NEURAL NETWORK BASED VOICED AND UNVOICED CLASSIFICATION USING EGG AND MFCC FEATURE

S.Bagavathi[1], S.I.Padma[2]

[1]*PG Scholar, Communication system*
PET Engineering College, Vallioor

[2] *Assistant Professor*
PET Engineering College, Vallioor

---***---

Abstract - *Speech recognition is a subjective phenomenon. This procedure still faces a considerable measure of issue. Different techniques are utilized for various purposes. In this work of project, it is shown that how the speech signals is perceived utilizing Fuzzy c-implies (FCM) Clustering Method in neural system. Voices of various people of different ages in a quiet and noise free condition by a good quality receiver are recorded. Same sentence of term 10-12 seconds is talked by these people. These talked sentences are then changed over into wave positions. At that point components of the recorded examples are removed via preparing these signals utilizing LPC. These systems are prepared to perform by the pattern recognition. Their significance to the order and portrayal of composed content is low; be that as it may, most viable speech recognition frameworks depend intensely on speech recognition to accomplish elite. For the vowel classification utilizing adaptive median filter with combination of speech and EGG information. The Mel-Frequency Cepstral Coefficients (MFCC) and Neural Network (NN) are utilized as components representing to the speech signal.*

Key Words: Adaptive median filter, FCM Clustering Method, MFCC, Neural Network, speech recognition, training algorithm.

1. INTRODUCTION

Speech recognition systems particularly to test accuracy of speech signal. Emotional condition of the speaker can cause the difference in the pronunciation of different persons. Surroundings can add noise to the signal. Some of the time speaker causes the expansion of noise itself. In speech recognition process, speech signal caught by microphone or telephone is changed over to an arrangement of characters. Hence for the interaction with machines human could utilize speech as a valuable interface. Human dependably need to accomplish characteristic, possessive and synchronous figuring the execution of Speech Recognition. Work is well done as a powerful classifier for vowel sounds with stationary spectra by those systems. Encourage forward multi-layer neural system are not ready to manage time differing data like time-varying spectra of speech sounds. To recognize the sound is short in time and low in energy. The two main steps that will produce such accurate results are either the feature extraction for the classification of speech data. To improved the method of classifier or feature this is not the objective of this paper. The difference of this paper from others is that we attempt to utilize an effectively created strategy as the classification scheme and the Mel-frequency Cepstral coefficients (MFCCs) as the speech feature. For their effectiveness, speech recognition technologies still need more work for individuals having speech signal. As the vowels are effectively and dependably recognized therefore they are utilized to recognize speech by describe the sounds of human speech.

2. SPEECH RECOGNITION OF COMBINING SPEECH AND EGG DATA

Generally, almost all speech recognition (SR) system consist the following steps: Signal pre-processing, feature extraction and classification. Speech recognition is utilized by two different methods from training and testing.

2.1. Signal pre-processing

The different sound is recorded by a microphone in such an environment where no noise is available. These speech signals are classified to many more way for speech recognition as such that pre- processing, filtering, Mel Frequency Cestrum Coefficient of used it. Samples are recorded with a microphone. First of all low and high frequency noise is eliminated by performing some digital filtering. The speeches signals are mainly between 300Hz to 750Hz. Pre-processing units can be used to the signal recognition in time domain before the feature extraction.

Normally, in the preprocessing stage the speech signal is utilized to analog to digital (A/D) conversion, enhancement, filtering and usually for SR applications silence removal.

2.2. Feature Extraction

It is extracting certain important information from the speech signal. Feature extraction could be seen as extracting certain mathematically parameterized information from the original source signal. There are many feature extraction techniques that may be utilized. Example includes Fast Fourier Transform (FFT) coefficients, and Mel-Frequency Cepstral Coefficients (MFCCs). In this investigation, we have opted to utilize MFCCs as the features.

2.3. Classification of speech and EGG signal

The input speech or test signal to determine the input speech uttered matches the desired targeted speech. Some of the categories of classification schemes are voice and unvoiced section using neural networks (NN) approach.

3. PROPOSED METHODOLOGY

In this paper Frame segmentation is normally used for decomposition the speech signal. It is not sufficient for short unvoiced consonants. The combination of information from EGG and speech signal is used.EGG gives information about vocal fold vibration. In the pre-processing stage median filter is used This combination of information from EGG and speech data is used for input signal. In pre-processing stage, envelop detection and adaptive median filter can be used for remove the noise. Thus it produces a vowels and consonants section, when a various time alignment of voice section detected. Voiced sections segmentation and unvoiced consonants detection is combine into the phonemes using neural network.

