Digit Recognition based on Euclidean and DTW

Sreeja Nair
EXTC Department.
FCRIT
Vashi-400703, Navi Mumbai, India
sreejan791@gmail.com

Milind Shah
EXTC Department.
FCRIT
Vashi-400703, Navi Mumbai, India
milind05in@yahoo.co.in

ABSTRACT
This paper describes the implementation of two isolated digit recognition techniques and is a comparison between the algorithms implemented. Any digit recognition comprises of mainly two stages feature extraction and similarity evaluation. Here, two feature extraction techniques, namely linear predictive cepstral coefficients (LPCC) and mel frequency cepstral coefficients (MFCC) are implemented and the similarity evaluation is done using Euclidean distance and Dynamic Time Warping (DTW). In DTW both single and averaged template matching is done. The results obtained for these algorithms are perused, compared and conclusions are drawn.

Keywords
Digit recognition, linear predictive cepstral coefficients, mel frequency cepstral coefficients, euclidean distance, dynamic time warping.

1. INTRODUCTION
Speech recognition is a process by which a computer recognizes a human speech and converts it into text. In particular, speech recognition for spoken digits finds a wide variety of applications. Some of them are banking by voice, data input to a computer, hands off and eyes off number dialing in mobiles, etc [1]. In practice speech recognition algorithms are complex due to inter speaker variations as well as intra speaker variations. Inter speaker variation is the difference in the same speech from person to person in terms of pronunciation, accent, etc. whereas intra speaker variability is the difference in utterance of speech by the same person. This is because humans can never produce words exactly the same way twice [2]. Moreover other factors such as slang, dialect, accent, etc are responsible for further variation of speech between speakers.

Speech recognition involves four steps namely, pre-processing, feature extraction, similarity evaluation and decision making [3]. Pre-processing is to prepare the signal for further processing. Pre-emphasis, end-point detection, etc are carried out in this stage. Feature extraction and similarity evaluation are the most important steps amongst all. Since speech is highly redundant, it is impractical to process, store and transmit the signal as it is. Hence a speech signal is represented in terms of a few number of parameters. There are different parameter or feature extraction techniques for speech recognition like LPC, LPCC, MFCC, PLP, etc. which are implemented by various researchers. In [1], Rabiner presents an initial implementation of digit recognition using parameters like LPC, log energy, zero crossing rate, etc. Atal in [4], has used LPCC for speaker recognition. He has also introduced the concept of frame wise averaging the coefficients of LPCC, which has slightly increased the accuracy of recognition [This averaging method has been used in this paper.] Similarly, MFCC based feature extraction has been carried out in [5] and Perception Linear Prediction (PLP) and Euclidean distance based speech recognition has been implemented in [6]. Once the features are extracted for a given signal, they have to be compared with the feature of the references stored which depends on the vocabulary of the recognition system. Similarity evaluation can be done using template based techniques like Euclidean distance and DTW or network models like Hidden Markov Models (HMM) and Neural Networks (NN).

There are two digit recognition techniques implemented in this paper. In the first method, the two feature extraction techniques are implemented and the feature vectors are compared using Euclidean distance whereas in the second method the same feature extraction techniques are compared using DTW. Section II describes the feature extraction techniques whereas Section III gives details about the similarity evaluation techniques. Section IV explains the implementation and results obtained are perused in Section V

2. FEATURE EXTRACTION TECHNIQUES

2.1 Linear Predictive Cepstral Coefficients(LPCC)
Linear prediction refers to predicting the present speech sample using the past samples. The predicted value is given by (1) [1,2]:

$$\hat{s}_n = \sum_{k=1}^{p} s_{n-k} a_k \quad (1)$$

where a_k are the prediction coefficients and s_{n-k} are the previous samples used to obtain the present sample $\hat{s}_n$ The prediction coefficients are obtained by minimizing the prediction error in the least squares sense using autocorrelation. The order or the total number of prediction coefficients is denoted by p. If G is the gain of LPC then, the cepstral coefficients C_m are obtained from the LPC parameters using (2), (3) and (4)[2]

$$C_0 = \ln(G) \quad (2)$$

$$C_m = -a_m + \frac{1}{m}\sum_{k=1}^{m}\left[-(m-k)a_k C_{(m-k)}\right], 1 \le m \le p \quad (3)$$

$$C_m = \sum_{k=1}^{p}\left[\frac{-(m-k)}{m} a_k C_{(m-k)}\right], p < m < n \quad (4)$$

