ISSN:0975-9646

Recognition and Classification Of Speech And Its Related Fluency Disorders

Monica Mundada , Bharti Gawali ,Sangramsing Kayte

Department of CS & IT,

Dr.Babasaheb Ambedkar Marathwada University
Aurangabad .MS (India)

Abstract— **Speech is an integral part of communication. Speech disorder is a problem with fluency, voice, and or how a person produces a speech sound. The main focus of this study is to identify the difference between normal and disordered speech. The proposed work classifies the normal and abnormal speech. The experimental investigation elucidated MFCC and DTW with the accuracy rate of 88 % and 75% respectively. The K-means classifier is used to distinguish the speech disorder with classification rate of 93% on basis of energy entropy and pitch values of the subject. The obtained results are justified using t-test.**

Keywords— **MFCC, Fluency disorder, DTW, KNN and feature classification.**

INTRODUCTION

Speech signal is the most natural method of communication between humans. Speech consists of various paradigms like articulation, voice and fluency. Lungs act as source of energy. Then, vocal folds chop the airflow from lungs to produce quasi periodic excitation. The Shape of the vocal tract determines nature of the sound unit to be produced. Speech is an outcome of time-varying vocal tract filter driven by a time-varying excitation. The state of the vocal cords, the positions, shapes and sizes of the various articulator changes slowly over the time which thereby results in producing the desired speech sounds [6]. However, there is 1% of the population have noticeable speech stuttering problem and it has been found to affect female to male with ratio 1: 3 or 4 times [1]. Stuttering is defined as a normal flow of which is disrupted by unintentionally of dysfluencies such as repetition, prolongation, interjection of syllables, sounds, words or phrases and involuntary silent pauses or blocks in communication. Stuttering cannot be completely cured; it may go into remission for some time [1, 2]. Speech pathology treatments help stutterers to shape their speech into fluent speech. Therefore a stuttering assessment is needed to evaluate performance of stutterers before and after therapy. Traditionally, speech language pathologist (SLP) counts and classifies occurrence of dysfluencies such as repetition and prolongation in stuttered speech manually. The evaluation of these treatment are subjective, inconsistent, time consuming and prone to error. Therefore, it might be good if stuttering assessment can be done automatically and thus having more time for the treatment session between stutterer and SLP. The work is beneficial to know the area of improvement in abnormal speech at early stages. The variations in normal and abnormal speech are studied with the parameter of distance matrix. This manuscript studies the difference in abnormal speech signal for Hindi and Marathi database. The experimental work in performed in Matlab with inclusion of MFCC and DTW to calculate the distance matrix. The K-means classifier is studied to distinguish the normal and abnormal speech signals. The remaining part of paper is organized as follows: The data acquisition is described in section 1. Feature extraction techniques are explained in section 2. Classification techniques described in section 3. Observation and results are followed in section 4. Conclusion and future work in section 5. References are elaborated in last concluding section.

I. EXPERIMENTAL SETUP AND DATA ACQUISITION:

The abnormal speech samples are collected from stammering patients and normal speech samples of respective parameters like age and gender is considered. The database included both Hindi and Marathi sentences .The age of subject is ranging from 17 — 26 years. The subjects include 4 male and 1 female speakers of both normal and abnormal groups. The speech recording is done using Visi Pitch. The main feature of using Visi Pitch is it's the quality of recording is excellent. The recording took place in the normal room without noisy sound and effect of echo. The sampling frequency for all recordings was 16000 Hz at the room temperature and normal humidity. The speaker were seating in front of the direction of the microphone with the distance of about 12-15 cm [12]. Table 1 displays the size of database which includes 150 utterances of both normal subjects. The same parameters like age and gender with same number of sentences and utterance are considered for collection of abnormal speech samples.

Table1. Details of database

Age	Gender	Sentence	Utterance
26	M	10	3
17	M	10	3
21	F	10	3
21	M	10	3
26	M	10	3

II. FEATURE EXTRACTION TECHNIQUES

Feature extraction is the process of retaining useful information of the signal while discarding redundant and unwanted information. Feature extraction is the parameterization of the speech signal. This is intended to produce a perceptually meaningful representation of the speech signal. Feature extraction typically includes the process of converting the signal to a digital from (i.e. signal conditioning), measuring some important characters of the signal such as energy or frequency response. Feature extraction may also involve transforming the signal into a form appropriate for the models used for classification.

