Isolated Word Recognition Based on VAD, MFCC and DTW

Ashraful Islam*, Jayanta Dey
Department of Electrical and Electronic Engineering,
Bangladesh University of Engineering and Technology,
Dhaka, Bangladesh
*asrafulashiq@gmail.com

***Abstract*—Speech recognition deals with developing methods and techniques to recognize speech by machine in different conditions. Most of the techniques developed so far work well on high SNR condition. This paper proposes a method for isolated word recognition based on Voice Activity Detection (VAD), Mel-frequency cepstral coefficients (MFCCs), Dynamic Time Warping (DTW). Applying Voice Activity Detection before extracting MFCC features,we removed unnecessary information from the speech signal, which made our algorithm faster and suitable for real-time speech processing. Our proposed method worked quite well in low SNR condition and in a situation when very little speech data is available for training.**

***Index Terms*—Speech Recognition,Dynamic Time Warping(DTW), MFCC, VAD**

I. Introduction

Speech processing is one of the most exciting areas of digital signal processing. It is a special case of pattern recognition. Speech recognition system can be classified as isolated word detection, connected word detection, continuous speech detection, spontaneous speech detection system. Generally, there are two phases in a speech recognition system - training and testing. In training phase, parameters are estimated from a large number of training data, and during testing phase , the features of test data are matched with the training data and decision based on the best matching is given.

Various methods have been proposed for speech recognition over the years. Segmented analysis, Mel frequency Spectral coefficients, Gaussian mixture model (GMM) [1] are used extensively for simple speech recognition. Hidden Markov Models (HMM) are also used a lot for its reliability [2]. For high level speech recognition, several classifiers based and channel compression methods have been proposed [3]. SVM has become a powerful tool in this area [4]. Artificial Neural Networks is another classifier in this domain with good accuracy [5]. But training a neural network requires huge data and training process is not that much easy. Although these methods may give better accuracy, they are complex in nature. They are not suitable for simple embedded application where the word detection job is far more easy and only challenge to is to do it in real time.

In our proposed method, we have proposed a simplified process to detect isolate words in simple embedded applications. Here we have used Voice Activity Detection(VAD) before extracting MFCC matrix from speech data. Applying VAD before feature extraction removes unnecessary components from the voice signal. MFCC matrices for different words are stored in a database. In testing period, voice signal is again passed through Voice Activity Detection algorithm, then MFCC matrix is calculated. For matching purpose, Dynamic Time Warping (DTW) is used which matches different MFCC matrices irrespective of their dimension. The word that gives shortest DTW distances is considered as the matched word.

II. Methodology

Our speech recognition system have a training phase and a test phase. In the training mode, different voice samples of a word are recorded by audiorecorder tool of MATLAB. This recorded voice is labeled, then preprocessed. The preprocessing stage starts with Voice Activity Detection algorithm, which removes noisy and inactive signal from voice, then MFCC features are calculated from the modified voice signal. The features are saved in a folder for comparison. In detection mode, a voice sample is recorded, then VAD is applied and MFCC features are calculated. The similarities between these features and the features of recorded voice signal are calculated - from which the best matched voice is decided.

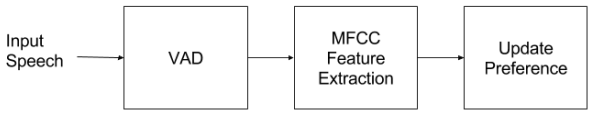

Fig. 1. Training Phase Block Diagram

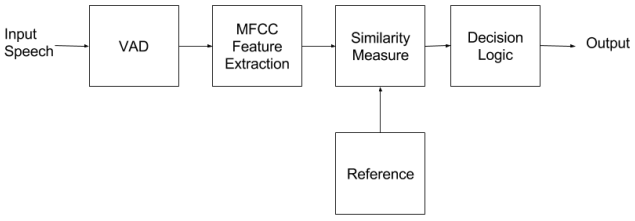

Fig. 2. Testing Phase Block Diagram

A. Pre-processing(Voice Activity Detection):

The first task in speech recognition is to extract the meaningful signal that actually contains the speech and discard the other. Voice Activity detection is a necessary step in this

process. It is a popular tool to detect unnecessary and silent part of a signal. VAD algorithm works by taking audio signal, processing the signal and extracting features from the signal, model fitting and making decision based on threshold.There are many algorithms developed in this literature.Most of them differ by the type and number of features used. The features that can be used in VAD are Fourier coefficients, periodicity, zero-crossing rate [6], spectral flatness and many more. In our method, we have used three features - short-term energy, spectral flatness, and dominant frequency component[7].

