ANN Based Assessment of Disfluent Speech

Sheena Christabel Pravin*, S.K. Ranganath, R. Anjana, T. Prabhu Pandiyan, Pradeep Rangarajan
Department of Electronics & Communication, Rajalakshmi Engineering College, Chennai, India
***Corresponding author: E-Mail: sheenachristabelpravin@gmail.com**

ABSTRACT

About 1% of theworld population suffer from stuttering, a continued involuntary repetition ofsound; especially initial constants. Another name for this is speech dis-fluency or Disturbed Speech. This includes word repetition, syllable repetition, prolongation, and interjection. The existing algorithm focuses wholly on either extraction or classification. This paper uses "Artificial Neural Network" or ANN and implements automatic analysis of dis-fluent speech by extracting Mel frequency cepstral coefficients (MFCC, ΔMFCC, ΔΔ MFCC) and prosodic features like pitch, energy, duration and the like. This paper aims at improving the fluency of the stuttered speech.

KEY WORDS: ANN, MFCC, ΔMFCC, ΔΔMFCC, Prosodic Features.

1. INTRODUCTION

Stuttering affects the fluency of speech. The disorders are characterized by its disruption in the production of speech sounds, also called as "dis- fluencies." Most people produce brief dis-fluencies from time to time. For instance, some of the words are repeated and other words are preceded by "um" or "uh." Dis-fluencies are not necessarily a problem; however, they can impede communicationwhen a person produces too many of them. Stuttered speech samples often include repetition of words or parts of words, as well as prolongation of speech sounds and a series of interjections. Stuttering can be diagnosed by an otolaryngologist, a neurologist, or a paediatrician who has knowledge about stuttering, although most of the actual assessment is carried out by a Speech Language Hearing Therapist (SLTH).

In our experimental study, a speech recognition system which is based on Artificial Neural Network Toolkit was built and tested. This paper mainly focuses on repetition and prolongation detection in stuttered speech signal. The acoustic and the pitch related featuressuch as Mel-frequency cepstral coefficients (MFCCs), formants, pitch, and Energy are used to test the effectiveness in recognizing repetitions andprolongations in stammered speech. The ANN or Artificial Neural Networks is used as the speech classifier. The results show that the ANN classifiertrained using MFCC feature achieves an average accuracy of 87.39 % for repetition and prolongation recognition (Awad, 1997).

Speech recognition is the process of identifying the spoken speech. Speech stuttering, also knownas stammering is a disorder thataffectsthe fluency of speech. It occurs in about1%of the population and has found to affect four timesas many males as females. Stuttering is a disorder in which the eloquent flow of speech is disintegrated by occurrences of dis-fluencies such as silent-pauses, prolongations, interjections, word repetitions and syllable repetitions. The objective of the work is to develop a software application which is feasible and is capable of finding the dis-fluencies in stuttered speech and identify the corrected speech. This helps people with speech disorder to easily communicate and exchange their ideas and emotions.

2. PROPOSED METHODOLGY

One of the most often used methods of judgement of the speech fluency is Artificial Neural Networks (ANNs). ANNs are trained on a set of parameterized samples that contains some kinds of disturbances and afterwards they are used as classifiers. ANNs are given unknown. Speech fragments and marks them as fluent or not. In general one ANN is trained to recognize one disturbance type (fricatives prolongations, plosives repetitions, blockades, etc.). Authors of the applied ANNs to classify speech utterances as fluent or not.

Building the Speech Recognition System: The Speech Recognition system is built with Artificial Neural Network as the classifier.

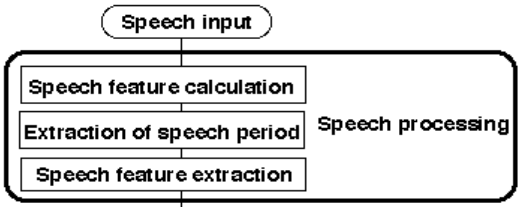

Figure.1. Automatic Speech Recognizer

Feature Extraction: Input signals were parameterized in the following way. Frames had 32 ms length (512 samples) and were taken with the step of 10 ms. Frames were treated by pre-emphasis (filter parameter 0.97) and Hamming filters. Next, Fourier analysis was performed and the results were filtered by 26 triangular filters (filters had equal width on the Mel scale and spread overthe whole frequency range up to the value of 8000 Hz). MFCC co-efficients, their first and second derivatives were calculated. Finally each frame feature vector consisted of 39 elements: 13 MFCC coefficients, 13 of the first and 13 of second derivatives (i.e. delta and acceleration parameters).

