# Isolated Tamil Digit Speech Recognition Using Template-Based and HMM-Based Approaches

S. Karpagavalli[1], R. Deepika[1], P. Kokila[1], K. Usha Rani[1], and E. Chandra[2]

[1] Department of Computer Science (PG), PSGR Krishnammal College for Women, Coimbatore
karpagam@grgsact.com,
{dpi.feb88,saikokila87,dhanalakshmi27}@gmail.com
[2] DJ Academy for Managerial Excellence, Coimbatore

**Abstract.** For more than three decades, a great amount of research was carried out on various aspects of speech signal processing and its applications. Highly successful application of speech processing is Automatic Speech Recognition (ASR). Early attempts to ASR consisted of making deterministic models of whole words in a small vocabulary and recognizing a given speech utterance as the word whose model comes closest to it. The introduction of Hidden Markov Models (HMMs) in the early 1980 provided much more powerful tool for speech recognition. And the recognition can be done for continuous speech using large vocabulary, in a speaker independent manner. Two approaches like conventional template-based and Hidden Markov Model usually performs speaker independent isolated word recognition. In this work, speaker independent isolated Tamil digit speech recognizers are designed by employing template based and HMM based approaches. The results of the approaches are compared and observed that HMM based model performs well and the word error rate is greatly reduced.

**Keywords:** Automatic Speech Recognition, Speaker Independent, Template-based approach, Hidden-markov model.

## 1 Introduction

Speech Recognition is also known as Automatic Speech Recognition (ASR), or computer speech recognition is the process of converting acoustic signal captured by microphone or telephone to a set of words.

Fundamentally, the problem of speech recognition can be stated as follows. When given with acoustic observation O = $o_1 o_2 \ldots o_t$, the goal is to find out the corresponding word sequence $W = w_1 w_2 \ldots w_n$ that has the maximum posterior probability $P\ (W|O)$ can be written as

P.V. Krishna, M.R. Babu, and E. Ariwa (Eds.): ObCom 2011, Part II, CCIS 270, pp. 441–450, 2012.

$$\hat{W} = \underset{w \in L}{\operatorname{argmax}} P(W|O) \tag{1}$$

Equation 1 can be expressed using Bayes rule as

$$\hat{W} = \underset{w \in L}{\operatorname{argmax}} \frac{P(O|W)\, P(W)}{P(O)} \tag{2}$$

Since the P (O) is the same for each candidate sentence W, thus equation 2 can be reduced as

$$\hat{W} = \underset{w \in L}{\operatorname{argmax}} P(O|W)\, P(W) \tag{3}$$

Where *P (W),* the prior probability of word *W* uttered is called the language model and *P (O|W),* the observation likelihood of acoustic observation O when word W is uttered is called the acoustic model [1].

The various components of the speech recognition system are Acoustic Front End, Acoustic Model, Pronunciation Dictionary, Language model and Decoder. The Schematic representation of the Speech Recognition Architecture is shown in figure 1.

## 2 Related Work

Speech recognizers for Spanish digit using template based and HMM based approaches are analyzed by Lucas D.Terrissi, Juan C. Gomez, University of Rosario, Argentina, [2005]. They have applied various template selection techniques and incorporated dynamic time warping for time aligning [2].

Arabic digit speech recognition System was developed by H Satori, M Harti, N Chenfour [2007]. This system is developed using the open source framework Sphinx-4, from the Carnegie Mellon University, a speech recognition system based on discrete hidden markov models (HMMs). In their work an in house Arabic speech corpus was developed and used for training and testing [3]. In the proposed work, a comparison between template-based and HMM-based approaches for isolated Tamil digit recognition is carried out. Template based approach with DTW time align algorithm implemented in Matlab environment. HMM based approach carried out using Sphinx Train and Sphinx-4 of Carniegie Mellon University. Both recognizers have been tested on a speech corpus generated from the Tamil speech utterances of digit from zero to nine spoken by 20 native speakers of Tamil Language.

