THE ASIAN BULLETIN OF BIG DATA MANAGMENT
Vol. 5. Issue 2 (2025)
https://doi.org/10.62019/abbdm.v5i2.333

ASIAN BULLETIN OF BIG DATA MANAGEMENT
http://abbdm.com/
ISSN (Print): 2959-0795
ISSN (online): 2959-0809

Automatic Speech Recognition by Using Neural Network Based on Mel Frequency Cepstral Coefficient

Muhammad Daud Abbasi, Zubair Sajid, Shahzad Karim Khawer, Syed Zain Mir, Abdul Basit, Muhammad Kashif

Chronicle

Article history
Received: Feb 12, 2025
Received in the revised format: March 24, 2025
Accepted: April 22, 2025
Available online: May 17, 2025

Muhammad Daud Abbasi, Zubair Sajid, Shahzad Karim Khawer & Syed Zain Mir is currently affiliated with the Department of computer science Iqra university, Karachi, Pakistan.
Email: daud.abbasi@iqra.edu.pk
Email: zubair.sajid@iqra.edu.pk
Email: shahzad.karim@iqra.edu.pk
Email: szainmir@iqra.edu.pk

Abdul Basit & Muhammad Kashif is currently affiliated with the Faculty of Computer and Information Technology, Indus University, Karachi, Pakistan.
Email: a.basit@indus.edu.pk
Email: m.kashif@indus.edu.pk

Corresponding Author*

Abstract

This paper deliberated and estimated the Neural Networks Automatic Speech Recognition (ASR) system based on an isolated small vocabulary speaker-independent manual cropping technique, from the training stage to the recognition stage. Besides this, the paper also examines three distinct blocks of speech recognition, i.e., Speech Preprocessor, Feature Extractor, and a Recognizer. Speech preprocessing involves windowing, framing, Short Term and Zero Crossing threshold energy, and End Point Detection calculation. Mel Frequency Cepstral Coefficients (MFCC) are extracted to represent the speech signal in frames and then passed through a Mel frequency filter. Multi-layer feed-forward network trained by the back-propagation method.

Keywords: Speech Recognition, Short Term Energy, Zero Crossing Rate, Mel frequency Cepstral coefficient, Neural Networks, Feed forward Neural Networks, Back Propagation.

INTRODUCTION

Speech is the natural and primary source of human communication. During early childhood, we learn relevant skills without getting instruction and depend on speech communication. Functionality of speech production organs like vocal tract and articulators is not just under conscious control, emotional states like anger, happiness, etc, or differences of genders also affect as well [1]. Automatic Speech Recognition (ASR) can be defined as an independent speech-driven system that stores some distinguishing characteristics of input source speech signals and then processes these stored features to match them to incoming speech words in an effort that allows the system to recognize the words spoken by a person. Speech recognition is a very popular and challenging research area among engineers and scientists around the world. The most important reason is the Human-machine interaction. The main goal is to achieve an efficient natural human-machine speech recognition interface [2]. The idea of such exciting applications has encouraged researchers in automatic speech

recognition (ASR) since the 1950's and because of these extensive and remarkable applications of ASR, Speech recognition has become an attractive domain of study in the areas of information, finance, and communication technology. Thanks to computers' overall advances and new technologies, the artificial neural network (simulating humans' perception) is an attention-grabbing approach [3].

Because of its simplicity and usability, the Hidden Markov Model (HMM) is quite significant in speech recognition, though it is significantly state-dependent and requires rigorous phoneme modeling. Another critical component of speech recognition is the recognition of past speech; specifically, this is one of the areas that are deficient in the HMM model. In addition, HMM demonstrates poor speaker adaptation, and it primarily deals with speaker-dependent systems requiring individual users' training [4]. To overcome the discussed problem, a Neural Network may offer a solution. Differing from HMM, the neural network approach is not much dependent on the speaker to a large extent, hence, the process involves model creation, while to create the network structure, the system is trained once only. It implies that the overall time to train the system decreases considerably compared to HMM. The entire output neurons in NN, individual neuron symbolizes one sort of recognition, a definite level of excitation; hence, no information loss occurs yet for the inaccurate output.

This article explores neural networks that seem to provide a practical solution for speech recognition inspired by research on the human nervous system, showing that the basis of most speech recognition systems is neural networks that have several advantages against the traditional methods, and discusses how to address the speech recognition challenge by designing a generative probabilistic model of speech. The next chapter focuses on the speech pre-processing phase required for extracting the significant information from the speech signal. Feature Extraction will be given in the third chapter. The implementation of the neural network was done in the fourth chapter. The result is discussed in the fifth chapter. The conclusion remarks are presented in the sixth chapter, and in the last chapter of the paper, future work is discussed.

Speech Recognition Pre-Processing

The underlying assumption made by feature extraction and recognizer techniques, which has been exploited for speech to be short-term stationary, is violated when applied to the speech signal blindly, and hence, no significant result can be obtained [6]. Before the speech signal is processed into the feature extraction and recognition phase, it must extract some valuable and interesting vectors from the raw information, called front-end processing. The speech boundaries are located during this step, and the acoustic signals of isolated words were segmented from the non-speech silence portion is predominantly done at front-end detection, influencing our preprocessing. The problem arises when there is noise from surrounding events of body actions, such as tongue clicks, gulps, coughs, or lip smacks. So that non-speech events or noise may interfere with speech waveform endpoint detection, this can be a challenge especially when the words are started and ended by low-energy phonemes (like weak first burst /p/ , /t/, fricative such as stuttering occurrence of ʊ sound in inhibit [ɪn̄hɪbəʔ] nasals like end-onset formant transitions).

A. **Normalization/ Mean Correction**

Normalize the mean correction of the digital signal to 0 (to avoid any offset caused by noise on the microphone) first. It provides the mean correction to remove any DC zero introduced by the microphone or AD converter. This operation enables the amplitude of data to be limited between -1 and +1 – that is, normalization´s main benefit. To be computed by division of the current value by the signal's absolute value.

B. **High-Pass Filtering**

The recorded speech signals had a sampling rate of 8 kHz and were digitized at 16 kHz, they also required pre-processing. Since the majority of the energy is contained in voiced sound, a lower range of under 4 KHz, extra bandwidth has been shown to degrade the performance of ASR systems [7]. Those extra filters helped a great deal: According to the experiment, the filters' clarity and high-pass produced 4% more output rate than the filters' band-passed data.

C. **Framing for Short-term Analysis**

The vocal process has dynamic and non-stationary properties with ongoing variations in speech waveform amplitude as a result of varying vocal tract and articulator parameters [8]. During speech analysis, it has been observed that during the nonstationary process, the statistics of the process change more gradually with the increase in time. While this is not completely true, it provides some way to treat a short-time portion of speech, about 10 ms- 40 ms long, as a stationary process. The procedure in the case of the speech signal involves breaking it up into sets of sequential overlapping frames, each consisting of N samples and starting a position m samples away from its neighbour, where normally m is fixed at half the frame. The length of each window is sufficient to extract useful stationary spectral information from the signal. While the first window is determined by the first N samples, the next window begins m samples later, having N-m samples in common with the prior one; the rate of the overlapping would ascertain the speed with which the vectors change, in comparison with the frame [9].

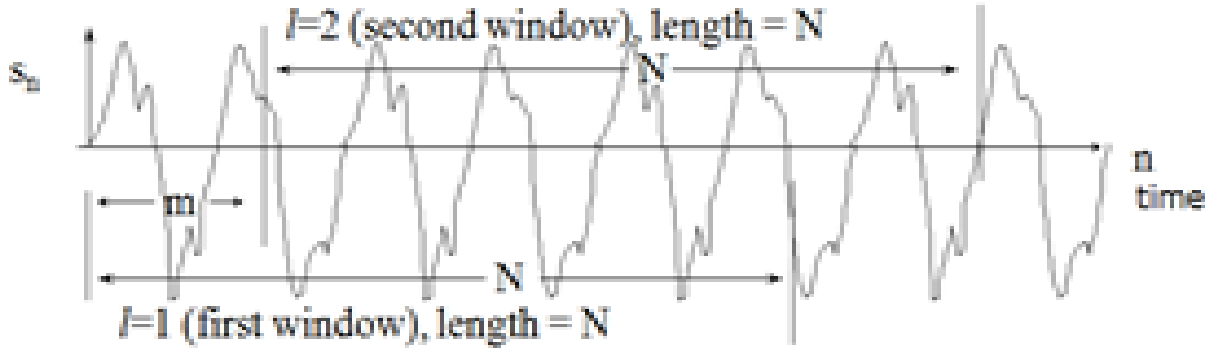

Figure 1.
Illustration of Framing [9]

Frame size is an important parameter in spectral analysis. A short frame size is necessary to capture any significant spectral characteristics, whereas long frame sizes allow for

processing frequency resolution. Experiments have shown that too short a frame duration, or equivalently, an increased frame rate, will indeed increase complexity and memory requirements. Generally speaking, frame sizes are taken at 20- 30ms with consecutive frames varying by a 5ms shift. In the proposed work speech signal is sampled at 8 KHz, each frame size equivalent to ~30 to 32 ms and consists of 256 samples, with 40% overlapping.

D. **Windowing for Short-term Analysis**

To conduct Fourier transform computations, we set small short-time stationary blocks in the frames, and since these blocks have discontinuities on their edges, we require windowing, perforation. When any function, or a speech signal s(n), is multiplied by a window function w(n) gives rise to zeros under categories by whose bounds they are defined, such that it results in zeros outside the range; thus, windowing tries to manage this. Applying a window that tapers at both ends is equivalent to the minimization of maximum spectral distortion caused by discontinuous differences at the endpoints of the window. Therefore, long windows would allow for high spectral resolution and capture small spectral variation, thus increasing spectral distortion. The duration of bracketing plays a crucial role in determining how much averaging will be done by any power or energy calculation. Window size is one of the critical parameters in short-term analysis. More precisely, the definition of a good frame and window duration eventually relies on the change of rate at which the shape of the vocal tract changes.

