International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 8 5080 - 5083

# Detection of Disguised Voice Using Probabilistic Neural Network

Abin Mathew George
M Tech Communication
Department of Electronics and Communication
BCCaarmel Engineering College,
Mahatma Gandhi University
Kerala, India
*mathewgeorge002@gmail.com*

Eva George
Assistant Professor
Department of Electronics and Communication
BCCaarmel Engineering College,
Mahatma Gandhi University
Kerala, India
*evageorge84@gmail.com*

***Abstract*—**Since voice disguise is the process of concealing one's identity, it is being widely used for illegal purposes. This has caused a negative impact on the audio forensics, so it is important to identify whether the voice is disguised or not. It is more difficult to identify whether the voice is disguised or not, if the voice is disguised using electronic scrambling devices or audio editing software tools. We know that voice disguise is the modification of frequency spectrum of speech signals, so we will be using mel-frequency cepstrum coefficients (MFCC's). In this paper, we will be extracting MFCC statistical moments including mean and correlation coefficients as acoustic features and then we will be using an algorithm based on these features and will be using probabilistic neural network (PNN) as classifier to distinguish whether the voice is disguised or not.

***Keywords*-***Electronic disguised voice; MFCC statistical moments; PNN*

*****

## I. INTRODUCTION

Voice is unique for every individual, so by voice we can able to verify the identity of a person. Voice identification and speaker recognition is done in various fields like audio forensics, biometric access control system etc.. But these identification systems suffers from the question of voice disguise. Voice disguise is a deliberate operation to conceal one's identity. There are two kinds of voice disguise: Intentional voice disguising and unintentional voice disguising.

Intentional voice disguising can be classified into electronic voice disguising and non-electronic voice disguising. Electronic voice disguising can be done using electronic software, which can be used to change the parameters of voice like pitch, speed duration, intensity etc.. Nowadays many audio editing softwares are available at the internet like Audacity, Cool Edit, Praat etc.. Non-electronic voice disguising is changing the tone of the voice mechanically like pinching nostrils, placing an object between the mouth etc..Unintentional voice disguising is changing the tone of a voice through emotions like excitement, sadness etc.. Electronic voice disguise and non-electronic voice disguise can be classified into voice conversion and voice transformation. Earlier studies revealed the classification of both electronic voice disguise and non-electronic voice disguise as voice transformation and voice conversion [3]. Voice transformation is done by changing the parameters of the speaker's voice. Voice conversion is the modification of the source speaker's voice so as to sound like the target speaker [2].

Since voice disguise can be achieved successfully with good disguising performance, it is being used for illegal purposes. It has raised serious challenges in the field of audio forensics to find the criminal suspect. Up to now, according to our best knowledge, few studies on identifying disguised voices have been reported. By using sophisticated algorithms, electronic methods can achieve much more natural disguise performance and present greater confusion on both automatic speaker recognition (ASR) systems and human beings than non-electronic ones. As a result, criminal cases using electronic disguise have been increasing in phone communications, online chatting, and other speech applications in recent years. Voices are often disguised to protect the privacy of interviewees in television and radio interviews. The most common disguising type adopted by criminals is change of pitch which introduces substantial variance of acoustic properties and results in poorer performance of speaker recognition. Pitch change resultsin corresponding change of other parameters anddegradation of speaker recognition by parameterdiscrimination, auditory perception and automatic speaker recognition. Some systematic changes of parameters provide clues for forensic voice comparison.

The acoustic feature which is widely used for identification of disguised voice and speaker recognition is MFCC. In sound processing, the mel-frequency cepstrum (MFC) is a representation of the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency.MFC is a collection of MFCC's. The scale of pitches judged by listeners to be equal in distance one from another is mel scale. The reference point between this scale and normal frequency measurement is defined by equating a 1000 Hz tone, 40 dB above the listener's threshold, with a pitch of 1000 mels. In previous work also, MFCC were used as acoustic features to get extracted and then support vector machine (SVM) classifier was used to identify whether the voice is disguised or not. Here we will be using PNN as classifier as the detection rate is high when compared to the SVM classifier.

## II. ELECTRONIC DISGUISED VOICE

Voice resampling, which is used to stretch or compress the waveform of a voice in time-domain, is an effective method to change the pitch of the voice. The principle of voice disguise

is to raise or to lower voice pitch by stretching or compressing frequency spectrum. In phonetics pitch can be raised or lowered by 12 semitones at most which indicates that pitch is always measured by 12-semitonesdivision. A scaling factor of pitch semitones is therefore the disguising factor. Suppose the pitch of a speech frame to be $p_0$, the disguising factor to be α semitones and the pitch of the modified speech frame to be p, we have

$$p = 2^{\alpha/12}.p_0, \quad (1)$$

If α is positive, pitch is raised and spectrum is stretched. Otherwise, spectrum is compressed and pitch is lowered. A disguising factor +k is used to denote a pitch-raise modification and − k is used to denote a pitch-lower modification with k semitones.

