Electronic disguised voice identification based on Mel-Frequency Cepstral Coefficient analysis

Shalate D'cunha, Shefeena P.S

Dept. of ECE, KMEA Engineering college, Edathala, Ernakulam Kerala, India

Abstract- In this paper, the proposed method is mainly based on analyzing the mel-frequency cepstral coefficients and its derivatives which varies as the voice is disguised. A classifier named the Support Vector Machine (SVM) classifier is used for identification of electronically disguised voice. MFCC statistical moments of training input and test input as a combination of original voice and disguised voice, are given as input to SVM classifier. Now the output obtained will be based on the matching between the training input and the test input. Inorder to provide further enhancement to the particular algorithm, probabilistic neural network(PNN) classifier is used and the performance is evaluated by comparing the accuracy of both classifiers. If the classifier output is matched, then the basic details of the particular person can be transmitted to another location through Email.

Index Terms- Automatic speaker recognition system, Electronic disguised voices, MFCC statistical moments, probabilistic neural network, Support vector machine classifier

I. INTRODUCTION

Voice changers change the tone or pitch of a voice, add distortion to the user's voice, or a combination of all of the above and vary greatly in price and sophistication. The usefulness of identifying a person from the characteristics of his voice is increasing with the growing importance of automatic telecommunication and information processing[1]. It is a worldwide growing tendency that in order to conceal their identities, perpetrators disguise their voices, especially in the cases of threatening calls, extortion, kidnapping and even emergency police calls. The voice disguise is defined as a deliberate action of speaker who wants to change his voice for the purpose of falsifying and concealing identity. The proposed method is based on analyzing the Mel-frequency cepstral coefficients (MFCC) which will be varying as the voice is disguised. The algorithm is mainly based on the MFCC coefficients which include the mean values and the correlation coefficients as the extracted acoustic features. A classifier named the Support Vector Machine classifier(SVM) is used for identification of electronic disguised voice. Existing method also uses the same principle of MFCC, but the dimension of acoustic features is greater than that of the proposed one which makes the existing method complex. The audio signal is divided into short segments by means of hamming window. The statistical moments of MFCC and its derivatives are computed. The acoustic features of training signal and test signal are given as input to SVM classifier. Now the output obtained will be based on the matching between the training input and the test input. Inorder to provide further enhancement to the particular algorithm, the accuracy of SVM classifier is compared with PNN. If the voice is found to be disguised, then the details of the particular person can be transmitted to another location through Email.

II. VOICE DISUISE

Voice disguising is the process of changing or altering one's own voice to dissemble his or her own identity. It is being widely used for many illegal purposes. Voice disguising can have negative impact on many fields that use speaker recognition techniques which includes the field of security systems, Forensics etc. The main challenge of speaker recognition technique is the risk of fraudsters using voice recordings of legitimate speakers. So it is important to be able to identify whether a suspected voice has been impersonated or not. The Mel Frequency Cepstral Coefficients (MFCC) is one of the most important feature extraction technique, which is required among various kinds of speech applications. Voice disguising will modifies the frequency spectrum of a particular speech signal and MFCC-based features can also be used to describe frequency spectral properties. The identification system uses the mean values and the correlation coefficients of MFCC and its regression coefficients as the leading acoustic features. Then Support Vector Machine (SVM) classifiers are used inorder to classify original voices and disguise voices based on the extracted features. Accurate detection of voices that are disguised by various methods was obtained and the performance of the algorithm is phenomenal.

III. TYPES OF VOICE DISGUISES

Disguise can be defined along two independent dimensions: Deliberate versus nondeliberate, and electronic versus nonelectronic[3]. Deliberate-electronic would be the use of electronic scrambling devices to alter the voice. This is often done by means of radio stations to conceal the identity of a person being interviewed. Nondeliberate-electronic disguise would include all of the distortions and alterations introduced by voice channel properties such as the bandwidth limitations of telephones, telephone systems, and recording devices. Deliberate nonelectronic disguise is the one what is usually thought of as disguise. It includes use of falsetto, teeth clenching, etc. Nondeliberate-nonelectronic are those alterations that result from some involuntary state of the individual such as illness, use of alcohol or drugs or emotional feelings. The project is proposed to

focus on deliberate-electronically disguised voices. Electronic-deliberate disguise is relatively uncommon, occurring in only one to ten percent of voice disguise situations.