Classification of speech signal is very important phenomenon in speech recognition process. Neural network is to be used for Different voice signals pre-processing, envelope detection and classification. A Number of processing units which are used for the processing of speech signals.

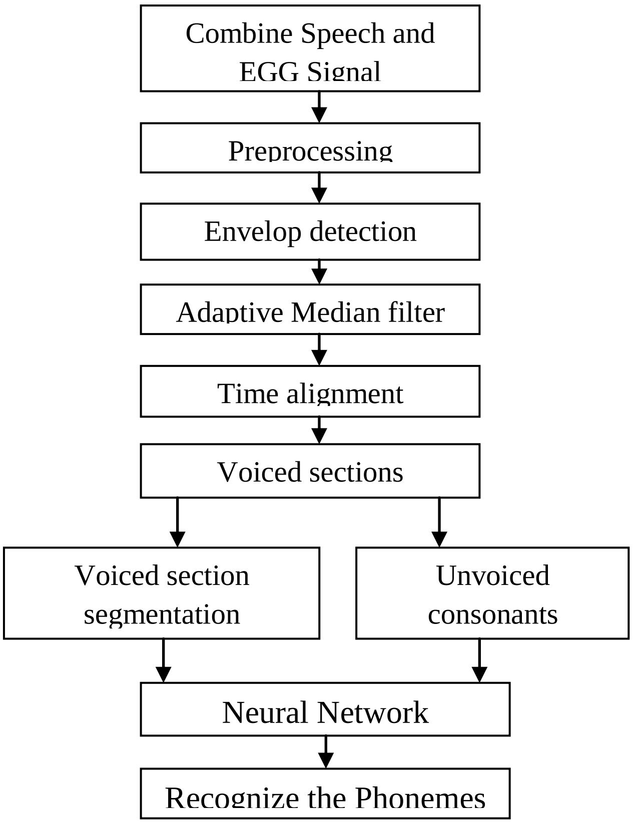


Fig -1: Flow Diagram of Proposed Method

The very simple techniques like preprocessing, filtering are processed by these types of units.

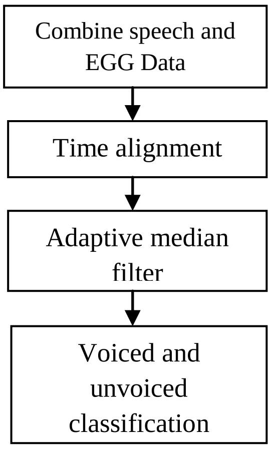

Combine speech and EGG Data
Time alignment
Adaptive median filter
Voiced and unvoiced classification

Fig- 2: Flow Diagram of Voiced Sections Detection

Combining speech and EGG data of defined as a waveform generated when the vocal folds vibration occurs. The adaptive median filter is used to remove the noise and classify the voiced and unvoiced. There are improving the Hit Rate (HR), Signal to Noise Ratio (SNR) and reduce the False Alarm Rate (FA).There is following that equation as,

$$\mathrm{SNR} = \log_{10}\left(\frac{\mu}{\sigma}\right)^{2} \text{............ (1)}$$

Where

μ = mean and σ = standard deviation .

$$\mathrm{HR} = \frac{\mathrm{NH}}{\mathrm{NR}} \text{.............................. (2)}$$

Where, NT is the total number of detected boundaries, NH is the number of correctly detected boundaries and NR is the total number of boundaries.

$$\mathrm{FA} = \frac{\mathrm{NT}-\mathrm{NH}}{\mathrm{NT}} \text{..................... (3)}$$

In order to assess the overall quality of a segmentation method, a global measure which simultaneously takes these scores into account is required. A well known measure is the F1 value as equation:

$$\mathrm{F1} = \frac{2\times(1-\mathrm{FA})\times\mathrm{HR}}{(1-\mathrm{F\ A})+\mathrm{HR}} \text{.......... (4)}$$

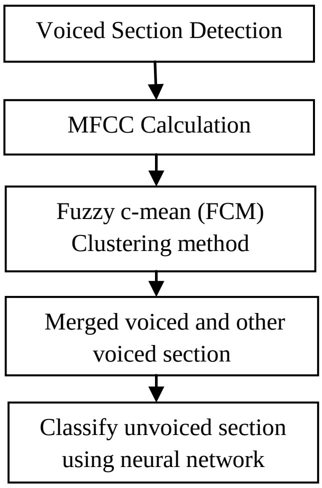

Voiced Section Detection
MFCC Calculation
Fuzzy c-mean (FCM) Clustering method
Merged voiced and other voiced section
Classify unvoiced section using neural network

Fig -3: Flow Diagram of Voiced Section Segmentation

The voiced section more than one vowel, semivowel or consonant. Mel frequency Cepstral Coefficient (MFCC) has been proved the speech data in each pitch-cycle have fixed length. The FCM clustering method used to avoid the spectrum leakage. Merge by the voiced sections and other sections are treated separately. As such that will be classified by the unvoiced section using neural network.

