Isolated Word Automatic Speech Recognition (ASR) System using MFCC, DTW & KNN

Muhammad Atif Imtiaz
Faculty of Electronics & Electrical Engineering
University of Engineering and Technology,
Taxila
atif.imtiaz@uettaxila.edu.pk

Gulistan Raja
Faculty of Electronics & Electrical Engineering
University of Engineering and Technology,
Taxila
gulistan.raja@uettaxila.edu.pk

Abstract— **Automatic Speech Recognition (ASR) System is defined as transformation of acoustic speech signals to string of words. This paper presents an approach of ASR system based on isolated word structure using Mel-Frequency Cepstral Coefficients (MFCC's), Dynamic Time Wrapping (DTW) and K-Nearest Neighbor (KNN) techniques. The Mel-Frequency scale used to capture the significant characteristics of the speech signals; features of speech are extracted using MFCC's. DTW is applied for speech feature matching. KNN is employed as a classifier. The experimental setup includes words of English language collected from five speakers. These words were spoken in an acoustically balanced, noise free environment. The experimental results of proposed ASR system are obtained in the form of matrix called confusion matrix. The recognition accuracy achieved in this research is 98.4 %.**

Keywords—ASR; MFCC; DTW; KNN

I. INTRODUCTION

Speech is propagation of periodic variations in the air from human lungs. The responsibility for the production and shaping of actual sound is done by the human vocal tract with the help of pharynx, nose cavity and mouth. Automatic Speech Recognition (ASR) system is the process of automatically interpreting human speech in a digital device and is defined as transformation of acoustic speech signals to words string. Generally goal of all ASR systems are used to extract words string from input speech signal [1]. In ASR process the input is the speech utterance and output is the in the form of textual data in association with given input. Some factors on which the performance of ASR systems mainly relies are vocabulary size, amount of training data and systems computational complexity. There are numerous applications of ASR like it is extensively used in domestic appliances, security devices, cellular phones, ATM machines and computers.

This paper describes an ASR System of English language experimented on small vocabulary of words. Rest of the paper is organized as follows: Section II describes the overall ASR System Overview, the major blocks used in ASR System. While implementation of ASR system using Feature Extraction and classification techniques are described in Section III. Section IV discuses the brief description of experimental setup, as well as some experimental results. Concluding remarks are discussed in section V.

II. ASR SYSTEM OVERVIEW

ASR system comprises of two main blocks i.e. Feature extraction block and a classification block as shown in Fig. 1.

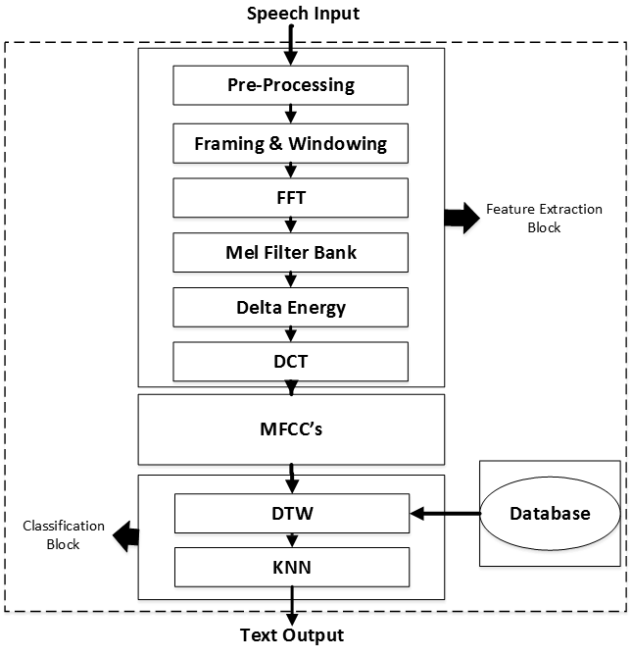

Fig. 1. Block Diagram of Proposed ASR System Design

The input to the block is speech and output of the block is textual data. The working of blocks is described below:

A. Feature Extraction Block

Feature Extraction is one of the most vital module in an ASR system. In ASR, speech signal is split up into smaller frames usually 10 to 25 msec. As there is redundant information, present in the speech signal. Therefore, to take out important and useful information feature extraction technique is applied. This will also help in diminution of dimensionality. Perceptual Linear Prediction (PLP) coefficients, Wavelet transform based features, Linear Predictive Coefficients (LPC), Wavelet packet based features and Mel Frequency Cepstral Coefficients (MFCC) are the widely used features in ASR [2]. MFCC is used in this research and is discussed in details in section III.

