Speech Recognition using SVM

R. Thiruvengatanadhan

Assistant Professor/Lecturer (on Deputation), Department of Computer Science and Engineering
Annamalai University, Annamalainagar, Tamil Nadu, India

Abstract:- *Speech Recognition approach intends to recognize the text from the speech utterance which can be more helpful to the people with hearing disabled. This paper describes a technique that uses support vector machines (SVM) to recognized speech based on features using Mel Frequency Cepstral Coefficients (MFCC). Modeling techniques such as SVM were used to model each individual word which is trained to the system. Each isolated word Segment using Voice Activity Detection (VAD) from the test sentence is matched against these models for finding the semantic representation of the test input speech. Experimental results of support vector machines shows good performance in recognized rate.*

Key Words**:** *Feature Extraction, Voice Activity Detection (VAD), Mel Frequency Cepstral Coefficients (MFCC) and support vector machines (SVM).*

1. INTRODUCTION

Speech recognition is a main core of spoken language systems. Speech recognition is a complex classification task and classified by different mathematical approaches: acoustic-phonetic approach, pattern recognition approach, artificial intelligence approach, dynamic dime warping, connectionist approaches and support vector machine. There have also been applications to speech recognition problems, namely phonetic classification [1] and post-classification of speech recognition hypotheses

Proposed work aims to develop a system which has to convert spoken word into text using SVM modeling technique using acoustic feature namely MFCC. In this work the temporal envelop through RMS energy of the signal is derived for segregating individual words out of the continuous speeches using voice activity detection method.

Features for each isolated word are extracted and those models were trained. During training process each isolated word is separated into 20ms overlapping windows for extracting 13 MFCCs features. SVM modeling technique is used to model each individual utterance. Thus each isolated word segment from the test sentence is matched against these models for finding the semantic representation of the test input dialogue. The frame work of the proposed system is shown in Fig.1.

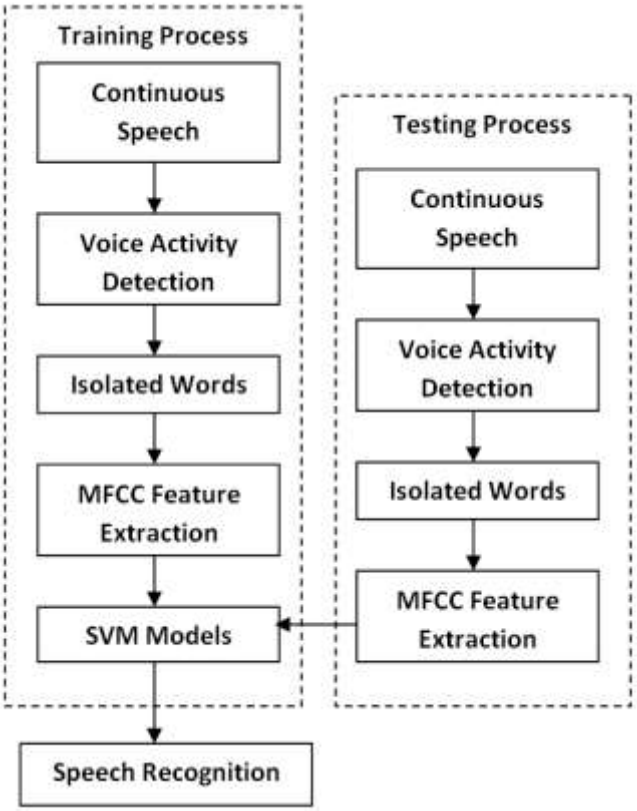

Fig -1: Framework of the proposed system

2. VOICE ACTIVITY DETECTION

Voice Activity Detection (VAD) is a technique for finding voiced segments in speech and plays an important role in speech mining applications [2]. VAD ignores the additional signal information around the word under consideration. It can be also viewed as a speaker independent word recognition problem. The basic principle of a VAD algorithm is that it extracts acoustic features from the input signal and then compares these values with thresholds usually extracted from silence. Voice activity is declared if the measured values exceed the threshold. Otherwise, no speech activity is present [3].

VAD finds its usage in a variety of speech communication systems like coding of speech, recognizing speech, hands free telephony, audio conferencing, speech enhancement and cancellation of audio [11]. It identifies where the speech is voiced, unvoiced or sustained and makes smooth progress of the speech process [12]. A frame size of 20 ms, with an overlap of 50%, is considered for VAD. RMS is extracted for each frame. Fig. 2 shows the isolated word separation.

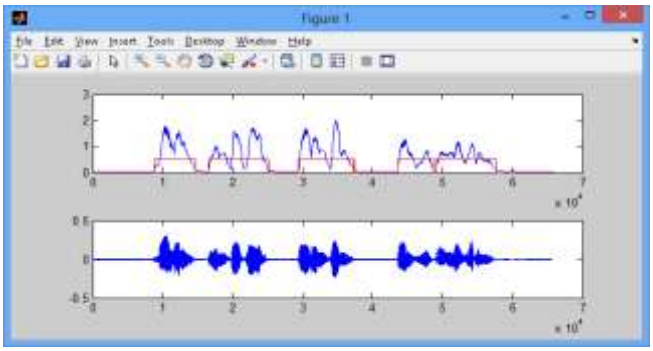

Fig -2: Isolated Word Separations.

