Experimental Results

The main objective of this paper is to develop speaker independent isolated speech recognition for Tamil language. These speech recognition systems were originally developed for English language. Tamil is a widely spoken language in India, and research done in the area of Tamil Speech Recognition is limited when compared to other similar languages. The experiment is done with 10 Tamil spoken digits (0-9) and 5 spoken names from 4 different speakers. The utterances consist of 10 repetitions from one male and three females between the age group of 20 to 35. The total size of the dataset is 15×4×10=600. To make the utterance variation, the speakers uttered the same word at different interval of time. The utterances were recorded at 16 kHz using audacity software at a silence environment and the experiments are done using MATLAB software. The same speech samples were used for different experiments with different algorithms. In this experiment, the dataset is divided into training and testing data, where 60% data is given for training and the remaining 40% data are given for testing. Here, same speaker's dataset are used for training and testing. The experiments are carried out for some significant findings about the performance evaluations of conventional techniques and machine learning techniques for Isolated Tamil speech recognition.

Figures 3 and 4 show the recognition accuracy achieved for the training and testing data respectively (s1, s2, s3 and s4 refers speaker 1, 2, 3 and 4). Since DTW does not need training, the experiments are shown only for testing process. Tables 1 and 2 illustrate the average word level accuracy achieved during training and testing process respectively. The highest word recognition accuracy achieved for the system is highlighted. The performance of the ASR system are measured based on the Word Error Rate (WER), Word Recognition Rate (WRR) and Real Time Factor. The equations are given below in Eqs.5-7 respectively.

$$Word\ Error\ Rate = \frac{Insertion(I) + Substituti\ on(S) + Deletion(D)}{Number\ of\ Reference\ Words(N)} \tag{5}$$

$$Word\ Recognition\ Rate = \frac{N - I - S - D}{N} \tag{6}$$

Table 3 illustrates the average time taken for training and testing the above system for all the speakers. The speed of a speech recognition system is commonly measured in terms of Real Time Factor. It takes time *P* to process an input of duration *I*. It is defined by the formula (7).

$$RTF \frac{P}{I} \tag{7}$$

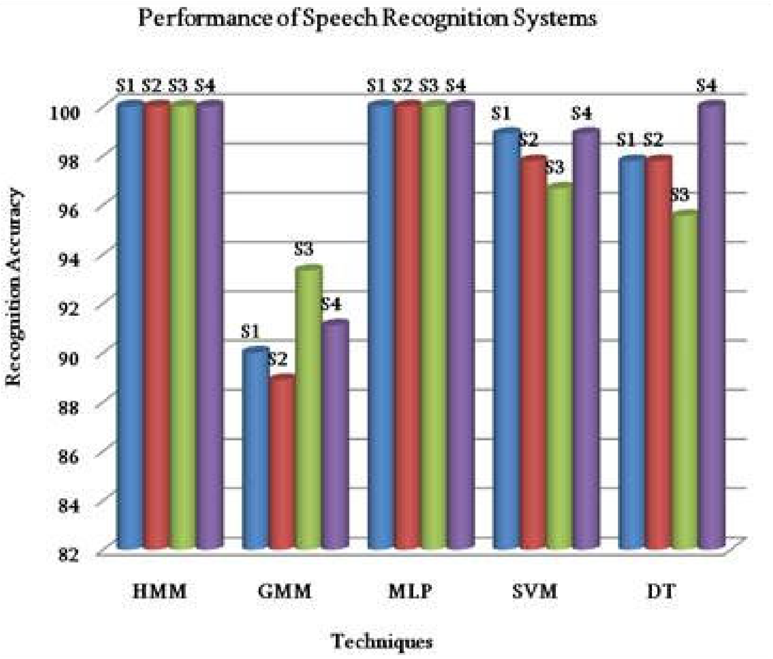

Fig. 3. Training Accuracy.