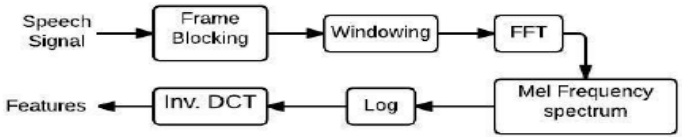

Figure.2. MFCC Feature Extraction

The feature extraction includes MFCC, ΔMFCC, ΔΔMFCC and Prosodic features. (Awad, 1997) There are two classifications in Emotional features viz. Phonetic and Prosodic. Prosodic features include understanding emotions, pitch, intonations and stress of the word spoken. The network is trained to recognize and classify the incoming words into the respective categories. The output from neural network is loaded into pattern classification.

Artificial Neural Network Architecture: The Network Architecture consists of 39 layers each of 3 13 layers. The input layer consists of 13, the hidden layer consists of 13 and the output layer consists of 13 coefficients. In speech processing, it is often a must to do framing and windowing. The width of the frames is generally about 30ms with an overlap of about 20ms (10ms shift). Each frame consists of N sample points of the speech signal. Overlap the rate of frames between %30 and %75 of the length of the frames. The speech is stationary for a short duration of a time. Frame the signal into short frames. For each frame, calculate the periodogram estimate of the power spectrum. Apply the Mel filter bank to the power spectra; sum the energy in each of the filter. Take the logarithm of all filter bank energies. Take the Discrete cosine transformation of the log filter bank energies.

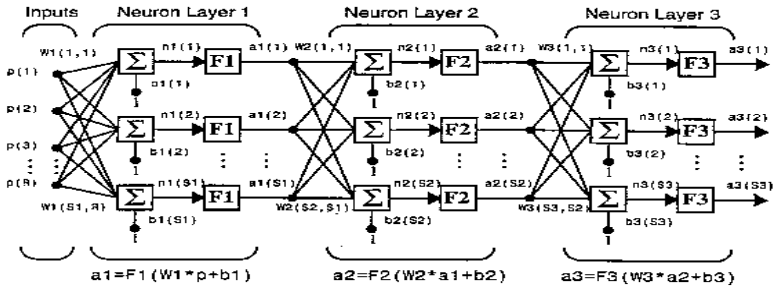

Figure.3. ANN Architecture

Training and Phase Training: To initiate the training process the initial weights are chosen arbitrarily. Then, the training or the process of learning begins. The neural network was trained on various samples namely "how" "are" "you" and "okay" from a number of speakers. The target file was created after training the set. There are four sets of words and were assigned a distinct category. Thus, the first set of frames to category one and the second set of category to another one and the like.

Reason for Opting ANN classifier: When compared to conventional computers, Artificial Neural Network or ANN is a self pre-programmed approach that can do uncertain actions. It is widely used in pattern recognition to conclude the unexpected inputs.ANN consists of several nodes that replicates the biological neurons of human brain. The main advantage in which we choose ANN is that it takes the information samples rather than the entire data for obtaining the necessary outputs. As the data are split into several samples, a complete study of a particular speech sample can be abstracted that saves both time and money. We can infer the unpredicted results from the observationsof the particular incoming inputs during ANN processing stage. It has the ability to recognise the unsynchronised relationship between the decision-making inputs and the pre-trained inputs.

3. RESULTS AND DISCUSSION

After training the network, the following results have been obtained and have been tabulated as follows. The data are from various speakers from both the gender to identify the type and severity level of stammering. The classification includes word repetition, syllable repetition, prolongation and pause/stop gaps. The table and the corresponding time domain graphs were also plotted. The table is formed and tabulated with the help of voice samples of two men and two women. Severity levels were categorized according to the stuttering percentages.The results were satisfactory and the training of Neural Network with additional speech samples further provided improved results.