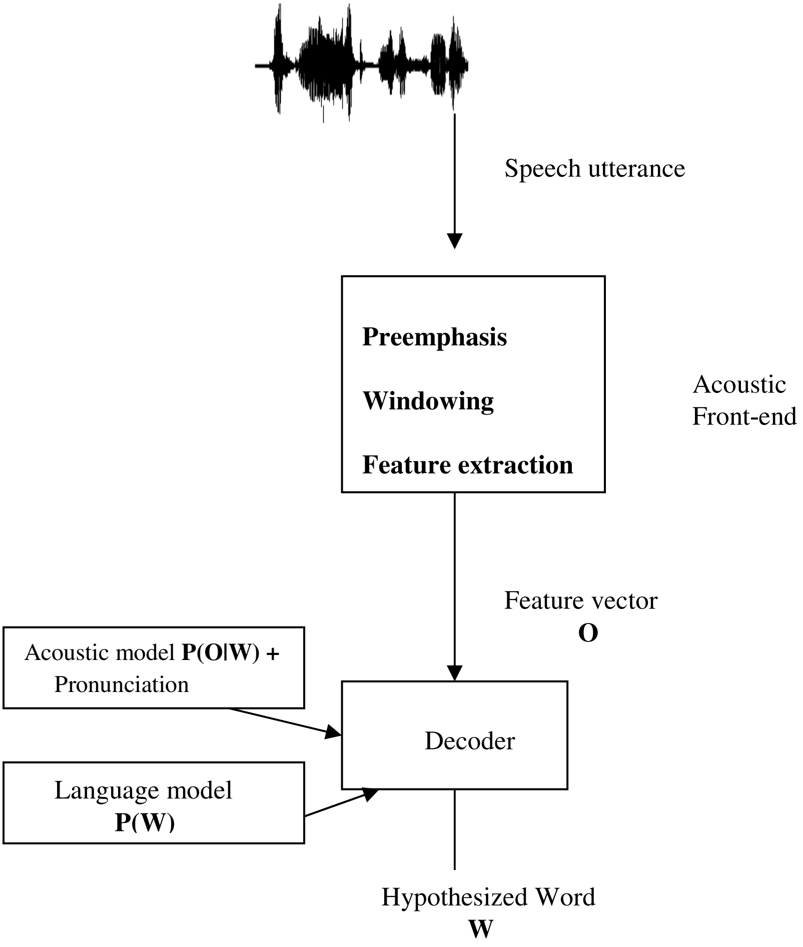


**Fig. 1.** Speech Recognition Architecture

## 3 Feature Extraction

The feature extraction is the first stage to extract feature vectors from the speech signals. Most speech recognition systems use the so–called Mel frequency Cepstral coefficients (MFCC) and its first and sometimes second derivative in time to better reflect dynamic changes. These are coefficients based on the Mel scale that represent sound. The word Cepstral comes from the word Cepstrum, which is a logarithmic scale of the spectrum (and reverses the first four letters in the word spectrum). It is shown in figure 2. First, the speech data are divided into 25 ms windows (frames).

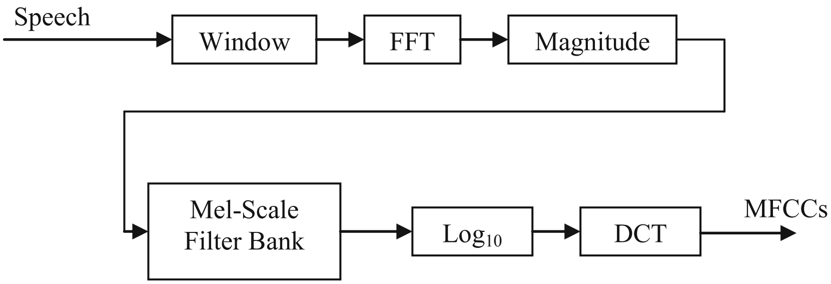


**Fig. 2.** Block Diagram of MFCC feature extraction method

A new frame is started every 10 ms making this the sampling period and causing the windows to overlap each other. Next, the fast Fourier transform is performed on each frame of speech data and the magnitude is found. The next step involves filtering the signal with a frequency-warped set of log filter banks called Mel-scale filter banks [4][5]. The log filter banks are arranged along the frequency axis according to the Mel scale, a logarithmic scale that is a measure of perceived pitch or frequency of a tone [6], thus simulating the human hearing scale. The Mel scale yields a compression of the upper frequencies where the human ear is less sensitive. Next, the logarithm is taken of the log filter bank amplitudes. Finally, the MFCCs are calculated using the discrete cosine transform (DCT). To further enhance speech recognition performance, an extra set of delta and acceleration coefficient features are sometimes calculated with MFCCs. These features are the first and second time derivatives of the original coefficients, respectively.