## III. PROBABILISTIC NEURAL NETWORK

A probabilistic neural network (PNN) is a feed forward network and predominantly a classifier to map any input pattern to a number of classifications. It has 3 layers of nodes. The figure below displays the architecture for a PNN that recognizes K = 2 classes, but it can be extended to any number K of classes. The input layer (on the left) contains N nodes: one for each of the N input features of a feature vector. These are fan-out nodes that branch at each feature input node to all nodes in the hidden (or middle) layerso that each hidden node receives the complete input feature vector **x**. The hidden nodes are collected into groups: one group for each of the K classes as shown in *Fig. 1* . Each hidden node in the group for Class k corresponds to a Gaussian function centered on its associated feature vector in the class. All of the Gaussians in a class group feed their functional values to the same output layer node for that class, so there are K output nodes.

At the output node for class k, the Gaussian functions of all the hidden nodes of that particular class will be summed to form probability density function (pdf). In that way, we will be following the same procedure for all other classes and will find pdf at the output node for each class. Then we will be selecting the maximum value among the pdf values obtained for each class and will map the input vector to that particular class whose pdf value is maximum. This is how PNN works for the input feature vector.

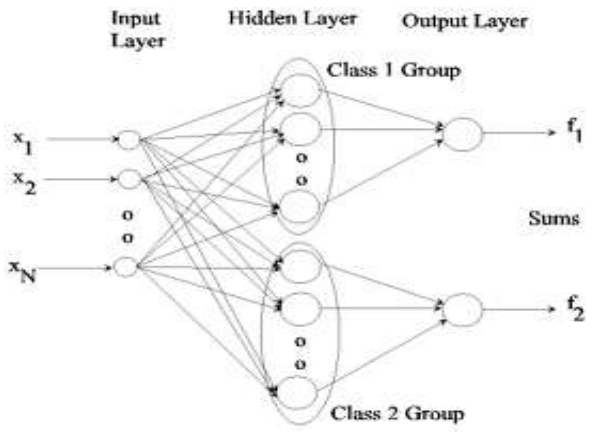


Fig.1.Probablistic Neural Network

## IV. EXTRACTION OF MFCC STATISTICAL MOMENTS

The following steps are used for extraction of acoustic features. Fig.2. shows the extraction of the MFCC statistical moments.

### A. *Pre-emphasis*

It is a technique used to enhance high frequencies of the signal in speech processing. It can spectrally flatten the signal. Speaker information is contained more in the higher frequencies than in lower frequencies, so it increases the energy of signal at higher frequencies.

### B. *Framing*

Since speech is a continuous time varying signal, it is important to frame the speech signal into short speech segments, as short speech segments are stationary.

### C. *Windowing*

Here speech frames and window is being multiplied. It is used to minimize the spectral distortion both at the start and at the end of each frame, as the framed signal results in discontinuity at the start and end of the frame. Here hamming window is used to obtain windowed frames ,where Z is the number of points in frame.

$$H(n) = 0.54 - 0.46cos\frac{2\pi n}{Z-1} \text{ ,n = 0,1,…,Z-1} \quad (2)$$

### D. *Fast Fourier Transform*

It is used to convert frames of N samples from time domain to frequency domain.

### E. *Mel Frequency Warping*

In this warping ,a set of 20 triangular band pass filters are multiplied with magnitude frequency response, so as to get smooth magnitude spectrum. The Mel frequency scale has linear frequency spacing below 1000 Hz and logarithmic spacing above 1000Hz. The formula to calculate mel-frequency $f_{Mel}$ for a given frequency $f$ in hz is:

$$f_{Mel} = 1127\ln(1 + \frac{f}{700}) \quad (3)$$

### F. *Discrete Cosine Transform*

It keeps only first few coefficients and avoids higher coefficients, as it represents fast changes in the filter bank energies which degrades the performance. In short, it is a compression step.

Consider a speech signal with N frames, assume $V_{ij}$ to be the $j^{th}$ component of the MFCC vector of the $i^{th}$ frame and $V_j$ to be the set of all the $j^{th}$ components.

$$V_j = \{v_{1j}, v_{2j}, v_{3j}, \dots, v_{Nj}\}, \quad j = 1,2,\dots,L \quad (4)$$

where L is the dimension of MFCC vectors based on each frame.

