International Journal of Computer Engineering & Technology (IJCET)
Volume 8, Issue 5, Sep-Oct 2017, pp. 1–10, Article ID: IJCET_08_05_001
Available online at
http://www.iaeme.com/ijcet/issues.asp?JType=IJCET&VType=8&IType=5
Journal Impact Factor (2016): 9.3590(Calculated by GISI) www.jifactor.com
ISSN Print: 0976-6367 and ISSN Online: 0976–6375

PERFORMANCE EVALUATION OF SPEAKER IDENTIFICATION SYSTEM FOR MALE SPEAKER DURING ADOLESCENCE

Sushma Bahuguna
BCIIT, Chandiwala Estate, Kalkaji, New Delhi

ABSTRACT

A voice mutation commonly refers to the deepening of voice of people from early to late adolescent stages. Males experience a more dramatic voice mutation than females during adolescence. This paper presents experiment for performance evaluation of speaker identification system for male speaker during adolescent stages. A text independent, closed set speaker identification system is implemented using Mel-Frequency Cepstral Coefficients (MFCC) and Dynamic Time warping (DTW) algorithms in MATLAB for training and testing of recorded short sample sentences of native Hindi male speaker during adolescent stages for speaker identification accuracy. The system achieved approx 98% speaker identification accuracy for the speech sentences trained against sample sentence of corresponding age but speaker identification performance was insignificant when trained speech sample sentence was of different age from the age of testing speech sentences. The results show strong influence of voice mutation on the speaker identification and the voice biometrics during adolescent stages may need careful sampling of speech data.

Key word: Speaker Identification System, Mel-Frequency Cepstral Coefficient, Dynamic Time Warping.

Cite this Article: Sushma Bahuguna, Performance Evaluation of Speaker Identification System for Male Speaker During Adolescence. *International Journal of Computer Engineering & Technology*, 8(5), 2017, pp. 1–10.
http://www.iaeme.com/ijcet/issues.asp?JType=IJCET&VType=8&IType=5

1. INTRODUCTION

The applications of speaker identification technology are quite varied and continually growing. Some broad areas include access control, transaction authentication, law enforcement, biometric banking, speech data management and personalization etc. The voice biometrics acts as a snapshot of certain characteristics of voices taken at a particular point in time. Much less attention has been paid to the effect of age related voice changes on speaker identification performance. For voice biometrics, dependence exists between the performance of speaker identification methods and the time lapse between the recording of reference samples and test speech signals [1]. In this paper we have evaluated performance of speaker

identification system based on recoded sample sentences of male native Hindi speaker during adolescent stages. Most of the voice change begins around early adolescence and adult pitch is reached three to four years later but the voice does not stabilize until the early years of adulthood. Males and females have roughly similar vocal pitch before early adolescent stage but during adolescence the male voice typically deepens while female voice deepens only by few notes [2]. The male voice goes through changes usually between the ages of twelve to fifteen. These voice changes have been classified by the scholars on the basis of different stages of development. [3] Classified male voice change as treble voice, cambiata voice, baritone range and developing baritone or emerging adult voice. Treble voice is unchanged voice having full, rich and soprano-like sound. Usually this stage happens to be till the age of eleven or twelve approximately. The cambiata voice is the first stage of change. At this stage the lower tones begins to be stronger and the voice may sound more like an unchanged voice. The second stage of change is baritone range around the age of thirteen when a males voice starts cracking. In this period voice unintentionally and suddenly enters a higher register for a brief period of time and voice crack lasts for only a moment and generally occurs less frequently as a person grows into maturity. Developing baritone or Emerging adult voice is the last stage of voice development when voice gradually moves towards vocal maturity. At this stage qualities such as expansion of range and more vocal consistency begins to appear but characteristics are still not full adult. Usually males attain adult voice at the age of fifteen. In present study twenty sample sentences of same speaker at each age of ten, thirteen and fifteen respectively, were recorded and stored in a database constituting sixty voice samples. Recording was done by electrets microphone in partially sound treated room. These recorded sentences were used for training and testing in speaker identification system.