For template-based approach, 20 speakers uttering 3 times each digit is recorded with the sampling rate16 kHz using Audacity tool and MFCC feature vector of 39 dimensions (12-MFCC, 12-ΔMFCC, 12-ΔΔMFCC, P, ΔP, ΔΔP where P is stands for raw energy of the speech signal) are extracted using Matlab code. For HMM based approach, SphinxTrain tool used the 13 dimensional MFCC and Sphinx-4 used 39 dimensions of MFCC feature vector for processing.

## 4 Methodology

Basically there exist three approaches for speech recognition that include Acoustic Phonetic approach, Pattern Recognition approach and Artificial Intelligence approach. The pattern-matching approach has become the predominant method for speech recognition in the last six decades.

The pattern-matching approach involves two essential steps namely, pattern training and pattern comparison. The essential feature of this approach is that it uses a well-formulated mathematical framework and establishes consistent speech pattern representations, for reliable pattern comparison, from a set of labeled training samples via a formal training algorithm. A speech pattern representation can be in the form of a speech template or statistical models like HMM and can be applied to a sound like a phoneme, a word, or a phrase. In the pattern-comparison stage of the approach, a direct comparison is made between the speech to be recognized and each possible pattern learned in the training stage in order to determine the identity of the unknown according to the goodness of match of the patterns [4][12]. A schematic representation of pattern recognition approach for the proposed work is presented in Figure 3.

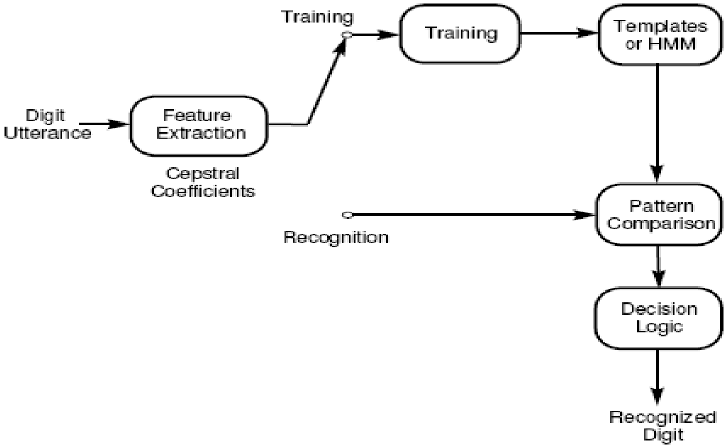


**Fig. 3.** Pattern Recognition Approach

## 5 Template-Based Approach

Template based approaches, in which unknown speech is compared against a set of prerecorded words (template) in order to find the best match. This has the advantage of using perfectly accurate word models; but also it has the disadvantages like suitable for small vocabulary and more computational time.

### 5.1 Time Alignment and Normalization

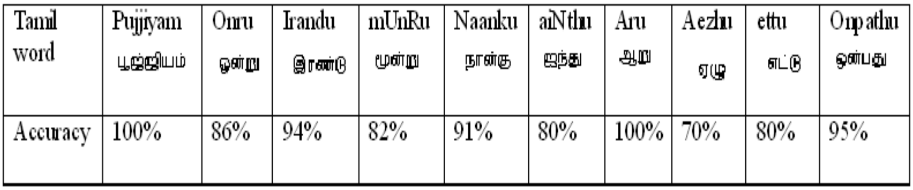
An utterance, which is to be recognized, however, is more complex than a steady sound, and thus a speech pattern almost always involves a sequence of short-time acoustic representations. The pattern-recognition approach for speech recognition compares the sequences of acoustic features. The problem associated with spectral sequence comparison for speech comes from the fact that different acoustic renditions, or tokens, of the same speech utterance are seldom realized at the same speed (speaking rate) across the entire utterance. Speaking rate variation as well as duration variation should not contribute to the linguistic dissimilarity score when comparing different tokens of the same utterance. Thus there is a need to normalize out speaking rate fluctuation in order for the utterance comparison to be meaningful before a recognition decision can be made. That can be achieved using Dynamic Time Warping (DTW), a dynamic programming algorithm that performs a non-linear time alignment to normalize the speaking rate fluctuations [2] [4].

### 5.2 Template Selections and Matching

The choice of the template will affect the performance of the recognizer. In this work, an utterance with medium duration from all the utterances of a particular digit is considered as template for that digit. Similarly template for each digit is selected and stored in the repository. The test data will undergo time alignment and will be matched against templates of each digit in the system's repository. The best matching template is the one for which there is the lowest distance path aligning the test input pattern to the template. A simple global distance score for a path is simply the sum of local distances that go to make up the path [7] [8].

