International Journal of Engineering Research and General Science Volume 3, Issue 2, Part 2, March-April, 2015
ISSN 2091-2730

Identification of disguised voices using feature extraction and classification

Lini T Lal, Avani Nath N.J,

Dept. of Electronics and Communication, TKMIT, Kollam, Kerala, India

linithyvila23@gmail.com, 9495052225

Abstract— Voice disguising is the process of altering or changing one's own voice to dissemble his or her own identity. It is being widely used for illegal purposes. Voice disguising can have negative impact on many fields that use speaker recognition techniques which includes the field of Forensics, Security systems, etc. The main challenge of speaker recognition 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. In this paper, we propose an algorithm to identify disguised voices. 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 modifies the frequency spectrum of a speech signal and MFCC-based features can be used to describe frequency spectral properties. The identification system uses mean values and correlation coefficients of MFCC and its regression coefficients as the acoustic features. Then Support Vector Machine (SVM) classifiers are used to classify original 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.

Keywords— Disguised voices, MFCC, regression coefficiets, mean value, correlation coefficients, SVM

INTRODUCTION

Voice is unique for every individual. So this voice can be used to verify the identity of a person. Voice identification and speaker recognition is used in many fields like Automatic Speaker Recognition Systems (ASRS), Audio forensics, Biometric access control systems etc. But such voice disguise identification systems often suffer from the question of disguised voices. Voice disguising is the process by which a speaker's voice tone gets changed and helps in hiding his/her identity. Voice disguising can be divided into two broad groups: Intentional voice disguising and unintentional voice disguising.

Unintentional modifications are caused by emotional conditions like excitement, stress etc. or by physical illness such as cold, sore throat etc. Intentional voice variations include the voice changes where people try to evade detection. Intentional variations can be further divided into two groups. Electronic voice disguising and non-electronic voice disguising. Electronic voice disguising modifies the voice electronically by using some electronic software. It modifies some specific parameters like the frequency, speaking rate, duration etc. in order to change the voice. Nowadays a wide variety of audio editing software such as Audacity, Cool Edit, PRAAT etc. are available. Non-electronic voice disguising on the other hand alters voice mechanically by hindering the speech production system itself. These include speaking with pinched nostrils, whispered speech, using a bite block or handkerchief over the mouth while speaking and so on. Voice disguising can be used for many useful purposes. This technique is used in television and radio interviews for the secure transmission of spoken information without revealing the identity of the speaker. The other applications of voice disguising include entertainment, speech coding, speech synthesis etc. But since voice disguising can be easily achieved using some electronic softwares and simply by altering voice naturally it is a common practice to use the voice disguising for illegal purposes nowadays.

Only a few studies have been reported yet on the identification of such disguised voices. Early studies on voice disguising classify both electronic and non-electronic voice disguising as voice conversion and voice transformation [3]. Voice conversion consists the modification of source speaker voice to sound like a target speaker voice, and voice transformation is the different possibilities to change one or more parameters of the voice [2]. Voice disguising can introduce great variations in the acoustic properties of voice such as fundamental frequency (F_0), intensity, speaking rate etc. Considering the two common voice disguising patterns of raising the voice pitch and lowering it, magnitude of F_0 change and its intensity is much greater in high pitched voices than that in low-pitched voices. Also for low-pitched voice

speakers show consistent tendency of decreasing the speaking rate by slowing down their speech [9]. The performance of Automatic Speaker Recognition Systems (ASRS) is greatly degraded by the presence of disguised voices. The effects of different non-electronic disguising patterns on SRS are different. Among the different disguising patterns available whispered speech, masking over the mouth and raised pitch highly degrades the performance of the Speaker Recognition Systems. FASRS is independent of language and dialect. Therefore it is resistant to foreign accent disguising [8]. Spectral analysis of speech signals provides interesting ways to describe the speech signal in terms of parameters or features. Among the different parameters available Mel Frequency Cepstral Coefficients (MFCC) are the most commonly used feature for speaker/speech recognition applications. MFCC well explains the frequency spectrum of a given voice signal and a disguised one. The identification system for disguised voices is based on the idea that the mean values and correlation coefficients ie, statistical moments of the MFCC, delta MFCC and double delta MFCC varies from that of the disguised voices. The feature extraction stage is one of the important stages. Given a learning problem and a finite training database, SVMs properly weight the learning potential of database and the capacity of the machine and so the classification of the voice as original and disguised is done using Support Vector Machine (SVM)[12].

