Technical journals online.com

Research Paper

SYLLABLE BASED CONTINUOUS SPEECH RECOGNITION FOR TAMIL LANGUAGE

C.Sivaranjani, B. Bharathi

Address for Correspondence

Department of CSE, SSN College of Engineering, Tamilnadu, India

ABSTRACT

Tamil is one of the ancient languages in the world. Recognition of Tamil speech would be beneficial to a lot of Tamil people. There are many speech recognition systems available for many languages. Especially recognition of Tamil language is a challenging task. In this paper, we propose to develop a system for continuous speech recognition in Tamil language. Basically it consists of two phases (i) Training phase (ii) Testing phase. During training phase, features are extracted from the input speech signal and acoustic model is built based on the sub-word units. The sub-word units can be word or phoneme or syllable. In this paper, syllable is used as sub-word unit. During testing phase, features are extracted from the test speech signal and compared with acoustic model to find matching pattern. Performance evaluation is based on word error rate that is comparing a reference transcription with the transcription output by the speech recognizer.

KEYWORDS – Speech Recognition, Segmentation, MFCC feature vectors, Hidden Markov Model.

1. INTRODUCTION

Speech is the vocalized form of human communication. It is based upon the syntactic combination of lexicals and names that are drawn from very large vocabularies. Each spoken word is created out of the phonetic combination of a limited set of vowel and consonant speech sound units. Speech processing is the study of speech signals and the processing methods of these signals. The signals are usually processed in a digital representation, so speech processing can be regarded as a special case of digital signal processing, applied to speech signal. Aspects of speech processing includes the acquisition, manipulation, storage, transfer and output of speech signals. Speech Recognition (SR) is the translation of spoken words into text. It is also known as Automatic Speech Recognition (ASR), computer speech recognition, or just Speech-To-Text (STT). People with disabilities can benefit from speech recognition programs. For individuals that are Deaf or Hard of hearing, speech recognition software is used to automatically generate a closed-captioning of conversations such as discussions in conference rooms, classroom lectures. Speech recognition is also very useful for people who have difficulty using their hands, ranging from mild repetitive stress injuries to involved disabilities that preclude using conventional computer input devices.

The organization of the paper is given below. Some of the work related with speech recognition for tamil language and methods for segmentation are discussed in Section3. Section 4 deals with the system description of the proposed system. Segmentation algorithm is explained in Section 4.1. MFCC feature extraction steps are described in Section 4.2. Section 4.3, briefly described the Hidden markov model. The experimental results are detailed in Section 5. Finally, Section 6 concludes the paper.

2. RELATED WORK

In general speech recognition system usually consists of two phases. They are called pre-processing and post-processing. Pre-processing involves feature extraction and the post-processing stage comprises of building an acoustic model, phonetic lexicon or the pronunciation dictionary and language model. In [1], they have implemented speaker independent isolated recognition system for tamil language using HMM. The Mel Frequency Cepstral Coefficients (MFCC) features are extracted. The database size used for this research work is 2,500 words and produces 88% of accuracy. The performance of speech recognition system is generally measured in terms of word error rate. Speech recognition is generally easy when vocabularies are small, but word error rate increases as the vocabulary size grows. In [2], they have developed Tamil speech recognition using semi-continuous speech that uses triphone based model to overcome the disadvantages in monophone. The disadvantage of monophone is co-articulation point is not considered since it is context independent. MFCC feature are extracted and HMM is used to build the acoustic model. In [3], An Efficient Method for Tamil Speech Recognition is implemented using MFCC and DTW. MFCC feature are extracted. In the next step feature vectors are compared with patterns in the database using Dynamic Time Warping (DTW) in order to find the exact spoken words. Five tamil words are trained and features are extracted. The accuracy achieved using this method is 95%. In [4], they have implemented a small vocabulary word based and a medium vocabulary triphone based continuous speech recognizers for Tamil language. MFCC feature are extracted and acoustic model is built using HMM. Word based model is suitable for small vocabulary and it is inefficient for larger vocabulary so by using triphone based model the larger vocabulary has been efficiently recognized. In [5], An Efficient Speech Recognition System is implemented. Speaker identification and speech recognition is done. MFCC feature are extracted and HMM is used to build acoustic model. Vector Quantization are used to make constant feature vector because MFCC feature vector is varying since speaker may speak same word at different times differently. Five words are trained by 4 persons, each word is repeated for 10 times. The accuracy obtained for speaker identification is 95\%and 98% for speech recognition. In [6], Two level language models are used to improve the performance of Tamil Speech Recognition that are Phoneme based model and Syllable based model. Syllable based is used to overcome the difficulties in phoneme based model. Phoneme based has ambiguity since occurrence of same phoneme in different words. In [7], A syllable based continuous speech recognizer for tamil language is implemented. A group delay based segmentation algorithm is proposed to extract accurate syllable units from the speech data. Forced viterbi algorithm is used to find most likely sequence. A rule based text segmentation algorithm is used to