E. **End Point Detection**

1) The main task of End Point Detection is separating speech data from background silence and noise resulting from the speech utterance. It takes in a framed signal as input and, as output, it gives truncation; this means that the signal before and after the speech declaration has been removed from the non-speech signal. Missed detection of boundaries with a speech segment during isolated word utterances leads to two adverse effects: increased computation and wrong recognition [10].

2) Short-term Energy Measure (STE Measure):

A more voiced-related section represents more energy than the unvoiced section, while silence has no energy and needs to be removed from the speech signal. Short-term energy is one of the common parameters that have been in use since the 1970s to distinguish between voiced sounds and unvoiced sounds or silence [11].

The short-term absolute energy was calculated by determining the absolute sum content of each frame as follows [12]:

$$E_s = \sum_{n=m-N+1}^{m} |s(n)w(m-n)| \qquad \textbf{(1)}$$

where "m" is the shift/rate of samples and w(m) represents the windowing function with the frame duration N-length terminating at n = m. The speech signals have been cropped to exclude the silence before and after each word is spoken

3) Zero-crossing Rate (ZCRate):

Zero-crossing represents the average number of times the signal varies its sign, used to distinguish between voiced or unvoiced speech [2]. Automatically, in a given signal, if

the Zero Crossing Rates are high or low, then the waveform changes quickly and hence the waveform contains high frequency or the waveform changes gradually and therefore the waveform contains low frequency, respectively. Hence, ZCR provides information about the average occurrence of energy concentration in the signal.

The relation of zero-crossing for non-stationary signals like speech was formally defined in 2000 by Deller as [13]:

$$z(x) = \frac{1}{2N} \sum_{m=0}^{N-1} s(m).w(n-m) \qquad (2)$$

4) Endpoint Detection Algorithm from STE and ZCR Measures:

The EPD algorithm's initial step calculates the signal's short-term energy STE for every speech sample at 10-ms intervals around each sample point. The energy thresholds are calculated based on the energy's minimum and maximum values. When estimating the spoken utterance endpoints, two energy thresholds, the lower and upper threshold energies, are taken into consideration. These are calculated in this way [14]:

TL = 8×MINSTE (3)

TU = 32×MINSTE (4)

MINSTE = MIN (IE,MEAN(STE) + STD(STE)) (5)

Where IE = 0.25, TU is the upper energy threshold, TL is the lower energy threshold, and *STE* stands for short-time energy.

Fig. 2 illustrates Short Term Energy for the word "seven"; the blue and green lines correspond to the upper TU and lower TL energy thresholds, respectively.

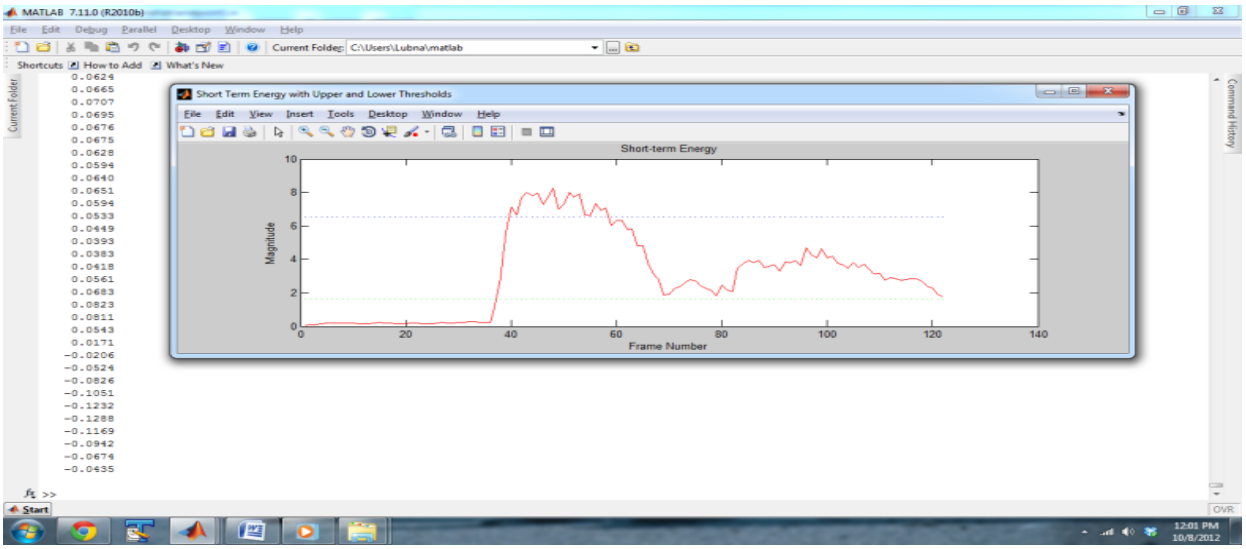

Figure 2.
Short Term Energy

By assuming, the noise of the speech signal is considerably lower than speech energy, the mean and the standard deviation are computed for the STE and ZCR duration of

the first 50 ms of recording, then another threshold, namely threshold zero crossing, is calculated as [15]:

$$TZC = \text{MIN}(IF, \text{MEAN}(ZCR) + \text{STD}(ZCR)) \quad (6)$$

$$IF = 0.25 \times N \quad (7)$$

Where *ZCR* zero crossing rate and N is the frame length;

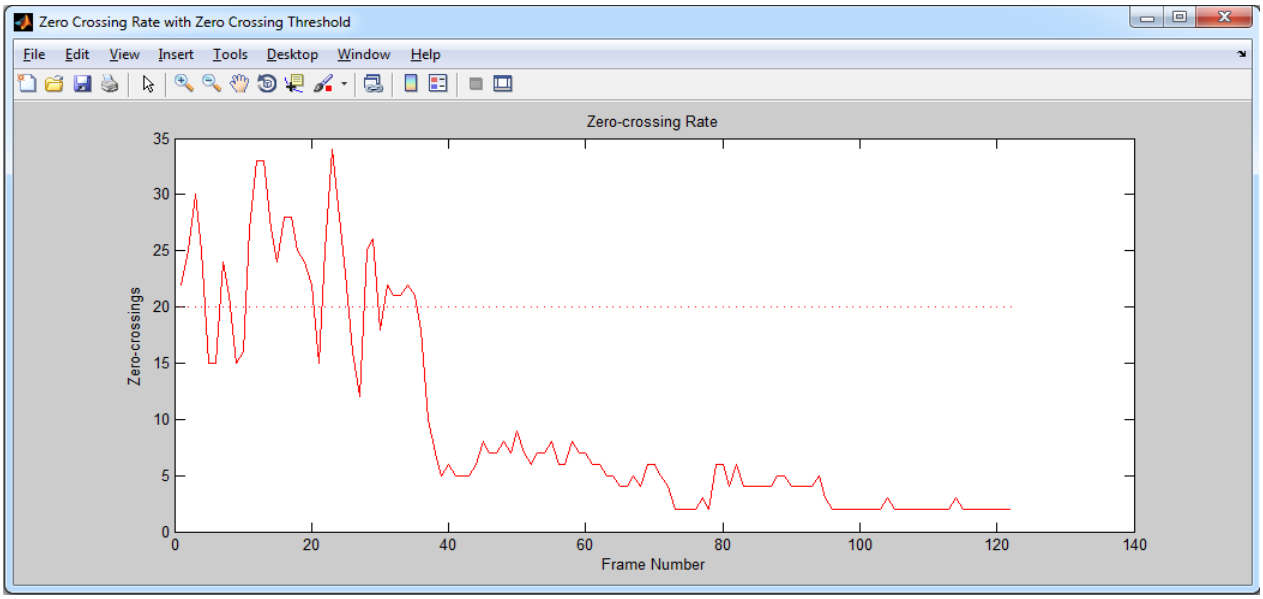

Figure 3.
Zero Crossing Rate with Zero Crossing Threshold

The illustration above in Figure 3 indicates that the zero-crossing rate is low in speech and fairly high in noise at the very beginning and end of speech events: the red horizontal line indicates the zero-crossing rate threshold. Upon calculating STE, ZRC, TL, and TU, the short-time energy curve is reviewed to determine the first and second sequential end points of the speech utterance concerning the lower threshold TL and upper threshold TU further analysis.

Estimate the onset of speech utterance by finding the sample points (t) with the largest index t so that the energy STE(t) is larger than the energy of the initial violation above the lower threshold ITL and smaller than the first breach above the upper threshold ITU. For establishing the endpoint of the speech utterance, the sample s(t) for which the index t is minimal is selected, though the energy E(t) of the selected sample must be above the energy of the last sample under the lower limit ITL and below the energy of the last sample over the upper limit ITU It is worth mentioning that, the initiation and termination locations referred to in the references are: [16] as follows:

$$N1 = \text{MIN}((STE(\tau) < ITL)) \cap ((STE(\tau) > ITU)) \quad (8)$$

$$N2 = \text{MAX}(STE(\tau) > ITU)) \cap (STE(\tau) < ITL)) \quad (9)$$

Where N1 and N2 are the speech signal's initial approximations for the start and end points, respectively.

It is reasonable to assume that the voice signal was contained inside the interval {N1, N2}. Examining the zero-crossing rates across a 50-ms interval before N1 and a 50-ms interval following N2 is how the endpoint detection process continues. The number of times the Zero Crossing Threshold (ZCT) rate was broken during the testing period before and following N1 and N2, respectively, is used to determine the speech signal's final endpoints.

