IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 39, No. 2, April, 2022, pp. 711-716
Journal homepage: http://iieta.org/journals/ts

Speaker Identification Based on Physical Variation of Speech Signal

Durgesh Nandan[1*], Mahesh Kumar Singh[2], Sanjeev Kumar[2], Harendra Kumar Yadav[3]

[1] Department of Electronics & Telecommunication, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune 412115, India
[2] Accendere Knowledge Management Services, CL Educate Ltd, New Delhi 110044, India
[3] Department of EEE, Raj Kumar Goel Institute of Technology, Ghaziabad 201003, India

Corresponding Author Email: durgeshnandano51@gmail.com

https://doi.org/10.18280/ts.390235

Received: 22 November 2019
Accepted: 18 December 2021

Keywords:
acoustic feature, classifier, disguised voice, speaker identification

ABSTRACT

Speaker identification for the speech signal processing request, determining the speaker is a challenge due to physical variation. This paper emphasizes a new algorithm based on acoustic feature analysis of text-dependent speech. In this proposed method text-dependent speech changed by ten physical variation methods. Acoustic feature of all types of voice is calculated by its arithmetical correlation coefficients and mean value. The audio characteristic is calculated with Mel-frequency cepstrum coefficient (MFCC), its derivatives and double derivatives. An acoustic characteristic is analysed by using normal voice and changed voice by different speakers, the mixed data used for test and training purpose. Passing all the training and test data through the various classifiers based on identification system. Speaker identification efficiency results are calculated from the different classifier.

1. INTRODUCTION

Speech communication is a simplest way of communication. In speech signal processing speech signal delivered the information about the message or text being spoken [1]. Speaker identification is a biometric sensation under speech signal processing application that identifies the speaker [2]. Speech signal not only delivers the information about the message but also about the speaker identity, speaker health, emotion, age, physical condition, gender information etc. So that speech signal has an independent role in research in the application of speech signal processing [3]. There is a dissimilar area of research in speech signal dispensation like voice recognition, broadcaster identification, and speech enhancement. Identification of speaker is to identify an unknown speaker depended on his voice or dialect [4]. Research in speaker identification is the wide area of application such that securities control, forensic casework, confidential information etc. [5]. In this study a method for the identification of textual speakers by physical variations of the voice are proposed [6]. The language ID based on text-related, necessary speaker says the ordinary phrase or message [7]. In considerable research for voice disguise by two methods one of them electronic disguised or electronic variation method second non-electronic disguised or physical variation method of voice [8]. In electronic disguised method voice disguised by electronically but in case of non-electronic disguised or physical variation method depends on the physical condition of speakers [9]. This proposed approach of identifying a speaker based on non-electronically disguised text-dependent or physical variation method of speech. In this work considered as speakers normal voice (NV) disguised by ten method fast speech (FS), slow speech (SS), raised pitch (RP), lower pitch (LP), mask on mouth (MM), bite off block (pencil) (BB), chew gum (CG), pinched nostril (PN), object in mouth (OM), foreign ascents (FA). For speaker identification, it is most important to extract the feature of speaker's voice [10]. There is a different method of feature extraction like MFCC, linear predictive cepstral coefficients, vector quantization etc. In this proposed method used MFCC feature extraction technique [11]. In existing research, that is based on forensic automatic speaker recognition (FASR) technique [12]. Here proposed a method for speaker identification by auditory feature analysis in addition to its arithmetical mean values and correlation coefficients. Analysed the existing model and designed using feature extraction technique [13]. For speaker identification selected our proposed work 20 candidates ranging between 20 to 25 years age. They converse the general text or message during a regular influence as well as a physical variation of their voice [14]. There are two type of feature extractions technique first one MFCC as discussed in this manuscript second feature extraction technique are Linear predictive cepstral coefficients (LPCC) techniques. The LPCC performances degrade in noisy environment. So, MFCC is better than LPCC.

By using the same approach like normal voice take the voice sample of all twenty speakers by ten types of physical variation method [15]. For identification of speaker from its disguised voice or physical variation of voice is identical. By evaluating all these things, this task segmented into two stages training stage as well as the testing stage [16]. By this, all the speakers statistical mean and correlation coefficients distributed at this stage with specific details [17]. With all the acoustic features details instruction and test data approved from side to side the classifier algorithms that is dependent on statisticals correlation coefficients and mean of regular speech feature and physical variation of voice [18].

