Isolated Digit Recognition Using MFCC AND DTW

MarutiLimkar[a], RamaRao[b] & VidyaSagvekar[c]

[a]Terna collegeof Engineering, Department of Electronics Engineering, Mumbai University, India
[b]Vidyalankar Institute of Technology, Department ofElectronics and Telecommunication Engineering, Mumbai University, India
[c]K.J.Somaiya Institute of Engineering & Information Technology, Department of Electronics Engineering Mumbai University, India
E-Mail : limkar_maruti@rediffmail.com, b.ramarao@vit.edu.in & vidyarsagvekar@gmail.com

Abstract **-** This paper proposes an approach to recognize spoken English words corresponding to digits zero to nine in an isolated way by different male and female speakers. The endpoint detection, framing, normalization, Mel Frequency Cepstral Coefficient (MFCC) and DTW algorithm are used to process speech samples to accomplish the recognition. The algorithm is tested on speech samples.The system is then applied to recognition of isolated word English digits, that is 'one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eight' and 'nine'. The algorithm is tested on speech samples that are recorded. The results show that the algorithm managed to recognize almost 90.5% oftheEnglish digits for all recorded words.

Keyword*-Speech Recognition, Dynamic Time Warping, Mel Frequency Cepstrum Coefficient.*

I. INTRODUCTION

Speech recognition is a popular and active area of research, used to translate words spoken by humans so as to make them computer recognizable. It usually involves extraction of patterns from digitized speech samples and representing them using an appropriate data model. These patterns are subsequently compared to each other using mathematical operations to determine their contents. In this paper we focus only on recognition of words corresponding to English numerals zero to nine.

In many speech recognition systems, end point detection and pattern recognition are used to detect the presence of speech in a background of noise. The beginning and end of a word should be detected by the system that processes the word. The problem of detecting the endpoints would seem to be easily distinguished by human, but it has been found complicated for machine to recognize. Instead in the last three decades, a number of endpoint detection methods have been developed to improve the speed and accuracy of a speech recognition system.

The main challenges of speech recognition involve modeling the variation of the same word as spoken by different speakers depending on speaking styles, accents, regional and social dialects, gender, voice patterns etc. In addition background noises and changing of signal properties over time, also pose major problems in speech recognition.

This paper proposes a speech recognition algorithm for English digit from 0 to 9.This system consists of speech processing inclusive of digit boundary and recognition which uses zero crossing and energy techniques. Mel Frequency Cepstral Coefficients (MFCC) vectors are used to provide an estimate of the vocal tract filter. Meanwhile dynamic time warping (DTW) is used to detect the nearest recorded voice.

The paper is organized as follows: section II describes the proposed approach, section III tabulates details of Experimentations done and results obtained section IV provides overall conclusions.

II. PROPOSED APPROACH

This paper proposes an approach to recognize automatically digits 0 to 9 from audio signals generated

by different individuals in a controlled environment. It uses a combination of features based on Voice Activity Detection, Mel Frequency Cepstral Coefficient (MFCC). A Dynamic Time Warping (DTW) is used to discriminate the speech data models into respective classes.

A. Feature Extraction:

The general methodology of audio classification involves extracting discriminatory features from the audio data and feeding them to a pattern classifier. Different approaches and various kinds of audio features were proposed with varying success rates. The features can be extracted either directly from the time domain signal or from a transformation domain depending upon the choice of the signal analysis approach. Some of the audio features that have been successfully used for audio classification include Mel Frequency Cepstral Coefficients (MFCC).

1. Mel Frequency Cepstrum Coefficients

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 spacing below 1000Hz and a logarithmic spacing above 1000Hz.As a reference point the pitch of a 1KHz tone 40dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore following approximate formula used to compute the mels for a given frequency f in Hz.

$$Mel(f) = 2595 * log10(1 + \frac{f}{700})$$

Here used one filter for each desired mel-frequency component. That filter bank has a triangular band pass frequency response and the spacing as well as the bandwidth is determined by a constant mel frequency interval. The mel scale filter bank is a series of 1 triangular band pass filters that have been designed to simulate the band pass filtering believed to occur in the auditory system. This corresponds to series of band pass filters with constant bandwidth and spacing on a mel frequency scale.Fig.1 Shows Block Diagram of Mel Frequency Cepstrum Coefficient.