The block diagram of LPCC computation [2] is as shown in Fig.1

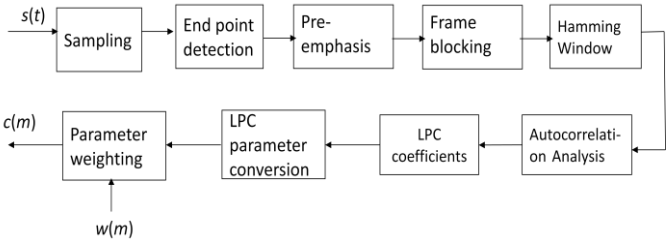

Fig.1 The block diagram to find Linear Predictive Cepstral Coefficients [2]

The speech signal recorded is sampled and the end point detection is done to remove the silence from the speech using both the short time energy and zero crossing rate as implemented

by Rabiner in [7]. The radiation loss affecting the higher frequencies is compensated with a pre emphasis filter. Once this pre processing is done the speech is divided into frames and windowed using Hamming window with 50% overlap. For each frame the LPC coefficients and consequently the cepstral coefficients are derived using the equations discussed before. After weighting the coefficients to give higher weights to lower frequency information, the features obtained are stored.

2.2 Mel Frequency Cepstral Coefficients (MFCC)

MFCC is another cepstral based feature extraction technique popularly used in recent times. It is based on human perception of speech. Humans are more sensitive to lower frequency sounds than higher frequencies. This factor is taken into consideration in mel filter banks. The mel scale is a non linear frequency scale [1] and is spaced logarithmically as shown in (5)

$$M(f) = 1125 \log_e (1 + f / 700) \quad (5)$$

where M(f) is the mel scale value for each frequency f. The frequency response of Mel filters as obtained from MATLAB is shown in Fig. 2. The x axis gives the frequency and y axis gives the amplitude of the mel filters normalized to 1.

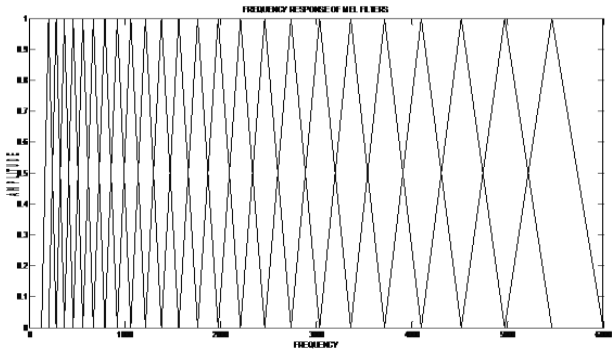

Fig.2. Frequency response of Mel filters

The calculation of Mel filter coefficients [8] involves few steps as shown in Fig. 3. The filter bank used here consists of 40 filters out of which outputs of 13 filters are taken which are adequate to represent the signal accurately.

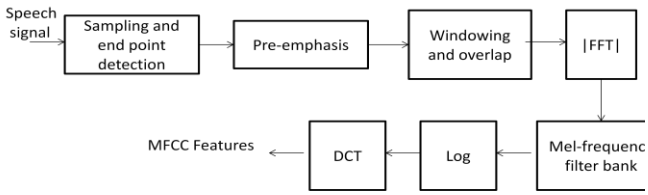

Fig 3. The block diagram to find Mel Frequency Cepstral Coefficients [8]

Similar to LPCC, the speech signal is initially sampled, pre emphasised and framed using overlapping Hamming windows. Next, the Fourier Transform of each speech segment is taken and multiplied with a mel filter bank consisting of 40 filters. After taking the logarithm of each output from each filter in the bank, its Discrete Cosine Transform (DCT) is computed to concentrate the energy on the lower frequencies. The first thirteen coefficients from the output of DCT is taken and stored as the MFCC features [8].