METHODOLOGY

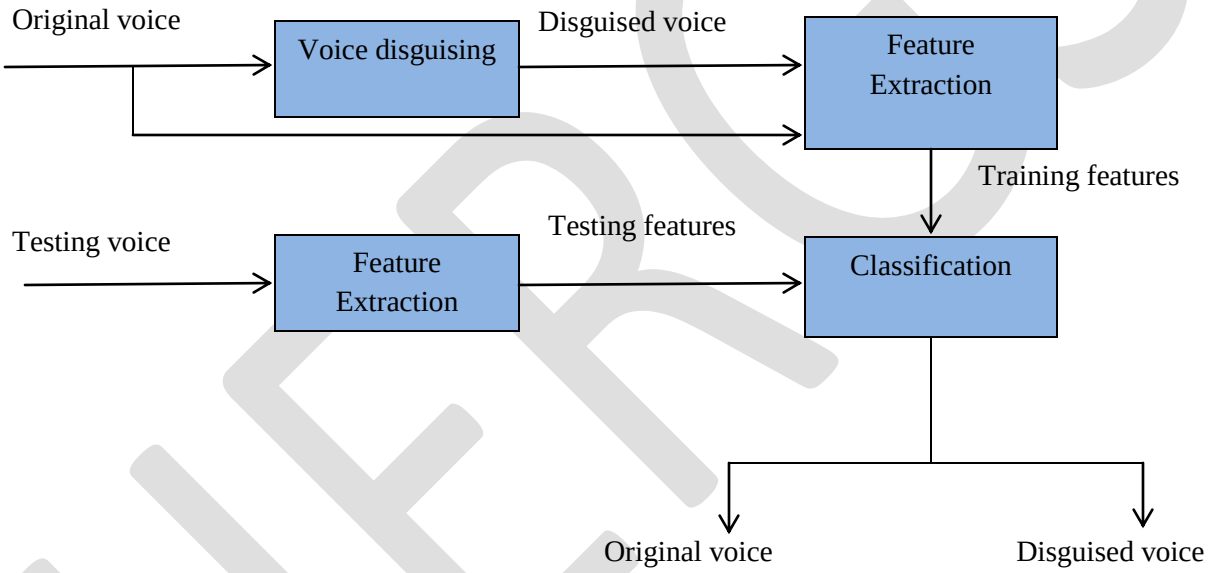


Fig. 1. Block diagram of disguised voice identification system

1) Database collection

Original and disguised voices are required as input speech signal for disguised voice detection. Speech recordings were collected from the students of TKM Institute of Technology. Kerala. Database of about 40 students were used for training which consists of 20 male and 20 female students. The speech recording was text and language independent. They were allowed to speak for more than 2s. The recordings were made at 16 kHz sampling rate and 16 bit quantization.

For electronic voice disguising the voice changing software Audacity was used. Semitones were used as the disguising factor. Disguising factor ranging from +1 to +11 and -1 to -11 were chosen and therefore 22 different kinds of disguised voices ere created from each of the original voice collected.

For non-electronic disguising three types of disguising patterns ie, speaking with pinched nostrils, covered mouth and bite block were selected. Each subject were asked to speak on their normal voice and using all the three above non-electronic disguising methods.

In the testing stage, database was collected from 20 speakers who were not included in the training stage. Electronically disguised database was created using Audacity from each original voice by choosing two or three different disguising factors from the total available 22 disguising factors. Non-electronically disguised database was selected from the speech recordings of the test subjects spoke using the above mentioned disguising patterns.

