English Digits Recognition using MFCC, LPC and Pearson’s Correlation

Mayur R Gamit¹, Prof. Kinnal Dhameliya²
¹M.Tech Research Student, ²Assistant Professor, E&C Department, C.G.P.I.T, Bardoli, India

Abstract—Automatic speech recognition is an important topic of speech processing. This paper presents the use of a Pearson’s correlation for English digits recognition. The Pre-processing is done and voiced speech is detected based on energy and zero crossing rates (ZCR). The proposed approach used in speech recognition is Mel Frequency Cepstral Coefficients (MFCC) and combine features of both MFCC and Linear Predictive Coding (LPC). The Pearson’s correlation is used as a classifier. The recognition accuracy is increased when combine features of both LPC and MFCC are used as compared to only MFCC approach using Pearson’s correlation as a classifier.

Keywords—Pre-processing, Mel frequency Cepstral Coefficient (MFCC), Linear Predictive Coding (LPC), Pearson’s correlation.

I. INTRODUCTION

Automatic speech recognition (ASR) has been the most investigated topic in speech processing since early 1960s [1]. 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 features from speech signal and representing them using an appropriate data model. ASR system involves two phases. Training phase and Testing phase [2]. In training phase, known speech is recorded and parametric representation of the speech is extracted and stored in the speech database. In the testing phase, for the given input speech signal the features are extracted and ASR system compares it with the reference templates to recognize the utterance.

This paper proposes an approach to recognize isolated words using Mel frequency Cepstral coefficient (MFCC) and Linear Predictive Coding (LPC) as a feature extraction techniques and Pearson’s correlation is used as a classification technique. This paper evaluates the recognition accuracy by using only MFCC and combination features of MFCC and LPC. The MFCC gives higher recognition accuracy in speech recognition systems as compared to other techniques. The main advantage of using MFCC techniques is because of less complexity and accurate results while LPC is suitable for speaker recognition.

The human voice is recorded and digitized. Then it is input to the pre-processing block. The main task of this block is to separate out the unvoiced speech samples from the voiced speech samples. The detection of voiced and unvoiced speech could be done on the basis of calculating energy and zero crossing rates. The frame work of speech recognition is shown below:

Figure 1. Block diagram of speech recognition system.

This paper proposes a novel and efficient way to recognize English digits from zero to nine by using advanced pre-processing strategy. Section 2, introduces a new pre-processing concept. Sections 3, describes the feature extraction techniques MFCC and combine features of both MFCC and LPC. Section 4, describes the Pearson’s correlation technique and section 5, and describes the experimental results.

II. PRE-PROCESSING CONCEPT

The recorded speech is fed to the pre-processing block. The speech is segmented into small chunks or frames. The determination of voiced and unvoiced speech samples can be done on the basis of Energy and Zero crossing rates (ZCR).

A. Energy

The speech signal is fragmented into small frames. Each frame is of \( \omega \) samples, where \( \omega < n \) as \( n \) is total number of samples. The energy of speech is calculated frame by frames. The square of each sample is done and finally summation of all squared samples is done. The equation to calculate energy is given below [3]:

\[
E = \sum_{i=1}^{n} x_i^2
\]
B. Zero Crossing Rate (ZCR)

The zero crossing rates means the number of times the transition of speech signal from positive to negative or vice-versa. The zero crossing rates are calculated frame by frames. If the zcr of speech samples having zero crossing rate more than it is a unvoiced otherwise voiced. According to the survey done by rabiner the ZCR of fricative speech samples are more than threshold then consider it as a voiced speech. Fricative has more zero crossing rates as compared to unvoiced speech.

The equation to calculate the ZCR is as follows:

\[ ZCR = \frac{\sum_{i=1}^{\infty} |\text{sign}(x_i) - \text{sign}(x_{i-1})|}{2} \]  

\[ ZCR = \sum_{i=1}^{\infty} \frac{|\text{sign}(x_i) - \text{sign}(x_{i-1})|}{2} \]  

C. Start point and end point detection

The calculation of the energy and zero crossing rates can be done on the basis of frames. The start point and end point can be accurately found out on the basis of both energy and zero crossing rates. First the start and end points can be determined based on energy only. Based on zcr of the start and end points are computed. If the zcr of the frame having start point is more than the start point is shifted towards left side and same as end point shifted towards right.

- If ZCR > (threshold) then start point is shifted towards left otherwise remains same.
- If ZCR > (threshold) then end point is shifted towards right otherwise remains same.

D. Removal of unvoiced parts between start point and end point

The recognition accuracy can be increased when the unvoiced part between the start and end points can be removed. There are certain samples of speech in between start and end points which have minimum energy. They do not contain any information, so by removing them makes recognition accuracy better. This advance concept has been used in this speech recognition system.

III. FEATURE EXTRACTION

In feature extraction the Mel frequency Cepstral coefficient (MFCC) and combine features of both MFCC and LPC are used. The both techniques are described below:

A. Mel-Frequency Cepstrum Coefficients (MFCC)

MFCC is one of the most commonly used feature extraction method in speech recognition [5][6][7]. The overall process is shown in figure below:

\[ \text{Input speech} \xrightarrow{\text{Pre-emphasis}} \text{Framing & windowing} \xrightarrow{\text{FFT}} \text{Mel Filter bank} \]

\[ \text{MFCC coefficient} \xrightarrow{\text{Energy}} \text{IDCT} \xrightarrow{\text{Log}} \]

\[ \text{Figure 2. MFCC block diagram [5][6]} \]

Step 1: Pre-emphasis

The signal is passed through a filter which emphasis a high frequencies. This process increases the energy of signal at high frequency. The equation used to denotes the pre-emphasis is shown below:

\[ S(n) = X(n) - a \times X(n-1) \]  

Where s(n) denotes the output sample, x(n) is present sample, x(n-1) is past sample and value of a is between 0.95 to 1.

