Stuttered Isolated Spoken Marathi Speech Recognition by using MFCC and LPC

Swapnil D. Waghmare
Department of Computer Science and Information Technology,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad-431004, India (M.S.).

Ratnadeep R. Deshmukh
Department of Computer Science and Information Technology,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad-431004, India (M.S.).

Pukhraj P. Shrishrimal
Department of Computer Science and Information Technology,
Dr. Babasaheb Ambedkar Marathwada University,
Aurangabad-431004, India (M.S.)

Vishal B. Waghmare
Department of Computer Science, Vivekanand College,
Tarabai Park, Kolhapur-416 003(MS) India

Ganesh B. Janvale
MGM's Institute of Biosciences and Technology,
Aurangabad -413 003 (MS) India,

Abstract- Stuttering is a speech disorder in which the person who stutters is unable to produce sounds, syllables, words or phrases as well as involuntary silent pauses or blocks. This paper presents the stuttered speech recognition system for Marathi Regional Language We collect the stuttered speech samples i.e. zero to nine spoken in Marathi language from Marathwada region. We also calculate the speech features i.e. Mel Frequency Cepstral Coefficents (MFCCs) and linear predictive coding (LPCs). We get the 75% recognition rate for MFCC and 82% for LPC. The different statistical methods will be applied to calculate the recognition rate.

Keywords – Stuttered Speech, MFCC, LPC, Confusion Matrix.

I. INTRODUCTION

The speech is not sound in a smooth manner; it is disrupted by disorder with unspecified etiology. It is commonly assumed that stuttering consequences from cohesion of biological [1], psychological and even social reactions. Voice carries various acoustic and linguistic characteristic like basic pitch subsequent formant frequency etc. Every individual involves different noticeable flow of a speech, rhythms [2] determined by structure and arrangement of larynx, pharynx, oral and nasal activity, paranasal sinuses, and thorax. The description of stuttering is still lacks of consensus to consistently constitutes sphere of stuttering in despite of carries large number of research. Spoken languages influence to communication among the human being. Speech has the ability of being used as mode of interaction with computer [3,4,5]. Human beings have been motivated to develop a system that can understand and recognize voice for normal speech. Based on research we found types of stuttering.

Types of Stuttering

(i) Developmental stuttering:

Developmental stuttering is very common in children, they are unable to get command on verbal skill as their speech and language processes are underdeveloped phase.

(ii) Neurogenic stuttering:

Neurogenic stuttering is caused by the impairment between motor control, nerves and muscle contradiction.

***(*iii) Psychogenic stuttering:**

Psychogenic stuttering is directly connected with patients' mental stress and speaking behaviors [6,7].

In the United Kingdom [UK] stuttering is identified as stammering, it can found to be a very serious and complex disorder in speech pathology. As per the global scenario it occurs in approximately 1% of the entire population and has found that it affects 1:3 or 4 times in female to male ratio. This dysfluency may cause difficulties in interpersonal language communication as well as reluctance to speak, sense of guilty, and low self-esteem 23. It is the oldest and most serious speech problems in the history of speech and language pathology.

World Health Organization (WHO) recognize stuttering with the code F98.5, and finalize stable definition as a speech having frequent repetition or prolongation of sounds or syllables or words, or by frequent hesitations or pauses that disrupt the rhythmic flow of speech [7,8]. It is evident from the past and concurrent literatures survey that, stuttering can be appraised a genetic disordered since 1930s [8] and employed several different dysfluencies are: interjection, revision, repetition, prolongations, and blocks. Figure 1 shows the general block diagram of stuttering recognition.

The rest of the paper is organized as follows. Proposed embedding and extraction algorithms are explained in section II. Experimental results are presented in section III. Concluding remarks are given in section IV.

II. PROBLEMS IN SPEAKING

(i) ***Interjection:***

These are negligible extraneous sound, word, syllable or phrase that doesn't alter the meaning of the original sentence. It depends on language, in English "um, uh, well, like" are frequently occurs. For example: The baby um- um- uh was um um hungry. Interjections are also known as filled pause or fillers.