2) Voice disguising

Voice disguising is done electronically and non-electronically

2.1) Electronic voice disguising

Electronic voice disguising, in effect, modifies the pitch of the voice. An effective time domain technique used for the pitch modification of a given voice is the voice resampling. Voice resampling is a mathematical operation that rebuilds a continuous waveform from its samples and then samples that waveform again at a different rate. Let the original short-time speech signal *x(n)*, of duration *D* and pitch *P* , be resampled by a factor of $1/\alpha$ to get the signal $x'(n)$. The resampled signal is of duration D' and pitch $P.'$ Then the relation between the original signal and the resulting resampled signal is given as:

$$D' = \frac{D}{\alpha} \quad (1)$$

$$X'(\omega) = 1/\alpha \, X(\frac{\omega}{\alpha}) \quad (2)$$

$$P' = \alpha P \quad (3)$$

where $X(\omega)$ and $X'(\omega)$ are the frequency spectrum of the original signal and the resampled signal respectively. When the value of $\alpha < 1$, the $X(\omega)$ is compressed in frequency domain and the pitch is lowered. Otherwise $X(\omega)$ is stretched and the pitch is raised. But during voice resampling the duration *D* of the original signal is changed to D' along with the change in pitch. Such duration alteration may result in that the speed of the voice signal $x'(n)$ is too fast or too slow when compared to the original signal *x(n)*. So in order to adjust the duration D' back to *D* time-scale modification can be used. The Synchronized Over-Lap Add (SOLA) algorithm is the mostly used technique for time-scale modification. Here the original voice signal is firstly decomposed into frames and several frames are repeated or discarded while leaving others unchanged. The idea of the SOLA technique is shown in figure 1.

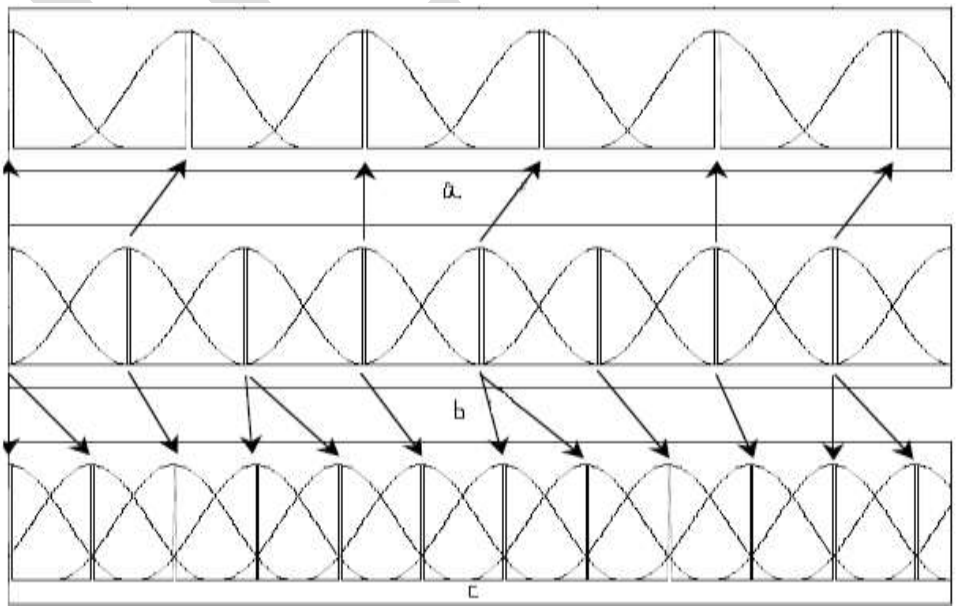


Fig. 2. Basic idea of SOLA algorithm. (a) Down-shifted signal. (b) Original signal (c) Up-shifted signal

During time-scale modification, the duration and speed of voices are changed without affecting the frequency contents and pitch. The duration D' of the resampled signal is adjusted back to the duration D of the original signal by using time-scale modification by a factor of α. The duration D'' and pitch P'' of the resulting time-scale modified signal $x''(n)$ is related to the original signal as:

$$D'' = \alpha D' = D \quad (4)$$

$$P'' = P' = \alpha P \quad (5)$$

The original signal $x(n)$ is disguised to $x''(n)$ by combining voice resampling by a factor of $1/\alpha$ with timescale modification by a factor of α. Then the disguising factor α is given as:

$$\alpha = \frac{P''}{P} \quad (6)$$

If $\alpha > 1$, P is raised. Otherwise, if $0 < \alpha < 1$, P is lowered. In phonetics, voice pitch is always measured by 12-semitones-division, implying that pitch can be raised or lowered by 11 semitones at most. So this semitone can also be used as a disguising factor to modify the pitch of the given voice. ie, the value of the disguising factor can range from ± 1 to ± 11. This algorithm forms the basis of the voice disguising method used in almost all disguising softwares.