Step 2: Framing and overlapping

The speech signal is split into several frames such that each frame can be examined in the short time instead of the entire signal. The frame size is of the range 0-20 ms. Then overlapping is applied to frames, hamming window is applied. The equation of hamming window is as follows [8]:

\[ S(n) = X(n) \times W(n) \]  

\[ W(n) = 0.54 - 0.46 \cos \left( \frac{2\pi n}{N-1} \right) \quad 0 \leq n \leq N - 1 \]

Step 3: Fast Fourier Transform

The Fast Fourier Transform (FFT) converts the frames from time domain to frequency domain. Therefore, FFT is executed to obtain the magnitude frequency response of each frame and to prepare the signal for the next stage.

\[ S(\omega) = \text{fft} (X(n)) \]

Step 4: Mel Filter bank

Human ear perception of frequency contents of sounds for speech signal does not follow a linear scale. Therefore, for each tone with an actual frequency f, measured in Hz, a subjective pitch is measured on a scale called the “mel scale”’. The Mel frequency scale is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000Hz [9]. To compute the Mel for a given frequency f in Hz, a following formula is used.
Each filter’s magnitude frequency 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.

**Step 5: Log and IDCT**

The output of Mel filter bank is given to log. This is the process to convert the log Mel spectrum into time domain using Inverse Discrete Cosine Transform (IDCT).

**Step 6: Energy**

The energy of all frames after IDCT is calculated. The result of the conversion is called Mel Frequency Cepstrum Coefficient. The set of coefficient is called acoustic vectors. Therefore, each input utterance is transformed into a sequence of acoustic vector.

**B. Combine features of MFCC and Linear Predictive Coding (LPC)**

The basic idea behind the Linear Predictive Coding (LPC) analysis is that a speech sample can be approximated as linear combination of past speech samples. The LPC provides a robust, reliable and accurate method for estimating the parameters that represent the vocal tract system. The autocorrelation analysis is done. Levinson Durbin’s algorithm is used to analyze the LP model. The LPC gives the best results for the speaker recognition rather than the speech recognition. Here, 10th order LPC is taken means 11 coefficients of each frames can be obtained. The 12 coefficients of MFCC and 11 coefficients of LPC are combined frame by frame basis to yields 23 coefficients of each frames.

**IV. PEARSON’S CORRELATION CONCEPT**

Correlation technique was invented by Pearson [10]. This technique is mainly used in recognition approaches. Correlation between sets of data is a measure of how well they are related. The most common measure of correlation in statistics is the Pearson Correlation. Correlation is a statistical technique that can show whether and how strongly pairs of variables are related. The Pearson correlation technique works best with linear relationships. The main result of a correlation is called the correlation coefficient. It ranges from -1.0 to +1.0.

\[
F(mel) = 2595 \times \log_{10} \left(1 + \frac{f}{700}\right) \quad (7)
\]

The closer \( r \) is to +1 or -1, the more closely the two variables are related. If \( r \) is close to 0, it means there is no relationship between the variables.

If \( r \) is positive, it means that as one variable gets larger the other gets larger than it is called positive correlation. If \( r \) is negative it means that as one gets larger, the other gets smaller or often called an inverse correlation. The equation of correlation is as follows [10]:

\[
r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}
\]

Where \( x \) and \( y \) are features vectors. The correlation coefficient ranges from -1 to 1. There may be positive, negative or zero correlation between two features vectors.

**V. MEASURING PARAMETERS**

The performance of the speech recognition system is often described in terms of accuracy. The result indicates the recognition accuracy by implementation of the proposed algorithm on a dataset consists of 28 speakers of which 14 are males and 14 are females. The obtained accuracy is compared with [4]. The sampling frequency is 8000 Hz. The both features extraction techniques which are described above are used and Pearson’s correlation is used as a classifiers. The measuring of recognition accuracy (RA) is done based on the equation

\[
RA = \left[\frac{\text{No. of times word recognized}}{\text{Total no. of trials}}\right] \times 100
\]

The table below indicates the recognition accuracy by using MFCC and Correlation concept.

<table>
<thead>
<tr>
<th>Number of sessions</th>
<th>Recognition Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>84%</td>
</tr>
<tr>
<td>2</td>
<td>78%</td>
</tr>
<tr>
<td>3</td>
<td>82%</td>
</tr>
<tr>
<td>Average</td>
<td>81.33%</td>
</tr>
</tbody>
</table>

The table below shows the recognition accuracy by using combine features of MFCC and LPC and correlation as classifiers. The 12 coefficients of MFCC and 11 coefficients of LPC are combine frames by frames. The accuracy at each session is shown below:
Table II
Recognition Accuracy Using MFCC+LPC And Correlation

<table>
<thead>
<tr>
<th>Number of sessions</th>
<th>Recognition Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>84%</td>
</tr>
<tr>
<td>2</td>
<td>78%</td>
</tr>
<tr>
<td>3</td>
<td>79%</td>
</tr>
<tr>
<td>Average</td>
<td>80.33%</td>
</tr>
</tbody>
</table>

The table below shows the comparative results of obtained recognition accuracy with the [4]. The proposed approach gives the more recognition accuracy than described in [4].

Table III
Comparative Results

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean Distance</td>
<td>30%</td>
<td>57.5%</td>
<td>81.33%</td>
<td>80.33%</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

The experimental result shows that by using the proposed MFCC and combination of both MFCC and LPC feature extraction techniques and Pearson’s correlation as classifier the results are higher as compared to [4]. The recognition accuracy may differ by using only MFCC, LPC and combination of both MFCC and LPC techniques as well as other classification techniques.