The calculation of the MFCC includes the speech data in each pitch-cycle is padding with trailing zeros to a fixed length n (128 is selected).The hamming window is used to avoid the spectrum leakage. The magnitude squared discrete Fourier transform (DFT) turns the windowed speech data into the frequency domain so that the short-term power spectrum P(f) is obtained. The spectrum P(f) is then filtered by a group of triangular band-pass filters along the Mel-frequency axis. The output is a set of sub-band energies E(d), d = 1,2,...D. The MFCC is calculated by the logarithm of E(d) equation as:

$$C(i) = \sqrt{\frac{2}{D}}\sum_{d=1}^{D}\left[\log\left(E(d)\right).\cos\frac{(2d-1)i\pi}{2D}\right] \text{...... (5)}$$

Where i = 1...D.

The Viterbi algorithm aims at finding the optimum segmental state sequence to achieve the re-assignment of a sequence to several clusters. To implement the FCM clustering method used for the voiced section detection and Merge by the voiced sections and other sections are treated separately. As such that will be classified into the unvoiced section. Unvoiced sections contain two components: unvoiced consonants and silence. Unvoiced consonants are produced by creating a constriction somewhere in the vocal tract tube and forcing air through that constriction, thereby creating non-linear and turbulent air flow. Voiced sections segmentation and unvoiced consonants detection is combine into the phonemes.

4. RESULTS AND DISCUSSION

Combining speech and EGG signal is used as an input in this project. A spectrogram is a visual representation of the spectrum of frequencies in a sound or other signal as they vary with time or some other variable. The Hilbert envelope is important in signal processing, where it derives the analytic representation of a signal. The adaptive median filter is used to remove noise. Finally, classify the voiced and unvoiced section using neural network.

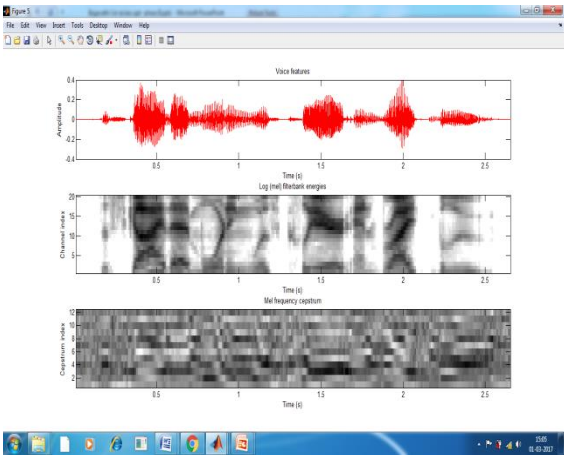
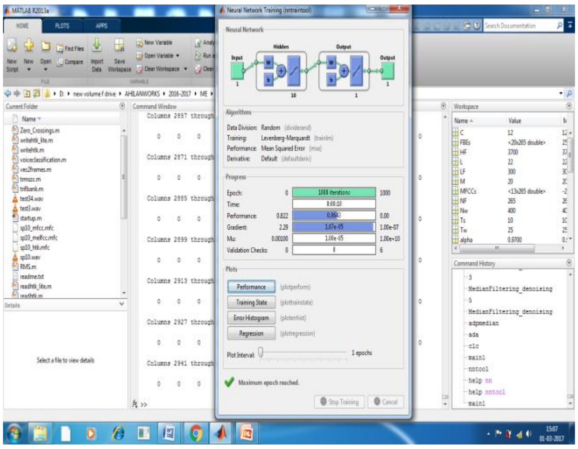


Fig -4: Voiced Section Detection

Fig 4 shown that the utterance is voiced section detection using the voice features have been aligned to the derivative of speech and EGG.


Neural Network Training (nntraintool)
Neural Network
Input
Hidden
Output
Output
1
10
1
1
Performance (plotperform)
Training State (plottrainstate)
Error Histogram (ploterrhist)
Regression (plotregression)
Stop Training
Cancel

Algorithms

Data Division: Random (dividerand)
Training: Levenberg-Marquardt (trainlm)
Performance: Mean Squared Error (mse)
Derivative: Default (defaultderiv)

Progress

Epoch:	0	1000 iterations	1000
Time:		0:00:10	
Performance:	0.822	0.0643	0.00
Gradient:	2.29	1.07e-05	1.00e-07
Mu:	0.00100	1.00e-05	1.00e+10
Validation Checks:	0	0	6

Plot Interval: 1 epochs

Maximum epoch reached.