The mean value of the speech signal can be calculated as

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

 ISSN: 2321-8169
5080 - 5083

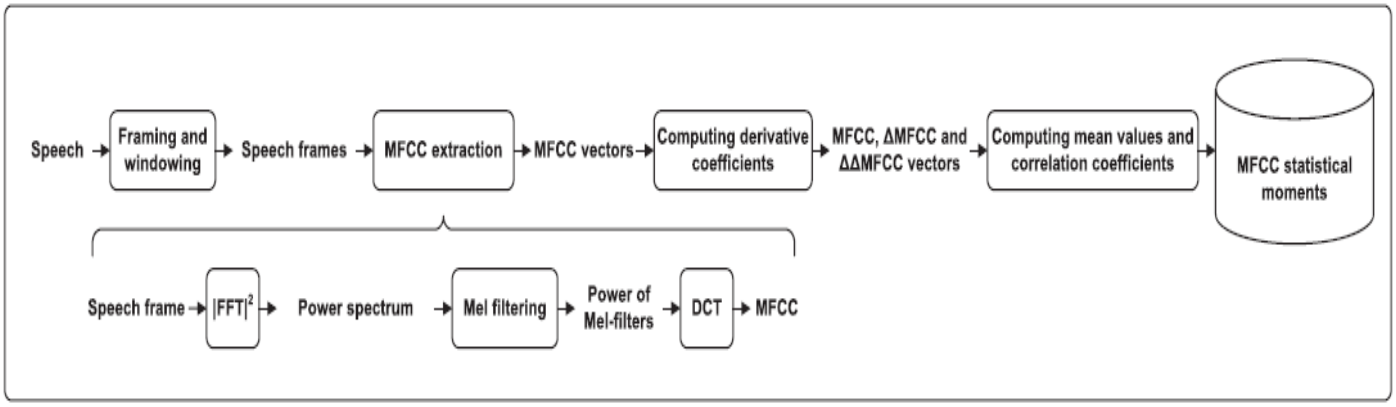


Fig. 2. MFCC Extraction

The correlation coefficient of the speech signal can be calculated as

$$CR_{jj'} = \frac{cov(V_j, V_{j'})}{\sqrt{VAR(V_j)}\sqrt{VAR(V_{j'})}}, 1 \leq j < j' \leq L \quad (6)$$

## V. Algorithm For Identification Of Voice Disguise

The proposed algorithm is based on MFCC and PNN classifier .Here audio wav file is given as an input, which can be either an original voice or disguised voice. In that way ,three databases are created, of which each contains 10 disguised voices and 10 original voices. Here disguising factor of -8 is used to disguise the voice. The features are extracted from each of these voices and is given as input to the PNN classifier in the form of vector and then classified into various classes. Then from the classes mentioned, PNN classifier will decide which class has maximum value and that particular value will be assigned to that input voice .These things will be done in training phase. An adaptive filter is used to remove the noise.

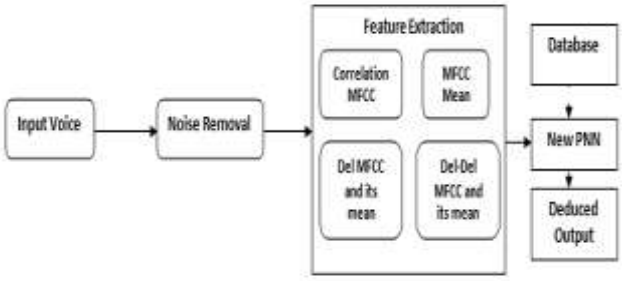


Fig. 3. System Block

In testing phase, the tested voice will be passing through the PNN classifier and will be compared with all the voices in the database, whose features has already been extracted in training phase and assigned a value to each voice. It will decide which class should be assigned to the testing voice depending on the maximum value and we will come to know whether the tested voice is original or not. I have defined two classes named as 1 and 2 for original and disguised voice respectively. The features extracted are correlation MFCC, mean of MFCC, del-MFCC, mean of del MFCC, del-del MFCC and its mean.

## VI. Simulation Result

Three databases are created in which each database contains 20 voice samples ,of which 10 voice samples are disguised and remaining 10 are not disguised. The voices are disguised by a factor of -8.In existing system, we used to compare the features extracted from the testing voice with the features extracted from the voices stored in the database using SVM classifier and was decided whether the voice is original or disguised. Fig.4. shows the simulation output of the detection rate of disguised voice for a disguising factor of -8.The voices are disguised using Audacity, Cool Edit and Praat software tools.