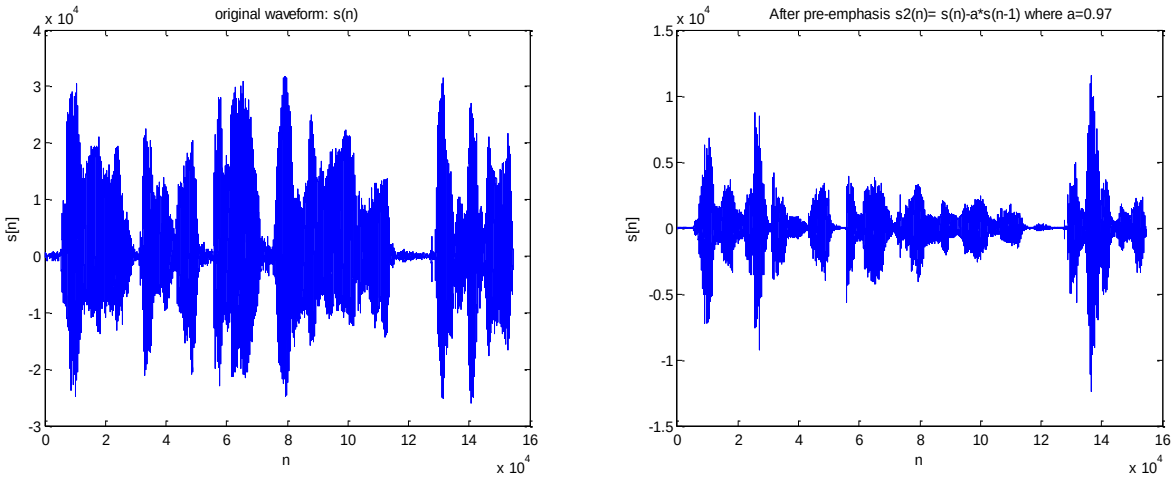
The feature exaction of speech signal is to amplified the spectrum where hearing becomes sensitive is processed in the front-end processing of speech signals. As a result, it only

amplifies the high-frequency portion of the spectrum because the lower frequency section has sufficient energy. As a result, pre-emphasis is crucial in spectrum analysis [19]. It equalises the natural slope before the spectral research, boosting the accuracy of the analysis. Figure 1 (a, b) shown before and after pre-emphasis, a time domains speech representations of a speech sentences were shown. The pre-emphasis filters are used as high-pass filter of first order. The equations of the filter in time domains are expressed: the input signal s[n] as well as pre-emphasising coefficients. The range was taken 0.9 to 1.0, the speech signal is expressed in time domain equation.

Because the speech signals are continually changing, it is assumed that it does not change considerably on short time scales for the reason of simplicity. As a result, the speech signal is divided into the multiple frames of 20-40ms durations, as shown in Figure 2. If the frames are substantially smaller, enough sample is required to achieve a reliable spectrum estimate; nevertheless, if the frame is much larger than typical, the signal changes greatly across the frame.

Windowing takes advantage of the windows to make the signal equal to zeros at both from the starting to the last. It decreases signal discontinuities and spectrum distortion. Frame size and frame shift affect the speech produced by each window frame. The numbers of millisecond between the left ends of consecutives window are known as frame size, whereas the numbers of millisecond between the right end of consecutives windows is known as frame shift. The original signal s[n] is multiplied with the windows w[n] at time n to produce a resultant signal.