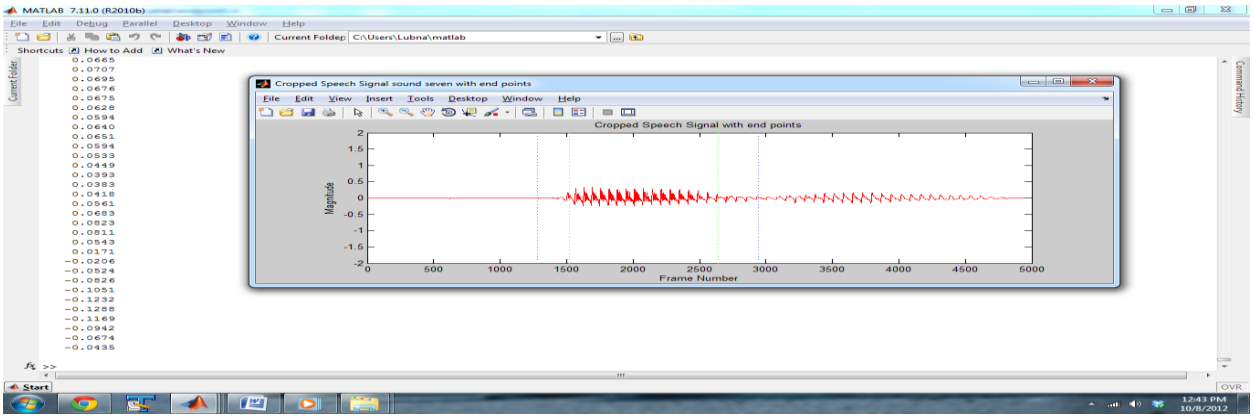

Figure 4.
Cropped Speech Signal sound 'Seven' with end points

The last and most specific boundaries of the signal are defined with regard to the count of zero-crossing rate ZCR(t) breaches the zero-crossing count IZCT on the 50 ms interval preceding N1 and the 50 ms interval preceding N2. If within the 50 ms interval preceding N1, the number of times the Zero Crossing Rate (ZCR) threshold breaches is more than 30, then the first count of Zero Crossing Rate (ZCR) breach is taken as the beginning bound of the signal. Otherwise, the speech start point remains N1. Similarly, for the 50 ms interval after N2, if the zero-crossing violation count exceeds 30, then that last zero-crossing rate violation is taken as the speech signal end, otherwise, the end sample of the speech signal stays at N2. In Fig. 4, EDP is End Point Detection. The image shows Blue vertical lines, which show approximation to the end bound of the speech signal, and green vertical lines show approximation to the commencement bound of the signal referred to as N1, N2, N3, and N4.

FEATURE EXTRACTION

Feature extraction is one of the vital elements of ASR, which extracts feature vectors or parameters containing relevant and useful information from the short-term frames speech signal under consideration.

The spectrogram has been used to compute Short-Time Fourier Transform, and is currently a very useful and efficient tool, such as the MFCC [17].

A. **MFCC Processing**

MFCC is a representation of real cepstral extracted features vector calculated from windowed short-time Fourier analysis based on frames [18]. It tentatively estimates the acoustic processing of the spectrum in terms of a nonlinear frequency scale. Significantly MFCC extraction process is applied independently, frame by frame, to reduce dimensionality, extract valuable and relevant information.

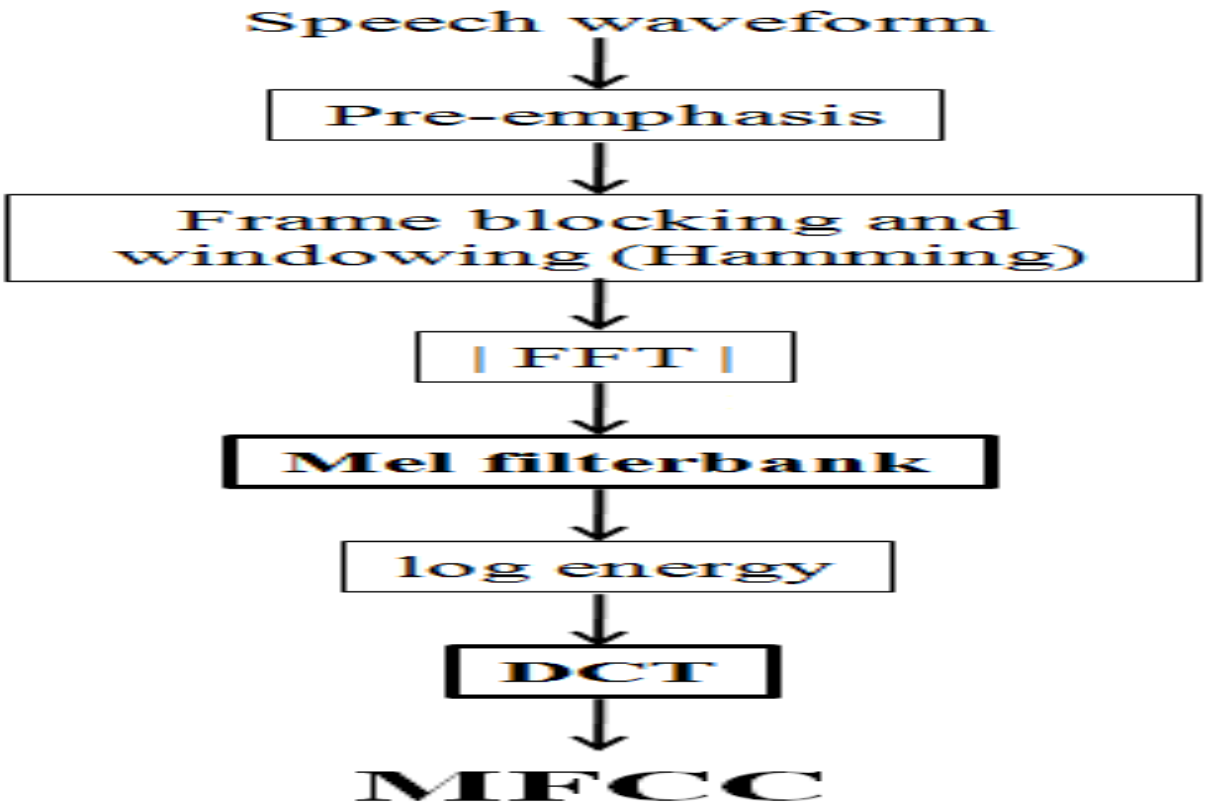

Figure 5.
Basic building blocks of MFCC

MFCC preserves the temporal and spectral qualities of the speech signal and expresses the speech amplitude spectrum concisely. The reason for the MFCC features' computation necessitated a process for efficient computation. The process for the extraction of MFCC feature vectors is shown in Fig. 5. Briefly, Mel-Cepstrum is computed by:

- Speech analysis is carried out over a short-term frame window
- The spectrum is obtained for each short-term frame analysis window by using the FFT
- Triangular overlapped windows are used for mapping the power spectrum obtained through the Mel-Filter onto the Mel-scale to obtain the Mel-Spectrum.
- Taking the logarithm of the energy of each mel frequency.
- By taking the DCT on the Mel-Spectrum, MFCC is obtained.
- Thus, the sequence of Cepstral vectors obtained, forwarded to the next recognition process

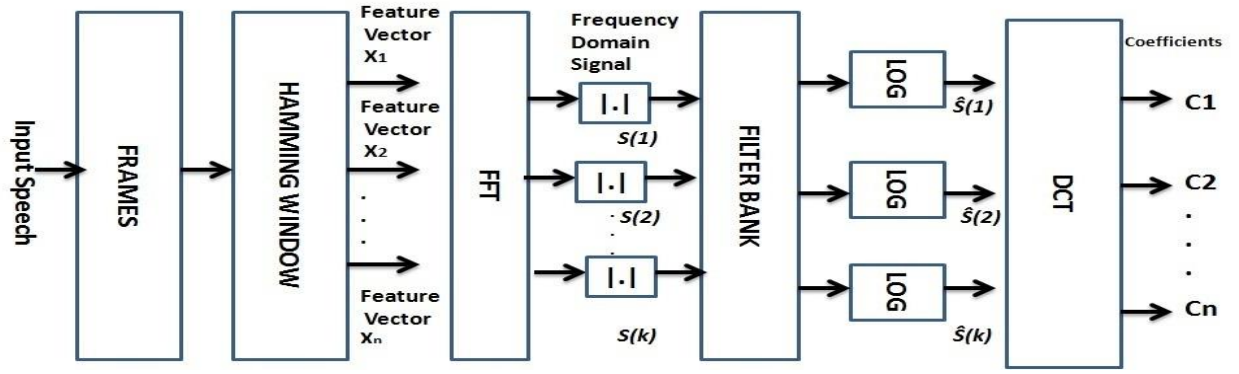

Figure 6.

Calculation of Mel Frequency Cepstral coefficients

1) Pre-emphasis: During the sound production, the high-frequency components of the speech signal are concealed, the pre-emphasis filter is used to compress the speech spectrum and to pay compensation on this frequency, illustrated as [19]:

$S_2(N) = S(N) - A*S(N-1)$ *(10)*

Where s(n) is the speech signal sent to a high-pass filter, $S_2(n)$ is the output signal, and the typical value of a ranges from 0.9 to 1.0

2) Framing After Pre-emphasis, framing must be done at fixed intervals in order to divide the signal into frames, so that we can model the signal into small sections. In order to reveal the dynamic variation in signals, it is obligatory to figure out speech parameters in small intervals [7].