IV. PROPOSED SYSTEM OF VOICE IDENTIFICATION

The electronic disguising is done using the voice changing software 'Audacity' by changing the pitch. The MFCC and its delta and double delta coefficients are extracted. The plots of MFCC, delta MFCC and double delta MFCC of the original and disguised speech samples are obtained. And these coefficients are used for the voice identification. Two groups or classes are available namely 'original' and 'disguised'.

A.Feature extraction

The first step of MFCC extraction process is to compute the Fast Fourier Transform (FFT) of each frame and obtain its magnitude. The next step will be to adapt the frequency resolution to a perceptual frequency scale which satisfies the properties of the human ears such as a perceptually mel-frequency scale. Then the power *Pm* of the *m*th Mel-filter $B_m(w)$ is calculated by:

$$Pm = \int_{f_{lm}}^{f_{um}} B_m(w)|F(w)|^2 dw, \quad m = 1,2,\dots\dots\dots,M \tag{1}$$

where f_{um} and f_{lm} are the upper and lower cut-off frequencies of $B_m(\omega)$. Next, the Discrete Cosine Transform (DCT) is applied to the Log-power {log *P*1, log *P*2, . . . , log *PM*} of the *M* Mel-filters to calculate the *L*-dimensional MFCC of *xi (n)*:

$$C_l = \sum_{m=1}^{M} \left[\log P_m \cdot cos\frac{l(m-0.5)\pi}{M}\right], \quad l = 1,2,\dots\dots\dots L \tag{2}$$

where C_l is the *l*th MFCC component, and *L* is less than the number *M* of Mel-filters. At this point, for the speech signal *x(n)* with *N* frames, *N L*-dimensional MFCC vectors have been extracted based on each frame. Derivative coefficients (ΔMFCC and ΔΔMFCC) reflecting dynamic cepstral features are computed from the MFCC vectors. Delta and delta-delta coefficients can be calculated as follows:

$$d_t = \left(\sum_{n=1}^{N} n\left(a_n(C_{t+n} - C_{t-n})\right)\right)/\left(2\sum_{n=1}^{N} n^2\right) \tag{3}$$

From frame t computed in terms of static coefficients C_{t+n} to C_{t-n}. Typical value for N is 2. Each of the delta feature represents the change between frames and each of the double delta features represents the changes between the frames in the corresponding delta features. Since the number of MFCC vectors varies with the duration of speech signals, statistical moments are used to obtain acoustic features with the same dimension. For the above *x(n)* with *N* frames, assuming v_{ij} to be the *j* th component of the MFCC vector of the *i* th frame, and *Vj* to be the set of all the *j* th components, V_j can be expressed as:

$$V_j = \{v_{1j}, v_{2j}, \dots\dots\dots v_{Nj}\}, \quad j = 1,2,\dots\dots\dots L \tag{4}$$

Here, two kinds of statistical moments, including the mean values E_j of each component set V_j , and the correlation coefficients $CR_{jj'}$ between different component sets V_j and $V_{j'}$, are taken into consideration. They are calculated by:

$$E_j = E(V_j), \quad j = 1,2,\dots\dots\dots L \tag{5}$$

$$CR_{jj'} = \frac{cov(v_j, v_{j'})}{\sqrt{VAR(vj)}\sqrt{VAR(vj')}}, \quad 1 \le j < j' \le L \tag{6}$$

