



A DEEP NEURAL NETWORK(DNN) FOR THE RECOGNITION OF SPEECH DISFLUENCY

Vinay N.A,¹,Vidyasagar K.N,²,

Bharathi S.H³

^{1,2,3} School of ECE, REVA University,
Bengaluru

July 17, 2018

Abstract

Speech has a unique physical characteristics varying from person to person and in recent years speech biometric is used in many real-time applications like Robotics, Search Engines, Mobiles, Tablets etc. For such application speech recognition plays a major role, which can be implemented by Automatic Speech Recognition (ASR) systems. The ASR extracts the information from the speech signal by using feature extraction and classification techniques. The ASR process involves converting the speech signal into words, but this transcription will be affected if there is any disfluency in the speech. In this proposed work, the recognition of such disfluent speech using features extracted by applying Mel-Hilbert Frequency Cepstrum Coefficient and classification by Deep Neural Network (DNN). To initialize DNN, a forward propagation and backward propagation models are used, which is having a large number of stochastic gradients. In this proposed methodology, DNN is used for classification and feature extraction is done by endpoint detection of each phrase using Mel-Frequency Cepstrum Coefficient based on the Hilbert spectrum (MHCC). The experiment is carried out on speech samples taken from UCLASS database

and the maximum accuracy from the dataset by using the proposed method is around 96%.

Index words: MHFCC,DNN,DBN,Hilbert Transform (HT),Disfluency,ASR, forward propagation, backward propagation.

1 Introduction

The rapid development of speech recognition in almost all the fields like mobiles, tablets, security systems etc. has given more insight for research. In these systems, ASR will recognize the speech accurately, but if there is any disturbance in speech due to environmental noise, style of speaking, emotion, stuttering causes disfluency in the discourse. This work is focused on recognizing the endpoint detection of disfluent speech using Mel-Hilbert Frequency Cepstrum Coefficient (MHFCC) and Deep Neural Network (DNN).The disfluent samples vary in their characteristics in terms of pitch and time duration, it will be separated based on their acoustic features and language model. These acoustic features and language model helps the ASR to give better transcription. The auditory features define the respective part of phones using featured vectors obtained by using MHFCC from the raw speech signal. For a set of interpretations (I), to find the same sequence of arguments (A), the Baye's rule is given as:

$$P\left(\frac{A}{I}\right) \propto P\left(\frac{I}{A}\right).P(A) \quad (1)$$

Where $P\left(\frac{A}{I}\right)$ —Probability estimated by acoustic model and $P(A)$ -probability estimated by language model.

Once all the features are extracted, they are used to train DNN using Hidden- Markov Model (HMM). Then each trained HMM connected to form larger units of speech.

After training the acoustic and language model, decoding is applied to generate the transformations. The decoding sequence of A^* for a set of interpretations is given by:

$$A^* = \operatorname{argmax}(P \frac{I}{A}).P(A)) \quad (2)$$

2 METHODOLOGY

The flow of proposed methodology is as shown in the figure 1, where the acoustic features of speech sample is implemented using Hilbert transform based Mel-Frequency Cepstrum Coefficient and to classifying of voiced and unvoiced section implemented by using DNN.

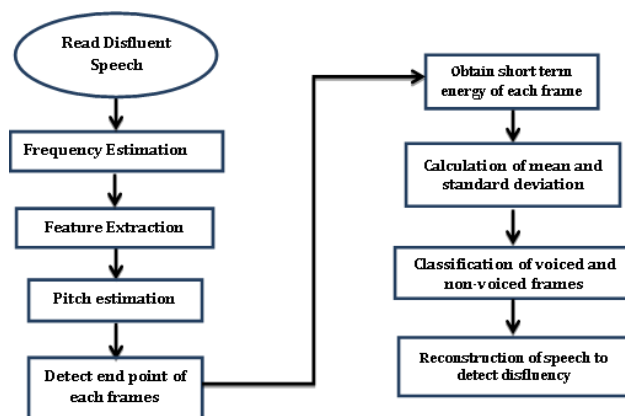


Figure 1: Proposed methodology flow chart

2.1 Mel Hilbert Frequency Cepstral Coefficients (MHFCCs)

In traditional MFCC technique, the Fourier transform considers the speech signal as a summation of predefined harmonics. But the disfluent speech is not having such predefined harmonics i.e, characteristics of disfluent speech are unknown. So, the analysis of such speech requires a robust and effective Time-Frequency Representation (TFR). To overcome this problem, the Hilbert transform is used.

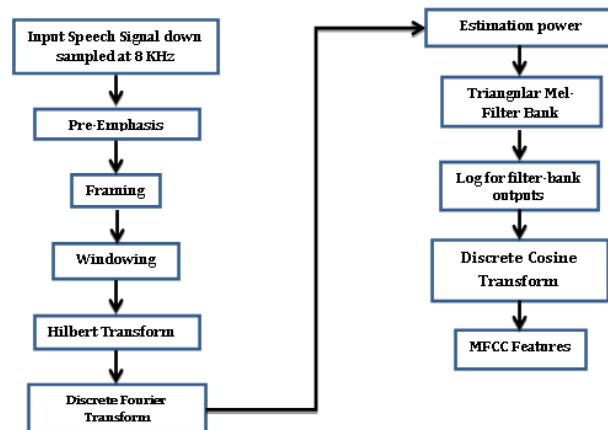


Figure 2: Steps to extract feature in MHFCC

The Hilbert Transform (HT) consists of two major parts: Empirical Mode Decomposition (EMD) and Hilbert Spectral (HS) Analysis. In EMD, any complex data can be decayed into fixed and smaller sum of Intrinsic Mode Functions (IMF). This IMF is well-defined as a function having the same number of zero crossings and extrema, also having symmetric envelopes defined by local maxima and minima. Since EMD based on local characteristic time scale data of non-stationary (speech) signals along with Hilbert transform results in energy-frequency-time distribution designated as Hilbert Spectrum (HS). The steps involved in feature extraction process using MHFCC are as shown in Figure 2.

