Removal of silent pauses and stuttered speech recognition by MFCC and K-means algorithm

K V Rashmi, Dr. V Udayashankara

Abstract— Speech is the basic means of communication which humans use to express their thoughts, feelings and ideas. Fluency measures the effectiveness of speech in delivering information while communicating with another person. Stuttering is a speech disorder which affects the normal fluency of speech. The individuals with this disorder find it difficult to speak fluently. Silent-pauses, repetitions, prolongations, interjections are some of the characteristics of dysfluency.The presence of silent pauses in speech affects the fluency of it.Mel frequency coefficients and K-means clustering algorithm provides the optimal solution in each and every aspects.

Index Terms— Dysfluency, Dynamic Time Warping, MFCC, Pre-processing, Silence removal, Speech recognition, Stuttering

1 Introduction

For human communication, speech plays the most important role. Fluency measures the effectiveness of speech in delivering information while communicating with another person. But not all the humans are blessed with good speaking capability. Stuttering is one of the fluency disorders which affect the life of stutterers and their surroundings as well. In stuttering, the normal speech flow is disturbed by repetitions and prolongations.

Speech stuttering is also known as dysphemia and stammering. It is one of the critical problems focused in speech pathology. It occurs in about 1% of the world's population and has found to affect males in comparison with females in the ratio of 4:1.

The major role of speech is to convey messages in a linguistic feature, and it consists of articulation, voice and fluency pattern. Sometimes, speech becomes unintelligible in some people due to a disorder. In order to benefit the people with such speech disorders, it is very much necessary to eliminate these disfluent speech characters, so that a clear and fluent speech can be obtained.

The necessity and need of communication with machines have provoked for the build and development of speech recognition components. Though enormous research work is going on in the speech recognition a lot of challenges are yet to be worked on. One such challenge is speech recognition for people with stuttered speech problem.

The dysfluencies in speech badly affects the performance of Automatic Speech Recognition (ASR) system and makes such system unusable to the user suffering from speech disorders.The research work mainly aims at correcting the stuttered speech and silent pause removal in stuttered speech available database. The study of dysfluencies for speech-based system and analysis has gained more attention in the field of healthcare, military, security and machine learning scenarios.

2 Database

The stuttering database was obtained from UCLASS (University College London Archive of Stuttered Speech) website. The database consists of recordings for monologs, readings and conversations. About 43 different speakers have contributed 107 readings. In this work, only a subset of the availabe sample that is 50 samples of speech were taken from the UCLASS archive for analysis. The speech samples were chosen to cover a broad range of gender, age and stuttering rate.

3 Proposed Methodology

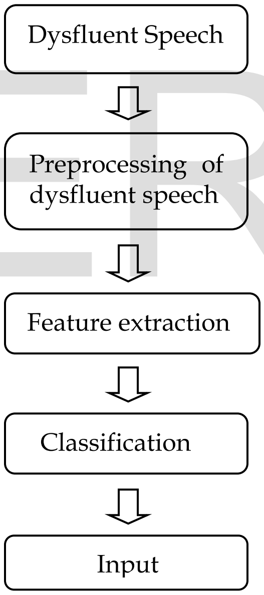

Fig.1 Block diagram of stuttering recognition process

Fig.1 shows the diagramatic representation of the proposed methodology. The detailed stuttering speech recognition process is as described below-

3.1 Pre-Emphasis

In general, the speech waveform suffers from additive noise. The performance of automatic speech recognition systems degrade greatly when speech is correpted by noise. In order to enhance the accuracy and efficiency of the feature extraction process, the speech signals are preprocessed. Pre-emphasis is performed by filtering the speech signal with a first order FIR filter, which takes the following form.

$$H(z) = 1-k*z^{-1} \qquad (0.9<k<1)$$

3.2 Stutter Removal

Stuttering is a speech disorder with many definitions characterized by certain types of speech dysfluencies. The different dysfluency classes are broken words; sound prolongations; word repetitions; syllable repetitions; interjections and phrase repetitions. This paper proposes the use of speech recognition technology to identify the silent paused stuttered speech.