2.2) Non-electronic voice disguising

Many different methods are available for non-electronic voice disguising. Changing one's own voice does not require special ability. This category of voice disguise alters the voice by using a mechanic system to hinder the speech production system, which includes pen in the mouth , handkerchief over the mouth, pinched nostrils, bite block etc. These also includes changing the prosody like dialect, accent or pitch register to get a low or high frequency voice modification, in order to trick the identity perception. Whispered speech, creaky voice, raised pitch, lowered pitch etc. are also examples of this category.

3) Feature extraction

MFCC is based on the known variations of the human ear's critical bandwidth with frequency. Human perception of frequency contents of sounds does not follow a linear scale. Perceptual analysis emulates human ear non-linear frequency response by creating a set of filters on non-linearly spaced frequency bands. Therefore for each voice tone with an actual frequency f, measured in Hz, a subjective pitch is measured on a scale called the 'Mel' scale. The name 'Mel' comes from the word melody to indicate the scale is based on pitch comparisons. Mel scale follows a linear scaling for frequencies less than 1 KHz and a logarithmic scaling for frequencies above 1 KHz. The Mel scale is a logarithmic mapping from physical frequency to perceived frequency and the cepstral coefficients extracted using this frequency scale are called MFCC. MFCCs are widely used in Automatic Speaker Recognition Systems (ASRS). The cepstral features obtained are roughly orthogonal since DCT is used and also MFCC is less sensitive to additive noise than some other feature extraction techniques such as Linear Predictive Cepstral Coefficients (LPCC). Delta and delta-delta coefficients of MFCC also known as differential and acceleration coefficients can also be used.

Following steps are used for MFCC extraction:

a) Pre-emphasis

Pre-emphasis is a technique used in speech processing to enhance high frequencies of the signal. In this step speech sample is passed through a filter which emphasizes higher frequencies. The speech signal generally contains more speaker information in the higher frequencies than in the lower frequencies. Pre-emphasis step will increase the energy of the signal at higher frequencies. Also it removes some of the glottal effects. Pre emphasis can spectrally flatten the signal.

b) Framing

Speech is a time varying signal. But on short time scale it is somewhat stationary. Therefore it is important to use short time spectral analysis. In framing the continuous time speech signal is broken into short time speech segments. The frames are of length 20-30 ms. The voice signal is divided into N=256 samples and adjacent frames are overlapped by M=100.

c) Windowing

This is the process in which the speech frames and the window is being multiplied. The framed signal results in discontinuity at the start and at the end of the frame. This spectral distortion is minimized by using window to taper the voice sample to zero at both the beginning and at the end of each frame. If the window being defined is *W(m)* with $0 \leq m \leq N-1$, where N stands for quantity of samples within every frame, then the output after windowing *Y(m)* is given as :

$$Y(m) = X(m).\ W(m) \quad (7)$$

X(m) is the input speech signal.

d) Fast Fourier Transform(FFT)

Fast Fourier Transform is used to convert each frame of N samples from time domain into frequency domain.

e) Mel Frequency warpping

The cochlea of the human ear performs a quasi-frequency analysis. The analysis in the cochlea takes place on a non-linear frequency scale. This scale is approximately linear up to about 1000 Hz and is approximately logarithmic thereafter. The Mel frequency scale has linear frequency spacing below 1000 Hz and logarithmic spacing above 1000Hz. In Mel frequency warping magnitude frequency response is multiplied by a set of 20 triangular band pass filters in order to get smooth magnitude spectrum. The formula to compute the Mel for a given frequency f in Hz is:

$$\text{Mel}(f) = 2595 * \log_{10}\left(1 + \frac{f}{700}\right) \quad (8)$$

f) Cepstral analysis

The basic human speech production model is considered as a source-filter model. Here the source represents the air expelled from the lungs whereas the filter gives shape to the spectrum of the signal. According to the speech production model the source *x(n)* and the filter impulse response *h(n)* are convoluted. This convolution can be represented in time domain as:

$$s(n) = x(n) * h(n) \quad (9)$$

which in frequency domain becomes

$$S(z) = X(z).\ H(z) \quad (10)$$

g) Discrete Cosine Transform(DCT)

DCT is a compression step. So it keeps only the first few coefficients. Higher coefficients represent fast changes in the filter bank energies and it can degrade the performance of the system. The advantage of taking the DCT is that the resulting coefficients are real valued, which makes subsequent processing easier.