2.2 Pre-emphasis:

The frequency of input speech sample is 41 KHz, but most of the speech parameters are not in this high frequency range, so the input speech frequency is scaled down to 8 KHz which can be used for further processing. Therefore by a passing, this high-frequency speech signal through a low pass filter, the higher frequency components and some DC distortion components are removed. The transfer function of such filter is given by [19].

$$I'(n) = i(n) - ai(n-1) \quad (3)$$

Where $I'(n)$ -output of pre-emphasis $i(n)$ -Input speech signal sampled at 8 KHz.

By taking Z- transform on both sides:

$$\frac{I'(Z)}{i(Z)} = H(Z) = 1 - \alpha Z^{-1} \quad (4)$$

The pre-emphasized signal is divided into frames of 5ms and each frame is having 256 samples.

2.3 Windowing:

In order to give continuity between first and last frames, each frame is convolved with a window function. To get low-side band attenuation hamming window is used [8].

$$Y(n) = v(n) * w(n) \quad (5)$$

Where Y(n)-Window output
v(n)-Frames of raw speech signal
w(n)-Window function

For hamming window: Window function is $w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right)$

Where N-Number of samples=256 and $0 < n \leq N-1$

2.4 Hilbert Transform:

The Hilbert transform based Mel-Frequency Cepstrum Coefficient is used for spectrum feature extraction process. It is an improved kind of the traditional MFCC based on spectral estimation. Extracting Hilbert spectrum envelope, can get instantaneous amplitude, instantaneous frequency and instantaneous phase, there by both the discrimination and anti-noise of MHCC are good. MHCC front-end feature extraction process is as shown in Figure 2.

Let y(n) be a windowed speech signal, the Hilbert transform of it is given by, [24]:

$$F[y(n)] = F[y(n)] * F\left[\frac{1}{\pi n}\right]$$

$$F[y(n)] = Y(\Omega), F[y(n)] = Y(\Omega), F\left[\frac{1}{\pi n}\right] = j \operatorname{sgn}(-w) \quad (6)$$

But Signum function is an odd function[24].

$$j \operatorname{sgn}(-w) = -j \operatorname{sgn}(w)$$

$$Y(\Omega) = Y(\Omega) * -j \operatorname{sgn}(w) \quad (7)$$

But $\operatorname{sgn}(w)$ is defined as:

$$\operatorname{sgn}(\Omega) = \begin{cases} 1, \Omega > 0 \\ -1, \Omega < 0 \end{cases} \quad (8)$$

$$Y(\Omega) = \begin{cases} -jY(\Omega)\Omega > 0 \\ jY(\Omega)\Omega < 0 \end{cases} \quad (9)$$

Similarly for discrete time signal Hilbert transform is given by [24]:

$$Y(\Omega) = S(\Omega) * Y(\Omega) \quad (10)$$

$$\text{But} \quad (11)$$

When N is odd:

$$S(\Omega) = \operatorname{sgn}(\Omega) \bullet \operatorname{sgn}\left(\frac{N}{2} - 2\right) S(\Omega) = \begin{cases} 1, \Omega = 1, \dots, \frac{N-1}{2} \\ 0, \Omega = 0 \\ -1, \Omega = \frac{N+1}{2}, \dots, N-1 \end{cases} \quad (12)$$

When N is even

$$S(\Omega) = \begin{cases} 1, \Omega = 1, \dots, \frac{N}{2} - 1 \\ 0, \Omega = 0, \dots, \frac{N}{2} \\ -1, \Omega = \frac{N}{2} + 1, \dots, N-1 \end{cases} \quad (13)$$

Thus Discrete Hilbert Transform is given by [24]:

$$Y(\Omega) = \begin{cases} -jY(\Omega)\Omega = 1, \dots, \frac{N}{2} - 1 \\ jY(\Omega)\Omega = \frac{N}{2} + 1, \dots, N-1 \end{cases} \quad (14)$$

For the recognition of disfluent speech, end of each frame should be detected. The end point of each frame is detected by Hilbert Transform specifically through Empirical Mode Decomposition (EMD). The principle operation in EMD is decaysignal $y(n)$ into a sum of bandlimited functions, $g_b(n)$, called intrinsic mode functions

(IMFs). Each IMF must satisfy two properties: (i) the number of extrema and the number of zero crossings are either equal or differ by one; (ii) the mean value of the envelope defined by the local maxima and the envelope defined by the local minima is zero. The Hilbert transform of a speech sample is as shown in the figure 8 (a) and figure 8(b) shows the endpoint detection using IMF.