(ii) ***Revisions:***

It occurs when speaker corrects the contents or grammatical structure of phrases. It can change the meaning of original message. Example of revision is broken words, sentence like "I'd like to chang... I'll modify…."

(iii) ***Repetition:***

It takes place if any part of sentence is involuntary said more than one. There are two repetition phenomenons, first is Syllable repetitions: A syllable or sound is repeated at the beginning of the words. For example: "I had a c-c-c-coffee". Another phenomenon is Word repetitions: Here example: "The baby-baby had the soup".

(iv) ***Prolongation:***

Unduly prolong the sounds or syllable for example: muuuuummmy has gone there.

(v) ***Broken Words:***

When speaker trying to pronounce syllable with too much force, and broke the whole word with pause. Example: "It was won [pause]derful"[9,10,11,12,13].

III. EXPERIMENTAL DATA

Speech corpus plays a key role in construction of ASR. Text corpus is a very crucial for language modelling, language synthesis and speaker recognition. The Text corpus are developed in such a manner that using minimum words and sentences that will cover the maximum phonetic variations of a language for which the speech application will be developed. For selecting the speech database we need the text corpus which is recorded from different speaker male and female. We select the speakers age from 15 to 55 years from Marathwada region. The speakers are classified into male and female speakers. The number of speakers are selected are about 12 and in which ten numbers is taken where a single word is repeated three times. Total number of database is about 324. The table 1 represents the words selected for the development of the isolated numerical word speech database along with the respective transliterations and IPA (International Phonetic Alphabet).

Table 1: Marathi spoken numerical with IPA

Devnagari	IPA	Digits
	/ɕəun̪əjə/	0
	/ekə/	1
	/d̪əon̪ə/	2
	/t̪əin̪ə/	3
	/tɕəarə/	4
	/pəatɕə/	5
	/səɦəa/	6
	/səat̪ə/	7
	/aʈʰə/	8
	/n̪əuː/	9

IV FEATURE EXTRACTION USING MEL-FREQUENCY CEPSTRAL COEFFICIENTS (MFCC)

The Mel Frequency Cepstral Coefficient is the well-known and most widely used feature extraction method in speech domain. The MFCC is based on the human auditory perception system. The human auditory perception system does not follow a linear scale of frequency. For each tone with actual frequency ‘f’ measured in Hz, a subjective pitch is calculated, is known as ‘Mel Scale’. The mel frequency scale is a linear frequency spacing below 1000 Hz and logarithmic spacing above 1000Hz. As a reference point, the pitch of a 1 KHz tone, 40 dB above the perceptual hearing threshold is defined as 1000 Mels [11]. The block diagram of MFCC feature extraction method is shown in figure 2. The various steps involved in the calculation of MFCC are described below:

(i) Pre-Emphasizing

The speech signal is first pre-emphasized with the pre-emphasis filter 1-az-1 to spectrally flatten the signal.

(ii) Framing and Windowing

A speech signal is assumed to remain stationary in periods of approximately 20ms. Dividing a discrete signal s[n] into frames in the time domain truncating the signal with a window function w[n]. This is done by multiplying the signal, consisting of N samples. The signal is generally segmented in frame of 20 to 30 ms; then the frame is shifted by 10 ms so that the overlapping between two adjacent frames is 50% to avoid the risk of losing the information from the speech signal. After dividing the signal into frames that contain nearly stationary signal blocks, the windowing function is applied.

(iii) Fast Fourier Transform

Fast Fourier Transform converts each frame N samples from time domain into frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of N samples {xn}, as shown in equation 1,

$$k=0, 1, 2 \ldots N-1 \qquad (1)$$

In general Xk are complex numbers and we only consider their absolute values (frequency magnitudes). The resulting sequence {Xk} is interpreted as follow: positive frequencies $0 \leqslant f < Fs/2$, correspond to values $0 \leqslant n < N/2-1$, while negative frequencies $-Fs/2 < f < 0$ corresponds to $N/2+1 \leqslant n \leqslant N-1$. Here, Fs denote the sampling frequency. The result after this step is often referred to as spectrum or periodogram. To obtain a good frequency resolution, a 512 point Fast Fourier Transform (FFT) is used.