978-1-4673-9791-9/16/$31.00 ©2016 IEEE

B. *Classification Block*

After extracting features from speech signal, the extracted features are given to the classification block for recognition purpose. In classification the input speech feature vector is used to train on known feature patterns and is tested on test dataset and the performance of classifier is evaluated on percentage recognition accuracy**.** In this research, DTW is used for feature matching and KNN is used for classification, both are discussed further in section III.

C. *Database*

In ASR system, the database is a group of speech samples. These samples of speech data are collected in a way to illustrate different changeable aspects of language. Selection of a dataset is of significant importance for successfully conducting ASR research. It provides a platform in comparing performance of different speech recognition techniques [3]. It also provides researchers a balance in different speech recognition aspects i.e. gender, age and dialect. A database comprises of large, medium or small sizes depending upon the word count. Data can be gathered from sources i.e. books, newspapers, magazines, lectures and TV Commercials. Due to issues of unavailability of volunteers and some identity issues, speech databases are not easily available. Some standard speech databases are available for few languages, like BREF for French, TIMIT for English and ATR for Japanese etc [4].

III. IMPLEMENTATION OF ASR SYSTEM

In this section implementation and description of feature extraction technique Mel frequency cepstral coefficient (MFCC), feature matching technique (DTW) and feature classification technique K-Nearest Neighbor (KNN) are discussed in detail:

A. *Mel frequency cepstral Coefficient*

Human Speech as a function of the frequencies is not linear in nature; therefore the pitch of an acoustic speech signal of single frequency is mapped into a "Mel" scale. In Mel scale, the frequencies spacing below 1 kHz is linear and the frequencies spacing above 1 kHz is logarithmic [5]. The Mel frequencies corresponding to the Hertz frequencies are calculated by using equation (1).

$$fmel = 2595\log(1+\frac{f}{700}) \quad (1)$$

The block diagram for Mel-Frequency Cepstral Coefficients (MFCC) computations is shown in Fig. 2.

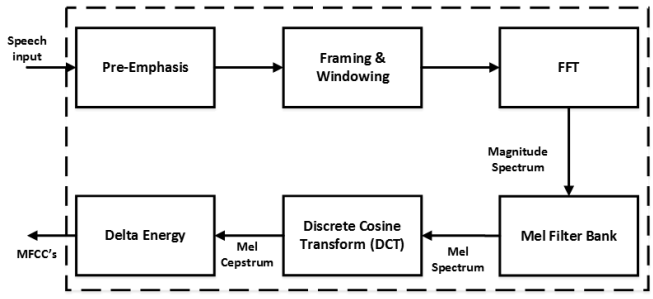

Fig. 2. Block Diagram for MFCC Computation

The inner blocks shown in Fig. 2 are individually described below in detail:

1) Pre- Processing: The audio signals in are recorded having a sampling rate of 16 kHz. Each word is stored in separate audio file. The pre-processing step includes the Pre-emphasis of signal to boost the energy of signal at high frequencies. The difference equation of Pre-emphasis filter is given by equation (2).

$$H(z) = \frac{B(z)}{A(z)} = \frac{b_0+b_1z^{-1}}{1} = 1 - 0.97z^{-1} \quad (2)$$

The Output response of pre-emphasis Filter is shown in Fig. 3.