2.5 Discrete Fourier Transform (DFT):

Spectral data implies the vitality levels at various frequencies in the given window. Time space information is changed over into frequency space to acquire the spectral data using DFT. The Discrete Fourier Transform (FFT) is given by [8]:

$$Y(K) = \sum_{k=0}^{N-1} X(k) e^{-j \frac{2\pi k i}{N}}; 0 \leq k \leq N-1 \dots \dots \dots (6)$$

Where X(K)-output of Hilbert Transform

To get frequency spectrum of speech signal 256 point FFT is applied. DFT comprises of complex numbers i.e. real and imaginary parts. Discourse acknowledgement frameworks manage real information. Subsequently, complex esteem is constantly disregarded [1]. In the result we expect the real and imaginary parts of V(k) as Re(V(k)) and Im(V(k)) respectively, at that point the phantom size of the speech standard can be obtained by utilizing condition given in equation (4). Inappropriate sizes of each frame are put away as lines in one single framework with a number of lines equivalent to the number of edges and number of segments equivalent to 256, which is additionally equivalent to the frame width [6].

$$|V(K)| = \sqrt{\text{Re}(V(k))^2 + \text{Im}(V(k))^2} \dots \dots \dots (7)$$

For the real part of DFT signal, power spectrum is determined by taking square of modulus value [6].

$$PS = |V(K)| \cdot |V(K)| \dots \dots \dots (8)$$

2.6 Mel-Scale Smoothing

In this step, the spectrum is rescaled to the range of mel-scale. This will be achieved by multiplying the power spectrum values with a set triangular band pass filters. The positions of these filters are equally spaced, as [6]:

$$B_m(k) = \begin{cases} 0 & k < f(l-1) \\ \frac{k-f(l-1)}{f(1)-f(l-1)} & f(l-1) \leq k \leq f(1) \text{(9)} \\ 0 & k > f(l+1) \end{cases}$$

Where $f(1)$ -center frequency, l - Number of filters, $k=0,1,\dots,\frac{N}{2}-1$

Each filter in filter bank as to satisfy: $\sum_{m=0}^{N-1} B_l(k) = 1$

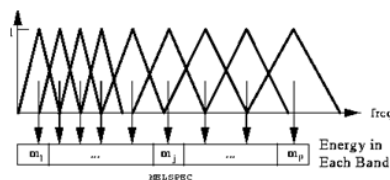


Figure 2: Triangular Band Pass Filter Bank

For each frame power spectrum values with triangular filter bank, Log has been taken to further reduce the noise and to estimate the errors as[5]. The log magnitude response of the speech is as shown in the figure 6(c):

$$L(m) = \ln(\sum_{k=0}^{N-1} |PS|^2 B_m(k)) \text{(10)}$$

Mel frequency is calculated as[6]

$$f_{mel} = 2595 \log_{10}(\frac{f}{1000} + 1) \text{(11)}$$

If 32 DFT coefficients are considered for logarithmic operation, it segmented to 6 bins. If mel-scale is between 0-6000 mel, then these 6 bins on a linear scale will be in the range of 0-1000, 1000-2000, 2000-3000, 3000-4000, 4000-5000, 5000-6000 mel. These 6 bins are converted to Hz to find out the starting and ending frequency bin. In this process, the lower frequency components become more important as the frequency bins at higher frequency become larger.

2.7 Discrete Cosine Transform (DCT):

A discrete cosine change (DCT) interconnects a limited arrangement of dataemphasesabout an sum of cosine dimensions wavering at various regularities. DCT is a Fourier-related change like the discrete Fourier change (DFT), yet utilizing just real numbers. The DCTs are for the majorportionrecognized with Fourier transform coefficients of an intermittently and evenly broadened succession while DFTs are identified with Fourier transform coefficients of an occasionally expanded grouping. DCTs are proportional to DFTs of generally twice the length, employed on a real data with even symmetry (since the Fourier change of a real and even size is real and smooth), while in a insufficientdissimilarities the information or potentially yield information are moved considerably a specimen. Its given by[9]:

$$D_i(m) = \sum_{n=0}^{M-1} L(n) \cos[\pi m(\frac{n-0.5}{M})] \dots\dots\dots (12)$$

Where $L(m)$ -logarithmic value , M -Number of filters and $0 < m \leq M - 1$

Discrete Cosine Transformdiscards the higher coefficients, the lower coefficients gives the smooth spectral shape.Nourishing the MFCCs as features to a machine learning algorithm, these lower-order coefficients will make good features, since they signifies some simple aspects of the spectral shape, while the higher-order coefficients are discarded are more noise-like and are not important to train on. Additionally, training on the Mel spectrum magnitudes themselves would probably not be as good because the particular amplitude at different frequencies are less important than the general shape of the spectrum.

Finally mean values of MFCC vectors were considered as feature component for each frame and MFCC vectors are constructed (column wise) for each frame.

2.8 Zero Crossing Rate:

It determines how many times a signal crosses zero amplitude in a frame, given by:

$$ZCR = \frac{1}{N} \sum_{n=0}^{N-1} \text{sgn}[s(n) - \text{sgn}[s(n-1)]] \dots \dots \dots (13a)$$

Short term energy of the signal is given by:

$$E_{dB} = 20 \log 10 \left(\frac{1}{N} \sum_{n=1}^N |s(n)|^2 \right) \dots \dots \dots (13b)$$

The energy in dB of a speech signal is as shown in the figure 7 and ZCR of disfluent speech is as shown in the figure 9.