The organization of the rest of the paper is as follows: Section II is all about the speaker identification system. Sections III and IV highlight details about Mel-Frequency cepstral Coefficient and Dynamic Time Warping. Sections V pertains to System Implementation and Results. Section VI concludes the paper.

2. SPEAKER IDENTIFICATION SYSTEM

Speaker identification system determines the speaker's identity on the basis of selection between a set of known voices and the user does not claim an identity. The training and testing are two distinct operational phases. In training phase the speech from known speaker is acquired to build the model for that speaker which is carried out before system deployment as a part of the system configuration. In testing phase the speech from an unknown utterance is compared against each of the trained speaker models. In closed-set identification, target speaker is assumed to be one of the registered speaker of the group and speaker's profile is stored in a database for which identification is performed. Text-independent speaker identification recognizes an individual without any constraints on what the individual is saying. To evaluate speaker identification accuracy the text-independent, closed-set, speaker identification system is implemented on Sample speech of native male speaker during adolescence.

The input speech samples passes through feature extraction and feature matching stages. Feature extraction plays a very important role in the identification process. This is basically a process of dimension reduction or feature reduction as this process eliminates the irrelevant data present in the given output, leaving behind the important information from the training data for classification. Mel- Frequency Cepstral Coefficients have been used for feature extraction giving a matrix having feature vectors extracted from all the frames.

For feature matching the popular method Dynamic Time warping based on dynamic programming has been used to compensate for speaking rate variability in template based system. Figure 1 depicts training and testing framework of the system [4].

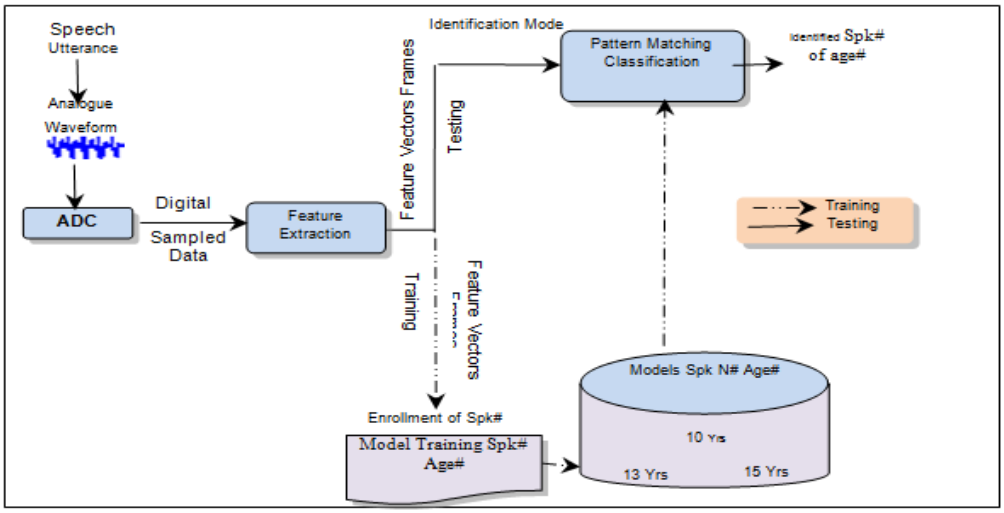

Figure 1 Training and testing framework

3. MEL- FREQUENCY CEPSTRAL COEFFICIENT

Mel-frequency Cepstral Coefficient (MFCC) is based on the human hearing characteristics, which uses a nonlinear frequency unit to stimulate the human auditory system. Low frequency component of speech signal carry more important information compared to high frequency components. MFCC are coefficients that frame the Mel-Frequency Cepstrum which is a representation of the short term power spectrum of sound. A detailed description of feature extraction steps can be found in [5].
The computational steps are described as follows.

3.1. Pre-emphasis

Pre-emphasis refers the passing of signal through filter to emphasize higher frequencies that increases the energy of signal at higher frequency which makes information from these higher formants available to the model.

3.2. Framing

Speech signals obtained from analog to digital conversion were segmented into small frame with the length 25ms and voice signal was divided into frames of N samples (N=256).

3.3. Windowing

Window function is needed to avoid unnatural discontinuities in the speech segment and distortion in the underlying spectrum. Hamming Window is used as window shape by considering the next block in feature extraction processing chain and integrates all the closest frequency lines.