In the proposed work for calculating MFCC, the cropped speech data is split into N = 256 samples in each, with 8000Hz Sampling Frequency corresponding to 32 ms, with 40% overlapping to better capture temporal changes from frame to frame.

3) Windowing: In the proposed work Hamming window is used for removing the edge effects, hence generating a cepstral feature vector for all frames. The Hamming window $W_H(n)$, defined as [20]:

$$W_H(n) = 0.54 - 0.46\cos\left(\frac{2n\pi}{N-1}\right) \quad (11)$$

, where 0<=n<=N-1

The use of Hamming windows is due to the fact that MFCC will be used, which involves the frequency domain, and Hamming windows will reduce the chance of high-frequency components in each frame due to such quick slicing of the signal.

4) Fast Fourier Transform FFT: Spectrogram is a good enough two-dimensional representation of speech signals for short-time Fourier analysis, demonstrating the frequency on the perpendicular axis where whereas on the x-axis, time is depicted. For mentioning energy point of time/frequency is depicted in gray scale, black represents energy in high, whereas low energy is symbolized by white.

Fourier transformation is usually done to execute cepstral and spectral analysis. One of the best performances of Fast Fourier Transform (FFT) can be achieved by Discrete Fourier Transform (DFT), which converts N-samples of frames into frequency spectrum and is better to use because it holds excellent properties of frequency localization.

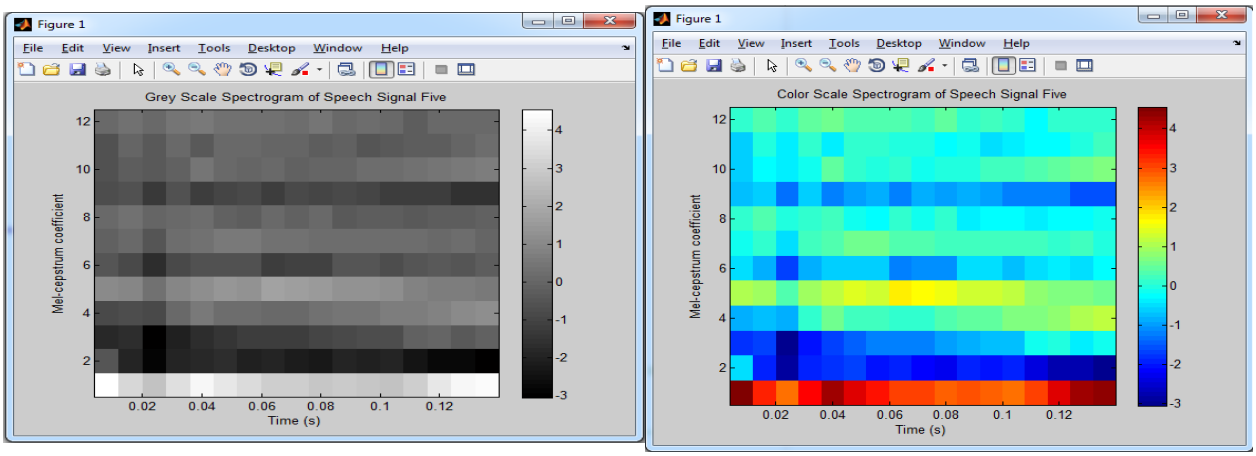

Figure 7.
shows Mel Frequency Cepstrum Coefficients MFCCs and Spectrogram of the word 'FIVE'.

5) Mel Scale Frequency: The frequency below thousand Hz, human hears tones by means of a linear scale as a substitute of logarithmic scale for those beyond thousand Hz. The speech signal whose frequency element is lower bears essential information in contrast with elements of higher frequency. Also, for low frequency, Mel scaling is done to emphasize it.

The subsequent step is adjusting the frequency resolution to a perceptual frequency scale and realizing the human ear properties [6], mels for a given frequency *f* in Hz can be computed by the formula [20]:

$$\text{MEL}(f) = 2595 * \log_{10}(1 + \frac{f}{700}) \quad (12)$$

The human ear can perceive frequency tones falling within the range of identical critical bands. The Mel frequency scale is logarithmic for frequencies exceeding thousand Hz, whereas linear below; also, samples are of identical numbers lower and higher than thousand Hz. Depicting the perceived tone's frequency.

Analysis carried on mel frequency is grounded on human perception trial, it has been experienced that ear operates like a filter, these are not evenly spaced over axis of frequency and are called as Mel filter banks; these pay attention to only particular elements of frequency, therefore extra filters are needed in the lower segment of frequency and vice versa because for spectrums' lower frequency elements the human ear possesses better frequency resolution and vice versa.

The inner functionality of an ear can be treated as a linearly spaced band-pass filter bank on lower frequencies, but for those higher ones, it is logarithmic. Energy lies at the lower frequency of a vocal signal. Hence, it is expected naturally to work with a mel-spaced filter possessing these attributes.

Collections of band pass filters along with the differing frequencies enveloping the frequency and formants; and these are pieces of concern within the spectra. At the time of frame, the results of filters are treated as features. Within the filter, middle frequencies are selected; these are located according to a perceptual scale. Middle frequencies, by the Mel frequency scale, are uniformly linearly spaced less than a thousand Hz and equally spaced logarithmically. Illustration highlights that middle frequencies are in correspondence with the Mel frequency domain's center frequencies, which are identically spaced. In the proposed work, triangular-shaped Mel Frequency Filter banks were calculated, as shown in Fig. 8, for every word having

- 24 filters in the filter bank
- 8000 sampling rate
- 256 sample frames

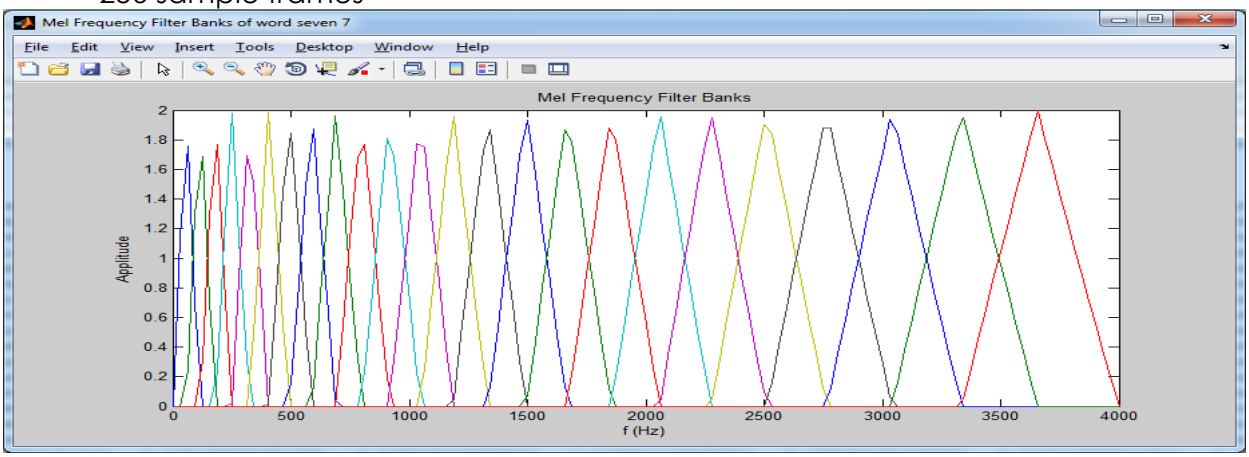

Figure 8.
Mel Frequency Filter Banks for word 'SEVEN'

6) Cepstral Analysis:

The compact representation of the speech spectrum is used for signal depiction. There are two components in S(f): the vocal tract response (H(f), linked with low frequency elements, and the excitation spectrum (E(f), which varies quickly. Smooth spectrum is vital, must represent H (f), whereas not signifying E (f) for the mission of speech recognition. To deal with this dilemma, making use of cepstral analysis, it is noted that the logarithm influences transforming the product into addition. Thereby, transforming the product of the magnitude of the Fourier transform into an addition.

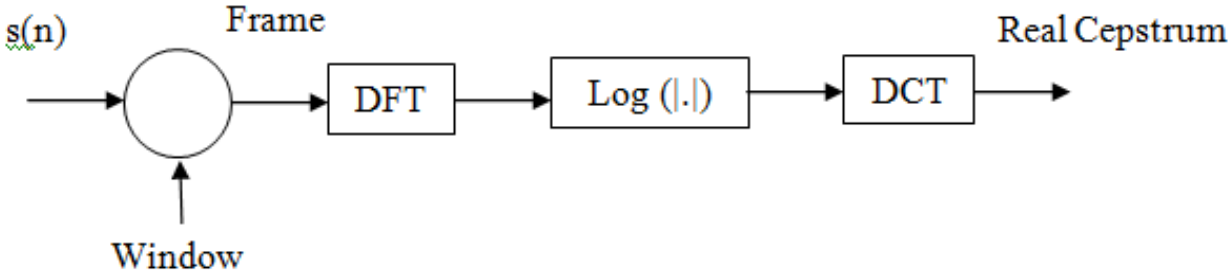

Figure 9.
Real Cepstrum Computation

Short-term real cepstrum computation represented in Fig. 9. The cepstrum can be specified as:

$C_S(N) = €^{-1}\{LOG|€\{S(N)\}|\}$ (13)

Where € and $€^{-1}$ are the Discrete Fourier Transform DFT and Inverse Discrete Fourier Transform or DCT respectively.

The breakdown of a speech signal s(n) into the excitation sequence e (n) and the vocal tract function h (n) [3]: The dividing of a speech signal s (n) into excitation sequence e (n) and a vocal tract feature h (n) [3].

