IETE 46th Mid Term Symposium "Impact of Technology on Skill Development" MTS- 2015
Special Issue of International Journal of Electronics, Communication & Soft Computing Science and Engineering, ISSN: 2277-9477

Comparison of Isolated Digit Recognition Techniques based on Feature Extraction

Sreeja R. Nair Milind S. Shah

Abstract- **With technological advancement telephonic and electronic transactions have increased manifold. Digit recognition plays an important role in such communication. It is also useful in security applications. This paper proposes to implement and compare some popular algorithms used for isolated digit recognition. Two feature extraction techniques namely linear predictive cepstral coefficients (LPCC) and mel frequency cepstral coefficients (MFCC) are implemented for a speaker dependent recognition system. The results obtained for these algorithms are perused and conclusions are drawn.**

Keywords- **Digit Recognition, Linear Predictive Cepstral Coefficients, Frequency Cepstral Coefficients, Euclidean distance.**

I. INTRODUCTION

With the advent of modern sophisticated technologies in telecommunication, it has been possible to significantly develop the speech recognition techniques. Speech recognition is a process by which a computer recognizes a human speech and then converts it to text. Digit recognition focuses on recognizing the digits, spoken by the speaker. 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 say exactly the same thing twice. The biggest challenge that any speech recognition system faces today is deciphering the various dialects and accents that people may have. Moreover, in natural speech, we often tend to use a lot of slang, which automated systems find difficult to understand. Hence a robust system aims at overcoming such drawbacks so that irrespective of these variations a highly accurate recognition rate is obtained.

Automatic Speech Recognition (ASR) uses two basic approaches based on whether the application is best served when the recognition starts with the speech signal or by using possible outcomes [1]. In Cognitive or bottom up approach features are extracted from the speech signal and based on these features the recognition system is designed. In the top down approach statistical signal analysis is used to derive information about the speech. In this approach, all possible outcomes depending on the application, are stored and compared with the input signal. In [2], Rabiner has implemented connected digit recognition using some preliminary parameters as features. These features include zero crossings, log energy, LPC coefficients, autocorrelation coefficients and LPC error. Several papers have implemented digit recognition based on LPCC like Atal has done in [3] where LPCC is used because of its various advantages. They are easy to determine, are independent of pitch and intensity and gives the combined information about the formant frequencies, bandwidth and glottal waveform [3]. Similarly digit recognition using MFCC has become very popular recently. It is a method modeling the human auditory perception system [4]. Though it is more complex to compute, it weighs the coefficients depending on the sensitivity of human ear.

In the next section implementation of digit recognition using both the techniques is explained. In section II and section III. The results obtained and the conclusions drawn are discussed respectively.

II. IMPLEMENTATION

A digit recognition system which is an example of ASR system mainly consists of the following stages as shown in the flow chart in Fig.1.

Fig.1. Block diagram of speech recognition system

Fig.1. Block diagram of speech recognition system

The first step that is normalization is needed to prepare the signal for further processing. Loses in the signal due to environmental and background noises are removed in this stage. The radiation loss at higher frequencies due to physiological factors is compensated using a pre emphasis filter. Also the end point detection is done in this stage to remove the silence and retain only the required parts of the signal. The pre emphasized and end point detected speech signal is parameterized i.e. its features like the linear predictive coefficients (LPC), linear predictive cepstral coefficients (LPCC), mel frequency cepstral coefficients (MFCC) etc. are extracted. This reduces the redundancy of the speech signal and makes it easier to store and process the speech. For the recognition system a reference digit

database is developed where the feature vectors of all the English digits from zero to nine are extracted and stored. When a test signal of one of these digits is given to the system it extracts the features of this new signal and compares them to the features stored already in the reference database using a similarity evaluation technique. The technique used here is Euclidean distance. Finally based on the comparison the reference digit which gives minimum error distance with the test signal is decided to be the test input given.

A. End Point Detection

End point detection is to determine the ends of the utterance and separates the digit spoken from the background noise or silence. End point detection done in this paper is based on the short time energy as well as zero crossing rate of the speech signal [5]. The short time energy helps to detect the voiced part from the signal. However if the digit spoken starts or ends with unvoiced phoneme then there is a chance of it not getting detected. As a result the zero crossing rate is used to determine the unvoiced part without going undetected. Initially the start and end points *N1* and *N2* respectively are determined by the lower and upper threshold values determined by the formula given by [6]:

$$I1 = 0.03 \times (I_{MAX} - I_{MIN}) + I_{MIN} \quad (1)$$

$$I2 = 4 \times I_{MIN} \quad (2)$$

$$ITL = \min (I1, I2) \quad (3)$$

$$ITU = 5 \times ITL \quad (4)$$

where *ITL* is the lower threshold and *ITU* is the upper threshold. The point at which the signal energy exceeds the lower threshold and then the upper threshold before falling below lower threshold is the start point *N*1 and the point at which it falls again below *ITL* is the end point *N*2. Next, the ZCR for 25 frames before the start and end points are checked to detect any unvoiced part that might have been missed. For unvoiced signal the ZCR would be in the range of 14 to 40 [3]. Hence a ZCR threshold *IZT* given by

$$IZT = \min(IF, IZC+2\sigma) \quad (5)$$

where *IF* is a fixed value that we define. Here *IF*=25, *IZC* is the mean ZCR during silence and σ is the standard deviation of ZCR during silence. If the ZCR crosses this threshold *IZT* before *N1* then a new starting point *N1'* is set where the threshold is exceeded. Similarly if the end point is shifted further to the right if the ZCR exceeds IZT after the pre determined end point *N2*. The new end point will now be *N2'*. This process is continued in steps of 25 frames till the beginning and end of signal is reached. Fig. 2 shows the plot of a signal before and after endpoint detection in MATLAB tool.