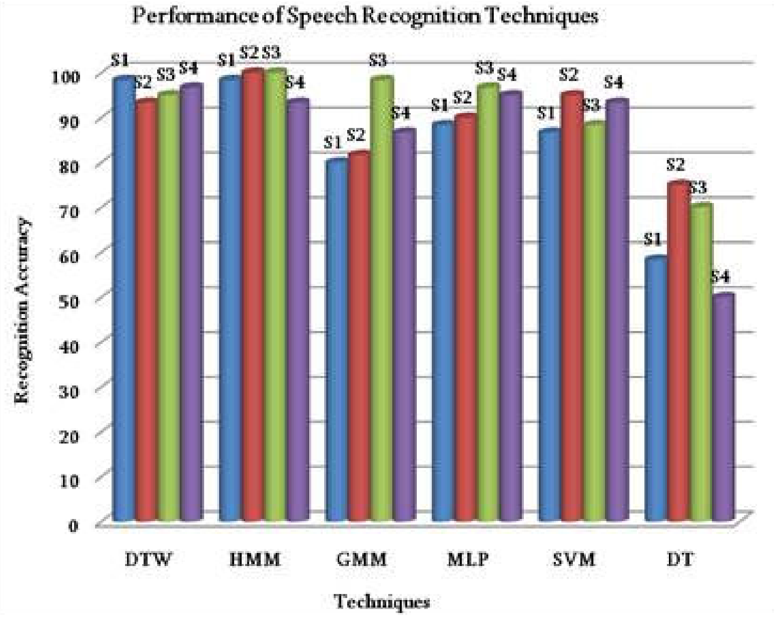

Fig. 4. Testing Accuracy.

Table 1. Average Word Level Accuracy Achieved for Training Data.

Words	HMM	GMM	MLP	SVM	DT
D0	**100**	**100**	**100**	95.83	**100**
D1	**100**	87.50	**100**	**100**	**100**
D2	**100**	91.67	**100**	**100**	91.67
D3	**100**	66.67	**100**	95.83	95.83
D4	**100**	83.33	**100**	**100**	**100**
D5	**100**	95.83	**100**	95.83	95.83
D6	**100**	**100**	**100**	**100**	**100**
D7	**100**	**100**	**100**	**100**	**100**
D8	**100**	**100**	**100**	**100**	**100**
D9	**100**	79.17	**100**	**100**	95.83
W1	**100**	**100**	**100**	95.83	95.83
W2	**100**	91.67	**100**	91.67	95.83
W3	**100**	**100**	**100**	**100**	95.83
W4	**100**	83.33	**100**	95.83	**100**
W5	**100**	83.33	**100**	**100**	**100**

Table 2. Average Word Level Accuracy Achieved for Test Data.

Words	DTW	HMM	GMM	MLP	SVM	DT
D0	**100**	**100**	81.25	93.75	87.5	68.75
D1	**100**	**100**	75	93.75	87.5	62.5
D2	81.25	87.5	**100**	87.5	81.25	62.5
D3	87.5	**100**	75	93.75	68.75	68.75
D4	81.25	93.75	68.75	87.5	93.75	50
D5	**100**	**100**	93.75	87.5	87.5	62.5
D6	**100**	93.75	93.75	**100**	**100**	56.25
D7	**100**	**100**	**100**	**100**	**100**	**100**
D8	**100**	**100**	81.25	**100**	**100**	43.75
D9	93.75	**100**	75	93.75	**100**	43.75
Words	**DTW**	**HMM**	**GMM**	**MLP**	**SVM**	**DT**
W1	93.75	**100**	**100**	81.25	75	50
W2	**100**	**100**	93.75	87.5	87.5	56.25
W3	**100**	93.75	93.75	81.25	93.75	68.75
W4	**100**	**100**	87.5	**100**	**100**	75
W5	**100**	**100**	81.25	**100**	**100**	81.25

(D- Digits, W- Words)

Table 3. Average Time Taken for Training and Testing the System.