$S(N)=E(N)*H(N)$ (14)

By transforming the time domain function into the frequency domain function of the speech signal, the convolution operator '*' turns into the multiplication.

$S(F)=E(F).H(F)$ (15)

Considering that the waveform is real-valued, the logarithm of Eq. (3.4) on both sides leads to

$LOG|S(F)| = LOG|E(F)|$. (16)
$H(F)|=LOG|E(F)|+LOG|$
$H(F)|=C_E(N)+C_H(N)$

By taking the logarithm of the spectrum, the multiplication turns into an addition.

7) Discrete Cosine Transform DCT: Next step is to compute the *real cepstrum* by applying the Inverse Discrete Cosine Transform (DCT). In short, under a cepstral transformation, the non-linear convolution of two signals $e(n)\otimes h(n)$ becomes equivalent to the linear sum of the cepstral representations of the signals, $c_e(n)+c_h(n)$

$C_s(n) = €^{-1}\{C_e(n)+C_h(n)\}= €^{-1}\{C_e(n)\}+ €^{-1}\{C_h(n)\}= c_e(n)+c_h(n)$ (17)

When a Discrete Cosine Transform (DCT) replaces inverse DFT, the procedure becomes simpler because the log magnitude spectral of the coefficients have real and symmetric attributes [21].
Due to the real and symmetric nature of the log magnitude spectral of the coefficients, use of Discrete Cosine Transform (Additionally, DCT is known to produce feature sets where the elements are barely or not correlated at all. The computation of DCT is superior to the inverse DFT computation.

The resulting MFCCs are calculated as follows:

$$c(k) = w(k)\sum_{n=1}^{N}\log\left(\left|x_k(n)\right|\right)\cos\left(\frac{\pi(2n-1)(k-1)}{2N}\right) \quad (18)$$

k=1,2,….L

$$w(k) = \begin{cases} \frac{1}{\sqrt{N}}, & k = 1. \\ \sqrt{\frac{2}{N}}, & 2 \le k \le L. \end{cases} \quad (19)$$

L = 24 (Number of filters in the mel-scale filter bank)

N=256(window Length)

And $x_k(n)$ stands for the output of the *kth* filter in the filter band

The zero-order coefficient in MFCC represents the logarithmic mean energy. This coefficient is usually removed from the feature space in the absence of confidence in absolute power measurements in speech recognition [22]. The data from the experiment shows that 12 coefficients are enough for a representation of the investigated speech data. Therefore, the zero order coefficient was ignored, and the first 12 MFCC coefficients were used for compact representation.

FEED FORWARD/MULTILAYER PERCEPTRONS NETWORK

The simplest arrangement of a Neural Network, multiple-layer perception, is mainly established for speech recognition [23]. In 1989, Cybenko and Funahashi described Multi Layer Perceptrons with a sufficient number of neurons in the hidden layer [24]. Input fed into the network is forwarded via the input layers toward the hidden layers and then towards the output. For this class of NN behaviour initiated the other name, feed-forward neural networks, signals travel only in one direction, that is, from input to output, and the same signal never affects its output, as that has no feedback loop, extensively used in pattern recognition.

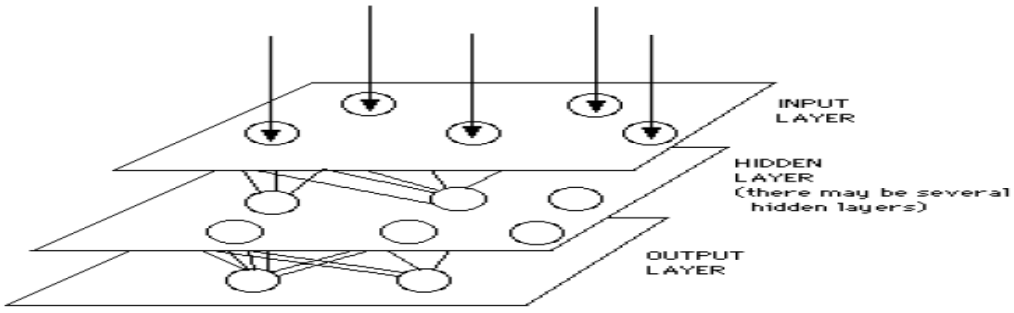

Figure10.
Feed Forward Neural Network

A. Back Propagation Learning

One of the most popular techniques, "Back Propagation" used by Multi-layer networks, is also proposed I this thesis. Error is calculated by comparing the correct answer with the output value, This learning is used to adjust the weight of neurons to minimize the error function. To downhill the derivative of the error function, also referred to as the generalized delta rule. Ending up of error function in local minima and speed convergence are the two major problems generally faced by back propagation. Back

propagation consists of the forward pass and the backward pass phases. In the *Forward Phase*, the input patterns are transmitted forward from the input layer of the network until they reach the output units. Input neurons, along with their weights, are summed to stimulate hidden layer neurons, thus producing a prediction sample.

Units i and j are typical units in the previous and output layer, respectively.

Activation function of an output unit can be computed as:

1. Total weighted input net_j computed as,

$$net_j = \sum_i (x_i * w_{ij}) \qquad (20)$$

2. Then the activity xj was calculated using the total weighted input function as shown in Eq. 5.2. Typically, use the sigmoid function

$$x_j = \frac{1}{1 + e^{net_j}} \qquad (21)$$

In *the Backward Phase*, computation of the derivatives flows in the backward direction, therefore entitled as back-propagation [25]. Errors at the output layers "back propagate" to the previous layer, thus used to update the weight of neurons of these layers, for the purpose of training in neural network algorithms.

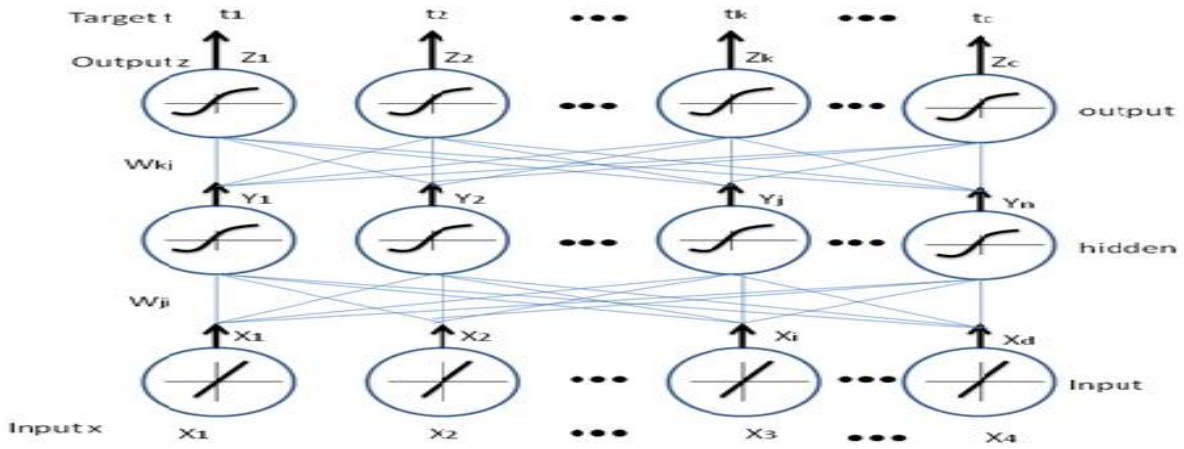

Figure 11.
Three Weight Layers Feed Forward Back Propagation Neural Network

Network calculates the Error E, after computing all the activity functions of the output unit's neuron. Error is computed by taking the difference between the output vector and the targeted output.

$$E = \frac{1}{2}\sum_j (x_j - t_j)^2 \qquad (22)$$

The weights need to be updated very smoothly such that it doesn't affect all the preceding knowledge, as collective information of weights is scattered throughout the

network. Thus a fundamental parameter called the learning rate (□), a small constant is thus introduced to control the amount of alteration of weight [26], suitable value for learning rate (□) is a crucial task, if (□)□is too large, it disturbs all the earlier learning but if it's too small, learning procedure remains everlasting.

B. Implementing the Back-Propagation Algorithm

The major steps involved in implementing the Back Propagation Neural Networks algorithm are described as follows:

1. The weights are initialized with random values.

2. From the training vector, select an input and corresponding output vector.

3. Compute the actual outputs from the network.

4. Calculate Error by taking the difference between the obtained output and the targeted output; adjust weights Wh and Wo of hidden and output layers, respectively, to minimize the error.

5. Repeat all steps from step 2 for all training vectors until the error is acceptably small.

Input vector X^T multiplied by the hidden layer weight matrix Wh to calculate the hidden layer output Yh as shown in Eq. 23

$$Yh, j = f\ (\Sigma Wh\ ji, *X) \quad (23)$$

Where *f* is the sigmoid function. It reduces the output value to lie in (-1, 1), and Wh,ji is the weight in the hidden neuron. The output of the network Yo, k, is calculated by applying the logsig sigmoid function at Yh.

$$Yo, k = f\ (\Sigma Wo,kj * Yh,j) \quad (24)$$

Where ΣWo,kj is the weight in the output neuron layer, and Yo, k is the output of the network. The error Ep is given by

$$Ep = 1/2\Sigma\ (dp,k - Yp,o,k)^2 \quad (25)$$

Hidden and Output delta error is computed by comparing the target and output vector to reduce the difference as follows [27]:

$$\delta h,j = Yh,j(1-Yh,j)\Sigma\delta o,k\ Wo,kj \quad (26)$$

$$\delta o,k = Yo,k(1-Yo,k)(dk-Yo,k) \quad (27)$$

The error is calculated based on the impact of output delta errors. To minimize the error weights are modify as:

$$\Delta\ Wo,\ kj = \square\ \delta o,\ k\ Yh,\ j \quad (28)$$

$$\Delta\ Wh,ji = \square\ \delta h,jXi \quad (29)$$

Where □□is the learning rate coefficient.