A. *Mel Frequency cepstral Coefficients MFCC*

MFCC is the most robust feature extraction technique as it is based on known variation of the human ear's critical bandwidth with frequency. The best values in the parametric representation of acoustic signals are an important task to produce a better recognition performance. MFCC has two types of filter which are spaced linearly at low frequency below 1000 Hz and logarithmic spacing above 1000Hz [7]. Mel frequency scale is used to study the phonetically characteristics of speech signal. MFCC is a spectral analysis method. The overall process of the MFCC is shown in Figure 1.

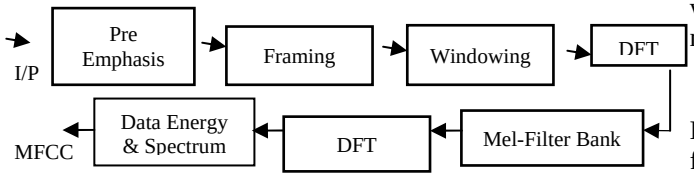

Figure1. Block diagram of MFCC

1. Pre-emphasis

Pre-emphasis refers to a system process designed to increase, within a band of frequencies. Hence, this step processes the passing of signal through a filter which emphasizes higher frequencies. This process will increase the energy of signal at higher frequency.

2. Framing

In framing, we split the pre-emphasis signal into several frames, such that we are analyzing each frame in the short time instead of analyzing the entire signal at once [5]. The frame length is set to 25ms and there is 10ms overlap between two adjacent frames to ensure stationary between frames.

3. Hamming Window

Hamming window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines. The Hamming window is represented as shown in "Eq. (1)". If the window is defined as W (n), the result of windowing signal is

$$Y[n] = X(n)*W(n) \quad \ldots\ldots (1)$$

Y[n] = Output signal
X (n) = input signal
W (n) = Hamming window

3.1.4 Fast Fourier transform

To convert each frame of N samples from time domain into frequency domain FFT is being used. The Fourier Transform is used to convert the convolution of the glottal pulse U[n] and the vocal tract impulse response H[n] in the time domain. This statement supports as shown in "Eq. (2)" below:

$$Y(w) = FFT[h(t)*X(t)] = H(w)*X(w) \quad \ldots\ldots (2)$$

If X (w), H (w) and Y (w) are the Fourier Transform of X (t), H (t) and Y (t) respectively.

3.1.5 Mel filter bank processing

The frequencies range in FFT spectrum is very wide and voice signal does not follow the linear scale. Each filter output is the sum of its filtered spectral components. After that the following equation as shown in "Eq. (3)" is used to compute the Mel for given frequency f in HZ:

$$F(Mel) = [2595 * \log 10[1+ f/700] \quad \ldots\ldots\ldots (3)$$

3.1.6 Discrete cosine transform

This is the process to convert the log Mel spectrum into time domain using DCT. The result of the conversion is called Mel Frequency Cepstrum Coefficient. Each input utterance is transformed into a sequence of acoustic vector. The voice signal and the frames changes, such as the slope of a formant at its transitions. Therefore, there is a need to add features related to the change in cepstral features over time. 13 delta or velocity features (12 cepstral features plus energy), and 39 features a double delta or acceleration feature are added. The energy in a frame for a signal x in a window from time sample t 1 to time sample t2, is represented as shown below, in "Eq. (4)".

$$\text{Energy } X = 2[t] \quad \ldots\ldots\ldots\ldots (4)$$

Where X[t] = signal

Each of the 13 delta features represents the change between frames corresponding to cepstral or energy feature, while each of the 39 double deltas. Features represent the change between frames in the corresponding delta features.

3.2 Dynamic time warping (DTW)

DTW is a template based feature extraction approach. It is a technique that calculates the level of similarity between two time series in which any of them may be warped in a non linear fashion by shrinking and stretching the time axis [3]. Dynamic Time Warping (DTW) is an efficient method for finding this optimal nonlinear alignment [13]. It is an instance of the general class of algorithms known as dynamic programming. Its time and space complexity is merely linear in the duration of the speech sample and the vocabulary size. The algorithm makes a single pass through a matrix of frame scores while computing locally. It is used to compute the best possible alignment warp between two speech samples and the associated distortion. The overall distortion is based on a sum of local distances between elements. In this study minimum distance is calculated from test speech signal to each of the training speech signal in the training set. This classifies test speech sample belonging to the same class as the most similar or nearest sample point in the training set of data. A Euclidean distance measure is used to find the closeness between each training set data and test data [4].