3 Deep-Neural Network (DNN):

A DNN is a feed-forward neural network, which allows the signal to flow only in one direction in the network. It has three important layers namely: input layer, hidden layer and output layer. Let us consider a neural network consisting of input layer having input vector $X = (x_1, x_2, \dots, x_Q)$, hidden layer of having N neurons $H = (h_1, h_2, \dots, h_N)$ and output layer $Y = (y_1, y_2, \dots, y_Q)$. The structure of DNN is as shown in the Figure 3:[13].

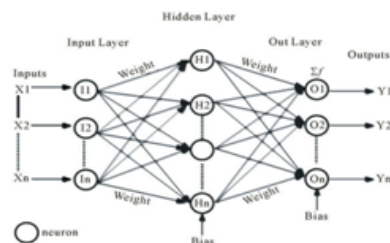


Figure 3: Structure of DNN

The output of an arbitrary neuron on hidden layer is [13]:

$$h_i = f(u_i) = f\left(\sum_{i=1}^Q W_{ki} X_i\right) \dots \dots \dots (14)$$

Where

W_{ki} -weight at k th element in $Q \times N$ weight matrix.

X_i -Input vector of i th element

The output of an arbitrary output neuron is [13]:

$$y_j = f(u_j) = f\left(\sum_{i=1}^Q W_{ij} h_i\right) \quad (15)$$

Where W_{ij} is the weight at i,j th element in $N \times M$ weight matrix h_i is the Hidden layer vector of i^{th} element. In forward propagation, the output vector is determined for a given input vector as:

1. The input layer is trained by input vectors (MFCC vectors).
2. The output of first is find out by weights, first hidden layer and bias value as shown in the structure of DNN in Figure 3.

The above steps are repeated for the next upcoming layers using nodes which are determined in the previous layers. The values for each layer are calculated as follows[13]:

$$X^{i+1} = f(w^i x^i + b^{i+1}) \quad (16)$$

Where x^i -input vector at i^{th} layer w^i -Matrix of weights between i & $i+1$ layer b^i -bases vector of i^{th} layer

In back propagation, the DNN must accomplish the required input vectors to determine the specified output values. These processes consist discovering the suitable weights and bias values to reduce the loss function in the network. The loss function is defined by[13]:

$$l(x) = \|y_j - f(x)\|^2 \quad (17)$$

Where $f(x)$ is a sigmoid function given by:

$$f(x) = \frac{1}{(1 + e^{-x})} \quad (18)$$

The procedure for back propagation as follows:

1. Weights as bases are declared for a small random variables.
2. Use the forward propagation to pass the input vectors through the network.

3. Calculate the faults at the output level as: $E = \frac{1}{2} \sum_{j=1}^N (y_j - l_j)$ where l_j -length of sequence
4. Chain rule is used compute the slopes of each errors with respect to weights and base values to send the errors through the network.
5. Apprise the weights and biases using the calculated slopes , ascended by learning rate to control the size of the update.
6. Recap steps (ii)-(v) until all the vectors are fed to network [13]. Chain Rule:

To determine the gradient w.r.t W_{ij} , first compute $\frac{\partial E}{\partial W_{ij}}$ as [13]

$$\frac{\partial E}{\partial W_{ij}} = \frac{\partial E}{\partial y_j} \cdot \frac{\partial y_j}{\partial u_j} \cdot \frac{\partial u_j}{\partial W_{ij}} \quad (19)$$

Error $E = \frac{1}{2} \sum_{j=1}^N (y_j - l_j)$ where l_j -length of sequence

Derivative of E with respect to y_j :

$$\frac{\partial E}{\partial y_j} = y_j - l_j \quad (20)$$

$$\text{Derivative of } y_j \text{ with respect to } u_j : \frac{\partial y_j}{\partial u_j} = y_j(1 - y_j) \quad (21)$$

$$\text{Derivative of } u_j \text{ with respect to } W_{ji} : \frac{\partial u_j}{\partial W_{ji}} = h_i \quad (22)$$

Then the new gradient is given by:

$$G^{new} = G^{old} - n(y_j - l_j)y_j(1 - y_j)h_i \quad (23)$$

The full gradient is given by:

$$\frac{\partial E}{\partial W_{ij}} = \sum_{j=1}^M \left(\frac{\partial E}{\partial y_j} \cdot \frac{\partial y_j}{\partial u_j} \cdot \frac{\partial u_j}{\partial W_{ij}} \right) \frac{\partial h_i}{\partial u_i} \cdot \frac{\partial u_i}{\partial W_{ij}} \quad (24)$$

3.1 Signal to Noise Ratio (SNR) :

The proposed algorithm results are measured by both Signal-Noise Ratio (SNR) and Segmental SNR (SSNR) of the enhanced waveforms indicates that the proposed method gives better performance compared to traditional method.