The hidden and output layer weights are changed as:

Wh, ji=Wh, ji+ □□δh, j*Xi (30)

Wo, kJ=Wo, kj+ □ δo, k Yh (31)

For each vector training remain continue until the error reduce to small value.

C. Back Propagation Neural Network Configuration

This paper present a 3-layer feed-forward neural network is used for both training and testing model of speech recognition.

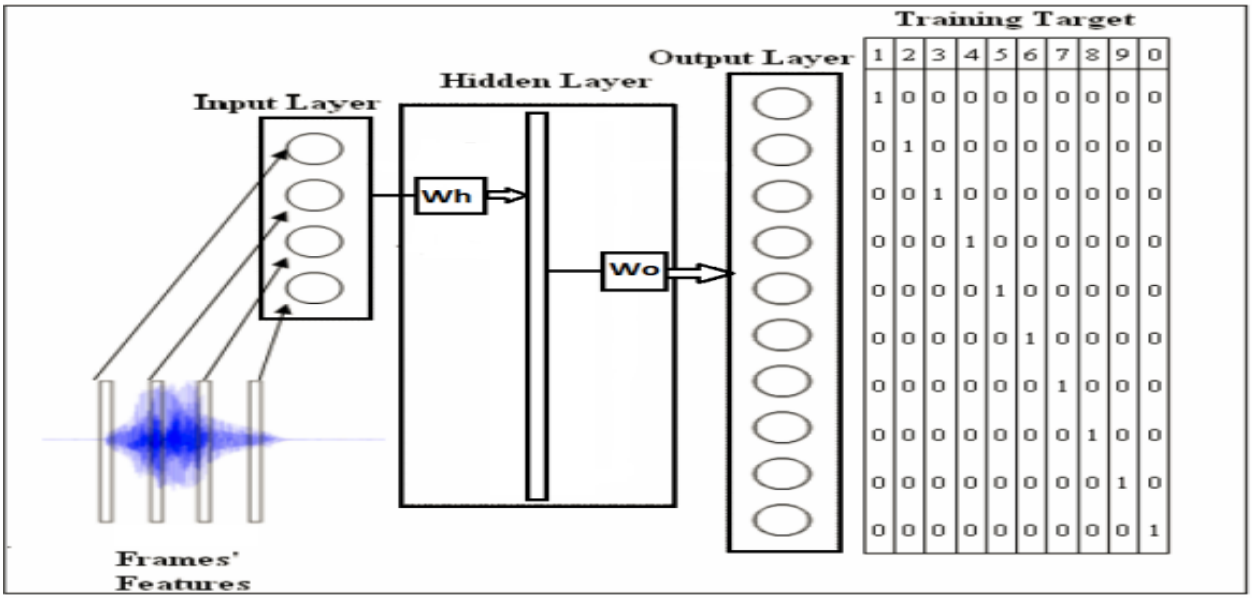

Figure12.
Complete and Simplified Neural Network Architecture

The first layer represents the input layer which contains 24 input neurons, as each speech signal is split into 24 frames orientation to the beginning and terminating point of the signal. During the training process and testing, Mel Frequency Cepstral Coefficient (MFCC) of each of the 24 frames are used as the input vector, furthermore each frame consists of 256 samples frequency.

The second layer is the hidden neurons layer, during the training process; the optimal numbers of hidden neurons are to be found out for efficient processing. The third layer is the output layer, consists of 10 neurons which represent 10 different digits to classify. In order to recognize *words*, a whole word at a time must be define in a neural network that is its outputs show the *N* words in the vocabulary and its inputs represent all the frames of speech in a whole word, and, so that we can compute back propagation error by defining the difference of the obtain output against the targets output is "1" for the correct word and "0" for all incorrect words, and back propagate error through the whole network.

The whole database consists of:

a. 10 experimental adult subjects uttered each word 5 times

b. Set of small vocabulary consists of ten words: {'zero', 'one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eight', 'nine'}

c. Total number of utterances: 10X 10 X 5 = 500 utterances

RECOGNIZER CLASSIFICATION RESULTS

For better consistency and direct comparison, the parameters of the BNN are standardized for all experiments described here. The applied optimization method is delta rule. The proposed thesis uses neural nets with one hidden layer of tansig activations, an output layer of logsig functions and a learning rate of 0.01, and epoch count of 100. The system has 10 output neurons while the hidden layers have between 10- 60 neurons. Each frame is represented.

A. Impact of Pre-processing on ASR

The performance of a network varies drastically depending on the pre-processing of the speech signal. Raw signal without pre-processing causes decrease in the efficiency of the network since it takes too long to train the network.

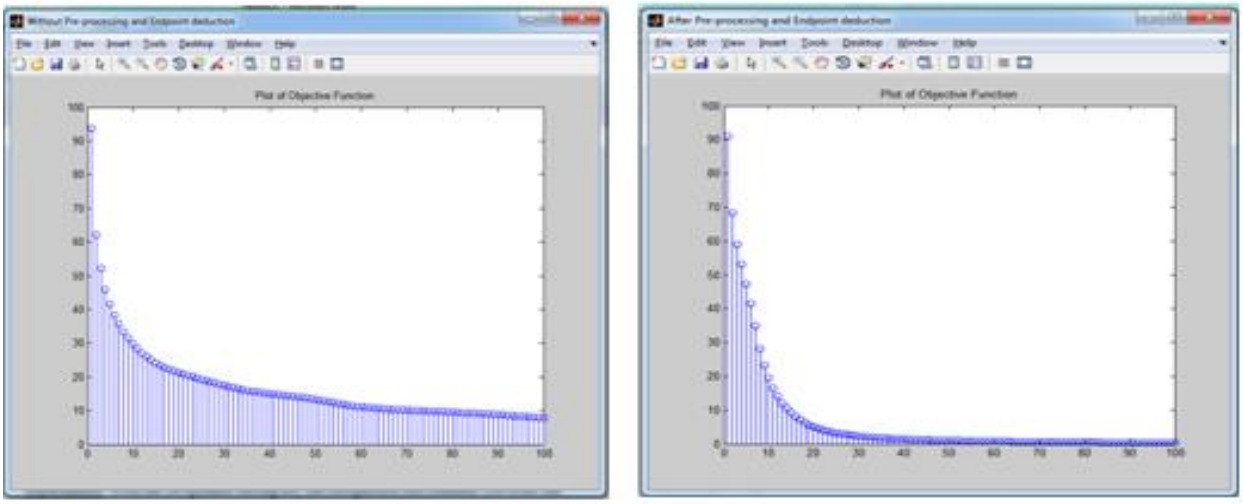

Figure13.
Accuracy Level (a) before Endpoint Deduction (b) after Endpoint Deduction

Experiment results from Fig. 13 (a) representing recognition results without pre-processing whereas Fig. 13 (b) representing speech recognition results after pre-processing shows the performance and recognizer rate of ASR decrease by 81% to 64%.

B. Vocabulary size

The size of vocabulary used in the training and testing set also affects the performance of the model. As the number of vocabulary size increases, the recognizing rate decreases rapidly, as shown in Fig. 14 with a 5-word vocabulary; the recognition rate is almost 89.5%.

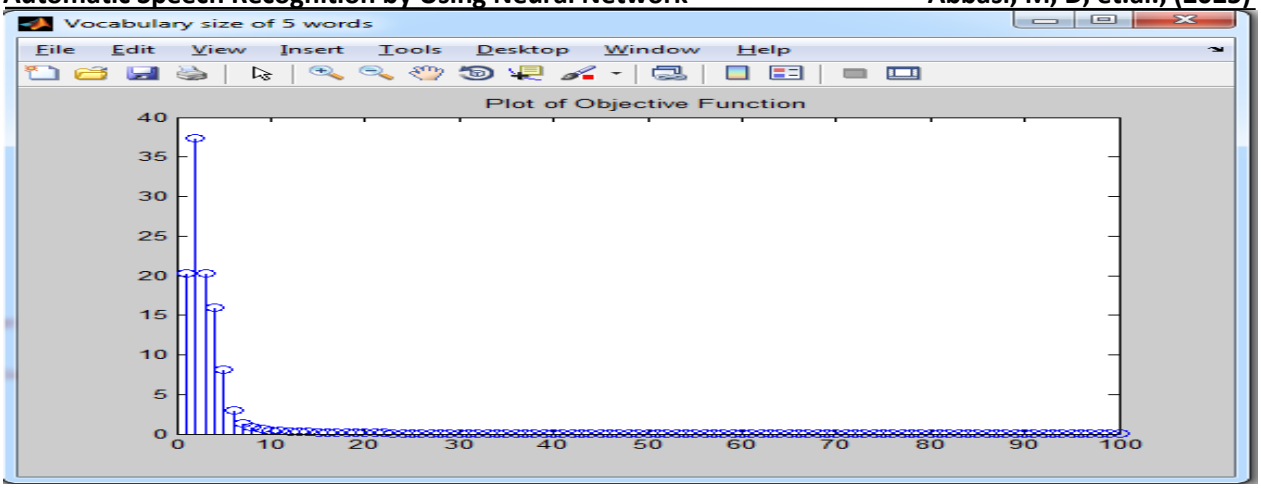

Figure11.
Accuracy Level of ASR of vocabulary size of 5 words

C. The number of Neurons in hidden layers

There is a need to say that the structure of a network plays a part in determining how well it performs. The performance of the network is better with an increase in neurons to a certain point, beyond which it eventually decreases. Through experimentation with several neurons in the hidden layer between 10 and 60, the highest accuracy was achieved when 40 neurons were used.