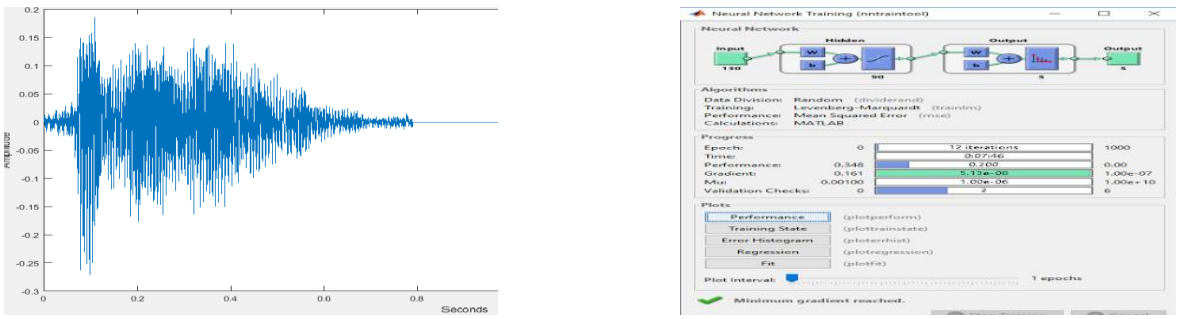

Figure.4. Stuttered Speech and NN training

The %SS denotes "syllable stuttered" which is total dis-fluency divided upon total words spoken. Mild, moderate and severe are the three severity levels observed. There were four classifications which were studied upon; namely syllable repetition, word repetition, prolongation and interjection. Mild level can be ignored while moderate severity level can be minimized using frequent feedback of the repeated words that the speaker has spoken. However, when the severity level is critical or severe it is extremely difficult to eradicate but over time it can be reduced or minimized. The results produced were satisfactory and were able to classify the level of severity in one's stuttered speech.

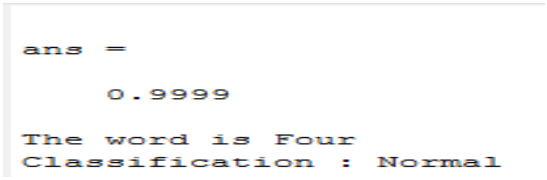

Figure.5. Speech recognised as normal

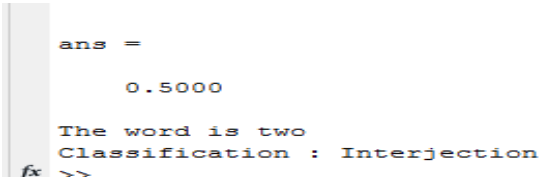

Figure.6. Speech recognised with Interjection

The silent pauses are identified as Interjections

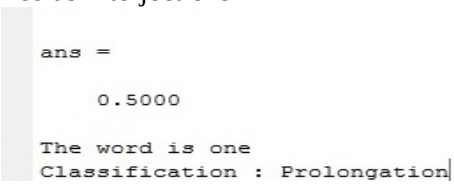

Figure.7. Speech Recognised as Prolongation

Table.1. Severity Assessment in Stuttered Speech

Voice Sample	Syllable Repetition	Word Repetition	Prolongation	Interjection	Total Disfluency	Total Words	%SS	Severity Level
M_0132_12y_1m_1	12	11	6	5	34	120	28.3	Mild
M_0991_08y4m_1	8	8	30	10	56	76	73.6	Severe
M1017_11y8m_1	10	10	20	1	41	96	42.7	Moderate
M1017_12y_5m_1	16	25	10	3	54	174	31.1	Mild

The Neural Network has successfully recognized the words and it has been trained to classify the speech samples.

4. CONCLUSION

Stuttering or stammering is a disorder of speech. In the last few decades, many researchers and pathologists have contributed to the research on stuttered speech recognition. There are three classifiers namely HMM, SVM and ANN. Each classifier has its own odds and evens. Out of all the three, ANN provides maximum accuracy of about 94.9%.

REFERENCES

Awad S.S, The application of digital speech processing to stuttering therapy, Instrumentation and Measurement Technology Conference, IMTC/97, 1997.

Marek Wisniewsk, Wieslawa Kuniszyk-Jozkowiak, Automatic Detection of Sututtering in Speech, Journal of Medical Informatics and Technology, 1997.

Prasad D Polur, Ruobing Zhou, Jun Yang, Fedra Adnani, Rosalyn Hobson S, Gold B and Morgan N, Speech and Audio Signal Processing, John Wiley & Sons, Inc, 2000,

Rodman R.D, Computer Speech Technology, Artech House Publishers, 1999, 126-128.

Swietlicka I, Kuniszyk JO, Zkowikak W, Smolka E, Artificial Neural Networks in the Disabled Speech Analysis, in Computer Recognition System, Springer Berlin / Heidelberg, 57, 2009, 347-354.

Szczurowska I, Kuniszyk JO, Zkowikak W, Smolka E, Design and Analysis of Algorithms- book by author- Parag Himanshu Dave, The Application of Kohonen and Multilayer Perceptron Networks in the Speech Non-fluency Analysis, Archives of Acoustics, 31, 2006, 205.