3.4. Fourier Transform

Fourier transform converts the convolution of the glottal pulse and the vocal tract impulse response from time domain into multiplication in the frequency domain [5]**.** The fast Fourier Transform (FFT) is performed to obtain complex spectral values from each frame. We

obtained 256 complex spectral values uniformly spaced from 0 to *Fs/2*, applying a 512-point FFT, where *Fs* represent sampling frequency.

3.5. Mel-Spaced Filter Bank Values

The *Mel-frequency* scale is linear frequency spacing below 1000Hz and a logarithmic spacing above 1000Hz. *Mels* for a given frequency *f* in Hertz can be computed as:

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

Complex spectral values (N = 256) represent too much spectral information and by smoothing of spectrum to only *K = 2*0, or so, more efficient representation may be achieved. Subjective spectrum can be stimulated by using a filter bank with one filter for each desired mel-frequency component. The filter bank has a triangular band pass frequency response and constant Mel-frequency interval determines the spacing and bandwidth by a constant mel-frequency interval. Filter bank values are obtained by Cross-wise multiplication of the *N* FFT magnitude coefficients by the *K* triangular filter bank weighting function and then accumulating the results from each triangle. The centers of the triangle filter banks are spaced according to the mel scale. Mel filter banks are non-uniformly spaced on the frequency axis, so we have more filters in the low frequency regions and less number of filters in high frequency regions. Mel spectrum can be computed as;

$$\tilde{S}(l) = \sum_{k=0}^{n/2} S(k) M_l(k) \quad (2)$$

Where:

$\tilde{S}(l)$: Mel spectrum,

$S(k)$: Original spectrum.

$M_l(k)$: Mel filter bank.

l=0, 1... *l* - 1, where *l* is the total number of Mel filter banks

$n/2$: Half FFT size.

Now, we proceed to the next stage to get the cepstrum or the mel-frequency cepstrum coefficients.

3.6. Cepstral Analysis

In the final step log Mel spectrum is converted back in to time domain using Discrete Cosine Transform (DCT) resulting Mel frequency cepstrum coefficients. The DCT does not needs complex arithmetic and implements the same function as the FFT more efficiently by taking advantage of the redundancy in a real signal. The following equation was used to calculate MFCC.

$$C_{\tilde{n}} = \sum_{k=1}^{K} \left(\log \tilde{S}_k\right) \left[n\left(k - \frac{1}{2}\right)\frac{\pi}{K}\right], \quad (3)$$

Where n=1, 2….K

The number of Mel cepstrum coefficients is chosen 20. By this process a set of Mel-frequency cepstrum coefficients is computed for each speech frame of 25 ms, which is called acoustic vector. These acoustic vectors are used to represent and recognize the voice

characteristic of the speaker and each input utterance is transformed into a sequence of acoustic vectors.

4. DYNAMIC TIME WARPING

Dynamic Time Warping Algorithm (DTW) calculates an optimal warping path between two time series [6]. The DTW algorithm is designed to exploit some observations about the likely solution to make the comparison between sequences more efficient. Theoretically, the major optimizations to the DTW algorithm arise from observations on the nature of good paths through the grid. These are outlined as Monotonic condition, Continuity condition, Boundary condition, Adjustment window condition and Slope constraint condition [9].

The algorithm calculates both warping path values between the two series and the distance between them. The technique is also used to find the optimal alignment between two times series if one time series may be "warped" non-linearly by stretching or shrinking it along its time axis. This warping between two time series can be used to determine the similarity between the two time series or to find corresponding regions between the two time series. The principle of DTW is to compare two dynamic patterns and measure its similarity by calculating a minimum distance between them. The classic DTW is computed as below [9].