III. CLASSIFICATION TECHNIQUES

Classification technique involves studying the distinguished features and creating them in different groups depending on their feature set. Classification is the property assigned to

new data set which is observed on basis of known training data set .In this work our focus is to perform two class classification i.e normal and abnormal speech samples using the K-means classifier. The main significance of K-means classifier is that there is always at least one item in each cluster. The clusters in K-means are non-hierarchical and they do not overlap. In this, every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

A. K-means: K-means classifies new instance query based on closest training examples in the feature space [9]. K-means is a clustering algorithm. Clustering algorithms are unsupervised techniques for sub-dividing a larger dataset into smaller groups. The main idea is to define k centroids, one for each cluster. k-means clustering aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster[10].In this study, the K-means classified the speech samples with pitch and energy entropy of respective subjects.

Algorithm

1. Place K points into the space represented by the objects that are being clustered. These points represent initial group centroids.
2. Assign each object to the group that has the closest centriod.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

B. t- test:

This test is performed for comparing the means of two samples (or treatments), even if they have different numbers of replicates [11]. The t-test compares the actual difference between two means in relation to the variation in the data. In t-test, initially the null hypothesis is designed .The mean of each subject including all utterances is calculated with their respective variance and standard deviation. T is given by formula.

$$t = \frac{\bar{x}1 - \bar{x}2}{\sigma}$$

Where, $\bar{x}1$ = mean of sample1

$\bar{x}2$ = mean of sample 2

σ = Standard deviation

If the calculated t value exceeds the tabulated value we say that the means are significantly different at that level of probability.

IV. Experimental Analysis

The experimental work is evaluated with statistical parameters like mean and standard deviation to study the variations associated with the original speech samples [8]. Table 2 shows the mean and standard deviation values for normal and abnormal speech samples.

Table2. Mean and Standard deviation values for Abnormal and Normal speech

Parameter		Abnormal Speech		Normal Speech	
Age	**Gender**	**Mean**	**Std.Dev**	**Mean**	**Std.Dev**
26	M	3.704	9.583	4.384	6.698
17	M	3.925	9.388	4.575	8.512
21	F	3.747	10.884	4.125	8.767
21	M	3.773	10.848	4.703	8.658
26	M	4.002	10.514	4.384	6.698

Table 3 shows the distance matrix measure of using DTW for all subjects for each sentence in the database. Table 4 shows the comparative recognition rate of MFCC and DTW.

Table 3: Distance matrix measure of Normal VS Abnormal Speech using DTW

	Subject1	Subject2	Subject3	Subject4	Subject5
S1	197.71	387.96	160.88	377.58	116.88
S2	180.10	200.77	146.05	346.99	94.12
S3	184.79	325.87	132.14	216.19	98.67
S4	207.32	413.09	206.10	394.47	142.60
S5	184.69	325.22	265.03	307.21	144.16
S6	279.95	373.17	173.79	303.98	183.05
S7	251.30	327.69	206.33	311.45	155.63
S8	137.45	174.91	129.12	304.45	107.06
S8	190.30	228.77	196.17	302.26	104.23
S10	227.29	377.84	193.18	301.45	103.34

Table4. Comparative recognition rate of MFCC and DTW

	Vector size	Recognition rate
MFCC	150	88%
DTW	150	75%

K-means K-means classifier distinguished the normal and abnormal speech samples. The classification is performed on basis of energy entropy and pitch of both the subject including all utterance. Figure2 depicts the plotting of all values of speech samples. Figure 3 plots the classification of normal and abnormal speech samples with 93% of classification rate.

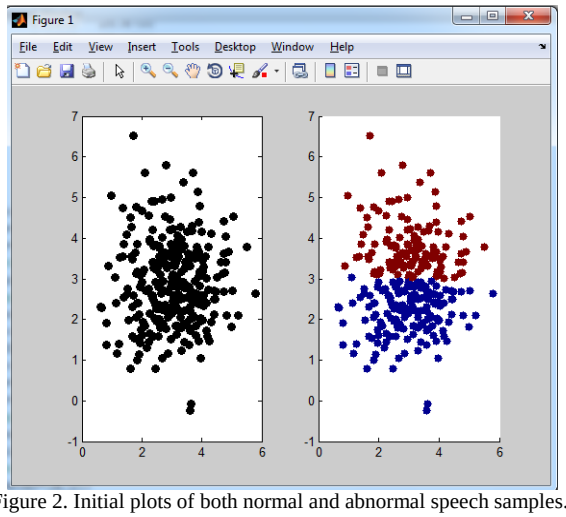

Figure 2. Initial plots of both normal and abnormal speech samples.