automatically annotate the text corresponding to the speech into syllable units. Text segmentation is based on linguistic rules. HMM is used to build acoustic model. The performance is better when compared to other segmentations. In [8], Prosodic Syllable Based Tamil Speech Recognition System is implemented. A group delay based segmentation algorithm is proposed to extract accurate syllable units from the speech data. Forced viterbi algorithm is used to find most likely sequence. Acoustic model is built using HMM. The system was trained with 20 words for 2 times by two speakers. The accuracy achieved using this method is 70%.

3. PROPOSED SYSTEM DESIGN

The first step in the proposed system is to collect speech data from different users. Then, the speech data is segmented based on the syllable. From this segmented speech data, MFCC feature vectors are extracted and using these feature vectors an acoustic model is built. During testing phase, the test speech input signal is recorded and features are extracted. Then extracted feature is compared with acoustic model to find the recognized text. The acoustic model uses language model and dictionary model.

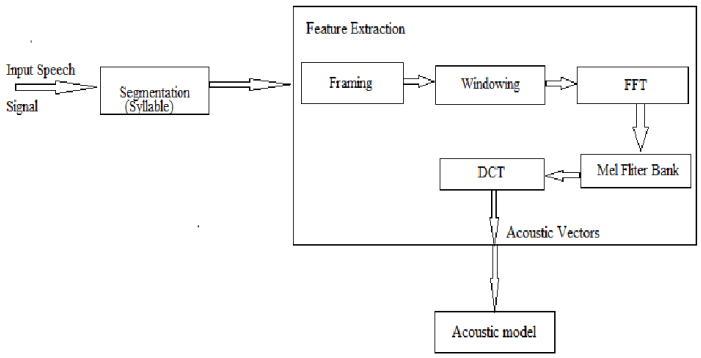

Figure 1: Training

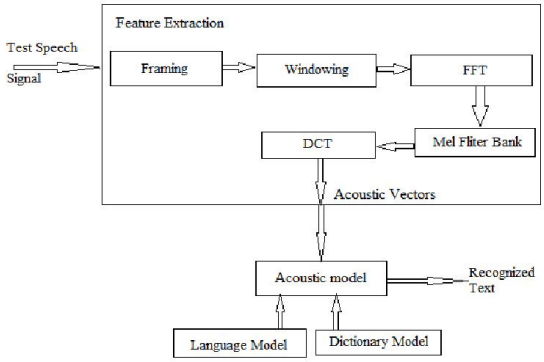

Figure 2: Testing

4.1 Segmentation

Forced Viterbi algorithm is used for segmentation, which is used for finding the most likely sequence of hidden states. Initially some speech data is segmented manually at syllable level. Then using manually segmented speech data the forced viterbi algorithm is used to segment the remaining speech data. For example, in speech recognition, the acoustic signal is treated as the observed sequence of events, and a string of text is considered to be the hidden state of the acoustic signal. The Viterbi algorithm finds the most likely string of text given the acoustic signal.