3. ACOUSTIC FEATURES FOR SPEECH RECOGNITION

An important objective of extracting the features is to compress the speech signal to a vector that is representative of the meaningful information it is trying to characterize. In these works, acoustic features namely MFCC features are extracted.

3.1 Mel Frequency Cepstral Coefficients

Mel Frequency Cepstral Coefficients (MFCCs) are short-term spectral based and dominant features and are widely used in the area of audio and speech processing. The mel frequency cepstrum has proven to be highly effective in recognizing the structure of music signals and in modeling the subjective pitch and frequency content of audio signals [4]. The MFCCs have been applied in a range of audio mining tasks, and have shown good performance compared to other features. MFCCs are computed by various authors in different methods. It computes the cepstral coefficients along with delta cepstral energy and power spectrum deviation which results in 26 dimensional features. The low order MFCCs contains information of the slowly changing spectral envelope while the higher order MFCCs explains the fast variations of the envelope [5].

MFCCs are based on the known variation of the human ears critical bandwidths with frequency. The filters are spaced linearly at low frequencies and logarithmically at high frequencies to capture the phonetically important characteristics of speech and audio.

To obtain MFCCs, the audio signals are segmented and windowed into short frames of 20 ms. Magnitude spectrum is computed for each of these frames using Fast Fourier Transform (FFT) and converted into a set of mel scale filter bank outputs.

The human ear resolves frequencies non-linearly across the audio spectrum and empirical evidence suggests that designing a front-end to operate in a similar non-linear manner improves the performance. A popular solution is therefore filterbank analysis since this provides a much more straightforward route to obtain the desired non-linear frequency resolution.

However, filterbank amplitudes are highly correlated and hence, the use of a cepstral transformation in this case is virtually mandatory. Fig. 3 describes the procedure for extracting the MFCC features.

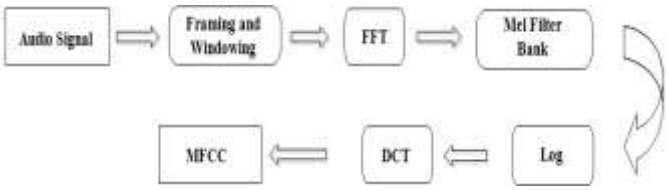

Fig -3: Extraction of MFCC from Audio Signal.

Mel frequency to implement this filterbank, the window of audio data is transformed using a Fourier transform and the magnitude is taken. The magnitude coefficients are then binned by correlating them with each triangular filter.

Here, binning means that each FFT magnitude coefficient is multiplied by the corresponding filter gain and the results are accumulated. Thus, each bin holds a weighted sum representing the spectral magnitude in that filterbank channel.

Logarithm is then applied to the filter bank outputs. Discrete Cosine Transformation (DCT) is applied to obtain the MFCCs. Since the mel spectrum coefficients are real numbers, they are converted to the time domain using the DCT.

In practice, the last step of taking inverse Discrete Fourier Transform (DFT) is replaced by taking DCT for computational efficiency. The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Typically, the first 13 MFCCs are used as features.

4. CLASSIFICATION MODEL

4.1 Support Vector Machine

A machine learning technique which is based on the principle of structure risk minimization is support vector machines. It has numerous applications in the area of pattern recognition [6]. SVM constructs linear model based upon support vectors in order to estimate decision function. If the training data are linearly separable, then SVM finds the optimal hyper plane that separates the data without error [7].

Fig. 2 shows an example of a non-linear mapping of SVM to construct an optimal hyper plane of separation. SVM maps the input patterns through a non-linear mapping into higher dimension feature space. For linearly separable data, a linear SVM is used to classify the data sets [8]. The patterns lying on the margins which are maximized are the support vectors.

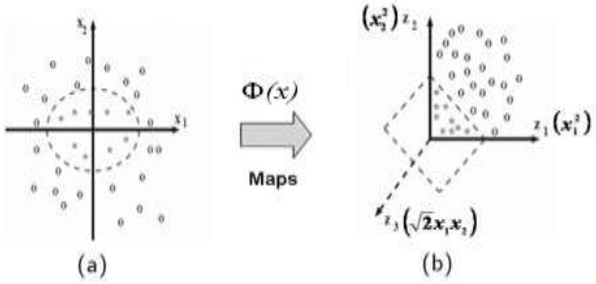

Fig -4: Example for SVM Kernel Function Φ(x) Maps 2-Dimensional Input Space to Higher 3-Dimensional Feature Space. (a) Nonlinear Problem. (b) Linear Problem.

The support vectors are the (transformed) training patterns and are equally close to hyperplane of separation. The support vectors are the training samples that define the optimal hyperplane and are the most difficult patterns to classify [9]. Informally speaking, they are the patterns most informative of the classification task. The kernel function generates the inner products to construct machines with different types of non-linear decision surfaces in the input space [10].