Recognizing speakers involves being able to receive a voice signal, recognise a speaker in a voice signal, and then recognise the speaker again. Waveform transformation is achieved by using a parametric representation to reduce the data rate while increasing the amount of processing and analysis that can be done. Good classification, in turn, comes from exceptional qualities [20]. The MFCC, a method of speech extraction, was described in this chapter. It has been shown in several contexts that these methods work, and are both effective and practical [21]. Previously described processes have been tweaked by researchers to increase their tolerance for noise, enhance their strength, and reduce the amount of time they take. For another thing, the issue of where to apply the technology is not settled [22].



a) b)

Figure 1. Speech sample waveforms a). Prior to pre-emphasis b). Post pre-emphasis [19]

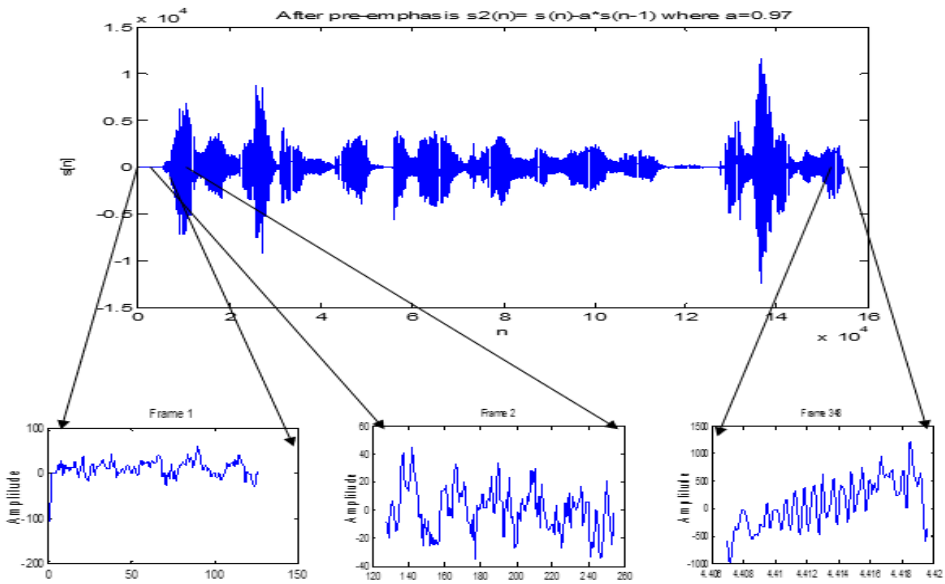


Figure 2. Framing of speech signal [19]

2. METHODLOGY

The normal voice waveform of a text message is seen in Figure 3. To design a database for training and testing purposes, we used the Audacity tool (http://audacity.sourceforge.net) to record twenty candidates' voices. The voice part provides ten varieties of the varied physical voice of the standard spoken text. Speech's spectral palette can be viewed as the MFCC spectrum.

For the frames analysis, the initial 12 cepstral coefficients are taken for the examination. The relationship between Cepstral coefficients and frequency bands is recognised.

The proposed method of speaker identification system shown in block diagram in Figure 4. Based on text-based speaker identification system. In this proposed model normal text speech signal as well as disguised speech applied to MFCC base feature extraction. Feature are extracted by MFCC, ΔMFCC, ΔΔMFCC coefficients. In this proposed structure main concern about its mean and correlation coefficients.

By this method calculate the auditory characteristic of standard voice compare with an acoustic feature of each and every one category of masquerading accent, for convenient these are taken 20 candidate's voice sample for this proposed model. For normal and disguised voices, the statistical mean and correlation values were determined, and results were kept in the database. A random combination of acoustic elements was taken from a dressed voice and a normal voice for this purpose. Set the text count to 13 and the training count to 27. The classifier uses the corresponding attributes, like those used in functional algorithms, to establish the speaker identity. This classifier employs a model to match various features to determine the voice of a speaker.