Energy is the most common features in speech detection. But in low SNR condition, only energy is not enough to detect speech signal. For this reason we have used a second feature called Spectral Flatness Measure(SFM), which detects the noisiness of a signal. The equation for this is :

$$SFM_{dB} = 10log(\frac{G_m}{A_m}) \quad (1)$$

Here G_m and A_m arithmetic and geometric means of speech spectrum respectively. The third feature is dominant frequency component of the speech frame.

To determine the threshold value, we have calculated the values of different features in active and inactive segments of different audio signal of single words. The values of are given below :

TABLE I
INACTIVE SEGMENT FEATURES

Energy	Dominant	Flatness
0.0593	0.0938	-inf
0.1631	0.1514	-1.7063
0.3214	0.1851	-2.4147
0.4845	0.2814	-2.6867
0.8602	0.4009	-2.6232
0.8553	0.3858	-3.1424

TABLE II
ACTIVE SEGMENT FEATURES

Energy	Dominant	Flatness
14.1493	1.7758	-5.7462
33.7600	3.2180	-5.5772
73.0898	4.0163	-7.1328
87.6643	4.1686	-6.5090
104.6514	5.0374	-2.8602
153.0714	v5.5791	-3.2466

By observing the features in different active and inactive segment, we have chosen the threshold values as energy : 5, flatness : -1, dominant : 0.6.

B. Feature Extraction(MFCC)

Human perception of frequency contents of sounds for speech signal does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the mel scale. The melfrequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000Hz[8]. As a reference point, the pitch of a 1 KHz tone, 40dB above the perceptual hearing threshold, is defined as 1000 mels.Therefore we can use the following approximate formula to compute the mels

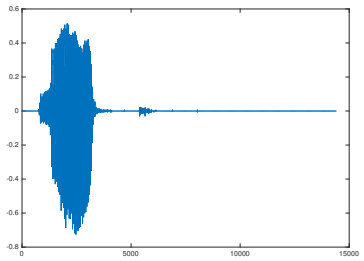

Fig. 3. Voice signal before VAD is applied

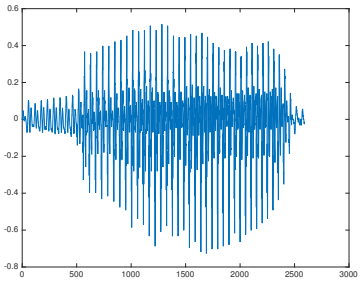

Fig. 4. Voice signal after VAD is applied

for a given frequency :

$$M(f) = 1125ln(1 + \frac{f}{700}) \quad (2)$$

The overall process of MFCC is shown in Figure 5 [9] :

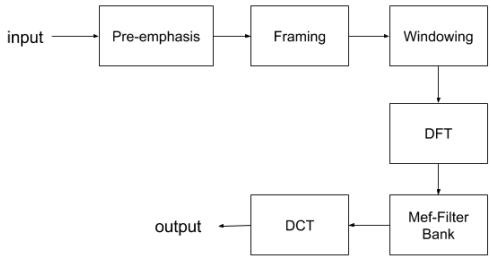

Fig. 5. MFCC Block Diagram

The speech signal is first preemphasised by passing the signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

$$y[n] = x[n] - 0.95x[n-1] \quad (3)$$

Then the voice signal is divided into frames of N samples. Adjacent frames are being separated by M (M ¡ N). Typical values used are M = 100 and N= 256. Hamming window is used by considering the next block in feature extraction processing chain and integrates all the closest frequency lines.After that, FFT is applied to convert each frame of N samples from time domain into frequency domain. As the frequencies range in FFT spectrum is very wide, bank of filters according to Mel scale is then performed. Each filters

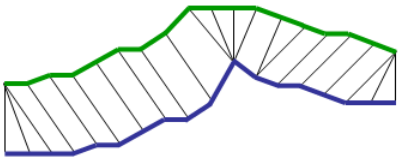

Fig. 6. Dynamic Time Warping

magnitude response is triangular in shape and equal to unity at the centre frequency and decrease linearly to zero at centre frequency of two adjacent filters [10]. Using Discrete Cosine Transform (DCT), log Mel spectrum is converted into time domain. The result of the conversion is called Mel Frequency Cepstrum Coefficient[10].

In our approach, we have divided our testing data in small windows of 25 ms, then we extracted 13 MFCC features from each window. The first MFCC feature denotes the intensity of the voice. As we do not want our algorithm to be intensity dependent, we ignored the first feature, and constructed a feature vector of 2-13 MFCC features.Number of feature in MFCC depends on the duration of voice activity. A word slowly uttered may result in larger number of MFCC features than the same word uttered fast.