1.1 Pre-Emphasis

This step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency. First order FIR filter is used.

$x'(n)=x(n)-a*x(n-1)$ $0 \leq a \leq 1$

typical value of 'a' is 0.95 (> 20 dB gain for high frequency).

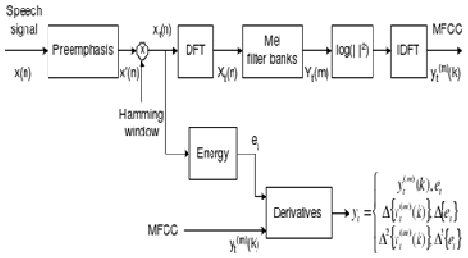


Fig.1 Block Diagram of Mel Frequency Cepstrum Coefficeints

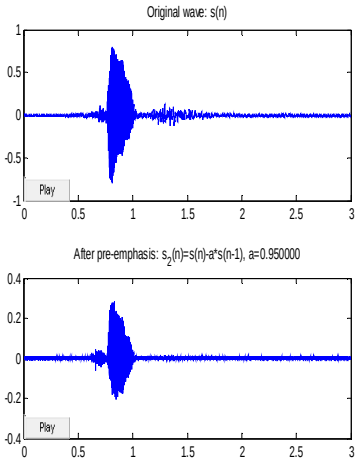


Fig. 2 Example of Pre-Emphasis

1.2 Windowing

A traditional method of spectral evaluation is reliable in case of stationary signal. Nature of signal changes continuously with time. For voice reliability can be ensured for a short time. Audio signal is continuous. Processing cannot wait for last sample.Processing complexity increases exponentially. It is important to retain short term features. Short time analysis is performed by windowing the signal.

Normally Hamming Window is used. The Hamming function is given by

- Hamming window:

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{L-1}\right) \qquad 0 \geq n \geq L-1$$

0 otherwise

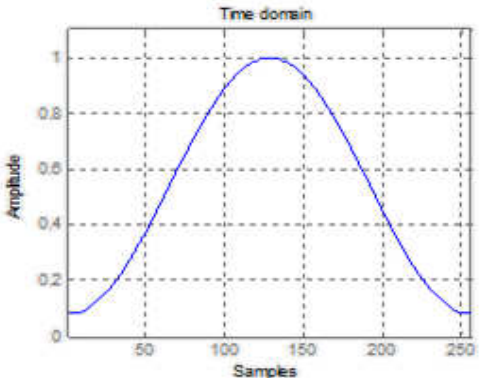


Fig.3 Hamming Window

1.3 Discrete Fourier Transform

To convert each frame of N samples from time domain into frequency domain. The Fourier Transform is to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain.

$$X[k] = \sum_{n=0}^{N-1} x[n] e^{-j2kn\frac{\pi}{N}}$$

Complexity of DFT is N^2 and Fast Fourier Transform (FFT) is $N*\log_2(N)$. In general, choose N=512, 1024 or 2^m. Magnitude and phase at N equidistant digital frequencies between 0 and 2π (rad/sec). Corresponding analog frequencies are kFs/N Hz, k=0, 1,..., N-1.

1.4 Mel Filter Banks

Human hearing is not equally sensitive to all frequency bands. It is less sensitive at higher frequencies, roughly greater than 1000 Hz. i.e. human perception of frequency is non-linear.

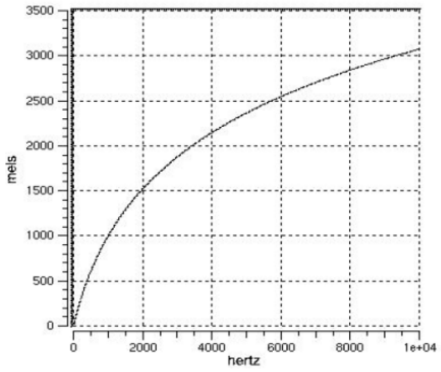


Fig.4 Mel Scale

A set of filters with triangular band pass frequency response believed to occur in the auditory system.One filter is assigned for each desired mel-frequency component. Spacing and bandwidth is determined by a constant mel-frequency interval.