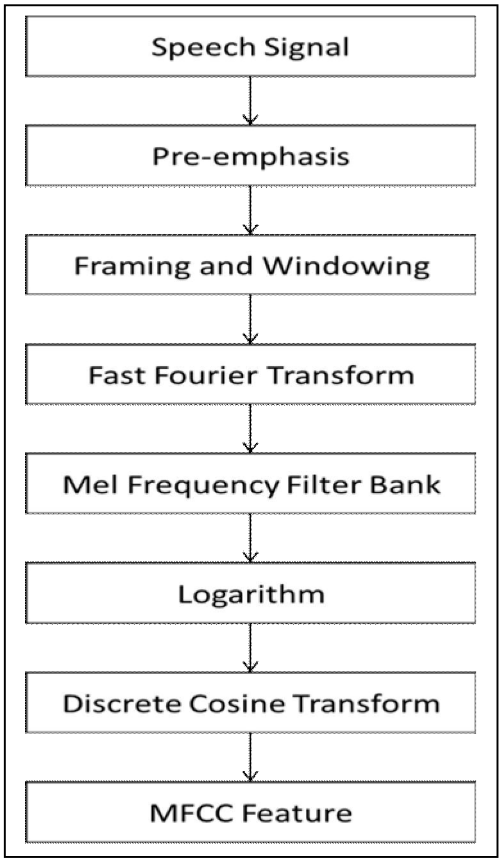

Figure 1 Block diagram of MFCC Feature Extraction method

(iv) Mel Frequency Filter Bank

A filter bank is created by calculating a number of peaks, uniformly spaced in the Mel-scale and then transforming back to normal frequency scale where they are used as peaks for the filter banks.

(v) Logarithm

The logs of the powers at each of the Mel frequencies are calculated. The new array/vector of Mel log power is generated.

(vi) Discrete Cosine Transform

Discrete Cosine Transform (DCT) is being used to achieve the mel- cepstrum coefficients. In a frame, there are 24 Mel Cepstral coefficients, out of 24 only 13 coefficients have been selected for the recognition system.

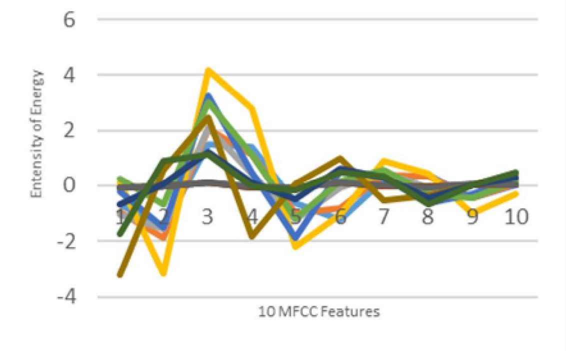

Figure 2: Extracted MFCC features for Male: Sample-1 with 10 frames, Word: Shunya(Zero)

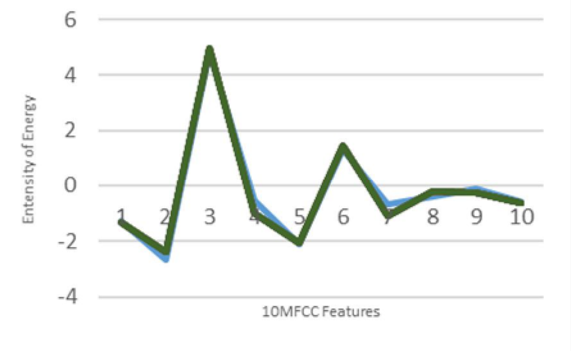

Figure 3: Extracted MFCC features for Male: Sample-1 with 10 frames, Word: Ek(One)