For a digit recognition system, the features sets for all digits i.e. from zero to nine are stored for desired speakers in the reference database. This is called training stage. Once this is done, an unknown input that has to be recognized, called the test signal is given to the system. The features of this signal are also computed the same way and compared with the features of each of the digits in the reference database using one of the two similarity evaluation techniques explained in the next section.

3. SIMILARITY EVALATION TECHNIQUES

3.1 Euclidean Distance

It is a primary distance measurement technique. It involves templates where each utterance is converted into a predetermined number of features called templates [1,9]. The similarity of two templates is inversely proportional to the distance between them. Euclidean distance is an L2 norm distance given by (6):

$$d(x, y) = \sqrt{\sum_{i=1}^{N} (x_i - y_i)^2} \quad (6)$$

$$\bar{c}_n = \frac{1}{Q} \sum_{n=1}^{Q} c_n \quad (7)$$

where x and y the two templates representing reference and test signal and Q is the total number of features in each template. In this paper, for Euclidean distance one feature vector is obtained for each digit by averaging the coefficients obtained across each frame using (7) as was done by Atal in [4]

3.2. Dynamic Time Warping

DTW is a traditional method used in speech recognition and is a popular method even in modern applications. In this paper single reference template method is used where the features of the shortest utterance of each digit spoken by each speaker is stored as reference. When a test digit spoken by the same speaker is given, its features are compared to those of the reference using DTW. In DTW when a test input is given to a recognition system, it computes the global distance between each of the references and the test signal. Next it finds the reference speech utterance for which the computed distance is least and decides that reference signal as the recognized speech. The global distance can be computed using the formula given in (8) [2-3] [13-14]

$$D(i, j) = d(i, j) + \min[D(i-1, j-1), D(i-1, j), D(i, j-1)] \quad (8)$$

where d is the local or distance error between two frames i and j of the test and reference signal respectively and D(1,1) = d(1,1). The DTW calculation can be represented using Fig. 4

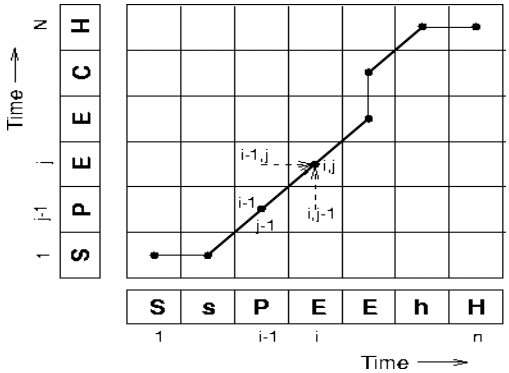

Fig. 4 An illustration of finding the optimum path [13]

The reference templates for DTW are created in two ways in this paper.

1. *Single Reference*: In this method one random reference from the speaker is considered as the reference template. Though this method is not very robust it gives comparatively better results for speaker dependent application.
2. *Average Reference:* In this method the features of four speech signals of each digit is averaged. It is based on Abdullah's paper [15] on cross word reference templates. To select the four speech signals that needs to be averaged, first the average of all the lengths of each digit is taken. The speech signal closest to the average length is taken as the main template. Now another three templates which are longer than that are time aligned to the length of the initial template using DTW. Once this is done the features of each of these four templates are averaged frame wise. The resultant template is the reference pattern. It can be shown in Fig. 5. Now during recognition all the templates are compared to this averaged template using DTW.