4.2 Feature Extraction

The second step of the proposed system is to extract features i.e. identify the components of the audio signal that are good for identifying the linguistic content and discarding all the other stuff which carries information like background noise, emotion etc. The main point to understand about speech is that the sounds generated by a human are filtered by the shape of the vocal tract including tongue, teeth etc. This shape determines what sound comes out. If we can determine the shape accurately, this should give us an accurate representation of the phoneme being produced. The shape of the vocal tract manifests itself in the envelope of the short time power spectrum, and the job of MFCCs is to accurately represent this envelope. The first step is to process of segmenting the speech samples obtained from analog to digital conversion into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples. The second step is Hamming Windowing. Hamming Windowing is used to eliminate discontinuities at the edges and integrates all the closest frequency lines. The third step is apply Fast Fourier Transform (FFT) which converts from time domain to frequency domain and extracts the frequency components of the signal. The fourth step is Mel Filter Bank, this is used when the signal may still contain unwanted information. The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. A set of triangular filters that are used to compute a weighted sum of filter spectral components so that the output of process approximates to a Mel scale. The Mel scale relates perceived frequency, or pitch, of a pure tone to its actual measured frequency. Humans are much better at discerning small changes in pitch at low frequencies than they are at high frequencies. Incorporating this scale makes our features match more closely what humans hear. The final step is to compute the Discrete cosine Transform, This is the process to convert the log Mel spectrum into time domain using Discrete Cosine Transform (DCT). The result of the conversion is called Mel Frequency

Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

4.3 Hidden Markov Model

Hidden Markov Model is used to build the acoustic model. Markov model (Markov process) is that which satisfies the Markov Property. Markov Property: The state of the system at time t+1 depends only on the state of the system at time t. HMM is Machine Learning Method. It uses state machines. HMM is useful in problem having sequential steps. It can only observe output from states, not the states. A hidden Markov model is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states. In a hidden Markov model, the state is not directly visible, however the output depends on the previous state is visible. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives some information about the sequence of states. Each state is described by the syllable recognized by the system. Each state branches off to different other states that are most likely to come next. In this model, each syllable is like a link in a chain, and the completed chain is a word. However, the chain branches off in different directions as the program attempts to match the digital sound with the syllable that's most likely to come next. During this process, the program assigns a probability score to each syllable, based on its built-in dictionary and user training.

Language Model

The language model provides context to distinguish between words and phrases that sound similar.

Dictionary Model

Different speakers may pronounce similar words differently. This requires building a dictionary that has multiple pronunciations for any given word.

4. Experimental Results

The speech data will be collected for about 6 minutes from 20 speakers. Recorded in laboratory environment (i.e noise free). Same channel is used. The text collected for speech data is from tamil stories for kids. The tamil story has been transliterated and manually segmented in to the corresponding syllable sequence. MFCC feature vectors are extracted, frame number followed by 39 coefficients which includes 13 cepstral coefficients, 13 delta coefficients and 13 acceleration coefficients respectively.

Example: Tamil Text and its Syllable Sequence

ஒரு நாள் இரண்டு தேவதைகளுக்கு சந்தேகம் வந்தது.

ஒ+ரு நா+ள் இ+ரண்+டு தே+வ+தை+க+ளு+க்கு சந்+தே+கம் வந்+த+து.