V Feature Extraction Using Linear Predictive Coding (LPC)

A speech sample can be approximated as linear combination of past speech samples by minimizing the sum of the squared differences between the actual speech samples and the linearly predicted ones over a finite interval, a unique set of predictor coefficients is determined. Speech sample is modelled as the output of linear, time-varying system excited by either quasi-periodic pulse during voiced speech, or random noise during unvoiced speech. The linear prediction method provides a robust, reliable, and accurate method for estimating the parameters that characterize the linear time-varying system representing vocal tract which will be helpful for emotion recognition. Most recognition systems assume all pole model known as auto regressive (AR) model for speech production.

$$s(n) = \sum_{k=1}^{p} a_k s(n-k) + G\,u(n) \qquad \text{... (2)}$$

the equation describing relation between speech samples s(n) and excitation u(n).

The experiment has been done after the development of stuttered speech databases in Marathi language. We have used the Linear Predictive Coding (LPC) for the feature extraction purpose of this work. Linear Predictive Coding (LPC) is having features carry certain information of dysfluencies.

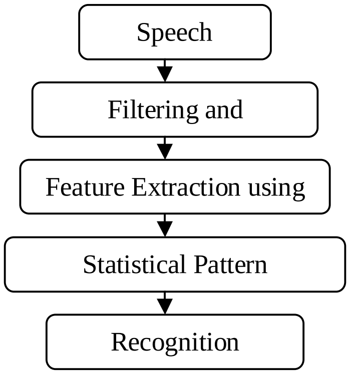

Figure 4: Block Diagram of Recognition System Using LPC

Human perception of speech is based on those frequencies that are strong, that have more power. Hence the vocal tract is often described in terms of its resonant frequencies, also known as formants. These resonances are because of the poles of the vocal tract transfer function. Theses formants are written as Fi; where i stands for the formant number, e.g., F1, F2,Fn. Usually there are four formants in the 0 to 4000Hz range of human speech. It was found that LPC is quit suitable for recognition in speech and that the combination of spectral features to improve the performance of the system using only LPC features. General method of recognition is shown in figure 4.

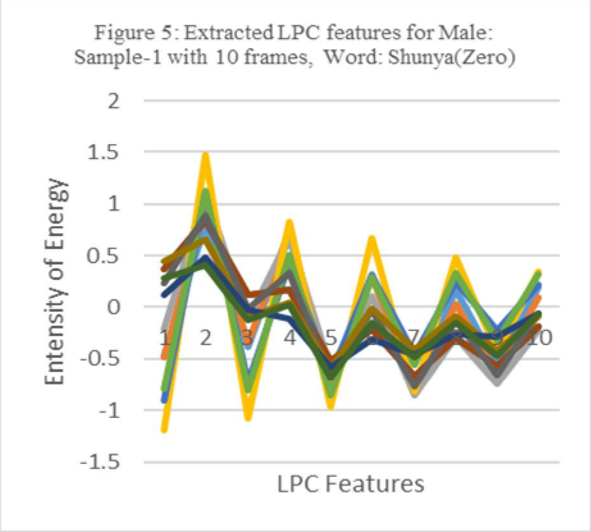

Figure 5: Extracted LPC features for Male: Sample-1 with 10 frames, Word: Shunya(Zero)

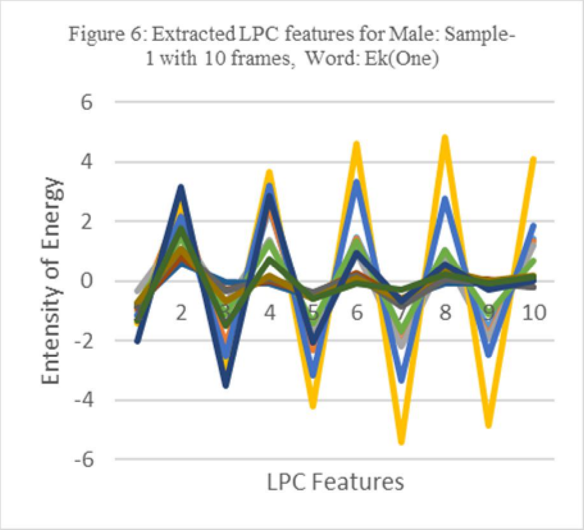

Figure 6: Extracted LPC features for Male: Sample-1 with 10 frames, Word: Ek(One)