Delta and delta-delta coefficients can be calculated as follows:

$$d_t = \frac{\sum_{n=1}^{N} n(c_{t+n} - c_{t-n})}{2\sum_{n=1}^{N} n^2} \quad (11)$$

From frame 't' computed in terms of static coefficients C_{t+n} to C_{t-n}. Typical value for N is 2. The MFCC feature vector describes only the power spectral envelope of a single frame. The information in the dynamics of the speech is given by its derivative coefficients. 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.

Mean values and correlation coefficients of the MFCC coefficients are calculated as follows:

Consider speech signal with N frames, assuming 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 such j^{th} components, V_j can be expressed as:

$$V_j = \{v_{1j}, v_{2j}, v_{3j} \dots v_{Nj}\} \qquad \text{for j=1,2,…L} \qquad (12)$$

Then the mean value of the speech signal can be calculated by using the equation:

$$E(j) = E(V_j) \qquad \text{for j=1,2,…L} \qquad (13)$$

and the correlation coefficients can be found out by the equation:

$$CR_{jj'} = \frac{cov(v_j, v_{j'})}{\sqrt{VAR(V_j)}\sqrt{VAR(V_{j'})}} \qquad (14)$$

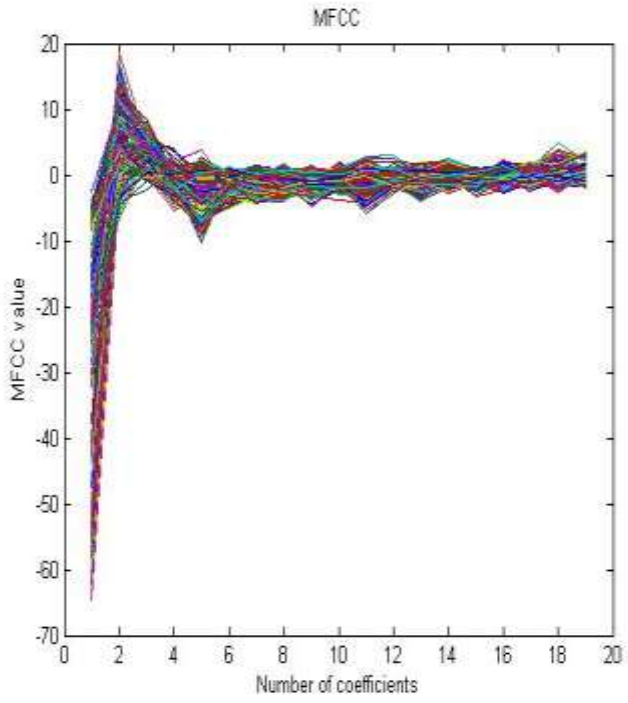
Using this method mean values and correlation coefficients of the derivative coefficients of MFCC can also be calculated.

4) Classification

The next step in the identification of disguised voices is the classification of the extracted features. Support Vector Machines (SVMs) are a useful technique for data classification. A classification task usually involves separating the available data into training and testing sets. Each instance in the training set contains one "target value" (i.e. the class labels) and several "attributes" (i.e. the features or observed variables). The goal of SVM is to produce a model (based on the training data) which predicts the target values of the test data given only the test data attributes. The feature vector is extracted from the input training database and is used to train SVM with linear kernel. Then the features are also extracted from the testing database. Based on the attributes the voice is classified to the two labels 'original' and 'disguised'. SVM classifies data by finding the best hyper plane that separates all data points of one class from those of the other class. The best hyper plane for an SVM means the one with the largest margin between the two classes. Margin means the maximal width of the slab parallel to the hyper plane that has no interior data points.

RESULTS AND DISCUSSIONS

The electronic disguising is done using the voice changing software 'Audacity' by changing the semitone. 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. From the plots we can find that ,the values of MFCC of original and disguised voices for the 19 coefficients in different frames varies.