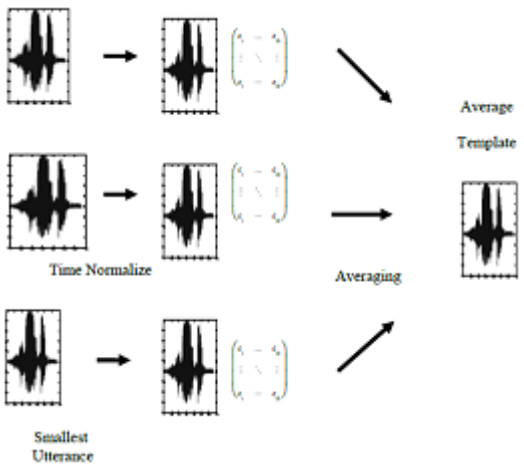

Fig. 5 Average template matching [15]

4. IMPLEMENTATION

In implementation the first step is to record the digits spoken by the different speakers. In this paper 100 utterances of each speaker is recorded i.e. each speaker repeats each digit 10 times. In this way recordings of 3 male and 3 female speakers were taken. It is recorded using Praat software at a sampling frequency of 12 kHz to satisfy the Nyquist criteria. The sampled signal is then subjected to end point detection. Next, it is segmented or framed with frame duration of 20 ms. It is then windowed using Hamming window with an overlap of 50%. Overlap is done so that there is no loss of data at the edges of the window. For each frame, LPCC as well as MFCC features are extracted and stored. The reference database is created with these features for each speaker. Now in the testing period, a test signal is given which is a digit recorded by the speaker whose reference is considered. The test signal undergoes the same process of endpoint detection, framing, windowing and feature extraction. These features are compared with those already stored. Euclidean distance and DTW are used for comparison. While the averaged feature vectors are compared in Euclidean distance, in DTW the features in each frame are kept intact and compared.

5. RESULTS

The parameter to find the accuracy of a recognition system is the Recognition Rate (RR) which is given by (9) where Ncorrect is the number of words recognized correctly and Ntotal is the total number of words in the vocabulary [5]

$$\text{Recognition Rate} = \frac{Ncorrect}{Ntotal} \times 100\% \qquad (9)$$

The Recognition Rate was calculated for each of the feature extraction technique and the two similarity evaluation technique for three male speaker and three female speakers each. The system implemented here is a speaker dependent recognition system which means the digits spoken by the same speaker are recognized. The results obtained by the experiment are summarized in Tables I , II and III.

Table 1. Recognition rates of the two feature extraction technique for male and female speaker using euclidean distance

Feature Extraction Technique	Speaker 1		Speaker 2	
	Male	***Female***	***Male***	***Female***
LPCC	74%	89%	75%	84%
MFCC	85%	70%	79%	67%

TABLE 2. Recognition rates of the two feature extraction technique for male and female speaker using dtw (single template)

Feature Extraction Technique	Speaker 1		Speaker 2	
	Male	***Female***	***Male***	***Female***
LPCC	85%	91%	81%	96%
MFCC	90%	87%	94%	94%

TABLE 3. Recognition rates of the two feature extraction technique for male and female speaker using dtw (averaged template)

Feature Extraction Technique	Speaker 1		Speaker 2	
	Male	***Female***	***Male***	***Female***
LPCC	92%	96%	95%	99%
MFCC	97%	97%	97%	96%

From the tables it is observed that for both male and female speakers the Recognition Rate is greater for Dynamic Time Warping than for Euclidean distance for both the feature extraction techniques. In DTW the averaging technique is superior to the single template matching technique. It is also observed that LPCC gives a better performance for female speakers while MFCC gives a superior performance for male speakers. This may be explained by the fact that formant and pitch frequencies of females are higher than males. Since LPCC gives more importance to higher formants as compared to MFCC, it seems to be a better option for females. Also, the difference between the accuracy is greater in Euclidean distance since it is not as efficient as DTW. The average RR for Euclidean distance is 83% and 92% for LPCC and MFCC respectively for male speakers and 86% and 68% for LPCC and MFCC respectively for female speakers. Similarly in DTW the average RR for male speakers is 83% and 92% for LPCC and MFCC respectively and for female speakers it is 94% and 91% in the same order. The accuracy is still better when averaging of

template is done. In this case the average RR for male and female is 97% and 96.5% respectively for MFCC and 93.5% and 97.5% for male and female respectively when LPCC is used.