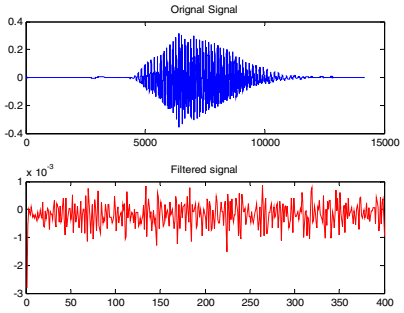

Fig. 3. Pre- Emphasis Filter Output

2) Framing and Windowing: The speech signal is not stationary in nature. In order to make it stationary framing is used. Framing is the next step after pre-processing; in this step speech signal is split up into smaller frames overlapped with each other. After framing windowing is used to remove discontinuities at edges of frames. The window method used in this research is Hamming Window. The Hamming Window is defined by equation (3).

$$w(n) = \begin{cases} 0.54 - 0.46\cos(\frac{2\pi n}{N-1}) & 0 \le n \le N-1 \\ 0 & otherwise \end{cases} \quad (3)$$

Where, N is total number of samples in a single frame. The output response of Original signal and windowed signal is shown in Fig. 4.

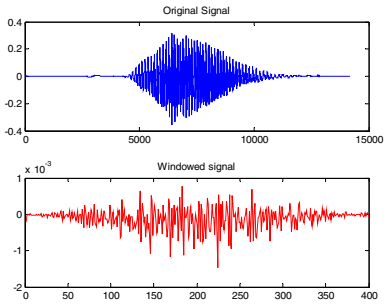

Fig. 4. Original Signal Vs Windowed Signal

3) Fast Fourier Transform (FFT): Fast Fourier transform is used for calculating of the discrete fourier transform (DFT) of signal, with size N=512 have been used [6]. This step is performed to transform the signal into frequency domain. The FFT is calculated using equation (4).

$$X[k] = \sum_{n=0}^{N-1} x[n]e^{-j2\frac{\pi}{N}kn} \quad (4)$$

Where, ***N*** is the size of FFT. The Magnitude spectrum of FFT is shown in Fig. 5.

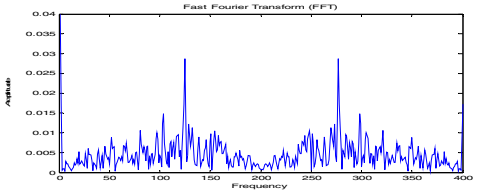

Fig. 5. Fast Fourier Transform Magnitude Spectrum

4) MelFilter Bank: The next step after taking FFT of the signal is the transformation from Hertz to Mel Scale, the spectrums power is transformed into a Mel scale [7]. The Mel filter bank comprises of triangular shaped overlapping filters as shown in Fig. 6.

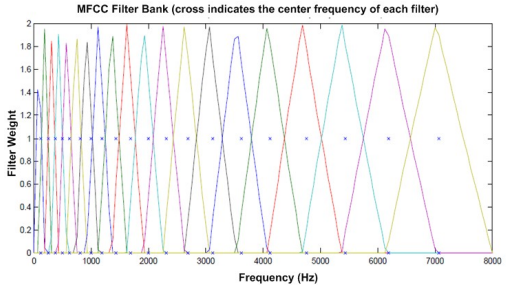

Fig. 6. MFCC Filter Bank Output

5) Delta Energy: In this step take base 10 Logarithm of output of previous step. The computation of Log energy is essential because of the fact that human ear response to acoustic speech signal level is not linear, human ear is not much sensitive to difference in amplitude at higher amplitudes. The advantage of logarithmic function is that it tends to duplicate behavior of human ear. Energy computation is calculated using equation (5). The graph for energy computation is shown in Fig. 7.

$$E = \sum_{t=t1}^{t2} x^2(t) \quad (5)$$

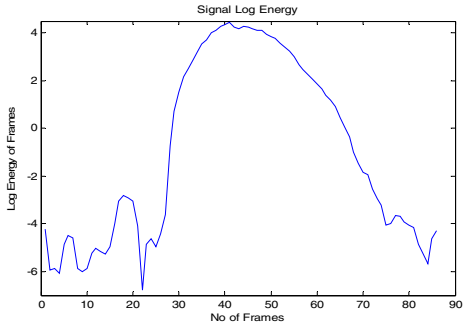

Fig. 7. Signal Log Energy Output

6) Discrete Cosine Transform (DCT): The Discrete Cosine Transform (DCT) is employed after taking logarithm of output of the Mel-filter bank. It finally produces the Mel- Frequency Cepstral Coefficients. In this research for an isolated word, 39 dimensional features are taken out i.e. 12-MFCC (Mel frequency cepstral coefficients), one energy feature, one delta energy feature, one double-delta energy feature, 12-delta MFCC features and 12-double delta MFCC features. An N-point DCT [8] is defined by equation (6).

$$X[k] = \sum_{n=0}^{N-1} 2x[n]\cos[\frac{\pi}{2N}k(2n+1)];\ \ k = 0,1,2,\dots N-1 \quad (6)$$

The MFCC's graph for a single word is shown in Fig. 8.