5. EXPERIMENTAL RESULTS

5.1 Dataset Collection

Experiments are conducted for speech recognition audio using Television broadcast speechdata collected from Tamil news channels using a tuner card. A total dataset of 100 different speech dialogue clips, ranging from 5 to 10 seconds duration, sampled at 16 kHz and encoded by 16-bit is recorded. Voice activity detection is performed to isolate the words in each speech file using RMS energy envelope.

5.2 Feature Extraction

In this work the pre-emphasized signal containing the continuous speech is taken for testing. Through VAD the isolated words are extracted from the sentences. Thus frames which are unvoiced excitations are removed by thresholding the segment size. Feature MFCC are extracted from each frame of size 320 window with an overlap of 120 samples. Thus it leads to 13 MFCCs respectively which are used individually to represent the isolated word segment. During training process each isolated word is separated into 20ms overlapping windows for extracting 13 MFCCs features.

5.3 Classification

Using VAD isolated words in a speech is separated.

N-SVMs are created for each isolated word. For training, isolated words from were considered. The training process analyzes speech training data to find an optimal way to classify speech frames into their respective classes. The derived support vectors are used to classify speech data. For testing 13 dimensional MFCC feature vectors were given as input to SVM model and the distance between each of the feature vectors and the SVM hyperplane is obtained. The average distance is calculated for each model. The text corresponding to the query speech is decided based on the maximum distance. The same process is repeated for different query speech, and the performance is studied. The performances of speech recognition for different SVM kernels are compared for MFCC acoustic features are shown in Table 1.

Table -1: Performance of speech recognition rate in different SVM kernel function.

SVM Kernels	Speech Recognition Rate
Polynomial	78%
Gaussian	95%
Sigmoidal	81%

6. CONCLUSION

In this paper, we have proposed speech recognition system using SVM. Voice Activity Detection (VAD) is used for segregating individual words out of the continuous speeches. Features for each isolated word are extracted and those models were trained successfully. SVM is used to model each Individual utterance. MFCC is calculated as features to characterize audio content. SVM learning algorithm has been used for the recognized speech by learning from training data. Experimental results show that the proposed audio support vector machine learning method has good performance in 95% speech recognized rate.

REFERENCES

[1] V. V apnik, The Nature of Statistical Learning Theory.

[2] Ivan Markovi, Sre´ ckoJuri´ Kavelj and Ivan Petrovi, "Partial Mutual Information Based Input Variable Selection for Supervised Learning Approaches to Voice Activity Detection," Applied Soft Computing Elsevier, vol. 13, pp. 4383-4391, 2013.

[3] Khoubrouy, S. A. and Panahi, I.M.S., "Voice Activation Detection using Teager-Kaiser Energy Measure," International Symposium on Image and Signal Processing and Analysis, pp. 388-392, 2013.

[4] O.M. Mubarak, E. Ambikai rajah and J. Epps, "Novel Features for Effective Speech and Music Discrimination," IEEE Engineering on Intelligent Systems, pp. 342-346, 2006.

[5] A. Meng and J. Shawe-Taylor, "An Investigation of Feature Models for Music Genre Classification using the Support Vector Classifier," International Conference on Music Information Retrieval, Queen Mary, University of London, UK, pp. 604-609, 2005.

[6] Chungsoo Lim Mokpo, Yeon-Woo Lee, and Joon-Hyuk Chang, "New Techniques for Improving the practicality of a SVM-Based Speech/Music Classifier," IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1657-1660, 2012.

[7] Hongchen Jiang, JunmeiBai, Shuwu Zhang, and Bo Xu, "SVM-Based Audio Scene Classification," IEEE International Conference Natural Language Processing and Knowledge Engineering, Wuhan, China, pp. 131-136, October 2005.

[8] Lim and Chang, "Enhancing Support Vector Machine-Based Speech/Music Classification using Conditional Maximum a Posteriori Criterion," Signal Processing, IET, vol. 6, no. 4, pp. 335-340, 2012.

[9] Md. Al Mehedi Hasan and Shamim Ahmad. predSucc-Site: Lysine Succinylation Sites Prediction in Proteins by using Support Vector Machine and Resolving Data Imbalance Issue. International Journal of Computer Applications 182(15):8-13, September 2018.

[10] Hend Ab. ELLaban, A A Ewees and Elsaeed E AbdElrazek. A Real-Time System for Facial Expression Recognition using Support Vector Machines and k-Nearest Neighbor Classifier. International Journal of Computer Applications 159(8):23-29, February 2017.

[11] Saleh Khawatreh, Belal Ayyoub, Ashraf Abu-Ein and Ziad Alqadi. A Novel Methodology to Extract Voice Signal Features. International Journal of Computer Applications 179(9):40-43, January 2018.

[12] Tayseer M F Taha and Amir Hussain. A Survey on Techniques for Enhancing Speech. International Journal of Computer Applications 179(17):1-14, February 2018.