The results of template based approach for each Tamil digit is summarized in Table 1. Isolated Tamil digit recognition accuracy is different for each word. The digit 0 (Pujjiyum) and digit 6 (ARu) has been correctly recognized for all the test data. The digit 7 (Aezhu) has the lowest accuracy, may be due to the pronunciation variation. The word 'Aezhu' has the special phone 'Zhu' which is not usually pronounced correctly by the native speakers of Tamil. The average accuracy rate is 87.8% while considering the results for all digits of Tamil.

**Table 1.** Results of template based approach for each Tamil digit



| Tamil word | Pujjiyam<br>பூஜ்ஜியம் | Onru<br>ஒன்று | Irandu<br>இரண்டு | mUnRu<br>மூன்று | Naanku<br>நான்கு | aiNthu<br>ஐந்து | Aru<br>ஆறு | Aezhu<br>ஏழு | ettu<br>எட்டு | Onpathu<br>ஒன்பது |
|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy | 100% | 86% | 94% | 82% | 91% | 80% | 100% | 70% | 80% | 95% |

## 6 Hidden Markov Model

HMM is very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of application. The introduction of Hidden Morkov Models (HMMs) in the early 1980 provided much more powerful tool for speech recognition. The elements of HMM is characterized by following:

1. Number of state N
2. Number of distinct observation symbol per state
3. State transition probability,
4. Observation symbol probability distribution in state
5. The initial state distribution

The Three Basic Problems for HMMs are,

Problem 1: Evaluation Problem -Given the observation sequence $O = O_1\ O_2 \cdot\cdot\cdot O_T$, and model $\lambda = (A,B,\pi)$, how do we efficiently compute $P(O|\ \lambda)$, the probability of observation sequence given the model.

Problem 2: Hidden State Determination (Decoding) -Given the observation sequence $O = O_1 O_2 \cdots O_T$ , and model $\lambda = (A,B,\pi)$ how do we choose corresponding state sequence $Q = q_1 q_2 \cdots q_T$ which is optimal in some meaningful sense.

Problem 3: Learning -How do we adjust the model parameter $\lambda = (A,B,\pi)$, to maximize $P(O| \lambda)$. Problem 3 is one in which we try to optimize model parameter so as to best describe as to how given observation sequence comes out.

Solution to three problems of HMM are Forward Algorithm for Evaluation Problem, Viterbi Algorithm for Decoding Hidden State Sequence $P(Q,O| \lambda )$ and Baum-Welch Algorithm for Learning [1] [9][10].

For HMM based approach training is carried out in SphinxTrain and model is implemented in Sphinx-4. Sphinx-4 is a flexible, modular and pluggable framework to help foster new innovations in the core research of hidden Markov model speech recognition systems. Automatic speech recognition involves many tasks that include speech corpus preparation, building pronunciation dictionary, acoustic model and language model.

### 6.1 Development of Speech Corpus

Tamil digit speech corpus was created in a noise free lab environment, as standard speech corpora are not available for Tamil language. For training 20 speakers uttering 20 times each digit is recorded with the sampling rate16 kHz using Audacity tool. For testing 10 speakers uttering 4 times each digit is recorded. And the necessary transcription files are prepared.

### 6.2 Building Language Model and Dictionary

Language models mainly describe the linguistic restrictions present in the language and to allow reduction of possible invalid phoneme sequences. Language model estimate the probability of sequences of words. Common language models are bi-gram and trigram models. These models contain computed probabilities of groupings of two or three particular words in a sequence, respectively. In this work, Statistical tri-gram language models were built using the CMU Statistical Language Modeling toolkit for word-based model.

### 6.3 Building the Acoustic Model

The HMM based acoustic model trainer from CMU, *SphinxTrain*, has been employed. Sphinx Train supports Mel frequency Ceptral Co-efficient (MFCC) features. The features are extracted from the training wav files recorded with sampling rate 16 kHz of 16 bits depth [11]. Pronunciation dictionary, filler dictionary, transcription files, MFCC feature files are used for training. Number of states in the HMM is 15 and context independent training was carried out. The training procedure comprises the following processes.

Flat-start monophone training: Generation of monophone or CI seed models with nominal values, and re-estimation of these models using reference transcriptions. This is also called flat initialization of CI model parameters.