Assume, we have two time series Q and C, of length *n* and *m* respectively, where:

$$Q = q_1, q_{2,} \ldots\ldots, q_i, \ldots\ldots q_n \quad (4)$$

$$C = c_1, c_{2,} \ldots\ldots, c_j, \ldots\ldots c_m \quad (5)$$

To align two sequences using DTW, an *n*-by-*m* matrix, where the (i^{th}, j^{th}) element of the matrix contains the distance $d(q_i, c_j)$ between the two points q_i and c_j is constructed. The absolute distance between the values of two sequences is then calculated using the Euclidean distance computation:

$$d(q_i, c_j) = (q_i - c_j)^2 \quad (6)$$

Each matrix element (*i*, *j*) corresponds to the alignment between the points q_i and c_j. The accumulated distance is measured by:

$$D(i,j) = min[D(i-1, j-1), D(i-1, j), D(i, j-1)] + d(i,j) \quad (7)$$

This is shown in [8]. The search for the minimum distance path can be done in polynomial time$P(t)$, using equation below [10].

$$P(t) = O[N^2 V] \quad (8)$$

Where, N is the length of the sequence, and V is the number of templates to be considered.

5. SYSTEM IMPLEMENTATION AND RESULTS

A pre-emphasis filter with Co-efficient 0.97 was applied from each 25ms frame at every 10ms and generated 20 MFCC. The feature vectors consisting of MFCC from each frame were subjected to DTW and resulting set of vectors was stored in the speaker database. Two speech samples (recorded sentences) from each age group were used for training. In testing mode sixty speech sentences were given as a voice input against trained data of varying age distinctly and the system made a decision about the speaker's identity.

The method used DTW algorithm as pattern classification technique for the comparison of two time dependent sequences. These are feature sequences sampled at equidistant point in time. Distance measure also called cost measure is calculated from these time series which is required for evaluation. For each comparison, the distance measure was calculated. Tables 1, 2 and 3 and corresponding figures 2, 4 and 6 show the difference of cost measure of each tested sample against the Distance measure of trained sample. A lower distance measure indicates a higher similarity.

Table 1 Distance measure between train data template of speech sample of age 10 and other test samples in database of varying age

Test against speaker's train data of age 10 with a cut-off of train sample 6.081900541						
S.No	Distance measure/Local cost measure					
	Age 10		**Age 13**		**Age 15**	
	Cuttoff	**Local cost measure**	**Cuttoff**	**Local cost measure**	**Cuttoff**	**Local cost measure**
1	2.0273	-4.0546	8.187	2.10489946	15.0584	8.9764995
2	2.0273	-4.0546	7.119	1.03709946	14.5052	8.4232995
3	4.7792	-1.3027	8.397	2.31459946	13.5123	7.4303995
4	4.5849	-1.4970	9.428	3.34619946	15.008	8.9260995
5	6.7298	0.6478	7.948	1.86609946	12.8985	6.8165995
6	4.4724	-1.6095	8.488	2.40559946	16.2455	10.163599
7	4.3225	-1.7594	9.791	3.70949946	15.3978	9.3158995
8	5.0981	-0.9838	8.637	2.55529946	13.1749	7.0929995
9	4.6797	-1.4022	7.31	1.22849946	18.5057	12.423799
10	4.8504	-1.2315	8.802	2.71989946	16.1727	10.090799
11	4.8182	-1.2637	9.565	3.48299946	14.3518	8.2698995
12	5.5781	-0.5038	7.445	1.36279946	14.1008	8.0188995
13	4.9217	-1.1602	10.08	3.99709946	14.9055	8.8235995
14	5.3734	-0.7085	9.336	3.25379946	17.4165	11.334599
15	4.827	-1.2549	10.14	4.06159946	14.1606	8.0786995
16	4.3919	-1.6900	9.381	3.29949946	17.4559	11.373999
17	5.6402	-0.4417	7.812	1.73049946	12.4559	6.3739995
18	5.7348	-0.3471	7.272	1.19009946	16.3306	10.248699
19	4.6206	-1.4613	8.536	2.45409946	13.2231	7.1411995
20	7.2939	1.2119	9.399	3.31669946	16.0152	9.9332995