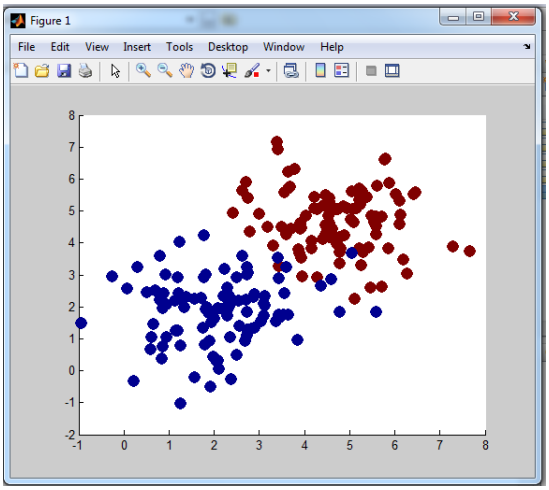

Figure 3. Classified the speech samples using K-means

t-test is performed to justify the classification accuracy. As discussed above, the steps are performed to calculate the mean and standard deviation values of both normal and abnormal speech samples. Then the t-test proved the hypothesis statement for classification of speech samples with accuracy rate of 95%.

V. CONCLUSION

The proposed study classifies the normal and the abnormal speech with K Means classifier with classification rate of 93%.t-test is carried out to justify the classification rate, which is proved to be 95% .The study also measures the distance matrix using MFCC and DTW in abnormal and normal speech samples with accuracy rate of 88% and 75%. In future, the work will be extended to study the accurate position of production system which produces the abnormal speech. The expert system will be designed to classify the speech and outlines the factors responsible for various speech disorders.

REFERENCES

1. M. Hariharan & Lim Sin Chee & Ooi Chia Ai & Sazali Yaacob." Classification of Speech Dysfluencies Using LPC Based Parameterization Techniques". Springer,LLC 2011. DOI 10.1007/s10916-010-9641-6.
2. Tian-Swee, T., Helbin, L., Ariff, A. K., Chee-Ming, T., and Salleh, S. H., Application of Malay speech technology in Malay SpeechTherapy Assistance Tools. Intelligent and Advanced Systems, 2007. ICIAS 2007. International Conference on, 2007, pp. 330–334.
3. halid A. Darabkh, Ala F. Khalifeh, Baraa A. Bathech, and Saed W. Sabah. " Efficient DTW Based Speech Recognition System for Isolated Words of Arabic Language". World Academy of Science, Engineering and Technology .Vol:7 2013-05-25.
4. Akanksha Singh Thakur, NamrataSahayarn " Speech Recognition Using Euclidean Distance" International Journal of Emerging Technology and Advanced ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 3, March 2013.
5. ANJALI BALA, ABHIJEET KUMAR, NIDHIKA BIRLA." VOICE COMMAND RECOGNITION SYSTEM BASED ON MFCC AND DTW". International Journal of Engineering Science and Technology. Vol.2 (12), 2010, 7335- 7342.
6. Rabiner, Lawrence R., and Ronald W. Schafer. Digital Processing of Speech Signals. Englewood Cliffs, NJ: Prentice-Hall, 1978.
7. Kashyap Patel, R.K. Prasad." Speech Recognition and Verification Using MFCC & VQ". International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319-6378, Volume-1, Issue-7, May 2013.
8. S.K.Mangal." Statistics in psychology iarlc4 education". PHI Learning Private Limited 2010.
9. Monica Mundada, Bharti Gawali." A review and study of influence factors associated with speech and its disorders". International Conference on Emerging Research in Computing, Information, Communication and Applications. (ERCICA-14).
10. Andrew Moore: "K-means and Hierarchical Clustering TutorialSlides"http://www2.cs.cmu.edu/~awm/tutorials/kmeans.html viewed on 09 /09/2014.
11. A tutorial on t-test. http://archive.bio.ed.ac.uk/jdeacon/statistics/tress4a.html cited on 09/09/2014.
12. Santosh Gaikwad, Bharti Gawali , Pravin Yannawar , Suresh Mehrotra." Feature Extraction Using Fusion MFCC for Continuous Marathi Speech Recognition".
13. Santosh Gaikwad, Bhart Gawali, Pravin Yannawar." Performance Analysis of MFCC & DTW for Isolated Arabic Digit" International Journal of Advanced Research in Computer Science, 2, Jan. –Feb, 2011,513-518.