Fig -5: Neural Network Classification

Fig 5shown that the utterance is neural network classification used to the network processes the records in the Training Set one at a time, using the weights and functions. They process records one at a time, and learn by comparing their classification of the record with the known actual classification of the record.

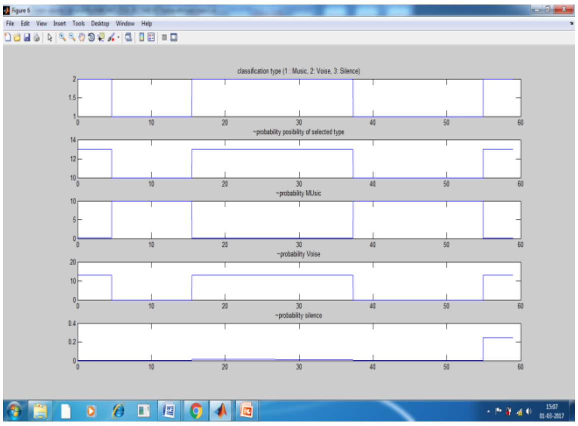


Fig- 6: Classification of Output Signal

Table -1: SNR and Accuracy Value

No of samples Signal	SNR Value	Accuracy Value
Test wave.1	1.73db	87.3%
Test wave.2	3.12db	90%
Test wave.3	6.23db	92%
Test wave.4	8.00db	96%

Fig 6 shown that the utterance is classification of output signal for different types of using music, voice and silence. Each section will be treated separately. Unvoiced section segmentation from classify the unvoiced section within voiced section detection and calculating RMS.

5. CONCLUSION

In this paper, the maximum average accuracy achieved for any network was 91.5%. EGG is adopted to design a precise and robust text-independent phoneme segmentation method. Unlike the traditional methods, phonemes are firstly classified into two categories named voiced (including vowels, semivowels and some consonants) and unvoiced (other consonants).There is comparison to the more accurate value of SNR and accuracy.

REFERENCES

[1] Park, S.S., Shin, J.W., Kim, N.S., "Automatic Speech Segmentation with Multiple Statistical Models", in Proceedings of Interspeech, 2006.

[2] Wouter Geuaert, Georgi Tsenav, Valeri Mladenov, "Neural Network used for Speech Recognition" Journals Automatic Control, volume.20.1.7, 2010

[3] Dr. R. L. K. Venkates, Dr. R. Vasantcha Kumari, G. Vani Jayasatu, "Speech Recognition using A Radial Basis Function Neural Networks" volume.3, PP 441-445, April 2011, 3rd INC on E computer technique IEEE.

[4] Khanagha.V et al (2014), "Phonetic segmentation of speech signal using local singularity analysis," Digital Signal Process., vol. 35, pp. 86–94.

[5] K. Daqrouq, "Wavelet Entropy and Neural Network for Text-Independent Speaker Identification," Engineering Applications of Artificial Intelligence, vol. 24, pp. 796-802, 2011.

[6] Amit.J and Carol.E.W (2003),"Speech segmentation using probabilistic phonetic feature hierarchy and support vector machines," in Proc. Int. Joint Conf. Neural Newts., vol. 1, pp. 675–679.

[7] S. Granqvist et al., "Simultaneous analysis of vocal fold vibration and transglottal airflow: Exploring a new experimental setup," J. Voice, vol. 17, no. 3, pp. 319–330, 2003.

[8] J. Neubauer et al., "Coherent structures of the near field flow in a selfoscillating physical model of the vocal folds," J. Acoust. Soc. Amer., vol. 121, no. 2, pp. 1102–1118, 2007.

[9] L. Chen et al., "Speech emotional features extraction based on electroglottograph," Neural Comput., vol. 25, no. 12, pp. 3294–3317, 2013.

[10] S. Nakagawa, K. Asakawa, and L. Wang, "Speaker recognition by combining MFCC and phase information," Spectrum, vol. 60, p. 76.4, 2007.

[11] Lijiang Chen, Member, IEEE, Xia Mao, Member, IEEE, and Hong Yan, Fellow, "Text-Independent Phoneme Segmentation Combining EGG and Speech Data," IEEE/ACM transactions on audio, speech, and language processing, vol. 24, no. 6,June 2016