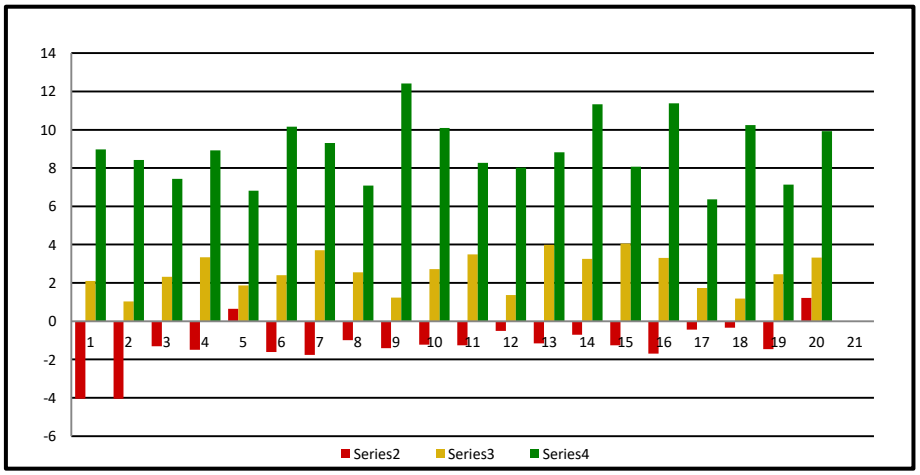

Figure 2 Histogram of distance measure between train data template of speech sample of age 10 and other test samples of varying age

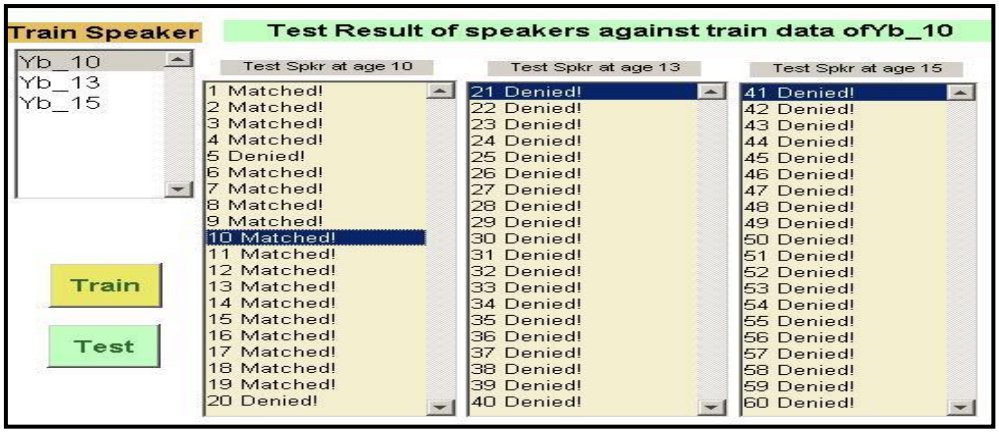

Figure 3 Speaker identification results against train sentences of age 10

Table 2 Distance measure between train data template of speech sample of age 13 and other test samples in database of varying age

Test against speaker's train data of age 13 with a cut-off of train sample 3.968388766						
S.No	Distance measure/Local cost measure					
	Age 10		**Age 13**		**Age 15**	
	Cut-off	**Local cost measure**	**Cut-off**	**Local cost measure**	**Cuttoff**	**Local cost measure**
1	9.095	5.1266112	2.7349	-1.2334888	4.8558	0.8874112
2	10.5489	6.5805112	3.5527	-0.4156888	4.4314	0.4630112
3	9.7796	5.8112112	3.0876	-0.8807888	4.995	1.0266112
4	9.1881	5.2197112	2.9135	-1.0548888	4.4353	0.4669112
5	9.0125	5.0441112	3.1625	-0.8058888	3.9168	-0.0515888
6	8.4492	4.4808112	2.7493	-1.2190888	5.2033	1.2349112
7	7.8521	3.8837112	3.1771	-0.7912888	5.3215	1.3531112
8	8.7597	4.7913112	3.3484	-0.6199888	4.2021	0.2337112
9	6.7519	2.7835112	3.8559	-0.1124888	6.1728	2.2044112
10	8.2638	4.2954112	2.774	-1.1943888	4.7042	0.7358112
11	7.5126	3.5442112	1.3228	-2.6455888	4.1746	0.2062112
12	7.9083	3.9399112	3.4826	-0.4857888	4.4609	0.4925112
13	9.3602	5.3918112	1.3228	-2.6455888	4.2261	0.2577112
14	7.4661	3.4977112	3.1948	-0.7735888	5.7739	1.8055112
15	7.6304	3.6620112	2.3236	-1.6447888	4.5101	0.5417112
16	8.3546	4.3862112	2.5702	-1.3981888	6.1085	2.1401112
17	7.2008	3.2324112	3.2015	-0.7668888	3.7033	-0.2650888
18	9.9643	5.9959112	3.3476	-0.6207888	5.7828	1.8144112
19	7.3083	3.3399112	3.1706	-0.7977888	4.3535	0.3851112
20	10.8945	6.9261112	4.06	0.0916112	5.276	1.3076112