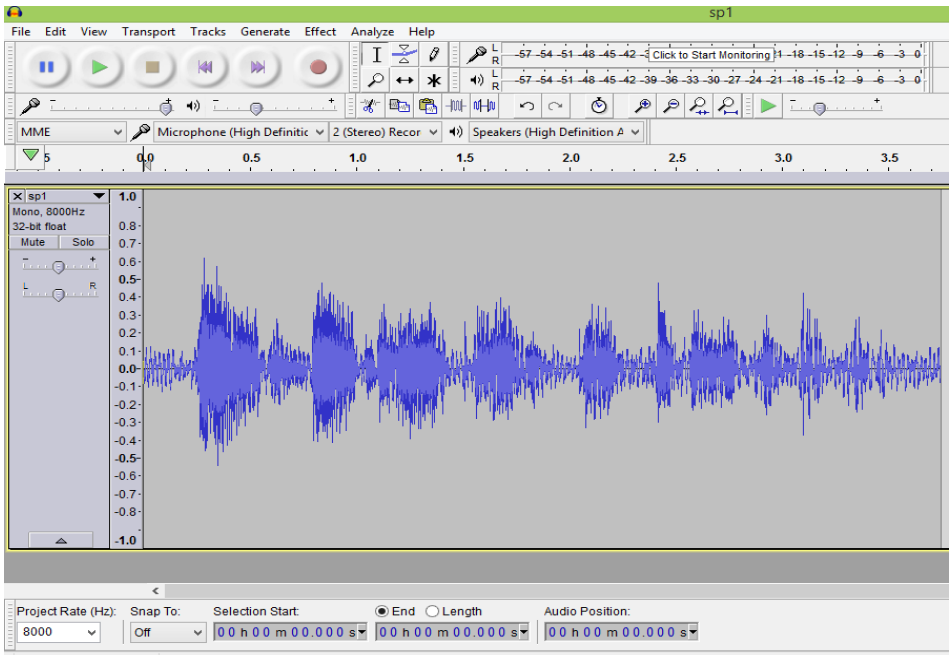


Figure 3. Normal voice sample

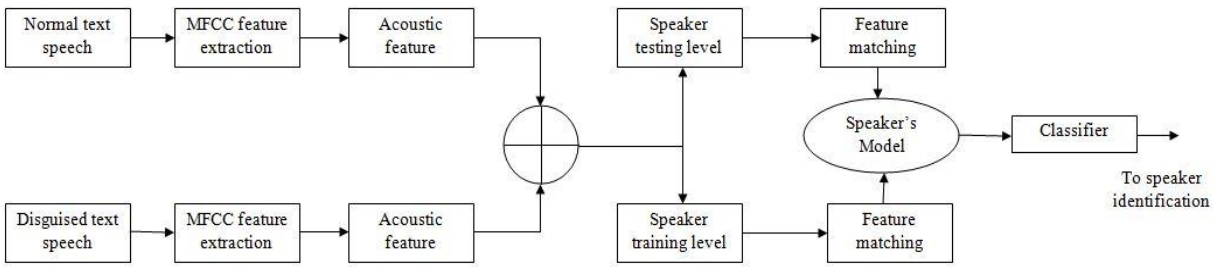


Figure 4. Proposed block diagram of text-based speaker identification system

3. STATISTICAL ANALYSIS USING MFCC

For Statistical analysis, first pre-emphasized technique used as a high-pass filter. Short-term speech features are represented by MFCCs, which are taken from their spectrum. It's frequently utilised in acoustic signal analysis. The MFCCs' scheme is well-known and well approved. It describes the relationship between the humans ear critical bandwidth as well as the Mel-frequency scale, in which filters are placed logarithmically above 1000 Hz and linearly below 1000 Hz. To eliminate discontinuities, The signals is first segmented into frame, and then each frame is multiplied by a Hamming window. The DFT of each frame are then calculates, and the log of the amplitude spectrums are measured. Finally, the speech spectra are smoothed using the discrete cosine transform (DCT), which aids in the generation of cepstral feature vectors for each frame. An analogous speech waveform is first converted to digital speech using MFCC. Its

sampling rate is larger than or equivalent to 10000 Hz to reduce aliasing problems in speech.

The process of MFCC generation is shown in Figure 5. The contribution voice signal v [n] as well as its postponement translation of voice gesture v [n-1] and used μ pre-emphasized coefficient. This Pre-emphasized coefficient be used range from 0.9 to 1.0, after this the filter equation can be articulated the same as:

$$y[n] = v[n] - \mu v[n-1] \tag{1}$$

By using the Hamming windowing techniques for dropping the supernatural dissimilarities and alteration of the voice signal and continue the non-interruptions of the voice gesture of every enclose, take 16ms-20ms where v [n] be the voice gesture and $w[n]$ is function of window [1, 15].

$$p[n] = v[n] * w[n] \tag{2}$$

where, $w[n]$ is the window functions.