The resulting E_j and $CR_{jj'}$ are combined to form the statistical moments *WMFCC* of the *L*-dimensional MFCC vectors:

$$W_{MFCC} = [E_1, E_2, \dots E_L, CR_{12}, CR_{13}, .., CR_{L-1L}] \tag{7}$$

Similarly, the statistical moments[6] *WΔMFCC* of the ΔMFCC vectors, and the statistical moments *WΔΔMFCC* of the ΔΔMFCC vectors are calculated[2]. Finally, *WMFCC*, *WΔMFCC* and *WΔΔMFCC* are combined to form the acoustic feature *W* of *x(n)*:

$$W = [W_{MFCC}, W_{\Delta MFCC}, W_{\Delta\Delta MFCC}] \tag{8}$$

B.Identifying disguised voices

The identification algorithm is based on MFCC statistical moments and SVM classifiers. Also probabilistic neural network could be used instead of SVM classifiers for better performance. Fig 1 illustrates the proposed system of voice identification in which first of all the recorded voice sets are disguised by means of Audacity software. Then the feature of these voice sets namely MFCC, Delta MFCC and double delta MFCC are extracted and their statistical coefficients are computed.

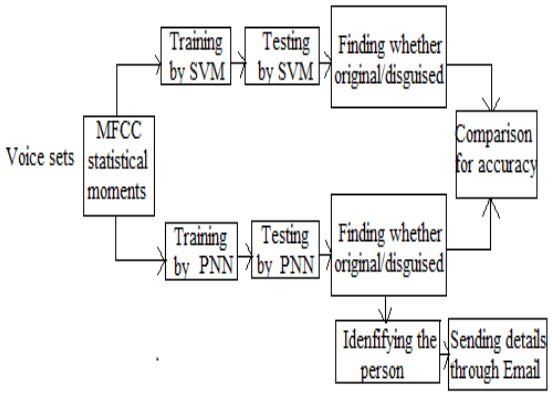

Figure 1. Proposed system of voice identification

Then training is done on SVM classifier by a set of original as well as disguised voice signals. After training, testing of the other voice signals in the database was done, so that each of the voice signals will be identified as either original or disguised. Then the accuracy of the two classifiers are plotted on using bar diagram to find which will be the better one. Now, from the voice signal, the name of the person should be identified. And the details of that particular person, particularly; name, gender, place and the voice will be transmitted to another location through Email.

V. RESULTS AND DISCUSSIONS:

A. Plot of MFCC

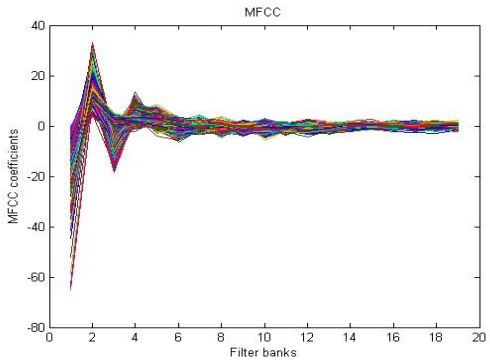

Figure 2. Output plot for MFCC

Fig 2 shown above is the plot between Mel frequency cepstral coefficients and filter banks. Thus a set of coefficients are obtained which could be plotted on the graph and these coefficients represent the auditory feature which for speech recognition.

B. Plot of delta MFCC

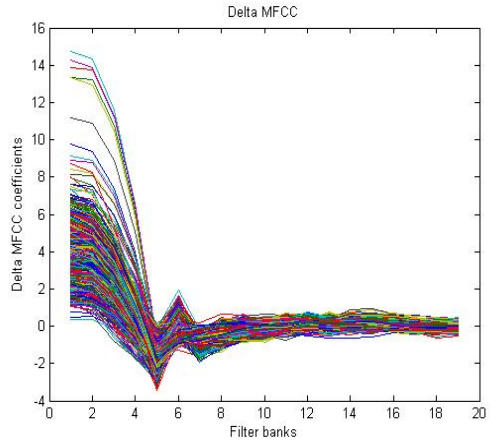

Figure 3: Output plot for Delta MFCC

Fig 3 shown above represents the plot between delta Mel frequency cepstral coefficients and the filter banks which is used to indicate dynamic characteristics of voice identification.

C. Plot of double delta MFCC

Fig 4 shown below represents Double delta coefficients, which are also used to obtain the dynamic characteristics of voice signal.