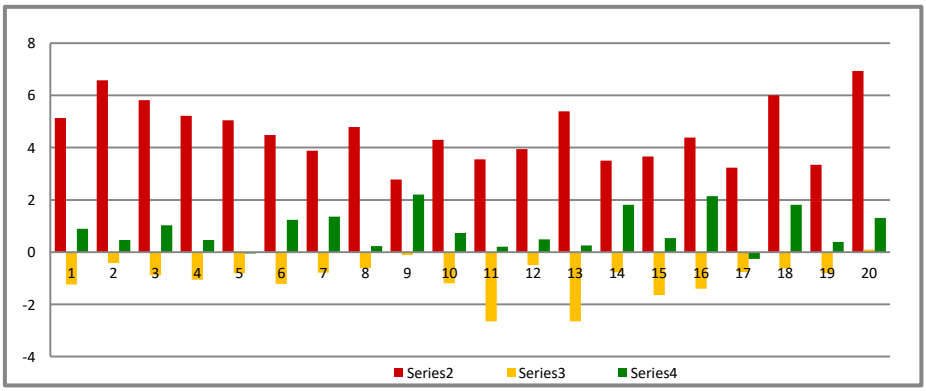

Figure 4 Histogram of distance measure between train data template of speech sample of age 13 and other test samples of varying age

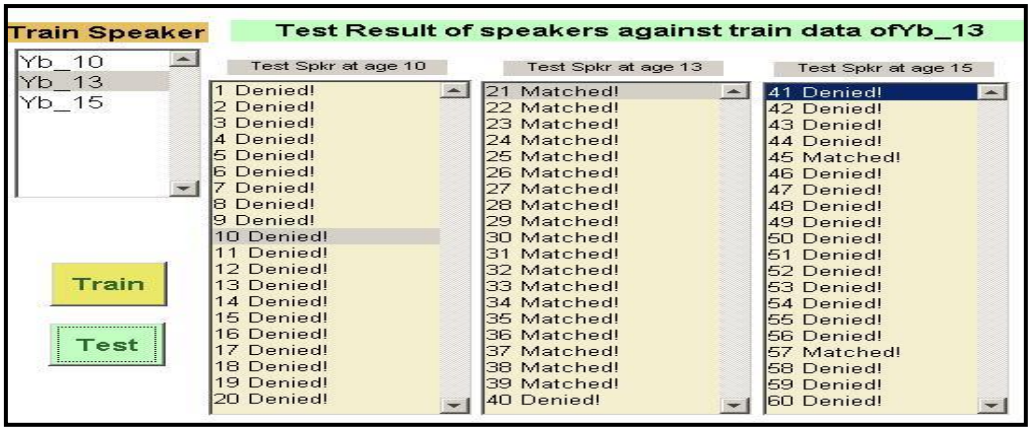

Figure 5 Speaker identification results against train speech sentences of age 13

Table 3 Distance measure between train data template of speech sample of age 15 and other test samples in database of varying age