$$w[n] = \begin{cases} 0.54 - 0.46\,cos\left(\frac{2\pi n}{L}\right) & 0 < L < n-1 \\ 0 & otherwise \end{cases} \tag{3}$$

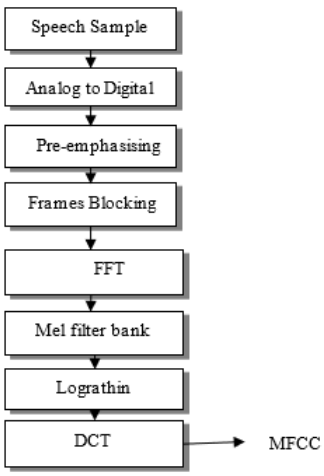


Figure 5. MFCC feature extraction

Frequency domain analysis of the voice signal after windowing:

$$P_i(k) = \sum_{n=1}^{N} p_i(n) e^{-j2\pi kn/N} \qquad 1 \le k \le N \tag{4}$$

Statistical moments extract acoustic vectors with similar durations as the number of MFCC characteristics changes with time. Here, x(n) is an N-frame speech signal whose MFCC vectors is vij, where j denotes the feature coefficients and it's denoted by the frame numbers, and it may be represented as:

$$V_j = \{v1j, v2j \dots, vNj\}; \tag{5}$$

$$where\ j = 1,2 \dots \dots L \tag{6}$$

In this manuscript different two methods of statistical coefficients are derived. First the means *Ej* of each MFCC's features coefficient of *Vj* is extracted. It is derived the correlation coefficients $CR_{jj'}$ of the different MFCCs features of V_j and V_j. The procedure is described in Eq. (7) and Eq. (8) respectively.

$$E_j = E(V_j);\ j = 1,2, \dots \dots L \tag{7}$$

$$CR_{jj'} = \frac{cov(V_j, V_{j'})}{\sqrt{var(V_j)}\sqrt{var(V_{j'})}};\ 1 \le j < j \tag{8}$$

Similarly, the mathematical moment of the delta MFCC ($K_{\Delta MFCC}$) and ΔΔMFCC ($K_{\Delta\Delta MFCC}$) are determined. At last, by combination of K_{MFCC}, $K_{\Delta MFCC}$ and $K_{\Delta\Delta MFCC}$ a statistical moments 'K' is created, which is given as:

$$K = [K_{MFCC}, K_{\Delta MFCC}, K_{\Delta\Delta MFCC}] \tag{9}$$

The most important classifier used for statically classification technique. The speaker's classification work can be categorized addicted to two types: test-dependent and training-dependent. Altered type of a classifier that be being used in the analysis are SVM classifier DT, NB, LDA and LR. The main benefit of these cepstral coefficients are derived from MFCC tend to be uncorrelated, making the systems easier to understand. Each frames MFCC features solely represent the power spectral envelope of the following frame. Derivative coefficients are used to extract speech-specific information in the dynamics. Each delta feature represents variance between frame, where each double delta feature represents variations between frame in the delta feature characteristics that follow.

4. RESULT

Using methodology performance of this proposed method compared with the existing methodology by taking the fundamental identification techniques of speaker in forensic automatic speaker recognition (FASR) method. The mean speaker recognition rate for speaker recognition was calculated using existing methodology. MFCC is used to calculate the statistical coefficient mean and correlation coefficient in the suggested technique. techniques that are shown in Figure 6.

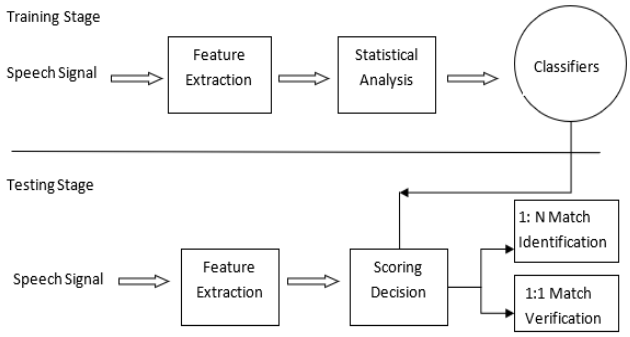


Figure 6. Speaker recognition system

The method is based on the typical speech and physical influence and uses statistical mean. To assemble the database, 20 volunteers' voices were recorded using an audacity application. This database features the typical voice and 10 different varieties of vibrated physical voice which are only used for common text as shown in Figure 7.