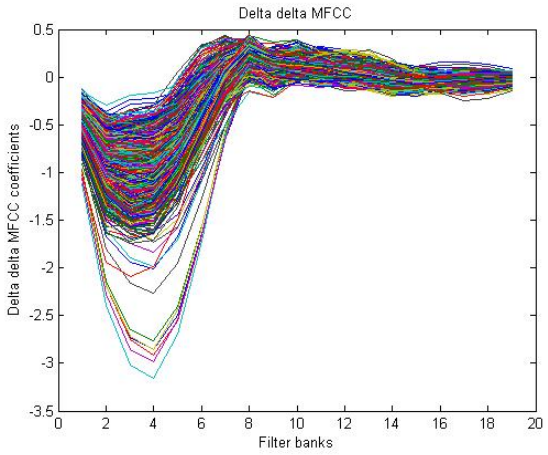

Figure 4: Output plot for Double Delta MFCC

D. Statistical moments

Fig 5 shows the statistical moments of MFCC, delta MFCC and double delta MFCC respectively. The statistical moments include the mean value and the correlation coefficients which is needed in the identification of voices.

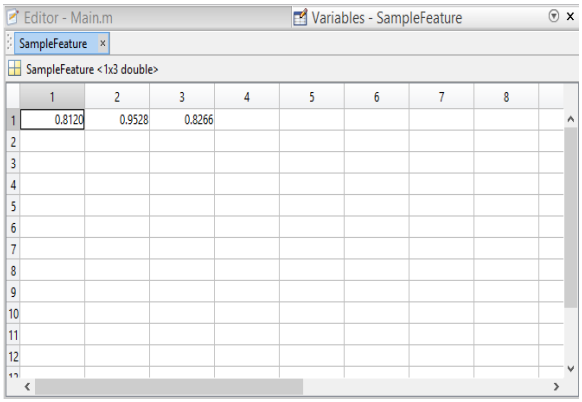

Figure 5. Statistical moments

E. Performance comparison based on accuracy

10 voice signals of 10 friends were chosen and they were disguised and a database was created. For training the SVM classifier, 8 voice signals of 10 friends are used. So SVM classifier was trained with a total of 160 voice signals. For testing purpose, the remaining 20 voice signals are used. The same procedure is repeated for probabilistic neural network. Then a graph indicating the accuracy of the two classifiers are plotted as below.

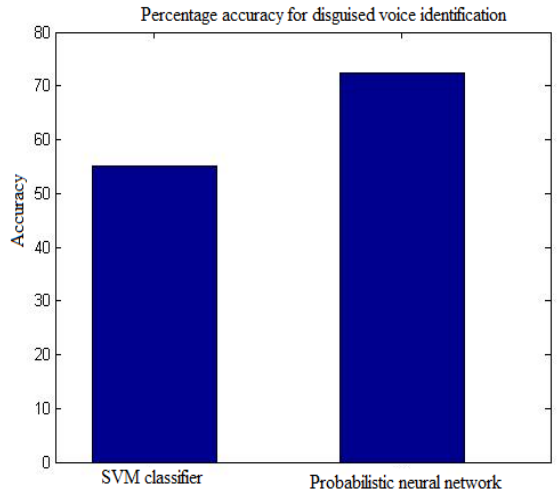

Figure 6. Comparison for accuracy

F. Result when identifying original voice

A GUI representation of the output of the proposed thesis is shown in Fig 7 and Fig 8. The input audio signal is selected by browsing the files and then it is checked to find whether the voice is disguised or not. Then the identification of voice is made by means of probabilistic neural network. If the input voice is found to be original, then the details of that person, mainly name, gender, place will be displayed.