6. CONCLUSION

From the results it can be concluded that though both the feature extraction techniques are cepstral based, and similarity evaluation is template based, the DTW method is superior to the Euclidean method. The reason for this increase in accuracy might be because DTW takes into account the alignment between the two sequences besides finding the lowest distance whereas in Euclidean though averaging of coefficients improves the recognition rate it still does not take into account the alignment between the words.

7. REFERENCES

[1] L. R. Rabiner and M. R. Sambur, "Some Preliminary Experiments in the Recognition of Connected Digits," IEEE Trans. on Acoust. Speech and Signal Processing, vol. ASSP-24, no. 2, pp. 170-182, April 1976.

[2] L. R. Rabiner, B. Juang and B. Yegnanarayana, "Fundamentals of Speech Recognition," 5th ed. Pearson, 2011.

[3] D. O'Shaughnessy, Speech Communications: Human & Machine, 2nd ed. Wiley-IEEE Press, 1999, pp. 367-435.

[4] B. S. Atal, "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," J. Acoust. Soc. Am., vol. 55, no. 6, pp. 1304-1312, 1974.

[5] A. A. M. Abushariah, T. S. Gunawan, O. O. Khalifa and M. A. M. Abushariah., "English Digits Speech Recognition Based on Hidden Markov Models," in International Conference on Computer and Communication Eng., Kuala Lumpur, Malaysia, May 2010.

[6] A. Revathi and Y. Venkataramani, "Speaker Independent Continuous Speech and Isolated Digit Recognition using VQ and HMM," Proc. IEEE, pp. 198-202, 2011.

[7] L. R. Rabiner and M.R. Sambur, "An Algorithm for Determining the Endpoints of Isolated Utterances" Bell Syst. Tech. J., vol. 24, no. 2,pp. 297-315, 1975.

[8] S. Savitha, "DSP Implementation of Isolated Digit Recognizer," M.Tech Dissertation, Dept. Elect. Eng., IIT, Bombay, India, 2008.

[9] A. S. Thakur and N. Sahayam, "Speech Recognition Using Euclidean Distance," International Journal of Emerging Technology and Advanced Engineering, vol. 3, no. 3, pp. 587-590, 2013.

[10] G. Nitin, "Implementation of Algorithms for Speaker Dependent Isolated Digit Recognition," M.Tech Dissertation, Dept. Elect. Eng., IIT, Bombay, India, 1997

[11] C. P. Lim, S. C. Woo, A. S. Loh and R. Osman, "Speech Recognition Using Artificial Neural Networks," Proc. IEEE, Malaysia, 2000, pp. 419-423.

[12] L. R. Rabiner and R. W. Schafer, "Digital Speech Processing for Man-Machine Communication by Voice " in Digital processing of Speech Signals, 3rd ed. Pearson Education,2009, pp. 505-516.

[13] L. R. Rabiner and C. E. Schmidt, "Application of Dynamic Time Warping to Connected Digit Recognition," IEEE Trans. ASSP, vol. -28, no. 4, pp. 377-388, Aug. 1980.

[14] H. Sakoe and S. Chiba, "Dynamic Programming Algorithm Optimization for spoken word recognition" IEEE Trans. ASSP, vol. 26, pp. 43-49, 1978.

[15] W. H. Abdulla, D. Chow and G. Sin, " Cross-words Reference Template for DTW-based Speech Recognition Systems," TENCON 2003,Conference on Convergent Technologies for Asia-Pacific Region, vol. 4, Oct. 2003, pp.1576 - 1579.

[16] L. R. Rabiner and S. E. Levinson, "Isolated and Connected word recognition- theory and application," IEEE Trans. Commun.,vol. 29, no. 5, pp. 621-658, 1981.

[17] L. Jalan and T. Palav, "Speech Recognition Based Learning System," International Journal of Engineering Trends and Technology, vol. 4, no. 2, pp. 165-169, 2013.