Table 1. Mean rate of standard speech and disguised by 10 types

Speech Condition	N.V.	R.P.	L.P.	F.S.	S.S.	P.N.	M.M.	W.H.	F.A.	B.B.	O.M.
FASR Method [1, 2] Existing	0.94	0.38	0.80	0.88	0.93	0.85	0.45	0.56	0.95	0.92	0.77
MFCC Method Proposed	0.93	0.96	0.93	0.96	0.93	0.93	0.93	0.90	0.95	0.94	0.95

For speaker identification for proposed method by MFCC feature extraction method is better than the FASR existing model which is shown the compared result of mean shown in Table 1.

The mean is the similarity index of twenty speakers by ten types of physical variation voice with their normal voice. After comparing both models existing and the proposed model it is shown with the intention of the projected method is improved than the existing model. In the proposed model for the mean value of RP is 0.96 but in the existing method is 0.38 and in MM proposed method is 0.93 but in case of the existing model is 0.45, at end overall it is shown that in by comparing the table, feature extraction by MFCC is better than FASR existing model. By this method, it be proved so as to the proposed technique is superior recognition rate than existing technique. Another feature is extracted by MFCC method is correlations coefficient, it is derivated from the MFCC, ΔMFCC and ΔΔ MFCC techniques.

In Table 2 represented the comparative results of all the MFCC correlations coefficient, correlation coefficient by MFCC values is very less than the ΔMFCC and ΔΔMFCC.

By analysing the statistical analysis, correlation coefficients help to establish a strong connection between the two values. You can use the statistical feature MFCC more effectively speaker identification. For speaker identification method is taken ΔΔMFCC for analysis purpose.

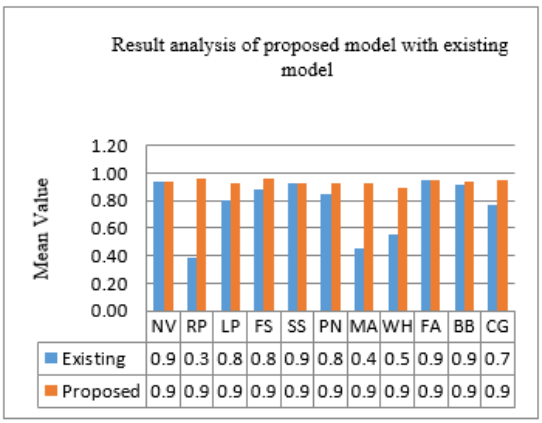


Figure 7. Mean value of proposed and existing model

Table 2. Compression table of correlation coefficient by a different MFCC method

S. No.	Observation Statics	MFCC	ΔMFCC	ΔΔMFCC
1	NV	0.092975	0.443952	0.608006
2	BB	0.078608	0.455734	0.619435
3	FA	0.072706	0.435045	0.570202
4	FS	0.120632	0.440903	0.612862
5	LP	0.119455	0.458542	0.643303
6	MM	0.068477	0.457214	0.599659
7	OM	0.044163	0.444922	0.62996
8	PN	0.081076	0.466028	0.603308
9	RP	0.042179	0.43151	0.6505
10	SS	0.108855	0.4876	0.661188
11	WP	0.087537	0.44989	0.567188

Different algorithm is used for the classifications phase for the speaker identification. Indirectly to recognise the various speakers utilising the DT classification algorithms proposed method. This method is utilised for categorising the speaker by a text-dependent speech signal. The text-based speech gesture has a succession of words or messages as a speaker is able to convey. DT classifier for speaker grading, the most robust identification technique currently developed.

Based on database training values, in Figure 8 shown acoustic analysis of MFCCs feature extraction techniques and Figure 9 shown the comparison result of correct recognition rate (CRR) method and proposed DT classification method for speaker identification dependent on text-dependent.