Speech	Training Time(in seconds)				Testing Time(in seconds)			
Recognition	S1	S2	S3	S4	S1	S2	S3	S4
Techniques								
DTW	-	-	-	-	26.55	35.76	27.12	26.27
HMM	137.26	82.24	69.68	73.72	91.79	56.60	49.10	48.91
GMM	104.26	110.74	68.56	72.45	76.20	74.29	46.74	48.38
MLP	2.63	2.97	2.61	3.37	3.01	2.77	2.37	2.68
SVM	4.62	2.32	2.39	2.81	2.92	1.75	2.56	2.24
DT	**0.06**	**0.03**	**0.01**	**0.01**	**0.06**	**0.02**	**0.02**	**0.01**

6. Findings and Discussions

In this paper, totally six popular algorithms were involved in the experiments namely, DTW, HMM, GMM, MLP, SVM and DT. Here, three types of significant analyzes were done for the isolated speech recognition in Tamil language. These are

1. Performance evaluation of the statistical pattern matching techniques and machine learning techniques for Tamil spoken isolated speech.
2. Time taken for processing the data.
3. Utterance level performance evaluation.

Based on the experiments, it was found that the HMM and DTW has provided the better results followed by MLP and SVM when compared to the other methods. High recognition rate was achieved for word/utterance level performance using HMM. By using HMM, 100% accuracy was obtained for all the words during training process. For test data also, HMM provided 100% accuracy for 11 words out of fifteen words for all the speakers involved. Next to HMM, DTW was also obtained 100% accuracy for 10 words out of fifteen words during testing process. Next to these two methods, MLP and SVM have offered considerable results for both training and test data. Like HMM, MLP has also reached a great improvement by offering 100% results for training data. Based on the time factor, it was observed that the statistical approaches are time consuming when compared to machine learning techniques. Processing time taken for DTW also found to be high when compared to MLP, SVM and DT. The decision tree algorithm has taken very less processing time, but produced very less accuracy for the above developed system. Based on the utterance level performance, the digit seven (D7) has achieved 100% accuracy for all the techniques and speakers involved in the experiments. Next to D7, the D6, D8, W4 and W5 gave high accuracy for all the speakers enrolled in the study. It was observed from the experimental results that, the performance of the above system was found to be good when compared with the results achieved for the existing system discussed here. The average word recognition accuracy achieved for test data is 95.83%, 97.92%, 86.67%, 92.50%, 90.83%, 63.33% for DTW, HMM, GMM, MLP, SVM and Decision tree algorithms respectively. Particularly, 97.92% is achieved by HMM which is higher than the existing results. By considering the above experiments and analysis factor, HMM and DTW followed by MLP and SVM gives better recognition rate for the above developed system. Specifically, statistical approach improves word level performance and machine learning approaches works better for speaker independent applications.

7. Conclusion and Future Work

Research on ASR has been attracted by many applications in the last two decades. This paper has discussed about developing speaker independent isolated speech recognition for the Tamil language. The six algorithms that are popularly used for ASR system were implemented for this work using MFCC feature vectors. Each technique was briefly explained and its merits and demerits were clearly specified. Among the adopted techniques HMM, DTW, MLP and SVM techniques were able to produce better results for all speakers involved in this study. Among them, the HMM and DTW offered extraordinary results for word level accuracy. The MLP and SVM techniques were also found to be good for speaker independent speech recognition systems. It was also found that, the machine learning approaches reduces the time complexity when compared with HMM and GMM techniques. The current system achieved satisfactory results for small vocabulary with limited speakers and the system will be extended for a medium or large vocabulary with more speakers. Based on the above results, the hybrid techniques will be developed and its results and findings will be presented in future publications.