A confusion matrix contains information about actual and predicted classification done by a classification system. Performance of such system is commonly evaluated using the data in the matrix. The following table shows the confusion matrix for a two class classifier.

The entries in the confusion matrix have the following meaning in the context of our study:

i) a is the number of correct prediction that an instance is negative.
ii) b is the number of incorrect prediction that an instance is positive.
iii) c is the number of incorrect of prediction that an instance negative and
iv) d is the number of correct prediction that an instance is positive.

Table1 entries in Confusion Matrix

		Predicted	
		Negative	Positive
Actual	Negative	a	b
	Positive	c	d

Several standard terms have been defined for the 2 class matrix:

i) The Accuracy (AC) is the proportion of the total number of predictions that were correct. It is determined using the equations.

$$AC = \frac{a+d}{a+b+c+d}$$

ii) The Recall or True Positive Rate(TP) is the proportion of positive cases that were correctly identified as calculated using the equation.

$$TP = \frac{d}{c+d}$$

iii) The False Positive Rate (FP) is the proportion of negative cases that were incorrectly classified as positive, as calculated using the equation.

$$FP = \frac{b}{a+b}$$

We are calculated the confusion matrix for the numbers using the formula for accuracy (AC) explained in above section. Hence we got the total accuracy calculated is about approximately 75.12 % for MFCC features and 82.23 % for LPC features shown in table 2 and 3 respectively.

Table 2: confusion matrix for MFCC features

	Shunya	Ek	Don	Tin	Char	Pach	Saha	Sath	Ath	Nau	Total No. of Samples for Testing	Recognition Rate
Shunya	16	0	0	0	1	0	0	3	0	0	20	80.00
Ek	1	10	0	2	0	0	0	0	1	2	16	62.50
Don	0	0	15	1	0	1	0	0	0	0	17	88.23
Tin	1	0	0	13	0	2	0	1	0	0	17	76.47
Char	2	0	0	0	8	0	0	0	0	1	11	72.72
Pach	0	3	0	0	0	12	0	0	0	0	15	80.00
Saha	1	0	1	0	3	0	10	0	0	0	15	66.66
Sath	2	0	0	0	0	2	1	11	0	0	16	68.75
Ath	1	2	0	0	0	0	0	0	16	0	19	84.21
Nau	1	0	1	1	1	1	0	0	0	13	18	72.22

Table 3: Confusion matrix for LPC features

	Shunya	Ek	Don	Tin	Char	Pach	Saha	Sath	Ath	Nau	Total No. of Sample for Testing	Recognition Rate
Shunya	15	0	1	0	0	0	0	0	1	0	17	88.23
Ek	0	14	1	0	2	0	0	0	0	0	17	82.35
Don	0	0	11	0	0	2	1	0	0	0	14	78.57
Tin	1	0	0	17	2	0	0	0	0	0	20	85.00
Char	0	1	0	0	12	0	1	0	1	0	15	80.00
Pach	1	0	2	0	0	12	0	0	0	0	15	80.00
Saha	0	1	2	0	0	0	13	0	0	0	16	81.25
Sath	0	1	0	0	0	2	0	13	0	0	16	81.25
Ath	1	0	0	0	0	2	0	0	14	0	17	82.35
Nau	2	0	0	0	1	0	0	0	0	15	18	83.33

VI CONCLUSION

The speech samples were recorded from various speakers who have the stutter disorder from the different hospitals and NGO’s. It contains the zero to nine isolated words spoken by fifty speakers from the age groups of 10 to 50 years. The MFCCs and LPCs are calculated as the speech features for the recognition system. The same features are given as an input to the confusion matrix algorithm which classify the samples. We got the 75 % average recognition rate for MFCC and 82% for LPC. We will implement the other classification techniques to improve the recognition rate.