The SNR and SSNR of a speech is measured by:

$$SNR = 10 \log_{10} \left[\frac{\sum_n x^2(n)}{\sum [x(n) - \hat{x}(n)]^2} \right] \dots\dots\dots (25a)$$

$$SSNR = \frac{1}{M} \sum_{j=0}^{M-1} 10 \log_{10} \left[\frac{\sum_{n=Nj+1}^{N(j+1)} x^2(n)}{\sum_{n=Nj+1}^{N(j+1)} [x(n) - \hat{x}(n)]^2} \right] \dots\dots\dots (25b)$$

Where M-Number of frames

N-Length of each frames

x(n)- Original Signal

-Enhanced signal

4 Results and Discussion

The experiments are carried on speech samples taken from UCLASS database; it consists of 12 female speech samples aged between 09 to 12 years and 12 male speech samples aged from 08-28 years. Speech signal is estimated through a simple energy based approach to identify silent and low energy regions, detecting action can be seen in the figure 5.

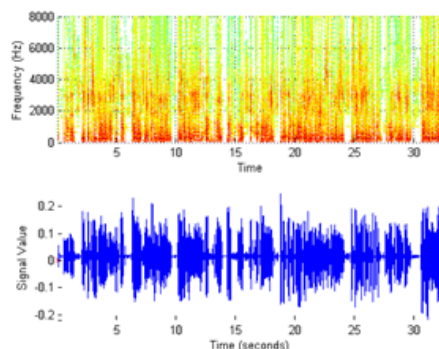


Figure 5: Detection for extracting voiced and unvoiced speech segments.

The standard deviation of fundamental frequency (F0) variations of the extracted voiced segments from speech sample is done by using the pitch of filler, or a filled pause, which is stable and lies around the median of the speakers. A small standard deviation represents a flat pitch contour that is indicative of a filled pause. Calculated the pitch and distance between the logarithms of the linear magnitude spectrum frequency and log magnitude spectrum frequency, spectra of each pair of the voiced sections around a long pause

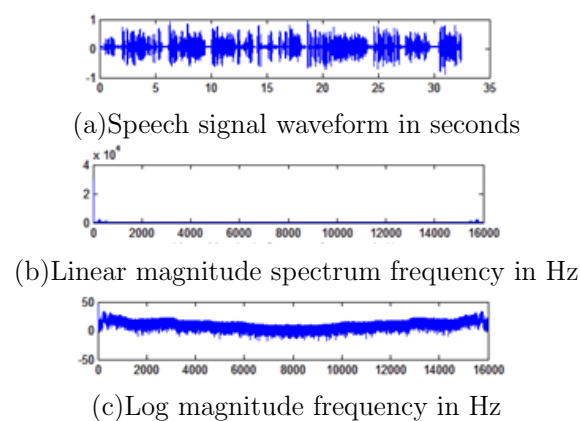


Figure 6: Frequency estimation is obtained

Hence, figure 6(a) shows the speech signal waveform. figure 6(b) shows the Linear Magnitude Spectrum frequency of the speech signal and figure 6(c) shows Log Magnitude Spectrum frequency, which is calculated as per equation 19.

A technique for silence elimination and separation of speech signals, is based on a simple audio feature: signal energy. This simple feature can be used for detecting silent periods in acoustic signals a set of particular features are selected for sample frames, where the energy of voiced regions is larger than the energy of the silent regions. The unvoiced regions simply contain ecological noises, and then the spectral energy for the voiced segments is again larger,

since these noisy sounds tend to have lower frequencies and therefore the spectral energy values are lower. The spectral energy of each frame is calculated as given in equation 22b.

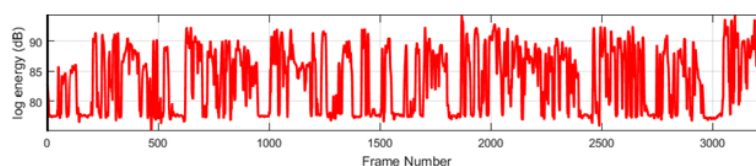


Figure 7: Spectral energy estimation of sound files

The endpoint of each frames helps to analyze the speech signal to detect disfluent part in each phrase, in this proposed work endpoint each frames is calculated by using Hilbert transform from equation 8 to equation 14. Figure 8(a) shows the endpoint of first frame which is from 0 to 5ms and figure 8(b) shows the variations of characters in first frame.

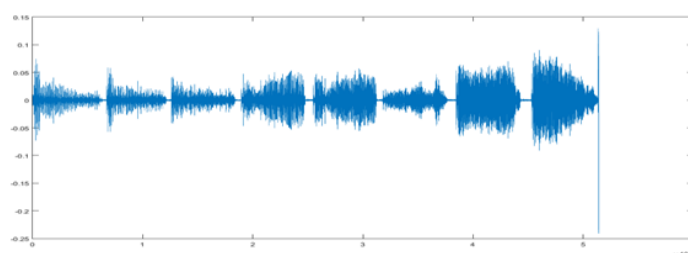


Figure 8(a): End point detection of speech

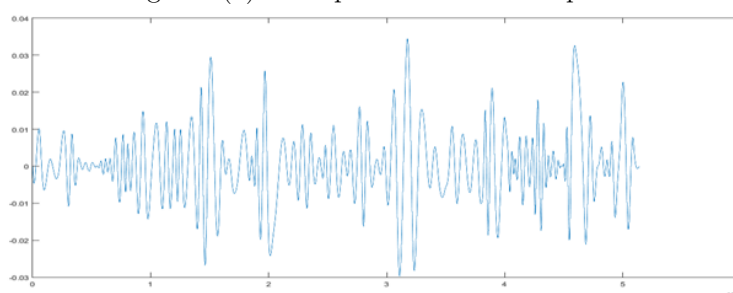


Figure 8(b): End point detection by IMF

In order to detect disfluent part effectively, Zero Crossing Rate is used and its calculated as per equation 22(a). In figure 9, the

ZCR line is crossed four times by speech signal, which indicates the presence of disfluent part in speech. The result of ZCR validate the disfluency.