Baum-Welch training of monophones: Adjustment of the silence model and reestimation of single-Gaussian monophones using the standard Viterbi alignment process.

After these processes, context independent word model was generated. *SphinxTrain* generates the parameter files of the HMM namely, the probability distributions and transition matrices of all the HMM models. The word model is implemented on Sphinx-4, which is a state-of-art HMM based speech recognition system.

## 7 Results and Discussion

The hypothesis word sequences from the decoder are aligned with reference sentences. The performance of the speech recognizers are measured in terms of Word Error Rate (WER) and Word Recognition Rate (WRR). Word errors are categorized into number of insertions, substitutions and deletions. Finally, the word error rate and word recognition rate are computed by the following equations (4) (5),

$$\text{Word Error Rate (\%)} = (100)\,\frac{\text{Insertion (I)+ Substitution(S)+Deletion(D)}}{\text{No of Reference Words (N)}} \tag{4}$$

$$\text{WRR=1-WER} = \frac{\text{N-S-D-I}}{\text{N}} \tag{5}$$

Other performance measures are speed and memory footprints. The results of HMM based approach are given in Table 2.

**Table 2.** Results of HMM based Approach for Tamil digits

| Words | 400 |
|---|---|
| Errors | 32 (Sub: 11 Ins: 16 Del: 5) |
| Accuracy | 92% |
| Time | Audio: 564.06s Processing 61.55s |
| Speed | 0.11 X real time |
| Memory | Average: 25.27MB Max: 35.38MB |

Comparison of the performance of Template based approach and HMM based approach is presented in Table 3. It is clearly observed that statistical model (HMM) outperforms the conventional template based approach for Tamil digit speech recognizer.

**Table 3.** Comparison of the performance of Template based approach and HMM based approach

| Model | Word Recognition Rate |
|---|---|
| Template | 87.8% |
| HMM | 92% |

## 8 Conclusion

The goal of automatic speech recognition research is to address the various issues relating to speech recognition. Various methodologies are identified and applied to ASR area which led to many successful ASR applications in limited domains. But in Tamil language, speech recognition applications are very less. In our work, we tried to design small vocabulary, isolated, speaker independent Tamil digit recognizers using template based approach and HMM based approaches. It is being observed that, in HMM based approach Tamil digit recognition rate is high with less computational time and memory.

## References

1. Jurafsky, D., Martin, J.H.: Speech and Language Processing - An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson Education (2002)
2. Terissi, L.D., Gomez, J.C.: Template-based and HMM-based Approaches for Isolated Spanish Digit Recognition. Intelligencia Artificial.Revista lberoamericana de Intelligencia Artificial 9(26) (2005)
3. Satori, H., Harti, M., Chenfour, N.: Arabic Speech Recognition System based on CMUSphinx. In: International Symposium on Computational Intelligence and Intelligent Informatics (March 2007)
4. Rabiner, L., Juang, B.-H.: Fundamentals of Speech Recognition. Prentice-Hall, Inc., Engelwood (1993)
5. Kamm, T., Hermansky, H., Andreou, A.G.: Learning the Mel-scale and Optimal VTN Mapping. In: Center for Language and Speech Processing, Workshop (WS 1997). Johns Hopkins University (1997)
6. Hornback, J.R., Lieutenant, S.: Speech Recognition Using The Mellin Transform, MS Thesis report, Air Force Instituite of Technology, Wright-Patterson Air Force Base, Ohio (2006)
7. Li, D., Strik, H.: Structure-Based and Template-Based Automatic Speech Recognition-Comparing parametric and non-parametric approaches. Microsoft Research, One Microsoft Way, Redmond, WA, USA, CLST, Department of Linguistics, Radboud University, Nijmegen
8. Hachkar, Z., Farchi, A., Mounir, B., El Abbadi, J.: A Comparison of DHMM and DTW for Isolated Digit Recognition System for Arabic Language. International Journal of Computer Science and Engineering 3(3) (March 2011); ISSN : 0975-3397

9. Jacob, B., Sondhi, M.M., Huang, Y.: Springer Handbook of Speech Processing, XXXVI (2008)
10. Rabiner, L.R.: A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
11. Thangarajan, R., Natarajan, A.M., Selvam, M.: Word and Triphone based approaches. Continuous Speech Recognition for Tamil Language 4(3) (March 2008)
12. Anusuya, M.A., Katti, S.K.: Speech Recognition by Machine: A Review. International Journal of Computer Science and Information Security 6(3), 181–205 (2009)