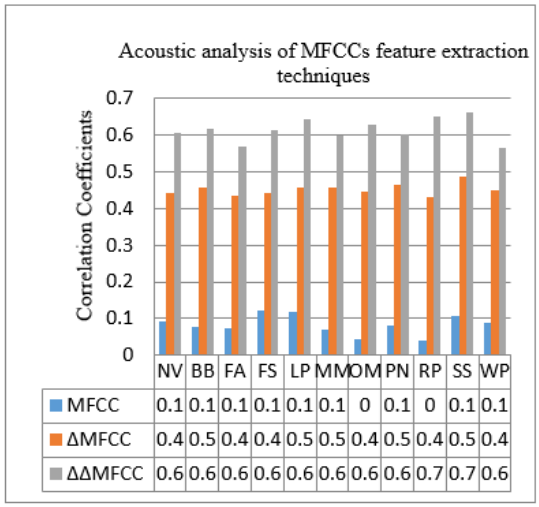


Figure 8. Graphical representation by MFCC feature extraction method

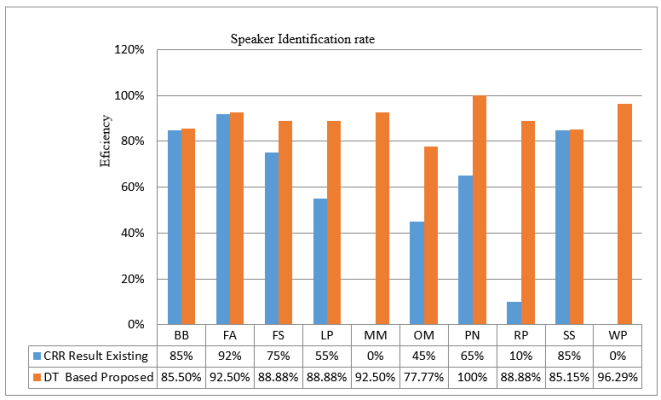


Figure 9. Comparison graph of proposed classifiers detection rate and existing CRR efficiency

From the above Figure 8 and Figure 9 showing the speaker identification by CRR method used in the existing model but in proposed method efficiency are calculated by DT based classification method that is shown the better detection rate than the existing model. In an existing model in case of voice is disguised by MM this gives the 0% efficiency for speaker identification, but in case of classification method, it gives the

better result 92.50% efficiency for speaker identification. Same result for WP method in this result existing model the CRR result is 0% but in case of the proposed model is 96.29% efficient.

5. CONCLUSION

For speaker identification improving the performance of physical variation method proposed is proposed. The speaker identification model using DT classifier has been sujjested. DT classifier has the better classification efficiency result. In proposed method acoustic feature computed by MFCC feature extraction method of normal and physical variation of speech. Its correspondence statistical coefficient of MFCC, ΔMFCC and ΔΔMFCC vectors and its acoustic coefficients pass through classifier and calculate the efficiency of DT based classifier. In this technique, a comparative analysis for the existing result with proposed classifier results are given.

REFERENCES

[1] Zhang, C., Tan, T. (2008). Voice disguise and automatic speaker recognition. Forensic Science International, 175(2-3): 118-122. https://doi.org/10.1016/j.forsciint.2007.05.019

[2] Tan, T. (2010). The effect of voice disguise on automatic speaker recognition. International Congress on Image and Signal Processing, pp. 3538-3541. https://doi.org/10.1109/CISP.2010.5647131

[3] Saloni, Sharma, R., Gupta, A.K. (2016). Estimation and statistical analysis of physical task stress on human speech signal. International Journal of Image, Graphics and Signal Processing, 8(10): 29-34. https://doi.org/10.5815/ijigsp.2016.10.04

[4] Sahoo, T.R., Patra, S. (2014). Silence removal and endpoint detection of speech signal for text-independent speaker identification. International Journal of Image, Graphics and Signal Processing, 6(6): 27-35. https://doi.org/10.5815/ijigsp.2014.06.04

[5] Wu, H., Wang, Y., Huang, J. (2014). Identification of electronic disguised voices. IEEE Transactions on Information Forensics and Security, 9(3): 489-500. https://doi.org/10.1109/TIFS.2014.2301912

[6] Wu, H., Wang, Y., Huang, J. (2013). Blind detection of electronic disguised voice. IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 3013-3017. https://doi.org/10.1109/ICASSP.2013.6638211

[7] Singh, M.K., Singh, A.K., Singh, N. (2018). Acoustic comparison of electronics disguised voice using different semitones. International Journal of Engineering and Technology (UAE), 7(2.16): 98-101. https://doi.org/10.14419/ijet.v7i2.16.11502

[8] Audacity 2.3.2. Audacity: Audio Editor and Recorder Software. http://audacity.sourceforge.net, accessed on July 15, 2019.