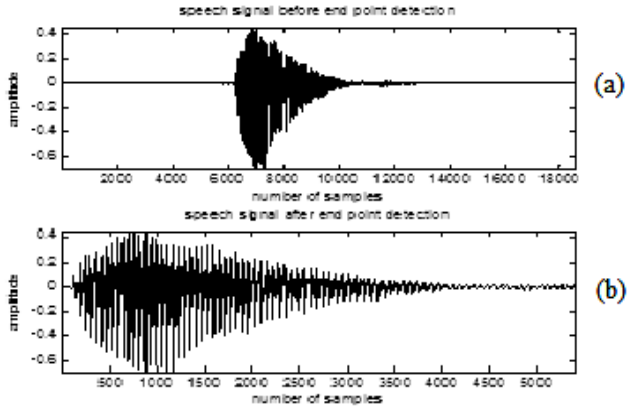

Fig 2. (a) The original speech signal (b) the signal after end point detection

B. Linear Predictive Cepstral Coefficients

Linear prediction refers to predicting the present speech sample using the past samples. The predicted value is given by [3]:

$$\hat{s}_n = \sum_{k=1}^{p} s_{n-k} a_k \quad (6)$$

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. The block diagram of LPCC computation can be shown in Fig. 3.

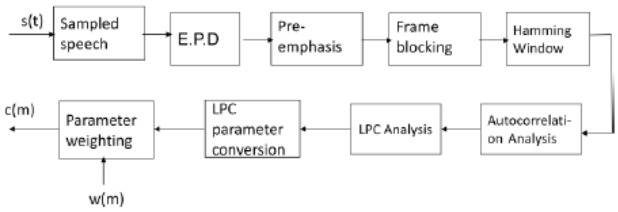

Fig.3. Block diagram to find Linear Predictive Cepstral Coefficients [7]

The input speech signal is sampled and end point detection (EPD) and pre emphasis are done to compensate for the loss in higher frequencies. It is then divided into frames of 20ms duration and multiplied by overlapping Hamming windows. After LPC coefficients are found by autocorrelation analysis these coefficients are converted to cepstral coefficients *Cm* using (7), (8) and (9). The number of cepstral coefficients *Q* should be 1.5 times the number of LPC parameters *p* [5]. The first coefficient *Co* represents the average energy obtained from the

gain G in each speech frame and hence is discarded i.e. amplitude is normalization is done[1].

$$C_0 = \ln(G) \tag{7}$$

$$C_m = -a_m + \frac{1}{m}\sum_{k=1}^{m}\left[-(m-k)a_k C_{(m-k)}\right], 1 \le m \le p \tag{8}$$

$$C_m = \sum_{k=1}^{p}\left[\frac{-(m-k)}{m}a_k C_{(m-k)}\right], p < m < Q \tag{9}$$

To achieve robustness for larger values of m i.e low weight near $m = Q$ and to truncate infinite computation an appropriate weighting in the form of band pass lifter which is a filter in the cepstral domain is used given by (10) where w_m is given by (11)

$$\hat{C}_m = w_m C_m \tag{10}$$

$$w_m = \left[1 + \frac{Q}{2}\sin\left(\frac{\pi m}{Q}\right)\right], 1 \le m \le Q \tag{11}$$

C. Mel Frequency Cepstral Coefficients

Mel Frequency scale is a non linear frequency scale which represents the sensitivity of human ears. Humans are more sensitive to lower frequency than the higher frequencies. This is replicated by the Mel filters which acts as linear filters for 1000Hz but as a log filter from 1000Hz to 3000Hz. The Mel scale can be obtained for each frequency by the given

$$M(f) = 1125 \log_e(1 + f/700) \tag{12}$$

Mel filters are triangular filters with equal area and placed logarithmically along the frequency [1],[4]. The frequency response of Mel filters is shown in Fig. 4

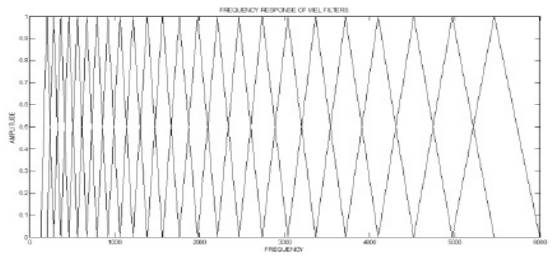

Fig.4. Frequency response of Mel filters

The calculation of Mel filter coefficients involves few steps as shown in Fig.5.