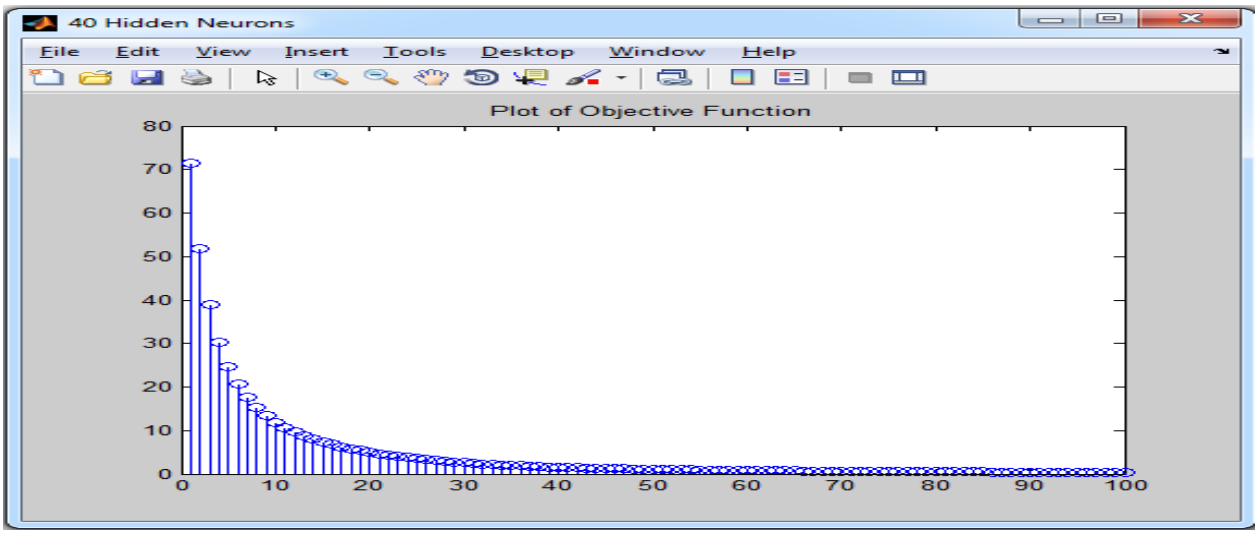

Figure12.
Accuracy Level with 40 hidden Neurons
Table 1.
Number of Neurons in the Hidden Layer Network

NO OF HIDDEN NEURON	RECOGNIZER RATE
10	68%
20	71%
30	78%
40	81%
50	74%
60	69%

D. Learning rate

For best learning rates we experimentally studied different learning rates and concluded from the following graphs that learning rate of 0.01 was found to give the best recognition rate and fast training.

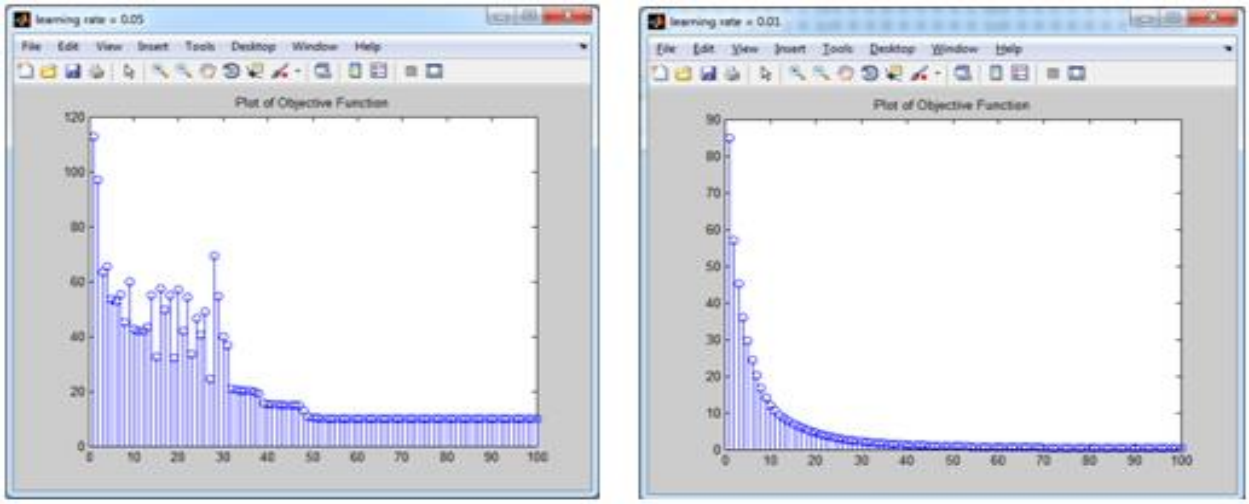

Figure 13.
Accuracy Level with (a) Learning rate 0.05 (b) Learning Rate 0.01

Table 2.
Effect of Learning Rate on Recognition Rate

LEARNING RATE	RECOGNIZER RATE
0.05	68%
0.01	81%

E. Confusion Matrix Result:

Finally, test build ASR model in three different scenarios and elaborate on their Confusion Matrix as follows.

1) Train and Test the Model by Same and Single Speaker: In the first scenario depicted in graph of Fig. 17, an experiment was conducted using the same and single speaker for training and testing. Overall accuracy level of the ASR model was 84%. This experiment yielded that if we use the same and single speaker for ASR model, we achieve the best results but this is not a practical solution for ASR models.

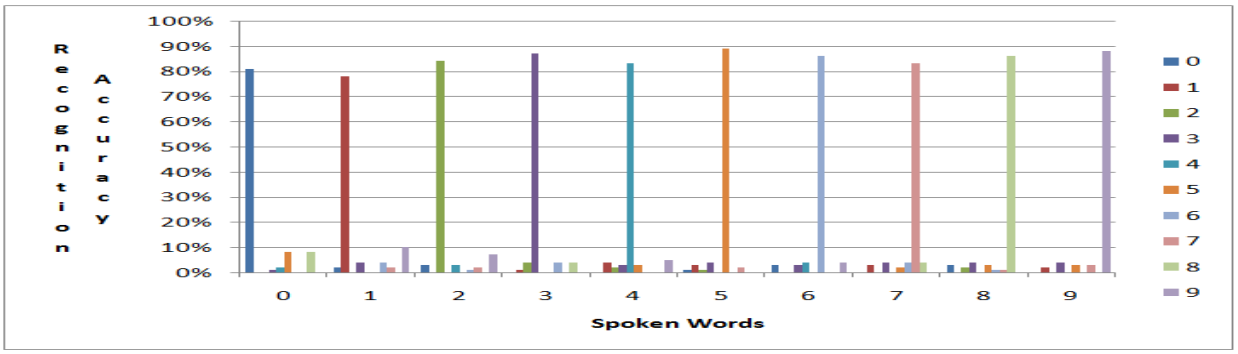

Figure17.
Accuracy Graph of model having same and single Speaker for Training and Testing the Model

1) Train the Model by Excluding a speaker and then Test Trained Model from the Excluded Speaker: In this next experiment we trained the system using 9 speakers and tested the trained system using a 10th speaker, who was excluded from the set of training speakers. The training database, consists 10 words spoken by 9 speakers uttered each word 5 times, 5*9*10 = 450 utterances and second set comprises same 10 words spoken by single excluded speaker, that was not part of training uttered each word 5 time, 5*1*10 = 50.

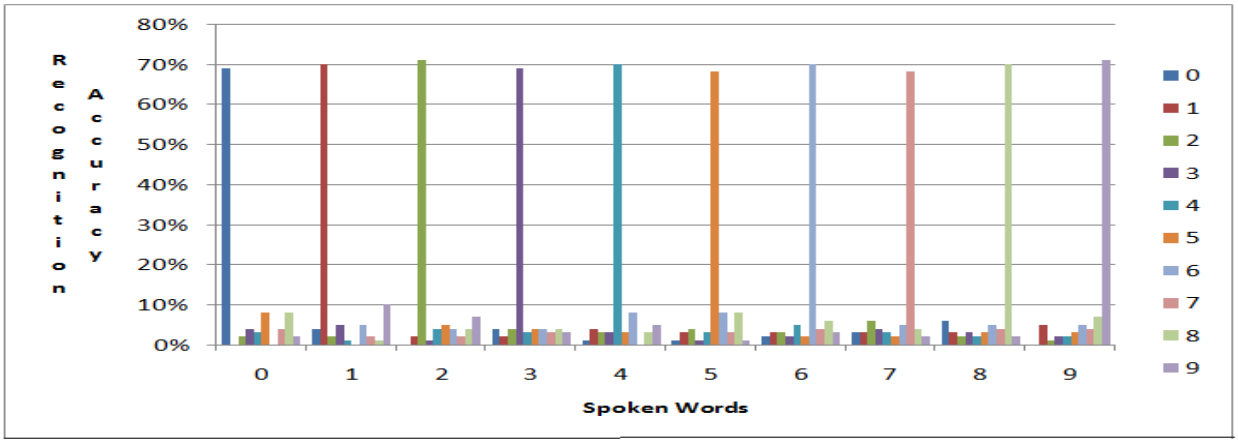

Figure18.