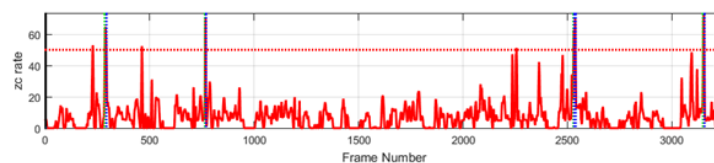


Figure 9: ZCR of speech

Between 0 to 500 frames the first prolonged word the legend is detected is shown in the figure 9(a) and (b) shows the segmented part of that word. In which X-axis represent time in ms and Y axis represent the amplitude.

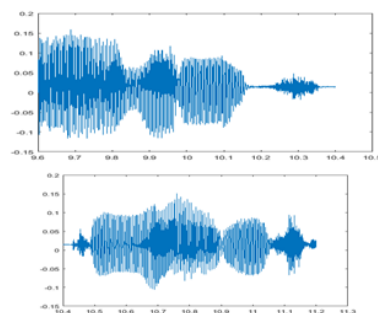


Figure 10 (a): Segmentation of detected disfluent word the legend.

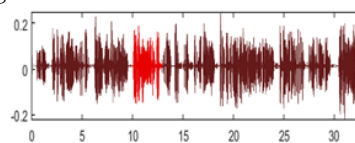


Figure 9(a): Detection first disfluency word.

Between 500 to 1000 frames the second repeted word one is detected is shown in the figure 10(a) and (b) shows the segmented part of that word. In which X-axis represent time in ms and Y axis represent the amplitude.

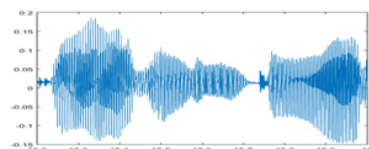


Figure 10 (a): Segmentation of word detected shown in the figure 10(b).

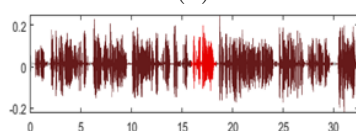


Figure 10(b) : Detection second disfluency word.

The SNR and SSNR of detected speech sample is obtained by using the equations (33) and (34) respectively. The values are:

Table 1: SNR and SSNR values for different female speakers

Age of the speaker	SNR in dB	SSNR in dB
9(30sec)	30.04	32.56
10(25sec)	28.25	29.95
11(20sec)	17.9	11.22
12(25sec)	22.01	28.5
13(35sec)	24.32	25.34
14(40sec)	22.25	30.8

In DNN, the distribution of weights and bias points for 32 MFCC feature vector is as shown figure 8.

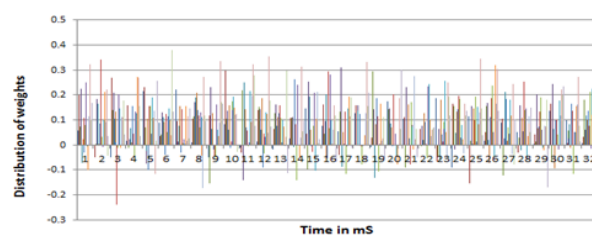


Figure 8: Distribution of weights

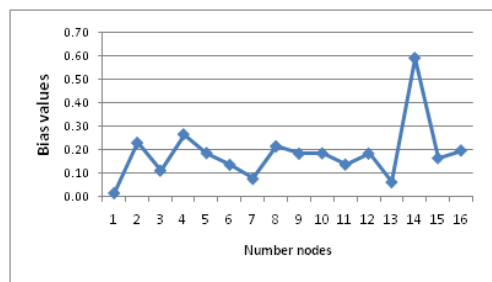


Figure 9: Distribution bias points

The chart shown in the figure , indicates percentage of detection of disfluency in female speakers aged between 9-14 years.

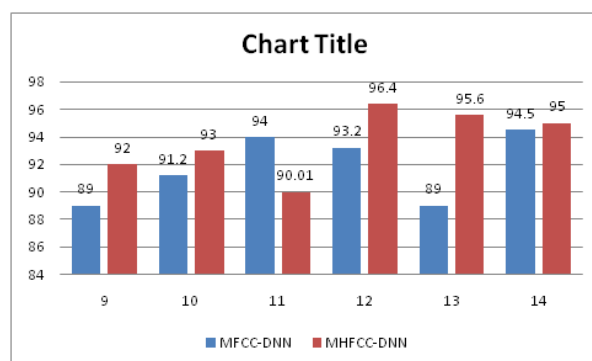


Figure 7: Percentage detection of disfluency words in female speech samples by using MFCC and MHFCC.