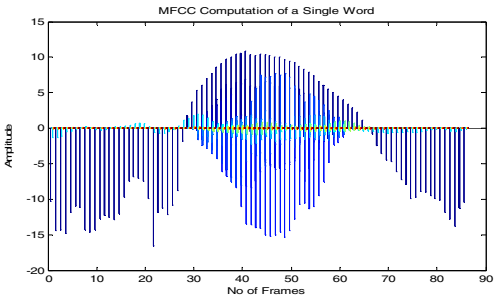

Fig. 8. MFCC's for Single Word

B. Classification & Recognition

In determining the performance of the system specifically ASR system, the role of classifier is very significant. In this research Dynamic Time Warping (DTW) and K-Nearest Neighbors have been used for Speech feature matching and Classification. DTW measures the resemblance in two time series, which are different regarding time or speed. In programming of DTW dynamic approach is taken in account in order to optimize the similarity between two time series. For continuous speech recognition case Hidden Markov Models (HMM) and Artificial Neural Networks (ANN) are considered suitable for classification. ANNs have a tendency to replicate the brain activity human. ANNs comprises of a set of neurons which are interconnected with each other. In ANN the output is measured by calculating the product of inputs weighted sum. One of the most popular classification techniques for continuous speech recognition is Hidden Markov Models (HMM). It is basically statistical classification technique and models a time series in the presence of two stochastic variables [9]. The proposed research focuses on ASR of words based on isolated word structure and it does not require any language model. In this research, Dynamic Time Wrapping (DTW) and K-Nearest Neighbor (KNN) techniques have been used for feature matching and classification based upon the MFCCs. The classification step includes two stages;

i) Training
ii) Testing

The results and percentage recognition accuracy are obtained in the form of Confusion Matrix. DTW and KNN are discussed further in next section

1) Dynamic Time Wrapping (DTW): DTW Algorithm calculation is in view of measuring closeness in two time series which might shift in time and speed. The comparison is measured in terms of position of two time arrangements if one time arrangement might be wrapped non-straightly by extending or contracting it along it's time pivot.

The wrapping in two time arrangements can further be utilized to discover relating regions in two time arrangements or to focus closeness between the two time arrangements. Numerically, DTW compares two time arranged patterns and measure the similarity between them with the help of minimum

distance formula. Consider two time series P and Q having length n and m i.e.

$$P = p_1, p_2, p_3 \dots \dots, p_i \dots \dots, p_n$$
$$Q = q_1, q_2, q_3 \dots \dots, q_j \dots \dots, q_m$$

In time series P and Q the i_{th} and j_{th} component of the matrix includes the distance $d(p_i, q_j)$ in the two matrix points p_i and q_j [10]. Then using Euclidean distance formula, in equation (7) measures the absolute distance between two points.

$$d(p_i, q_j) = \sqrt{(p_i - q_j)^2} \quad (7)$$

Every matrix element i and j is belongs to the alignment in points pi and q_j. Then, using equation (8) accumulated distance is calculated.

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

2) K- Nearest Neighbor (KNN): The working of KNN classifier in this research is dicussed below.

- KNN method consists of assigning the index of the feature vector that is nearest to given score in the feature space.
- Minimum score indices from DTW are processed in KNN method.
- It Converges the Current feature on to respective feature of feature Space.
- Same numbers of features are returned by KNN but these features are from feature Space.
- Mode of the KNN returned Features gives the most frequent feature lies in and it would be the Recognized Word.