Test against speaker's train data of age 15 with a cut-off of train sample 2.573607326						
S.No	Distance measure/Local cost measure					
	Age 10		**Age 13**		**Age 15**	
	Cut-off	**Local cost measure**	**Cut-off**	**Local cost measure**	**Cut-off**	**Local cost measure**
1	12.5283	9.9546927	4.8331	2.2594927	0.8579	-1.7157073
2	15.7532	13.179593	5.8532	3.2795927	1.2873	-1.2863073
3	13.1542	10.580593	5.6246	3.0509927	1.3144	-1.2592073
4	12.5894	10.015793	5.126	2.5523927	1.4265	-1.1471073
5	11.3136	8.7399927	5.2691	2.6954927	1.2836	-1.2900073
6	10.5153	7.9416927	4.4227	1.8490927	1.495	-1.0786073
7	11.9924	9.4187927	5.3506	2.7769927	1.1878	-1.3858073
8	12.6176	10.043993	4.7808	2.2071927	1.3512	-1.2224073
9	9.6889	7.1152927	5.8522	3.2785927	1.3271	-1.2465073
10	10.7734	8.1997927	4.8789	2.3052927	1.3341	-1.2395073
11	10.6146	8.0409927	4.782	2.2083927	1.199	-1.3746073
12	10.3156	7.7419927	3.1582	0.5845927	1.3916	-1.1820073
13	13.2661	10.692493	4.4273	1.8536927	1.3123	-1.2613073
14	11.1795	8.6058927	4.4639	1.8902927	1.1959	-1.3777073
15	11.0301	8.4564927	4.2136	1.6399927	1.2945	-1.2791073
16	12.4344	9.8607927	5.1117	2.5380927	1.5763	-0.9973073
17	10.0473	7.4736927	3.9429	1.3692927	1.3575	-1.2161073
18	13.255	10.681393	4.9499	2.3762927	1.6169	-0.9567073
19	10.6743	8.1006927	5.1541	2.5804927	0.8579	-1.7157073
20	15.6126	13.038993	5.9292	3.3555927	1.4632	-1.1104073

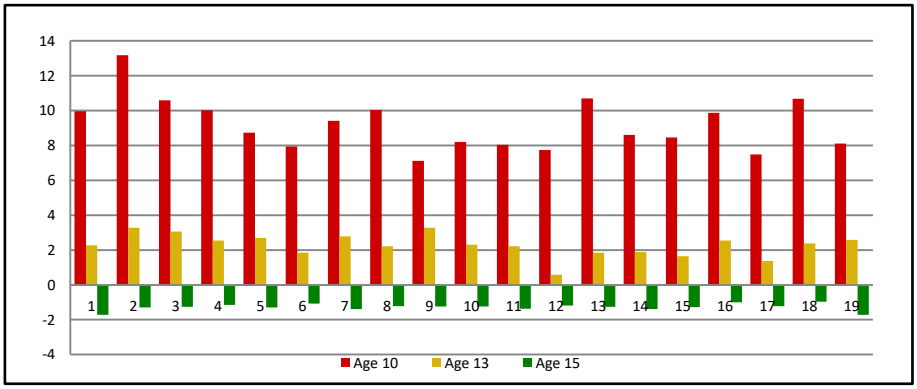

Figure 6 Histogram of distance measure between train data template of speech sample of age 15 and other test samples of varying age