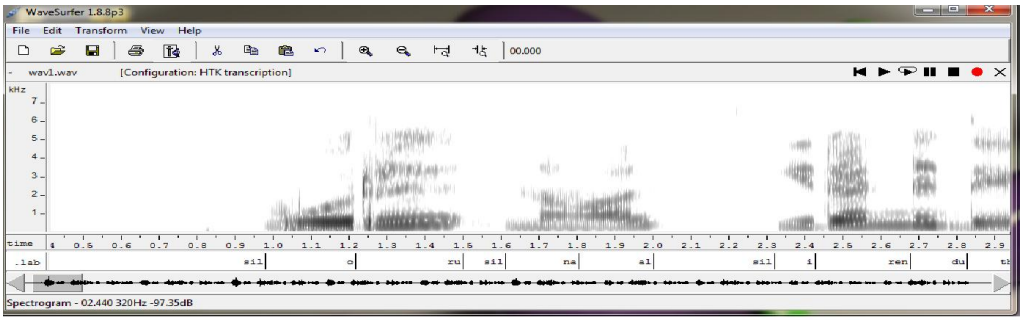

Figure 3: Sentences labelled at syllable level

The sentences have been segmented based on syllable units. For example sil represents silence and o, ru, na represents the syllables shown in Figure: 3. The initial proto model have been generated using 16 mixture component and 3 states. The proto type acoustic model is trained using the training utterences of the speaker. The trained syllable model "o" is shown in Figure: 4

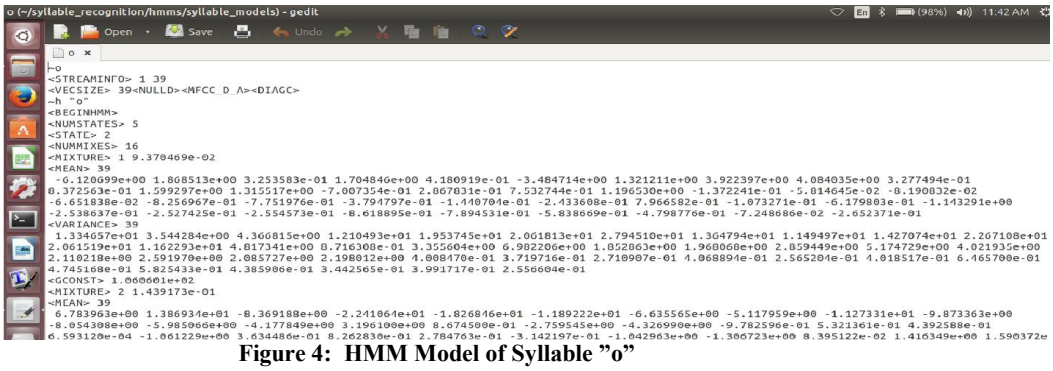

Figure 4: HMM Model of Syllable "o"

CONCLUSION

The proposed work is to develop a continuous speech recognition system for tamil language. It has two phases training and testing. In training phase, the first step is, the speech utterences are collected from 20 persons and segmented the speech utterences at syllable level using forced Viterbi alignment algorithm. In the second step, MFCC features vectors are extracted and an acoustic model have been built. In testing phase speech utterences will be collected, features are to be extracted and then the extracted feature will be compared with acoustic model along with language model and dictionary model thereby the tamil text will be recognized. Finally performance evaluation is to be carried out based on word error rate.

REFERENCES

1. Vimala,Radhab V, "Speaker Independent Isolated Speech Recognition System for Tamil Language using HMM", International Conference on Communication Technology and System Design. Volume 30, pp 1097-1102, 2011.

2. Hanitha Gnanathesigar, "Tamil Speech Recognition using Semi Continuous Models "International Journal of Scientific and Research Publications, Volume 2, Issue 6, ISSN 2250-3153, June 2012.
3. 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.
4. Thangarajan R,Natrajan A M, Selvam M, "Word and Triphone Based Approaches in Continuous Speech Recognition for Tamil Language", Transactions on Signal Processing, pp 76-85, 2008.
5. Suma Swamy and K.V Ramakrishna, " An Efficient Speech Recognition Sytem ", An International Journal (CSEIJ), Vol. 3, No. 4, pp 21-27, August 2013.
6. Saraswathi1 S,Geetha T V, " Two level language models for improving the performance of Tamil Speech Recognition," International Conference on Computational Intelligence and Multimedia Applications, pp 87-91, 2007.
7. Lakshmi A, Hema AMurthy "A syllable based continuous recognizer for tamil language", Interspeech, pp 1878-1881, Sept 2006.
8. Akila.A.Ganesh, Chandra Ravichandran, "Prosodic Syllable Based Tamil Speech Recognition", System International Conference Signal Processing and Communication (ICSC) 2013, pp 401-406, 2013.