A verbal speech signal can be categorized into two as voiced speech and unvoiced speech. Being able to distinguish between voiced and unvoiced speech is very important for speech signal analysis, which can be determined by characteristic features like energy and zero crossing rate. Energy feature of the speech signal is employed for determining voiced and unvoiced speech.

3.3 Framing

Analyzing a stationary signal is simple and easy compared to continuously varying signal. The speech signal is continuously varying but from a short time point of view it is stationary, this is from the fact that glottal system cannot change immediately and research states that speech is typically stationary in the window of 20ms. Therefore, the signal is divided into frames of 20ms which corresponds to n samples:

$$n=t_s t_{fs}$$

In speech processing it is often advantageous to divide the signals into frames to achieve stationary.

3.4 Feature Extraction

To identify a speech signal, the speech features should be matched with the previous signal or upcoming signal. Hence feature extraction is performed to convert speech signal to some types of parametric representations for further analysis. There are several feature extraction techniques namely Linear Predictive Coefficients (LPC), Linear Predictive Cepstral Coefficients (LPCC), Mel Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction cepstra (PLP).

3.5 MFCC based feature extraction process

MFCC is one of the successful feature extraction methods in speech dysfluency classification. MFCC is used as it is based on the known variations of the human ear's critical bandwidths, with frequency filters spaced linearly at low frequencies and logarithmically at high frequencies to capture important characteristics of speech. This is expressed in the mel-frequency scale; which is a linear spacing below 1000Hz and a logarithmic spacing above 1000Hz. The approximate formula to compute the Mel's for a given frequency f in Hz is given by:

$$\text{Mel}(f) = 2595 * \log10 (1+ (f/700)) \quad (1)$$

The MFCC computation process is as shown in Fig. 2.

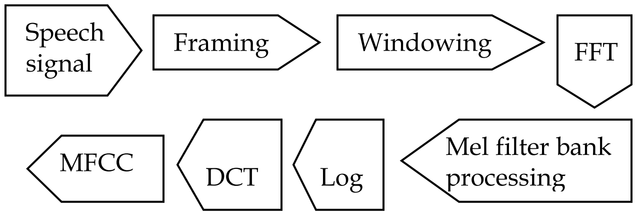

Fig.2 MFCC Computation

3.5.1 Windowing:

The next step in the processing is to window each individual frame so as to minimize the signal discontinuities by using the window to taper the signal to zero at the beginning and end of the frame. Windowing is a point wise multiplication between framed signal and the window function. A good window function has a narrow main lobe and low side lobe levels in their transfer functions. The hamming window is applied to minimize the spectral distortions and discontinuities.

The hamming window coefficients are estimated as:

$$W(n) = 0.54- 0.46 \cos (2\prod (n/N)) \; (0 \le n \le N) \quad (2)$$

3.5.2 Fast Fourier Transform (FFT):

The speech signal can be analysed much better in frequency domain. Thus, FFT is applied on the windowed signal which is essentially still a DFT for transforming discrete time domain signal into its frequency domain. The difference is that FFT gives more efficient and faster computations which are given by the equation:

$$Y(w) = FFT(h(t) * X(t)) = H(w) * X(w) \quad (3)$$

3.5.3 Mel Frequency Wrapping:

One way to more concisely characterize the signal is through filter banking. The frequency ranges of interest are divided into N bands and measure the overall intensity in each band. Intensity in each band is measured by simply adding up all the values in the range, or compute power measure by summing the squares of the values. To agree better with the human perceptual capabilities mel-frequency scale is used which follows a linear spacing below 1000Hz and a logarithmic spacing above 1000Hz.

3.5.4 Discrete Cosine Transform (DCT):

The last process in Mel-Filter feature extraction is to apply inverse transform to obtain the enhanced speech signal. Since speech signal is not present in the entire transform coefficient and to obtain original signal DCT is applied. DCT provides higher energy compaction as compared to DFT. Unlike DFT the DCT coefficients are real and there is no phase component. Hence DCT is a good choice for speech enhancement. With the values from each filter band given, cepstrum parameter in Mel scale can be estimated and MFCC features are obtained.