5 Conclusion

The recognition of disfluency is complex because it will hamper the transcription of Automatic Speech Recognition process, in this paper recognition of such disfluency is achieved with a accuracy of 96% using MHFCC and Deep Neural Network. Future this algorithm is used train to detect other disfluency problem like blockage and also it can be tested for different samples.

References

- [1] S.M. Thampi et al. (eds.), Tamil Speech Emotion Recognition Using Deep Belief Network(DBN)Advances in Signal Processing and Intelligent Recognition Systems, Advances in Intelligent Systems and Computing 678, DOI 10.1007/978-3-319-67934-1 29, Springer International Publishing AG 2018.
- [2] J. Cao et al., Excavation equipment classification based on improved MFCC features and ELM, Elsevier, Neurocomputing (2017), <http://dx.doi.org/10.1016/j.neucom.2016.03.113>.
- [3] Niko Moritz, Birger Kollmeier, Jörn Anemüller, "Integration of Optimized Modulation Filter Sets Into Deep Neural Networks for Automatic Speech Recognition", Audio Speech and Language Processing IEEE/ACM Transactions on, vol. 24, pp. 2439-2452, 2016, ISSN 2329-9290.
- [4] Sunday Iliya and Ferrante Neri, Towards Artificial Speech Therapy: A Neural System for Impaired Speech Segmentation, International Journal of Neural Systems, Vol. 26, No. 6 (2016) 1650023, World Scientific Publishing Company, DOI: 0.1142/S0129065716500234.
- [5] Jing Su, Qing Liu, Meilin Wang, Jiangzhong Cao, Wing-Kuen Ling, "Design of convolution neural network with frequency selectivity for wearable camera embed glasses based image recognition systems via nonconvex functional inequality constrained sparse optimization approach", Industrial Electronics (ISIE) 2016 IEEE 25th International Symposium on, pp. 1090-1093, 2016, ISSN 2163-5145.
- [6] Mishra, S., et al., Hindi vowel classification using QCN-MFCC features. Perspectives in Science Direct (2016), <http://dx.doi.org/10.1016/j.pisc.2016.01.010>
- [7] Hardik B. Sailor, Hemant A. Patil, "Filterbank learning using Convolutional Restricted Boltzmann Machine for speech recognition", Acoustics Speech and Signal Processing (ICASSP) 2016 IEEE International Conference on, pp. 5895-5899, 2016, ISSN 2379-190X.

- [8] Takumi Kobayashi, Jiaying Ye, "Discriminatively learned filter bank for acoustic features", Acoustics Speech and Signal Processing (ICASSP) 2016 IEEE International Conference on, pp. 649-653, 2016, ISSN 2379-190X.
- [9] D. G. Bhalke , C. B. Rama Rao, & D. S. Bormane, Automatic musical instrument classification using fractional fourier transform based- MFCC features and counter propagation neural network, J Intell Inf Syst- Springer Science+Business Media New York 2015 , DOI 10.1007/s10844-015-0360-9.
- [10] D. Yu and L. Deng, Automatic Speech Recognition using Deep Neural Network-Hidden Markov Model Hybrid Systems, Signals and Communication Technology, DOI 10.1007/978-1-4471-5779-3_6. Springer Chapter 6 -Verlag London 2015.
- [11] D. Yu and L. Deng, Deep Neural Network Sequence-Discriminative Training , Signals and Communication Technology, DOI 10.1007/978-1-4471-5779-3_6. Springer Chapter 8 -Verlag London 2015.
- [12] Adam Jacksa and Katarina L. Haleya, Auditory masking effects on speech fluency in apraxia of speech and aphasia: comparison to altered auditory feedback, Journal of Speech, Language, and Hearing Research Vol. 58 16701686 December 2015
- [13] Jessica Ray, Brian Thompson, Wade Shen, MIT Lincoln Laboratory, Comparing a High and Low-Level Deep Neural Network Implementation for Automatic Speech Recognition, 2014 First Workshop for High Performance Technical Computing in Dynamic Languages, IEEE Computer Society,. DOI 10.1109/HPTCDL.2014.12.
- [14] Juan Rafael Orozco-Arroyave, Elkyn Alexander Belalcazar-Bolanos, Julian David Arias-Londono,, Characterization Methods for the Detection of Multiple Voice Disorders: Neurological, Functional, and Laryngeal Diseases, IEEE Journal Of Biomedical And Health Informatics, Vol. 19, No. 6, November 2015.