C. Similarity Measure(Dynamic Time Warping)

Dynamic time warping is a popular method to find an optimal alignment between two given time series. The two sequences are aligned by warping the time axis iteratively until an optimal match is found [11]. Two sequences can be arranged on the sides of a grid, with one on the top and the other up the left hand side. Both sequences start on the bottom left of the grid. Inside each cell a distance measure can be placed, comparing the corresponding elements of the two sequences. To find the best match or alignment between these two sequences one need to find a path through the grid which minimizes the total distance between them. The procedure for computing this overall distance involves finding all possible routes through the grid and for each one computes the overall distance. The overall distance is the minimum of the sum of the distances between the individual elements on the path divided by the sum of the weighting function.

In our work, we have modified slightly the warping function. Here we are warping between the test and train feature vectors instead of the traditional temporal data of two instances. If $\mathbf{x_{train}}$ and $\mathbf{x_{test}}$ are the feature vectors extracted from the train and test data respectively, then we define their warping distance in the following manner:

$$d = ||\mathbf{x_{test}} - \mathbf{x_{train}}||_{\mathbf{2}} \quad (4)$$

Now this distance calculation is used to align the test and train data for optimal distance.

III. Result

We have calculated our result for five words. They are - water, stop, medicine, left, food. Five samples of each word were recorded with MATLAB audio-recorder tool. For testing, voice was recorded with the same toolbox. After applying VAD, distance was measured with DTW algorithm for every sample of word in database, and then average cost was calculated and assigned as the final cost of that word. In this way, costs were calculated for every word. The word which gives the lowest cost is the matched word. From our experiment, we have seen that our algorithm matched all voices perfectly in a low SNR condition.

TABLE III
DTW-BASED DISTANCE SCORE

Test \ Train	water	stop	medicine	left	Food
Water	803.44	926.64	987.4	868.08	848.36
Stop	961.05	442.63	876.43	606.53	685.34
Medicine	1218.1	1278.3	861.51	989.84	1117.8
Left	1135.5	1227.9	1010.0	920.82	1036.0
food	1101.6	1144.2	1014.5	1029.0	899.07

IV. Conclusions

Though our system was trained with the voice of a single person, it can detect the same word of another person with good accuracy. In other words, it shows some tolerance. Its accuracy increases considerably with inclusion of training data from several persons. In a very low SNR condition, this simple method sometimes gives erroneous results. But our algorithm works accurately when only one person's voice is used in the recognition system and even when small number of training samples are available. The slowest block in our proposed system is DTW block. Implementing DTW algorithm in C programming language instead of MATLAB, we have tested our algorithm in real-time and have seen that it produces the same accuracy.

References

[1] Campbell, William M., Douglas E. Sturim, and Douglas A. Reynolds. "Support vector machines using GMM supervectors for speaker verification." IEEE signal processing letters 13.5 (2006): 308-311.

[2] L. Rabiner, A tutorial on Hidden Marcov Model and selected applications in Speech Recognition, Proceedings of the IEEE, Vol.77,No.2, 1989, pp. 257-286.

[3] Hohmann, Volker, and Birger Kollmeier. "The effect of multichannel dynamic compression on speech intelligibility." The Journal of the Acoustical Society of America 97.2 (1995): 1191-1195. APA

[4] Zahi N.Karam,William M.Campbell A new Kernel for SVM MIIR based Speaker recognition MIT Lincoln Laboratory, Lexington, MA, USA.

[5] Trentin, Edmondo, et al. "Neural networks for speech recognition." Spoken Dialogues with Computers (1998): 311-361. APA

[6] Muda, Lindasalwa, Mumtaj Begam, and I. Elamvazuthi. "Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques." arXiv preprint arXiv:1003.4083 (2010).

[7] M. H. Moattar and M. M. Homayounpour, "A Simple But Efficient Real-time Voice Activity Detection Algorithm," in *EUSIPCO*, 2009.

[8] Kola, Jonathan, Carol Espy-Wilson, and Tarun Pruthi. "Voice activity detection." Merit Bien (2011): 1-6.

[9] Zaidi Razak,Noor Jamilah Ibrahim, emran mohd tamil,mohd Yamani Idna Idris, Mohd yaakob Yusoff,Quranic verse recition feature extraction using mel frequency ceostral coefficient (MFCC),Universiti Malaya.

[10] Jamal Price, sophomore student, Design an automatic speech recognition system using maltab, University of Maryland Estern Shore Princess Anne.

[11] Tiwari, Vibha. "MFCC and its applications in speaker recognition." International Journal on Emerging Technologies 1.1 (2010): 19-22.

[12] Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P., 2006. The HTK Book (for HTK Version 3.4.1). Engineering Department, Cambridge University.