Accuracy Rate Graph of a Model trained by Excluding a Speaker and Tested by Excluded Speaker

The result shows that trained system gives an accuracy level of 68% as shown in Accuracy Rate Graph of Fig. 6.18

2) Train and Test Model by All Speakers: The final experiment contains the overall conclusion of the most optimized finding throughout the study; all 10 speakers are used for both the training and testing phase

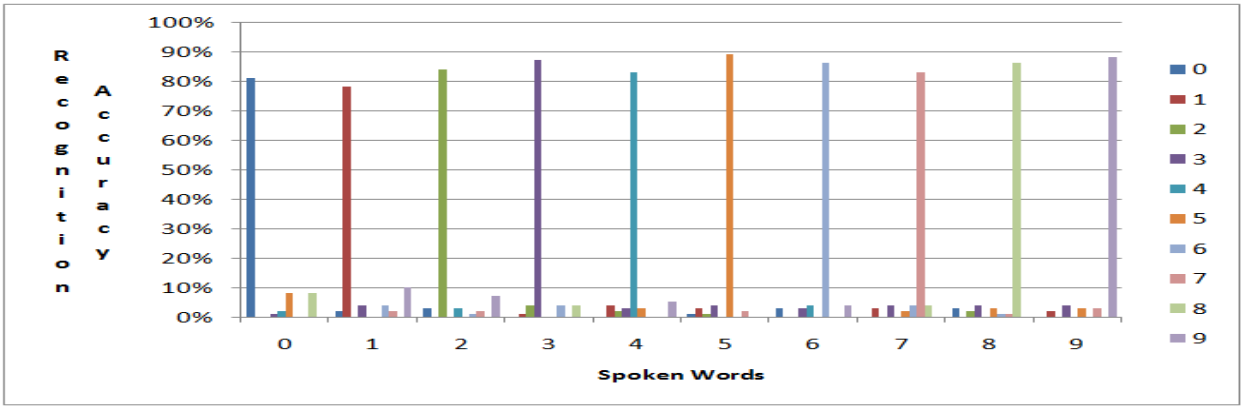

Figure19.
Accuracy Rate Graph of Optimized Recognition Model

Confusion matrix of Final Optimized ASR model having following characteristic, achieving overall 81% of accuracy after pre-processing phase and vocabulary size of 10 words.

CONCLUSION

This paper demonstrated that a neural network is a constructive base for a small vocabulary, noise-robust, speaker independent and isolated speech recognition system. The zero-crossing rate has been shown to provide a relatively good measure for distinguishing between voiced and unvoiced speech. Transformation and compression of speech signal is done by cropping which made the following process simpler. MFCC signal analysis technique proved to be simple and efficient to extract relevant features vector and compact the data with preserving the important information. Exhibits an originating isolated word recognition model that uses a feed forward neural network with a Back Propagation technique, BNN is an effective approach for small vocabulary ASR. The recognition rate is 89.5% in most cases for the 5-word vocabulary systems, and 81% for the 10-word system. Find the relationship between layers and also find optimized number of hidden neurons used for the best result. The performance of a network is directly proportional to the training set, means that performance increase when more data are provided.

FUTURE WORK

As examined in this study, several of the practical applications of these individual word-based speech recognition methods lie outside of the realm of speech recognition, including sorting undersea acoustic waves, improving the performance of missile guidance and tracking systems, and detecting targets using sonar technology. At present, it is impossible to determine a set of definitive procedures to teach these neural networks. A lot of regulation on training emanates from particular experimentation. The need for more flexible learning and testing approaches is necessary for improving recognition performance, and to accommodate larger vocabularies. Going forward, Predictive approach makes use of isolated and independent network for every word whereas integrated network is used all along the classification approach. For that reason, at any instant newer words can be brought and trained without putting impact of the rest of the system. Predictive approach presents possibility for parallelism whereas on introducing more classes to the classification network, the system is to be retrained in its entirety. HMM model is based on supposition, deals with temporal model and therefore restricts the effectiveness whereas NN is free from such suppositions, efficient NN-HMM hybrids model can be developing where temporal analysis deal by HMM and acoustic analysis by NN.

DECLARATIONS

Acknowledgement: We appreciate the generous support from all the contributor of research and their different affiliations.

Funding: No funding body in the public, private, or nonprofit sectors provided a particular grant for this research.

Availability of data and material: In the approach, the data sources for the variables are stated.

Authors' contributions: Each author participated equally to the creation of this work.

Conflicts of Interests: The authors declare no conflict of interest.

Consent to Participate: Yes

Consent for publication and Ethical approval: Because this study does not include human or animal data, ethical approval is not required for publication. All authors have given their consent.

REFERENCES

Abbasi, F. L., Ebrahim, M., Abro, A. A., Daniyal, S. M., & Younus, I. (2025). CNN-Based Face Detection Focusing on Diverse Visual Variations.

Abbasi, M. M., Daniyal, S. M., Abro, A. A., Hussain, D., Amjad, U., & Zahid, N. B. (2024). Applying Neural Networks to Predict Ventilator Demand: A Study of Pakistan's Healthcare Sector. *VFAST Transactions on Software Engineering*, *12*(3), 217-229.

Ahmed, H., Haque, M. F. U., Khan, H. R., Nadeem, G., Arshad, K., Assaleh, K., & Santos, P. C. (2024). Selecting the best compiler optimization by adopting natural language processing. IEEE Access.

Ali, A., Ajaz, S., & Daniyal, S. M. (2025). Optimized Artificial Neural Network-based Approach for Task Scheduling in Cloud Computing. *Annual Methodological Archive Research Review*, *3*(5), 22-34.

Al-Karawi, K. A. (2025). Convolutional neural network-based detection of audio replay attacks in speaker verification systems. *International Journal of Speech Technology*, *28*(1), 175-184.

An Introduction to Speech Recognition B. Plannerer March 28, 2005

C. Becchetti and L. P. Ricotti, Speech Recognition Theory and C++ Implementation, John Wiley & Sons, West Sussex, England, 1999.

D.E. Rumelhart; G.E. Hinton and R.J. Williams, Learning internal representations by error propagation, Rumelhart, D.E. et al. (eds.): Parallel distributed processing: Explorations in the microstructure of cognition (Cambridge MA.: MIT Press, 1986), 318-362.

Daniyal, S. M., Amjad, U., Khaliq, A., Zahid, N. B., Abbasi, F. L., & Hussain, S. M. T. (2024). Analyzing Student's Emotions in the Classroom: A Deep Learning Approach to Facial Expression Recognition. *International Journal of Artificial Intelligence & Mathematical Sciences*, *3*(1), 11-19.

Daniyal, S. M., Hussain, S. M. T., Abbasi, F. L., Hussain, D., Abbasi, M. M., & Amjad, U. (2024). A hybrid deep learning model for precise epilepsy detection and seizure prediction. *Spectrum of engineering sciences*, *2*(3), 62-77.

Dash, Y., Abraham, A., Gupta, S., Rathore,

G. S. Ying, C. D. Mitchell, L. H. Jamieson, "Endpoint detection of isolated utterances based on a modified Teager energy measurement," *Proceedings of 1993 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-93),* Vol. 2, pp. 732-735, April 1993.

H Qiang and Z. Youwei, "On prefiltering and endpoint detection of speech signal," *Proceedings of 1998 Fourth International Conference on Signal Processing, (ICSP '98),* Vol. 1, pp. 749-752, October 1998.

H. Bourlard and N. Morgan, Connectionist speech recognition: A hybrid approach., Kluwer Academic Publishers, Boston, USA, 1994.

http://documents.wolfram.com/applications/neuralnetworks/NeuralNetworkTheory/2.5.1.html

J. W. Picone, "Signal modeling techniques in speech recognition," *Proceedings of the IEEE,* Vol. 81, No. 9, pp. 1215-1247, September 1993.

L. Deng and D. O'Shaughnessy, *Speech Processing A Dynamic and Optimization-Oriented Approach,* Marcel Dekker, New York, 2003.

L. Rabiner and B. H. Juang, Fundamentals of speech recognition. Englewood Cliffs, NJ, 1993.

Masood, A., Daniyal, S. M., & Ibrahim, H. (2025). Enhancing Skin Cancer Detection: A Study on Feature Selection Methods for Image Classification.

Molau, S., Pitz, M., Schlüter, R. & Ney, H. (2001), Computing Mel Frequency Cepstral Coefficients on the Power Spectrum, IEEE International Conference on Acoustics, Speech and Signal Processing, Germany, 2001: 73-76.

Nadeem, G., Rehman, Y., Khaliq, A., Khalid, H., & Anis, M. I. (2023, March). Artificial Intelligence based prediction system for General Medicine. In 2023 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET) (pp. 1-6). IEEE.

Nikmah, A., Damayanti, A., & Winarko, E. Voice-Based Emotion Identification Based on Mel Frequency Cepstral Coefficient Feature Extraction Using Self-Organized Maps and Radial Basis Function.

S. V., & Patil, H. (2025). Enhancing Respiratory Monitoring by CNN Using Mel Frequency Cepstral Coefficients. In *International Conference on Soft Computing and Pattern Recognition* (pp. 572-580). Springer, Cham.

Tomar, S., & Koolagudi, S. G. (2025). Blended-emotional speech for Speaker Recognition by using the fusion of Mel-CQT spectrograms feature extraction. *Expert Systems with Applications*, *276*, 127184.

WolfFarm Research, "Neural Networks Documentation", from

Yousuf, W. B., Talha, U., Abro, A. A., Ahmad, S., Daniyal, S. M., Ahmad, N., & Ateya, A. A. (2024). Novel Prognostic Methods for System Degradation Using LSTM. *IEEE Access*.

Zhang, Y., Ma, L., & Li, Y. (2025). Fuzzy Speech Recognition Algorithm Based on Continuous Density Hidden Markov Model and Self Organizing Feature Map. *International Arab Journal of Information Technology (IAJIT)*, *22*(2).