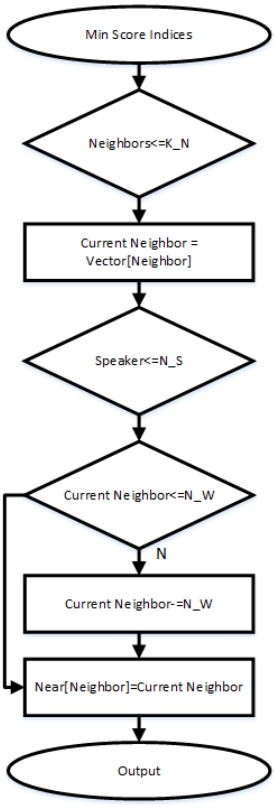

Fig. 9. Flow diagram of KNN

Fig. 9. Shows the flow diagram of KNN classifier, here K_N is the number of nearest neighbors, N_S is the number of speakers and N_W is the number of words in vocabulary.

3) Confusion Matrix: In order to check the efficiency of the system i.e. recognition accuracy and percentage of error, a confusion matrix is formed. In case of N words, it will contain N×N matrix. In confusion matrix all diagonals entries, state A_{ij} for i=j, showed the no of time a word i is matched correctly [11]. Similarly non-diagonal entries, state $\mathbf{A_{ij}}$ for $\mathbf{i \neq j}$, showed the number of times a word i is is confused with the word j

$$\begin{matrix} A_{11} & A_{12} & A_{13} & \dots & A_{1N} \\ A_{21} & A_{22} & A_{23} & \dots & A_{2N} \\ A_{31} & A_{32} & A_{33} & \dots & A_{3N} \\ . & . & . & \dots & . \\ . & . & . & \dots & . \\ A_{N1} & A_{N2} & A_{N3} & \dots & A_{NN} \end{matrix}$$

4) Percentage Error: The calculation of percentage of error is very important in order to check the overall system performance and it is calculated in the form of confusion Matrix. For this purpose a single isolated word is tested and check how many time it is recognized successfully and stated in diagonal entry in row i. percentage is calculated by dividing successfully entries divided by the total no of entries. Thus, Correct Match C and percentage error E, for a particular word, can be represented as in equation (9) & (10); The results obtained from confusion matrix are further dicussed in section IV.

$$Correct\ Match\ C = \frac{A_{ij}}{A_{i1}+A_{i2}+A_{i3}+\cdots A_{iN}}; where\ i = j, j = 1,2,3, \dots N \quad (9)$$

$$\%\ \text{of error E= (1-C) x100} \quad (10)$$

IV. EXPERIMENTAL RESULTS AND DISCUSSION

The experiments were performed on a small size vocabulary of English. The setup includes words spoken from five different speakers. These words were spoken in an acoustically balanced, noise free environment. The implementation and experimental results were analyzed with the help of MATLAB R2014b. The testing and training results of ASR are obtained in the form of matrix called confusion matrix as shown in Fig. 10.

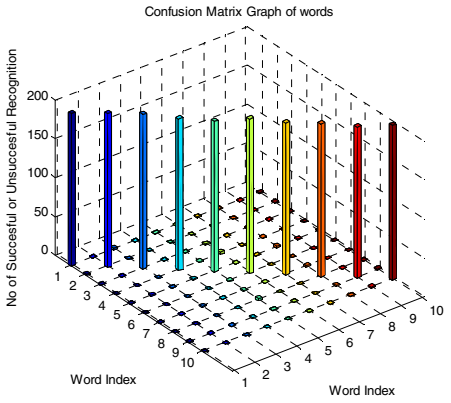

Fig. 10. Confusion Matrix Graph of Words

In Fig. 10 of confusion matrix graph the x-axis and the y-axis are showing the indices of the words. The z-axis shows the height i.e. it shows the total number of times, an individual Word is successfully recognized or it confused with any of other word. The diagonal slots show heights as

successful recognition rate. The maximum possible attained possibility of height in this case is 200. The total number of times a word is tested in this case is 200. The values of correct match *C* and error % *E*, for words, are summarized in Table I.