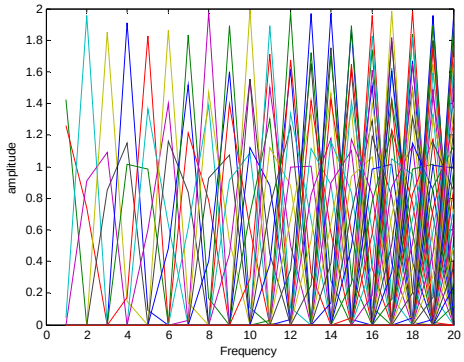


Fig.5 Mel Filter Banks.

1.5 The Cepstrum

In this final step, the log mel spectrum is converted back to time. The result is called the mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients (and so their logarithm) are real numbers, we can convert them to the time domain using the Discrete Cosine Transform (DCT).

Human response to signal level is logarithmic. Human response is less sensitive to slight differences in amplitude at high amplitudes than low amplitudes. Logarithm compresses dynamic range of values making feature extraction less sensitive to dynamic variation. Makes frequency estimates less sensitive to slight variations in input (power variation due to speaker's mouth moving closer to mike). Magnitude of DFT discards unwanted phase information.

1.6 Mel Frequency Cepstrum Computation

IDFT converts frequency samples to time

Generally J = 12 or 16

$$y[k] = \sum_{m=1}^{M} \log\left(|Y(m)|\right) \cos\left(k(-0.5)\frac{\pi}{M}\right),$$

$$k = 0, \ldots . J$$

2. Feature Vector Matching

The features of the speech signal are in the form of *N* dimensional feature vector. For a segmented signal that is divided into *M* segments, *M* vectors are determined producing the *M x N* feature matrix. The *M x N* matrix is created by extracting features from the utterances of the speaker for selected words or sentences during the training phase. After extraction of the feature vectors from the speech signal, matching of the templates is required to be carried out for speaker recognition. This process could either be manual (comparison of spectrograms visually) or automatic. In automatic matching of templates, speaker models are constructed from the extracted features. There after a speaker is authenticated by comparison of the incoming speech signal with the stored model of the claimed user.

2.1 Dynamic Time Warping

A many-to-many matching between the data points in time series x and the data point in time series y matches every data point *xi* in *x* with at least one data point *yj* in *y*, and every data point in *y* with at least a data point in *x*. The set of matches *(i; j)* forms a warping path.We define the DTW as the minimization of the lpnorm of the differences over all warping paths. A warping path is minimal if there is no subset of forming an warping path: for simplicity we require all warping paths to be minimal. In computing the DTW distance, we commonly require the warping to remain local. For time series *x* and *y*, we align values *xi* and *yj* only if *w* for some locality constraint w(0) . When *w* = 0, the DTW becomes the local distance whereas when *w(n)*, the DTW has no locality constraint. The value of the DTW diminishes monotonically as w increases.

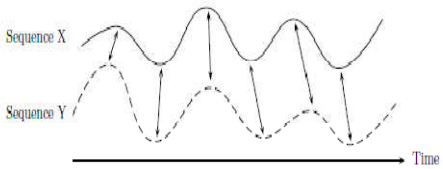


Fig. 6 Time Alignment of two time independent sequences.

2.1 DTW Constraints

The warping path is typically subject to several constraints:

Boundary conditions: *w*1 = (1,1) and *w*K = (*m*,*n*), simply stated, this requires the warping path to start and finish in diagonally opposite corner cells of the matrix.

Continuity: Given *w*k = (*a*,*b*) then *w*k-1 = (*a'*,*b'*) where *a*–*a'* £1 and *b*-*b'* £ 1. This restricts the allowable steps in the warping path to adjacent cells (including diagonally adjacent cells).

Monotonicity: Given *w*k = (*a*,*b*) then *w*k-1 = (*a'*,*b'*) where *a*–*a'* ³ 0 and *b*-*b'* ³ 0. This forces the points in *W* to be monotonically spaced in time.