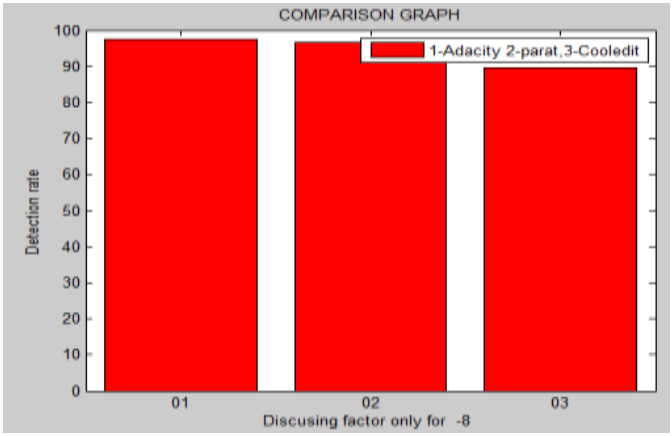


Fig. 4. Existing System Output

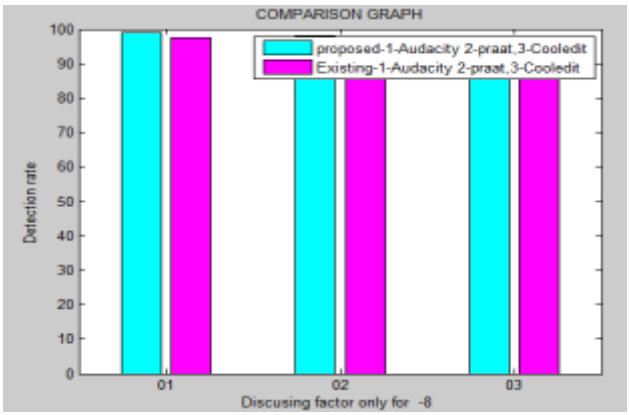


Fig. 5. Comparison of existing and proposed system

The detection rate for audacity, cool edit and praat for existing system are 97.57%, 89.51%, 96.51% respectively. In the proposed system, we will be extracting six features from each voice, as mentioned in section v. So in total, we will be obtaining 120 features from each database, as there are 20 voice samples in each database. In 120 features, we will be assigning value 1 for 60 features and value for the remaining using PNN classifier. These features will be compared with the features extracted from testing voice and selects the maximum value and decides whether voice is disguised or not. Here also voice is disguised using Audacity, Praat and Cool edit software tools. The voice is disguised by a factor of -8. Fig.5. shows the comparison of detection rate of disguised voice of existing and proposed system. It shows that detection rate is high for proposed system than existing system.

## VII. CONCLUSION

The algorithm for identification of disguised voice, which is based on PNN classifier is mentioned in this paper. Here MFCC vectors, Del-MFCC vectors and Del-Del MFCC vectors are extracted from the input voice and are taken as acoustic features. We also have briefly discussed about the classification of MFCC. The steps required for extraction of MFCC has also been discussed. The identification system based on PNN is designed in simulation experiment, voice was disguised using software tools like Audacity, Cool Edit, Praat. It is disguised by a factor -8.Then we will be classifying these voices using PNN classifier. The comparison of detection performance of PNN classifier for various software tools is shown. Then comparison of the output formed using SVM classifier and PNN classifier is shown in Fig.5.From the graph, we could notice that the detection performance of PNN classifier is high when compared to the SVM classifier.

## REFERENCES

[1] P. Perrot, G. Aversano, and G. Chollet, "Voice disguise and automaticdetection: Review and perspectives," in *Progress in Nonlinear SpeechProcessing* (Lecture Notes in Computer Science). New York, NY, USA:Springer-Verlag, 2007, pp. 101–117.

[2] R. Rodman, "Speaker recognition of disguised voices: A program for research," in *Proc. Consortium Speech Technol. Conjunct. Conf. SpeakerRecognit. Man Mach., Direct. Forensic Appl.*, 1998, pp. 9–22.

[3] H. J. Künzel, J. Gonzalez-Rodriguez, and J. Ortega-García, "Effect ofvoice disguise on the performance of a forensic automatic speaker recognitionsystem," in *Proc. IEEE Int. Workshop Speaker Lang. Recognit.*,Jun. 2004, pp. 1–4.

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

[5] C. Zhang and T. Tan, "Voice disguise and automatic speaker recognition,"*Forensic Sci. Int.*, vol. 175, no. 2, pp. 118–122, 2008.

[6] T. Tan, "The effect of voice disguise on automatic speaker recognition," in*Proc. IEEE Int. CISP*, vol. 8. Oct. 2010, pp. 3538–3541.

[7] 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.