VII ACKNOWLEDGEMENT

This work is supported by University Grants Commission under the scheme Major Research Project entitled as "Development of Database and Automatic Recognition System for Continuous Marathi Spoken Language for agriculture purpose in Marathwada Region". The authors would also like to thank the speaker who gave therir samples for experiments and the Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad for providing the infrastructure to carry out the research.

REFERENCE

[1] Szabelska, E., Kruczyńska, A.Computer-based speech analysis in stutter. *Applied Computer Science*,2013, 9(2), pp. 34-42.

[2] Czyżewski, A., Kostek, B., Skarżyński, H. Technika komputerowa w audiologii, foniatrii i logopedii. A. L. Akademicka Oficyna Wydawnicza Exit (Ed.), 2002.

[3] Chen, W. Y., Chen, S. H., Lin, C. J. A speech recognition method based on the sequential multi-layer perceptrons. *Journal of Neural Networks*,1996; 9(4), pp.655-669.

[4] Shriberg, E. E. Phonetic consequences of speech disfluency. *SRI INTERNATIONAL MENLO PARK CA*, 1999.

[5] Shrishrimal P. P., Deshmukh R. R.,Waghmare V. B. Development of Isolated Words Speech Database of Marathi words for Agriculture purpose. *Asian Journal Of Computer Science & Information Technology*, 2013; 2(7) pp.217-218.

[6] Ward, D., Sudden onset stuttering in an adult: Neurogenic and psychogenic perspectives. *Journal of Neurolinguistics*,2010;23(5), pp.511-517.

[7] Krishnan, G., Tiwari, S. Revisiting the acquired neurogenic stuttering in the light of developmental stuttering. *Journal of Neurolinguistics*,2011,24(3), pp.383-396.

[8] Ravikumar, K., Reddy, B., Rajagopal, R., Nagaraj, H. Automatic detection of syllable repetition in read speech for objective assessment of stuttered disfluencies. *Proceedings of world academy science, engineering and technology*,2008, 36, pp.270-273.

[9] Subramanian, A., Yairi, E. Identification of traits associated with stuttering. *Journal of communication disorders*, 2006, 39(3), pp.200-216.

[10] Howell, P., Sackin, S., Glenn, K. Development of a Two-Stage Procedure for the Automatic Recognition of Dysfluencies in the Speech of Children Who Stutter. Psychometric Procedures Appropriate for Selection of Training Material for Lexical Dysfluency Classifiers. *Journal of Speech, Language, and Hearing Research*,1997,40(5), pp.1073-1084.

[11] Ravikumar, K. M., Rajagopal, R., Nagaraj, H. C. An approach for objective assessment of stuttered speech using MFCC features. ICGST *International Journal on Digital Signal Processing DSP*, 2009, 9, pp.19-24.

[12] Howell, P., Sackin, S., Au-Yeung, J. Assessment procedures for locating stuttered events,*Proceedings of the Second World Congress on Fluency Disorders*,1998.

[13] Ravikumar, K. M., Rajagopal, R., Nagaraj, H. C. An approach for objective assessment of stuttered speech using MFCC features. ICGST *International Journal on Digital Signal Processing DSP*, 2009, 9, pp.19-24.

[14] Chee, L. S., Ai, O. C., Hariharan, M., Yaacob, S. (2009, November). MFCC based recognition of repetitions and prolongations in stuttered speech using k-NN and LDA. *IEEE Student Conference on Research and Development (SCOReD*), 2009, pp. 146-149.

[15] Ai, O.C., Hariharan M., Yaacob, S., Chee, L.S. Classification of Speech Dysfluencies with MFCC and LPCC features. *Journal of Expert Systems with Applications*,2012,39(2),pp.2157-2165.