TABLE I: RECOGNITION & ERROR PERCENTAGE OF WORDS

Word	Value of Correct Match C	Recognition Accuracy (%)	Error (%) = (1-C)x100
"Dark"	0.98	98	2
"Wash"	0.99	99	1
"Water"	0.995	99.5	0.5
"Year"	0.975	97.5	2.5
"Don't"	0.97	97	3
"Carry"	0.995	99.5	0.5
"Greasy"	0.98	98	2
"Like"	0.985	98.5	1.5
"Oily"	0.975	97.5	2.5
"That"	0.995	99.5	0.5
Accumulative Average	0.984	98.4	1.6

Table I. describes the recognition and error rates of a dataset. Firstly each word is evaluated on individual basis and then accumulative average of the dataset is calculated. The data is obtained in the form of confusion matrix as a result of testing the ASR system. The accumulative average success rate obtained for the dataset given above is 98.4 % with 1.6 % error rate.

V. CONCLUSION

The proposed research on an ASR system delineates MFCC, DTW and KNN techniques. The extraction of features is performed using MFCC, DTW is used for speech features matching and KNN is used for classification. Minimum score indices acquired from DTW are processed in KNN. The experimental results are obtained in the form of confusion matrix. It is observed during the whole research that the proposed ASR System shows good recognition performance when MFCC, DTW and KNN are used jointly. The recognition accuracy achieved in this research is 98.4 % with an error of 1.6 %.

REFERENCES

[1] J.M. Gilbert∗, S.I. Rybchenko, R. Hofe, S.R. Ell, M.J. Fagan, R.K. Moore, P. Green, "Isolated word recognition of silent speech using magnetic implants and sensors," *International Journal of Medical Engineering and physics*, vol. 32, pp. 1189-1197, August 2010.

[2] Vimala.C and Dr.V.Radha "A Review on Speech Recognition Challenges and Approaches" World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 1, pp. 1-7, 2012.

[3] J. Clear, and N. Ostler S. Atkins, "Corpus design criteria," *Oxford Journal of Literary and linguistic computing*, vol. 7, no. 1, pp. 1-16, 1992.

[4] L. F. Lamel, and M. Eskenazi J. L. Gauvain, "Design Considerations and Text Selection for BREF, a large French Read-Speech Corpus," in *1st International Conference on Spoken Language Processing, ICSLP*, 1990, pp. 1097-1100.

[5] M Murugappan, Nurul Qasturi Idayu Baharuddin, Jerritta S "DWT and MFCC Based Human Emotional Speech Classification Using LDA" *International Conference on Biomedical Engineering (ICoBE)*, Penang, 27-28 February 2012, pp. 203-206.

[6] Michael Pitz, Ralf Schl¨uter, and Hermann Ney Sirko Molau, "Computing Mel-Frequency Cepstral Coefficients on the Power Spectrum," in *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2001. Proceedings. (ICASSP '01)*, USA, 2001, pp. 73-76.

[7] Ibrahim Patel and Dr. Y. Srinivas Rao "Speech Recognition using HMM with MFCC-AN analysis using frequency spectral decomposition technique" *Signal & Image Processing : An International Journal (SIPIJ)* Vol.1, No.2, pp.101-110, December 2010.

[8] AMilton, S.Sharmy Roy, S. Tamil Selvi "SVM Scheme for Speech Emotion Recognition using MFCC Feature" International Journal of Computer Applications (0975 – 8887) Volume 69– No.9, pp.34-39, May 2013.

[9] Areg G. Baghdasaryan and A. A. (Louis) Beex "Automatic Phoneme Recognition with Segmental Hidden Markov Models*" IEEE 2011 Conference on Signals, Systems and Computers*, ASILOMAR, 2011, pp. 569-574.

[10] Anjali bala, Abhijeet kumar, Nidhika birla. "Voice command recognition Systemm Based on MFCC and DTW" *International Journal of Engineering Science and Technology*, vol. 2, no. 12, pp. 7335-7342, Jan 2010.

[11] Hua-Nong Ting, Boon-Fei Yong, Seyed Mostafa Mirhassani, "Self-Adjustable Neural Network for speech recognition," *International Journal of Engineering Applications of Artificial Intelligence*, vol. 26, pp.2022-2027,July2013.