- [15] Leimin Tian, Johanna D. Moore, Catherine Lai, Emotion Recognition in Spontaneous and Acted Dialogues, 2015 International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE computer society.
- [16] Dong Yu, Senior Member, IEEE, LiDeng, Fellow, IEEE, and Frank Seide, Member, IEEE, The Deep Tensor Neural Network With Applications to Large Vocabulary Speech Recognition, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 21, No. 2, February 2013.
- [17] Yulia Tsvetkov, Zaid Sheikh, Florian Metze, Identification And Modeling Of Word Fragments In Spontaneous Speech, 978-1-4799-0356-6/13/31.00 2013 IEEE International Conference on Acoustics, Speech, and Signal processing .
- [18] T. Lakshmi Priya, N.R Raajan, Preethi P , Speech and Non Speech Classification using kNN algorithm, International Conference on Modelling Optimization and Computing Science Direct- Elsevier, doi: 10.1016/j.proeng.2012.06.120
- [19] Anatoly D. Goldstein and Robert E. Hillman, Integration, Reuse and Sharing of Data on Voice Disorders, IEEE IRI 2012, August 8-10, 2012, Las Vegas, Nevada, USA 407, 978-1-4673-2284-3/12/31.00 2012 IEEE.
- [20] Ooi Chia Ai, M. Hariharan , Sazali Yaacob, Lim Sin Chee, Classification of speech dysfluencies with MFCC and LPCC features, Expert Systems with Applications- Elsevier- 2011 doi:10.1016/j.eswa.2011.07.065
- [21] M. Hariharan , Lim Sin Chee , Ooi Chia Ai & Sazali Yaacob, Classification of Speech Dysfluencies Using LPC Based Parameterization Techniques J Med Syst (2012) 36:18211830 DOI 10.1007/s10916-010-9641-6.
- [22] Wen Wang, Gokhan Tur, Jing Zheng, Necip Fazil Ayan, Speech Technology and Research Laboratory, SRI International, Menlo Park, CA, USA, Automatic Disfluency Removal For Improving Spoken Language Translation, 978-1-4244-4296-6/10/25.00 2010 IEEE 5214 ICASSP 2010.

- [23] Victoria Tumanova a,* , Patricia M. Zebrowski a,1, Rebecca N. Throneburg b,2,Mavis E. Kulak Kayikci , Articulation rate and its relationship to disfluency type, duration, and temperament in preschool children who stutter, *Journal of Communication Disorders*-2010 Elsevier Inc. doi:10.1016/j.jcomdis.2010.09.001.
- [24] Zeinab Mahmoudi, Saeed Rahate, Mohamad Mahdi Ghaseme, Classification Of Voice Disorders In Children With Cochlear Implantation And Hearing Aid Using Multiple Classifier Fusion, 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010).
- [25] Gheorge Todoran, Rodica Holonec and Ciprian Iakab, Discrete Hilbert Transform.Numeric Algorithms, Vol.49, Nov.2008, Mediamira Science Publisher .
- [26] Serdar Yildirim and Shrikanth Narayanan, Senior Member, IEEE, Automatic Detection of Disfluency Boundaries in Spontaneous Speech of Children Using AudioVisual nformation, *IEEE Transactions On Audio, Speech, And Language Processing*, Vol. 17, No. 1, January 2009.
- [27] Lim Sin Chee, Ooi Chia Ai, Sazali Yaacob, Overview of Automatic Stuttering Recognition System, *Proceedings of the International Conference on Man-Machine Systems (ICoMMS)*, 11-13 October 2009, Batu Ferringhi, Penang, MALAYSIA.
- [28] Ronny A. Boey a,* , Paul H. Van de Heyning b, Floris L. Wuyts b, Louis Heylen b, Reinhard Stoop c, Marc S. De Bodt , Awareness and reactions of young stuttering children aged 27 years old towards their speech disfluency, *Journal of Communication Disorders* 42 (2009) 334-346- 2009 Elsevier Inc. Science Direct doi:10.1016/j.jcomdis.2009.03.002.
- [29] Lotfi Salhi Talbi Mourad Adnne Cherif Signal Processing Laboratory - Sciences Faculty of Tunis - Tunis ElManar Tunisia, Voice Disorders Classification Using Multilayer Neural Network, 978-1-4244-2628-7/082008 IEEE International Conference on Signals, Circuits and Systems, 2008.

- [30] Jui-Feng Yeh and Chung-Hsien Wu, Senior Member, IEEE, Edit Disfluency Detection and Correction Using a Cleanup Language Model and an Alignment Model, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 14, No. 5, September 2006.
- [31] Matthew Lease, Student Member, IEEE, Mark Johnson, and Eugene Charniak, Recognizing Disfluencies in Conversational Speech, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 14, No. 5, September 2006.
- [32] Chung-Hsien Wu, Senior Member, IEEE, and Gwo-Lang Yan, Speech Act Modeling and Verification of Spontaneous Speech-With Disfluency in a Spoken Dialogue System, IEEE Transactions On Speech And Audio Processing, Vol. 13, No. 3, May 2005
- [33] Dagen Wang, Shrikanth S. Narayanan, A Multi-Pass Linear Fold Algorithm For Sentence Boundary Detection Using Prosodic Cues ,0-7803-8484-9/04 ICASSP-2004,IEEE.
- [34] MatthiasHonal, and Tanja Schultz, Automatic disfluency removal on recognized spontaneous speech -Rapid adaptation to speaker-dependent disfluencies, 0-7803-8874-7/05 IEEE-ICASSP 2005.
- [35] Che-Kuang Lin, Student Member, IEEE, and Lin-Shan Lee, Fellow, IEEE, Improved Features and Models for DetectingEdit Disfluencies in Transcribing Spontaneous Mandarin Speech, IEEE Transactions On Audio, Speech, And Language Processing, Vol. 17, No. 7, September 2009.T. N.
- [36] Sainath, B. Kingsbury, A. r. Mohamed and B. Ramabhadran, "Learning filter banks within a deep neural network framework," 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, 2013, pp. 297-302. doi: 10.1109/ASRU.2013.6707746