[9] Soong, F., Rosenberg, A., Rabiner, L., Juang, B. (1987). A vector quantization approach to speaker recognition. IEEE International Conference on Acoustics, Speech, and Signal Processing, pp. 387-390. https://doi.org/10.1109/ICASSP.1985.1168412

[10] Reynolds, D.A., Quatieri, T.F., Dunn, R.B. (2000). Speaker verification using adapted Gaussian mixture models. Digital signal Processing, 10(1-3): 19-41. https://doi.org/10.1006/dspr.1999.0361

[11] Liu, Q., He, X., Guan, F.W., Zhao, Y.C., Jiang, F., Tian, F.X., Wang, S.X. (2019). Method and implementation of improving the pointing accuracy of an optical remote sensor using a star sensor. Traitement du Signal, 36(4): 311-317. https://doi.org/10.18280/ts.360403

[12] Dehak, N., Kenny, P., Dehak., R., Dumouchel, P., Ouellet, P. (2011). Front end factor analysis for speaker verification. IEEE Transactions on Audio Speech and Language Processing, 19(4): 788-798. https://doi.org/10.1109/TASL.2010.2064307

[13] Singh, M.K., Singh, A.K., Singh, N. (2018). Disguised voice with fast and slow speech and its acoustic analysis. International Journal of Pure and Applied Mathematics. 118(14): 241-246.

[14] Padilla, M.T., Quatieri, T.F., Reynolds, D.A. (2006). Missing feature theory with soft spectral subtraction for speaker verification. International Conference on Spoken Language Processing, pp. 913-916. https://doi.org/10.1109/INFL.2006.195623

[15] Singh, M., Nandan, D., Kumar, S. (2019). Statistical analysis of lower and raised pitch voice signal and its efficiency calculation. Traitement du Signal, 36(5): 455-461. https://doi.org/10.18280/ts.360511

[16] Künzel, H.J., Gonzalez, R.J., Ortega, G.J. (2004). Effect of voice disguise on the performance of a forensic automatic speaker recognition system. IEEE International Workshop Speaker Lang. Recognition, 23(4): 1-4.

[17] Singh, M.K., Singh, A.K., Singh, N. (2019). Multimedia utilization of non-computerized disguised voice and acoustic similarity measurement. Multimedia Tools and Applications, 79: 35537-35552. https://doi.org/10.1007/s11042-019-08329-y

[18] Crochiere, R.E., Rabiner, L.R. (1981). Interpolation and decimation of digital signals—A tutorial review. Proceedings of the IEEE, 69(3): 300-331. https://doi.org/10.1109/PROC.1981.11969

[19] Grimaldi, M., Cummins, F. (2008). Speaker identification using instantaneous frequencies. IEEE Transactions on Audio, Speech, and Language Processing, 16(6): 1097-1111. https://doi.org/10.1109/TASL.2008.2001109

[20] Seresht, H.R., Ahadi, S.M., Seyedin, S. (2017). Spectro-temporal power spectrum features for noise robust ASR. Circuits, Systems, and Signal Processing, 36(8): 3222-3242. https://doi.org/10.1007/s00034-016-0434-0

[21] Singh, M.K., Singh, A.K., Singh, N. (2018). Multimedia analysis for disguised voice and classification efficiency. Multimedia Tool and Applications, 78: 29395-29411. https://doi.org/10.1007/s11042-018-6718-6

[22] Algabri, M., Mathkour, H., Bencherif, M.A., Alsulaiman, M., Mekhtiche, M.A. (2004). Automatic speaker recognition for mobile forensic applications. Mobile Information Systems, 11(5): 96-100. https://doi.org/10.1155/2017/6986391