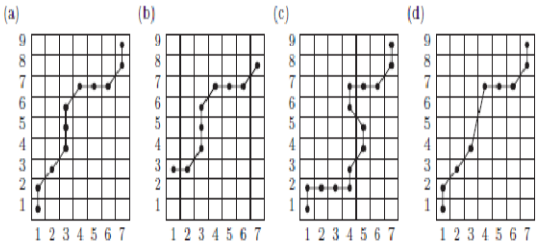
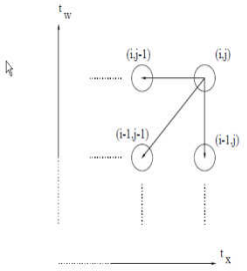


Fig.7 DTW constraints (a) Admissible warping paths satisfying the conditions. (b) Boundary Condition is violated (c) Monotonicity condition is violated (d) Step Size Condition is violated.

2.1.1 The Algorithm



$$\delta(i,j) = \min \begin{cases} \delta(i,j-1) + d(\vec{w}_i,\vec{x}_j) \\ \delta(i-1,j-1) + 2\cdot d(\vec{w}_i,\vec{x}_j) \\ \delta(i-1,j) + d(\vec{w}_i,\vec{x}_j) \end{cases}$$

Fig.8 Local Distance

1. Start with the calculation of g(1,1) = d(1,1).

Calculate the first row g(***i***, 1) =g(***i***–1, 1) + d(***i***,1).

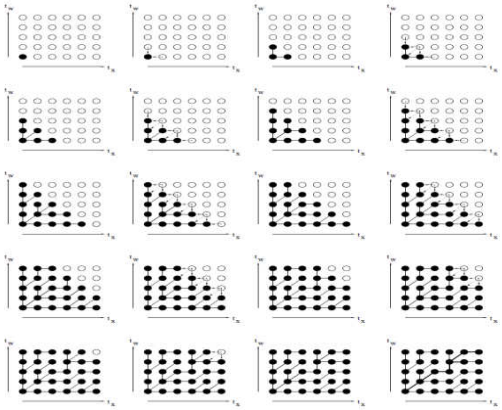


Fig. 9 Optimal Path Assignment

2. Calculate the first column g(1, ***j***) =g(1, ***j***) + d(1, ***j***).
3. Move to the second row g(***i***, 2) = min(g(***i***, 1), g(***i***–1, 1), g(***i*** – 1, 2)) + d(***i***, 2). Book keep for each cell the index of this neighboring cell, which contributes the minimum score (red arrows).
4. Carry on from left to right and from bottom to top with the rest of the grid g(***i***, ***j***) = min(g(***i***, ***j***–1), g(***i***–1, ***j***–1), g(***i*** – 1, ***j***)) + d(***i***, ***j***).
5. Trace back the best path through the grid starting from g(***n***, ***m***) and moving towards g(1,1) by following the red arrows.

Hence the path which gives minimum distance after testing with the feature vectors stored in the database is the identified speaker.

III. RESULT

After running the algorithm, the results are obtained. The system requires the user to record numbers 0 until 9 in the English language. After that the system saves the recorded voice into a English digit directory. Then, the user is required to record any single number between 0 and 9 as shown in Figure 12. The input word is then recognized as the word corresponding to the template with the lowest matching score. The recognition is implemented using DTW where the distance calculation is done between the tested speech and the reference word bank.