References

1. Maruti, L.; Rama, R.; and Vidya, S. (2012). Isolated digit recognition using MFCC and DTW. *International Journal on Advanced Electrical and Electronics Engineering,* (*IJAEEE*), 1(1), 59-64, ISSN (Print): 2278-8948.
2. Vimala, C.; and Radha, V. (2012). A review on speech recognition challenges and approaches. *World of Computer Science and Information Technology Journal (WCSIT)*, 2(1), 1-7.
3. Kimberlee, A.K. *An Introduction to Speech Recognition*. Voice Systems Middleware Education, IBM Corporation.
4. Bharathi, B.; Deepalakhmi, V.; and Nelson, I. (2006). A neural network based speech recognition system for isolated Tamil words. P*roceedings of International Conference on Neural Networks and Artificial Intelligence*, Brest, Belarus, June 2006.
5. Sigappi.; and Palanivel. (2012). Spoken word recognition strategy for Tamil language. *International Journal of Computer Science Issues*, 9(3), 227-233.
6. Rojathai, S.; and Venkatesulu, V. (2012). A novel speech recognition system for Tamil word recognition based on MFCC and FFBNN. *European Journal of Scientific Research*, 85(4), 578-590.
7. Ibrahim, M.; El-emary, M.; Mohamed, F.; and Hamza, A. (2011). Hidden Markov model/Gaussian mixture models (HMM/GMM) based voice command system: A way to improve the control of remotely operated robot arm TR45. *Scientific Research and Essays*, *Academic Journals*, 6(2), 341-350.
8. Do, V.H. (2011). *Hybrid architectures for speech recognition. Conformation.* Report submitted to the School of Computer Engineering, Nanyang Technological University.
9. Michael, P.; Bhuvana, R.; and Stanley, F.C. (2012). *Lecture* 3 - *Gaussian mixture models and introduction to HMM'*i. IBM T.J. Watson Research Center, Yorktown Heights, New York, USA.
10. Salvador, S.; and Chan, P. (2004). FastDTW: Toward accurate dynamic time warping in linear time and space. 3*rd Workshop on Mining Temporal and Sequential Data*, *ACM KDD* '04,Seattle, Washington (August 22-25).
11. Chunsheng, F. (2009). *From dynamic time warping (DTW) to hidden Markov model (HMM)*. University of Cincinnati.
12. Lawrence, R.R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 257-286.
13. Juang, B.H.; and Rabiner, L.R. (1991). Hidden Markov models for speech recognition. *Technometrics*, 33(3), 251-272.

14. Damousis, I.G.; and Argyropoulos, S. (2012). Four machine learning algorithms for biometrics fusion: a comparative study. *Hindawi Publishing Corporation Applied Computational Intelligence and Soft Computing*.
15. Pekka, T. (2004). Bayesian classification using Gaussian mixture model and EM estimation: Implementations and comparisons. *Lappeenranta*.
16. Md, S.; Dzulkifli, M.D.; and Sheikh, S. (2011). Malay isolated speech recognition using Neural Network: a work in finding number of hidden nodes and learning parameters. *The International Arab Journal of Information Technology*, 8(4), 364-371.
17. Vikramaditya, J. *Tutorial on Support Vector Machine* (*SVM*). School of EECS, Washington State University, Pullman 99164.
18. Multilayer perception. Retrieved October 1, 2013 from http://en.wikipedia.org/wiki/Multilayer_perceptron.
19. Hsu, C.-W., H.; Chang, C.-.C.; and Lin, C.-J. (2010). *A practical guide to support vector classification*. National Taiwan University, Taipei 106, Taiwan.
20. Shady, Y.El.M.; Mohammed, I.S.; and Hala, H.Z. (2009) *Speaker independent Arabic speech recognition using support vector machine.* Department of Electrical Engineering, Shoubra Faculty of Engineering, Benha University, Cairo, Egypt.
21. Paavo, N. (2010). *Data mining course* (*TIES445*)*, Lecture of Nov* 23. Department of Mathematical Information Technology, University of Jyväskylä.
22. Decision tree. Retrieved October 1, 2013 from http://en.wikipedia.org/wiki/Decision_tree.