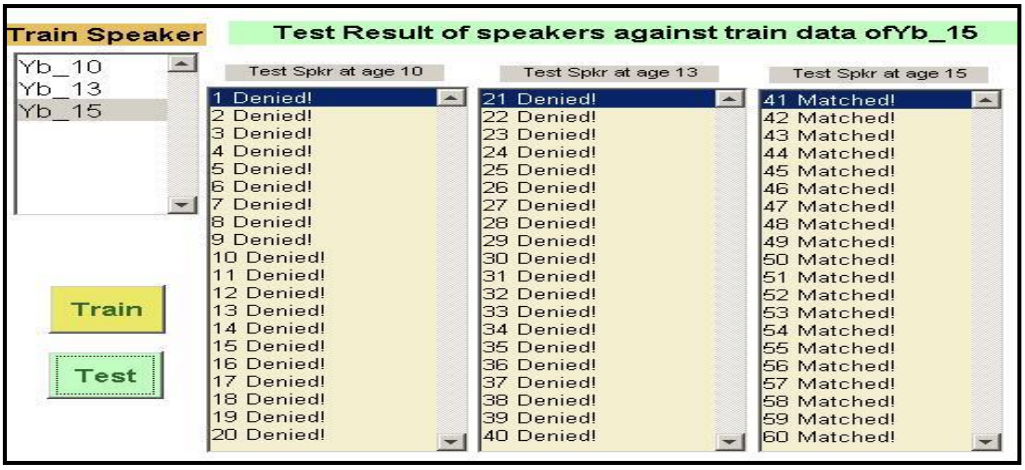

Figure 7 Speaker identification results against train sentences of age 15

The results shown in the figures 3, 5 and 7 reveal a strong degradation of speaker identification performance when test speech sentences are of different age from the age of train speech sentences. Speaker identification accuracy of 98% was achieved with testing speech sentences against trained speech of same age whereas speaker identification accuracy was deteriorated and was insignificant when the test speech in consideration was from a different age of same speaker as against the train speech sample. It is found that biometric data have been deviated from enrollment at different stages of voice mutation and performance deterioration of identification system is resulted. One way of addressing the problem is to renew the trained data within a certain periodicity. However, this is time consuming and resource intensive. On the other hand if train data is not updated, an individual may not be recognized.

6. CONCLUSIONS

In this paper we attempted to evaluate performance of speaker identification system implemented using MFCC and DTW algorithms for short Hindi speech sentences of male speaker during adolescent stages. The results achieved speaker identification accuracy of 98% with the trained and test speech samples of same age whereas the performance of the speaker identification system was insignificant when trained and testing speech sentences were of different ages. So, voice mutation has effectively influenced the speaker identification system and performance of speaker identification system might not be at usable level for particular task is case of training and testing speech sentences are of different ages during adolescent stages. The results are limited to recognize the speaker based on the devices used for recording the corresponding speech and there is lack of ample databases. Further study with other modeling approaches within speaker identification community and ample databases is necessary to assess effects of voice mutation on speaker identification system.

REFERENCES

[1] Yuri Matveev "The Problem of Voice Template Aging in Speaker Recognition System" : in M. Zelezny et al. (eds.) : SPECOM 2013, LNAI 8113, pp. 345-353. Springer International Publishing, Switzerland 2013.

[2] "Voice changing", The Lowdown. Retrieved 7 January 2012.

[3] Collins, Don L "The cambiata concept", The Cambiata Press. http://www.cambiatapress.com /CVMIA/ TheCambiataConcept2.html

[4] Sushma Bahuguna,Y. P. Raiwani "Emotion Based Text Independent Speaker Identification Using Gaussian Mixture Model for Hindi Speech" IFRSA's International Journal Of Computing, Volume: 5, issue 2, April 2015

[5] Reynolds, D. A. and Rose, R. C. "Robust Text Independent Speaker Identification using Gaussion Mixture Speaker Models", IEEE Trans, Speech Audio Process (3) 1995, pp.72-83.

[6] Sushma Bahuguna and Y. P. Raiwani, "A Study of Acoustic Features Pattern of Emotion Expression for Hindi Speech", International Journal of Computer Engineering & Technology (IJCET) 2013.Volume:4, Issue: 6, Pages: 16-24.

[7] J. Walker and P. Murphy, "A Review of Glottal aveform Analysis": Y. Stylianou, M. Faundez-Zanuy, A. Esposito (Eds): WNSP 2005, LNCS 4391, pp 1-21, 2007.

[8] Hiroaki Sakoe and Seibi Chiba, "Dynamic Programming Algorithm Optimization for Spoken Word Recognition", IEEE Transaction on Acoustic Speech and Signal Processing, February 1978.

[9] Clarence Goh Kok Leon, "Robust Computer Voice Recognition using Improves MFCC Algorithm", Politeknik Seberang Perai, Malaysia.

[10] Toni M. Rath and R. Manmatha, "Word Image Matching Using Dynamic Time Warping" , University of Massachusetts, Amherst, MA 01003 2002 .

[11] Anagha S. Bawaskar and Prabhakar N. Kota. Comparative Study of LPCC and Fused Mel Feature Sets For Speaker Identification Using GMM - UBM, International Journal of Electronics and Communication Engineering & Technology, 6 (9), 2015, pp. 82 - 96.

[12] Viplav Gautam, Saurabh Sharma,Swapnil Gautam and Gaurav Sharma, Identification a nd Verification of Speaker Using Mel Frequency Cepstral Coefficient, International Journal of Electronics and Communication Engineering & Technology, 3 (2), 2012, pp. 413-423.