3.6 Vector Quantization

Vector Quantization (VQ) is an efficient and simple approach for data compression. It is used to preserve the prominent characteristics of data. VQ is one of the ideal methods to map huge amount of vector from a space to a predefined number of clusters, each of which is defined by its central vector or centroid. One of the keypoint of VQ is to generate a good codebook such that distortion between the original signal and the reconstructed signal is the minimum. Various techniques to generate codebook are available. The method most commonly used to generate codebook is the K-means algorithm.

The K-means algorithm is a straightforward iterative clustering algorithm that partitions a given dataset into user specified number of clusters K.

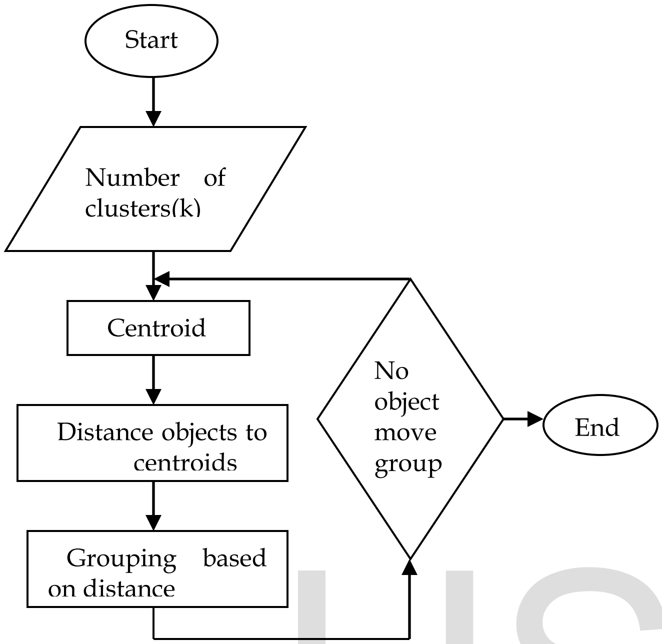

Fig.3 Steps in K-means Algorithm

In brief, the K-means algorithm is composed of the following steps:
1. Clusters the data into k groups where k is predefined.
2. Selects k points at random as cluster centers.
3. Assigns objects to their closest cluster center according to the Euclidean distance function.
4. Calculate the centroid or mean of all objects in each cluster.
5. Repeats steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds.

3.7 DTW Score Matching

Comparing the template with incoming speech might be achieved via a pair wise comparison of the feature vectors in each. The total distance between the sequences would be the sum or the mean of the individual distances between feature vectors. The problem with this approach is that if constant window spacing is used, the lengths of the input and stored sequences are unlikely to be the same. The Dynamic Time Warping algorithm achieves this goal; it finds an optimal match between two sequences of feature vectors which allows for stretched and compressed sections of the sequence. In time series analysis, dynamic time warping (DTW) is an algorithm for measuring similarity between two temporal sequences which may vary in time or speed.

4 Results and Discussions

A speech recognition system capable of finding the dysfluencies in a silent paused stuttered speech and recognition of the corrected speech has been developed. MATLAB software is used for developing stuttered speech recognition system. MATLAB is a high-performance language for technical computing. It integrates computation, visualization, and programming environment. It has powerful built-in routines that enable a very wide variety of computations. It also has easy to use graphics commands that make the visualization of results immediately available. Specific applications are collected in packages referred to as toolbox. There are toolboxes for signal processing, symbolic computation, control theory, simulation, optimization, and several other fields of applied science and engineering.

With the utilization of data from UCLASS, the raw speech signals or samples are subjected to the pre-processing stage. From each pre-processed signal different stuttered samples like repetitions, prolongations and interjections are identified and segmented manually and MFCC features are extracted.

The database is divided into two subsets: training set and testing set based on the ratio 80:20 respectively. The Table 1 shows the distribution of speech segments for training and testing. To analyze speech samples first we extract MFCC feature, afterwards two training database is constructed for dysfluent and fluent speech samples. Once the system is trained, test set is employed to estimate the performance of classifiers.