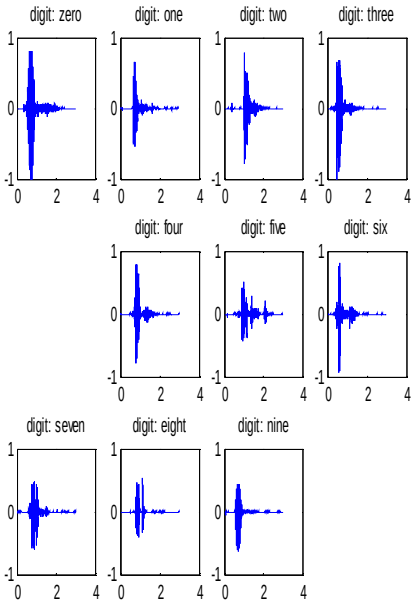


Fig. 10 Screenshot output after recording ten digits in the English language

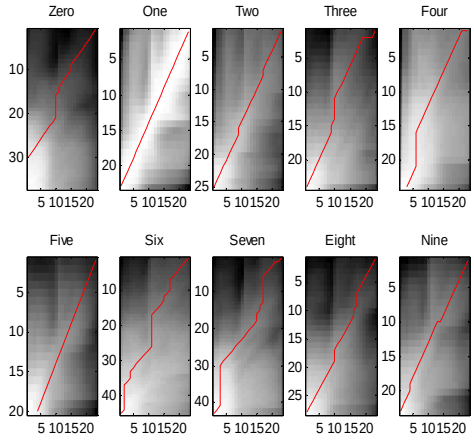


Fig. 11 Screenshot output of DTW path taken for recognition of utterance of ‘one’ with respect to other digit

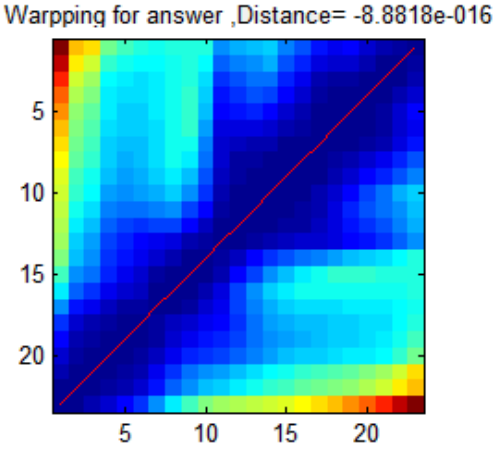


Fig.12Screenshot of DTW path taken for recognition of utterance of ‘one’

The recognition algorithm is then testedfor accuracy. The test is limited to digits from 0 to 9. Random utterance of numbers is doneand the accuracy of 100 samples of numbersis analyzed. The results obtained from theaccuracy test are about 90.5% of accuracy. Theresults obtained are as displayed in Table 1.Most of the time, the inaccuracy of recognition is due to sudden impulses of noise or a sudden drastic change in the voice tone.

Table 1. Accuracy test results

Word	Accuracy
Zero	80.0
One	95.0
Two	80.0
Three	100.0

Four	90.0
Five	100.0
Six	80.0
Seven	100.0
Eight	100.0
Nine	80.0
Average	90.5

IV. CONCLUSIONS

This paper has shown a speech recognition algorithm for English digits using MFCC vectorsto provide an estimate of the vocal tract filter. Meanwhile, DTW is used to detect the nearest recorded voice with appropriate constraint. The results showed a promising English digit speech recognition module. Recognition with about 90.5% accuracy can be achieved using this method, which can be further increased with further research and development.

REFERENCES

[1] Gold, B. and Morgan, N. (2000). Speech and Audio Signal Processing.1st ed. John Wiley and Sons, NY, USA, 537p.

[2] Deng, L. and Huang, X. (2004). Challenges in adopting speech recognition. Communications of the ACM, 47(1):69-75.

[3] Milner, B.P. and Shao, X. (2002). Speech reconstruction from mel-frequency cepstral coefficients using a source-filter model.Proceedings of the7th International Con- ference on Spoken Language Processing (ICSLP) 2002; Sept 16-20, 2002, Denver, Colorado, USA, p. 2,421-2,424.

[4] Rabiner, L.R. and Sambur, M.R. (1975). An algorithm for determining the endpoints of isolated utterances. The Bell System Technical J., 54(2):297-315.

[5] Rabiner, L.R. and Schafer, R.W. (1978). Digital Processing of Speech Signals.1 st ed. Prentice-Hall Inc., Englewood Cliffs, NJ, USA, 509p.

[6] Deng, L. and Huang, X. (2004). Challenges in adopting speech recognition. Communications of the ACM, 47(1):69-75.

◈◈◈