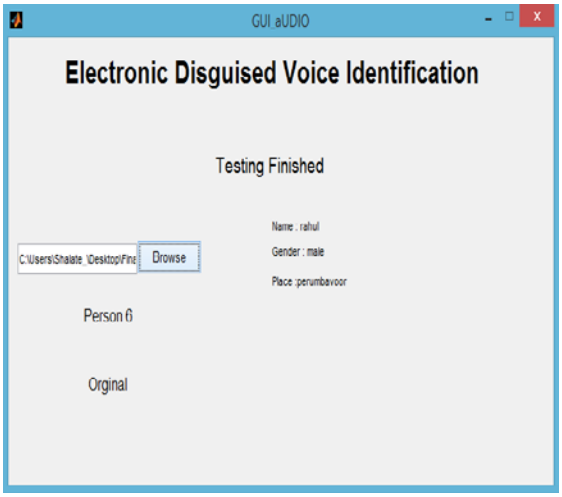

Figure 7. Result when input voice signal is original

G. Result when identifying disguised voice

Figure 8. Result when input voice signal is disguised

If it is recognized that the input voice signal is disguised, then the details of the particular person, and the disguised voice which could be used for further investigation is send via an Email. If the mail is sent successfully, then a dialogue box indicating 'Mailed successfully' will be displayed.

VI. CONCLUSION

The thesis is mainly based on the statistical moments of MFCC and its derivatives. A classifier named the Support Vector Machine classifier is used for this algorithm for identification of electronic disguised voice. Inorder to provide further enhancement to the particular algorithm, the SVM classifier is substituted by a probabilistic neural network classifier which had provided a better output. A comparison is made between the SVM classifier and probabilistic neural network and it is found that probabilistic neural network posses more accuracy than that of SVM classifier. Typically probabilistic neural network posses 22.5 percentage more accuracy than that of SVM classifier. Also if the output obtained from the classifier is found to be disguised, the details of that particular person is send via an Email.

ACKNOWLEDGEMENT

Authors wish to thank all those people who had helped in successfully completing this project.

REFERENCES

[1] Haojin Wu, Yong Wang and Jiwu Huang, "identification of electronic disguised voice", IEEE transactions on information forensics and security, vol. 9. No. 3, March 2014.

[2] S. S. Kajarekar, H. Bratt, E. Shriberg, and R. de Leon, "A study of intentional voice modifications for evading automatic speaker recognition," in Proc. IEEE Int. Workshop Speaker Lang. Recognit., Jun. 2006, pp. 1–6.

[3] H. Wu, Y. Wang, and J. Huang, "Blind detection of electronic disguised voice," in Proc. IEEE International Conference on acoustics, speech and signal processing, vol. 1. Feb. 2013, pp. 3013–3017.

[4] Patrick Perrot, Celine Preteux, Sophie Vasseur, Gerard Chollet, "Detection and Recognition of voice disguise" Proceedings, IAFPA 2007.

[5] P. Perrot and G. Chollet, "The question of disguised voice," J. Acoust. Soc. Amer., vol. 123, no. 5, pp. 3878-1–3878-5, Jun. 2008.

[6] Lini T Lal, Avani Nath N.J,International "Identification of disguised voices using feature extraction and classification" Journal of Engineering Research and General Science Volume 3, Issue 2, Part 2, March-April, 2015,

[7] T. Tan, "The effect of voice disguise on automatic speaker recognition," in Proc. IEEE Int. Certified Information Security Professional, vol. 8. Oct. 2010, pp. 3538–3541.

[8] Surbhi Mathur, Choudhary S. K and Vyas J. M "Speaker Recognition System and its Forensic Implications" Open Access Scientific Reports, Vol 2, Issue 4, 2013.

[9] Sitanshu Gupta, Sharanyan S and Asim Mukherjee "Performance Analysis of Support Vector Machine as Classifier for Voiced and Unvoiced Speech" Int'l Conf. on Computer & Communication Technology IEEE 2010.

[10] K. Z. Mao, K. C Tan, and W. Ser "Probabilistic neural-network structure determination for pattern classification" IEEE transactions on neural networks, vol. 11, no. 4, July 2000.

AUTHORS

First Author - Shalate D'cunha, pursuading Mtech in Communication Engineering, KMEA Engineering college, Edathala., Email: shalate.dcunha@gmail.com
Second Author – Shefeena P.S, Assistant Professor, KMEA Engineering college, Edathala., Email: shefeenaps@gmail.com