In the testing process stuttered speech is given to the system. The system eliminates the silent pause from the speech and extracts MFCC features and compares it with the training database.The stuttered speech is recognized after comparing it with the database using dynamic time warping.

	Speech segments	Training	Testing
Repetition	50	40	10
Prolongation	50	40	10
Interjections	50	40	10

Table.1 Data distribution

Classification of the system is done based on three different stuttering types ie. repetition, prolongation, interjection. Each of the system is based on the variation of the code book size in different numberslike 16, 64 and 256.
For codebook size of 16, about 64.66% of accuracy in repitition, about 81.2% of accuracy in prolongation type and about 85.7% of accuracy in interjection is obtained.It has been found that the accuracy of the system increases with the increase in codebook size. The experiment was repeated three times, each time different training and testing sets were built randomly.The acoustic model parameters of the speech units are estimated using training data. Language models are obtained from the collected large database with script files.

4 Conclusion

In this paper an approach for recognition of silent paused stuttered speech is presented. Stuttering is eliminated by considering the fact that voiced speech has more energy than the unvoiced speech. The feature extraction was performed using MFCC algorithm. The VQ code book is generated by clustering the training features vectors of the dysfluent speech and then stored in the database. In this method, the K-means algorithm is used for clustering purpose. DTW algorithm was used to match the dysfluent speech with the database. Finally the silent paused stuttered speech is corrected and recognized. Stuttered speech recognition and correction system is successfully developed. And this system is used to clearly understand the words uttered by a person with speech disorder. The current system is employed only for isolated silent pause stuttering word. The system can be further improved for complete sentences and also for multi modal stuttering.

REFERENCES

[1] Szczurowska I, Kuniszyk-Jozkowiak W and Smolka E. “The application of Kohonen and multilayer perception networks inthe speech nonfluency analysis” Archives of Acoustics, 31, Vol. 4,205,2006.

[2] Stephen. So, and Paliwal. K .K, “Efficient product code vector quantization using switched split vector quantizer”, Digital Signal Processing Journal, Elsevier, Vol 17, pp. 138-171, Jan 2007.

[3] L S Chee, O C Ai, M Harihan and S Yaacob, “ Automation detection of prolongations and repetitions using lpcc,” in International Conference for Technical postgraduates 2009, TECHPOS 2009, pp.1-4, IEEE, 2009.

[4] S wietlica, W. Kuniszyk-Joz’kowiak and E Smo, “Hierarchical ann system for stuttering identification,” Computer Speech and Language, Elsevier, vol. 27, no.1, pp.228-142, 2013.

[5] P Mahesha and D SVinod, “ An Approach for classificatio of dysfluent and fluent speech using K-NN and SVM,” International Journal of Computer Science, Engineering and Applicatios(IJCSEA) Vol. 2, No. 6, December 2012.

[6] G. Jhawar, P. Nagraj and P Mahalakshmi, “Speech disorder recognition using mfcc”,in International Conference on Communication and Signal Processing, ICCSP 2016, pp. 246-250.IEEE,2016.

[7] P.B Ramteke, S.G. Koolagudi and F Afroz, “Repetition detection in stuttered speech,” in Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics, Springer, vol. 43, pp. 611-617,2016.

[8] Aharonson E, Aharonson K, Raichlin-Levi, A Sotzianu, O. Amir and Z. Ovadia Blechman, “A real-time phoneme counting algorithm and application for speech rate monitoring,”Journal of Fluency Disorders, Elsevier, vol. 51, pp. 60-68,2017.

[9] U. T. Petronas and U. T Petronas, “Towards the development of crteria to assess stuttering mobile apps assess stuttering mobile apps,”2017.

- *K V Rashmi is currently pursuing masters degree program in biomedical signal processing and Instrumentation in JSS Science and Technology University,Mysuru,Karnataka, India. E-mail: rashmikv05@gmail.com*
- *Dr. V Udayashankara is currently working as a professor in JSS Science and Technology University,Mysuru,Karnataka, India.*