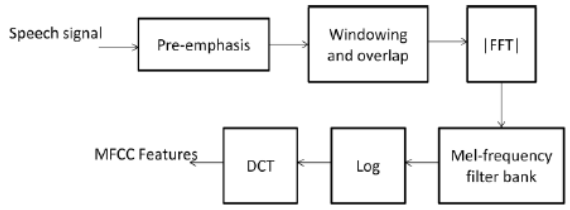

Fig. 5. The block diagram to find Mel Frequency Cepstral Coefficients [4]

Here, the speech signal, after framing and windowing is converted into short segments. The Fourier Transform of the speech segment is calculated and multiplied with the transfer function of 40 mel filters. The logarithm of the 40 coefficients obtained thus is then calculated. The time domain cepstral coefficients are obtained by taking the Discrete Cosine Transform (DCT) of the input. The DCT also concentrated the energy in the lower frequencies. Out of the 40 coefficients, 13 cepstral coefficients are enough to represent each speech frame.

D. Averaging

Using the above two techniques, a predetermined number of LPCC and MFCC are obtained for each frame. Now to further reduce the number of coefficients involved each coefficients in each frame are averaged using (13) [3]

$$\hat{c}_n = \frac{1}{N}\sum_{n=1}^{N} c_n \tag{13}$$

where N is total number of frames in each speech utterance and $\hat{c}_n$ is the averaged value of the coefficients. Thus a feature vector for each digit is obtained.

E. Similarity Evaluation

Once the features are extracted and reference database is created, features of the test input digit are determined using the techniques discussed above and these features are compared using Euclidean distance given by (14)[1],[9].

$$d(x,y) = \sqrt{\sum_{i=1}^{N}(x_i - y_i)^2} \tag{14}$$

where x and y are the feature vectors obtained for test and reference signals respectively using the averaging method explained previously

III. RESULTS

The parameter to find the accuracy of a recognition system is the Recognition Rate by (15) where Ncorrect is the number of words recognized accurately and *Ntotal* is the total number of words in the vocabulary [10].

$$\text{Recognition Rate} = \frac{N_{correct}}{N_{total}} \times 100\% \quad (15)$$

An experiment was conducted where the Recognition Rate was calculated for each of the feature extraction technique for two male speaker and two 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 Table I and Table II.

TABLE I. RECOGNITION RATES OF THE TWO FEATURE EXTRACTION TECHNIQUE FOR MALE SPEAKER

Feature Extraction Technique	Speaker 1	Speaker 2
LPCC	85%	78%
MFCC	93%	83%

TABLE II. RECOGNITION RATES OF THE TWO FEATURE EXTRACTION TECHNIQUE FOR FEMALE SPEAKER

Feature Extraction Technique	Speaker 1	Speaker 2
LPCC	83%	85%
MFCC	85%	87%

From the tables it is observed that for both male and female speakers the Recognition Rate is greater for MFCC feature extraction method as compared to LPCC method. An increase of 2% to 7% has been observed.

IV. CONCLUSION

From the results it can be concluded that though both the feature extraction techniques are cepstral based, the MFCC method is superior to the LPCC method. The reason for this increase in accuracy might be the result of the property of MFCC which takes into account the sensitivity characteristics of human ear. The lower frequencies which contain more important information are resolved well and given more weight here.

REFERENCES

[1] D. O'Shaughnessy, Speech Communications: Human & Machine, 2nd ed. Wiley-IEEE Press, 1999, pp. 367-435.
[2] 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
[3] 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
[4] S. Savitha, "DSP Implementation of Isolated Digit Recognizer," M.Tech Dissertation, Dept. Elect. Eng., IIT, Bombay, India, 2008.
[5] 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
[6] 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
[7] L. R. Rabiner, B. Juang and B. Yegnanarayana, "Fundamentals of Speech Recognition ," *5th ed.* Pearson, 2011.
[8] S. D. Apte, "Spectral Parameters of Speech," *Speech and Audio Processing*, Wiley-India ed.,2013, pp. 97-320.
[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] A. A. M. Abushariah, T. S. Gunawan, et. al., "English Digits Speech Recognition Based on Hidden Markov Models," *ICCCE,* Kuala Lumpur, Malaysia, May 2010
[11] G. Nitin, "Implementation of Algorithms for Speaker Dependent Isolated Digit Recognition," M.Tech Dissertation, Dept. Elect. Eng., IIT, Bombay, India, 1997.
[12] 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.
[13] L. F. Lamel, L. R. Rabiner and A. E. Rosenberg, "An Improved Endpoint Detector for Isolated Word Recognition," *IEEE Trans. ASSP,* vol. 29, no. 4, pp. 777- 785, 1981
[14] 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
[15] Ahmad A. M. Abushariah, Teddy S. Gunawan, et. al., "English Digits Speech Recognition Based on Hidden Markov Models," *ICCCE,* Kuala Lumpur, Malaysia, May 2010.