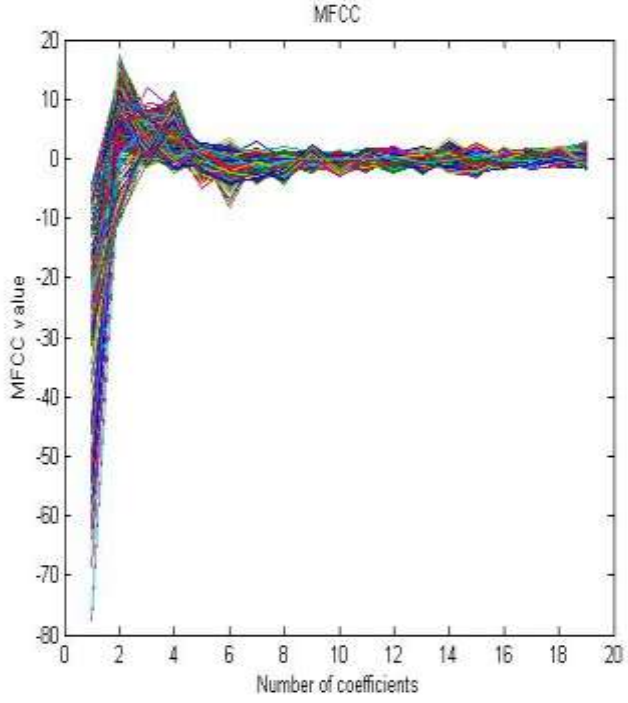


Fig. 3. (a) Plot of MFCC values of original signal (b) Plot of MFCC values of disguised voices

From figure 4 (a) and (b) and figure 5 (a) and (b) it can be shown that the values of delta and double delta coefficients also varies in original and disguised voices

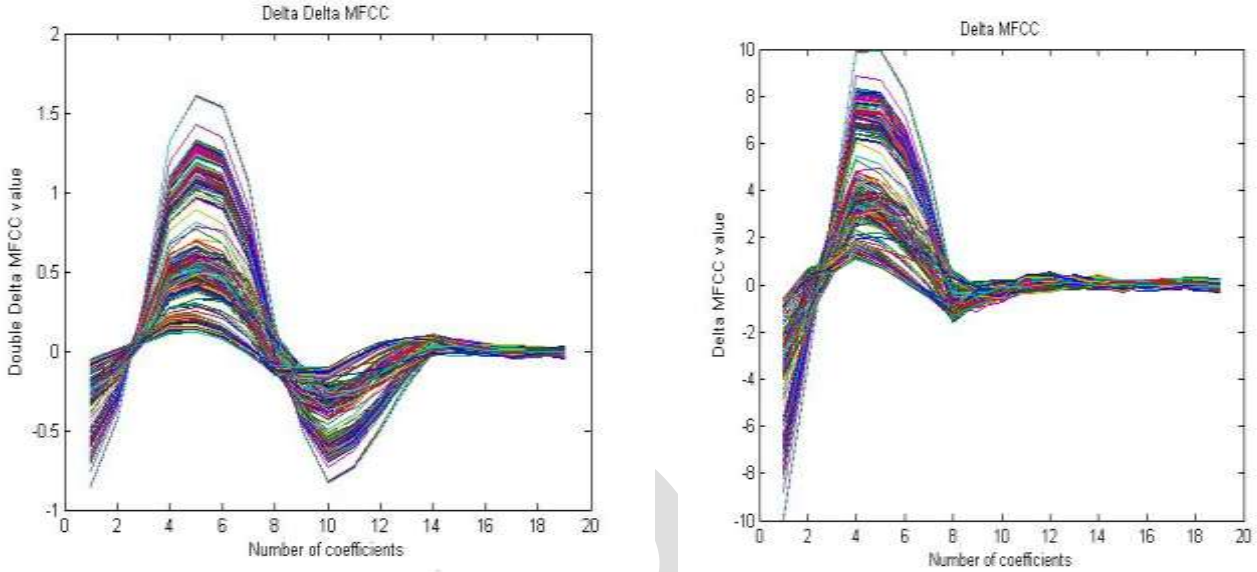


Fig.4 . (a)Plot of delta MFCC values of original signal (b) Plot of delta MFCC values of disguised voices

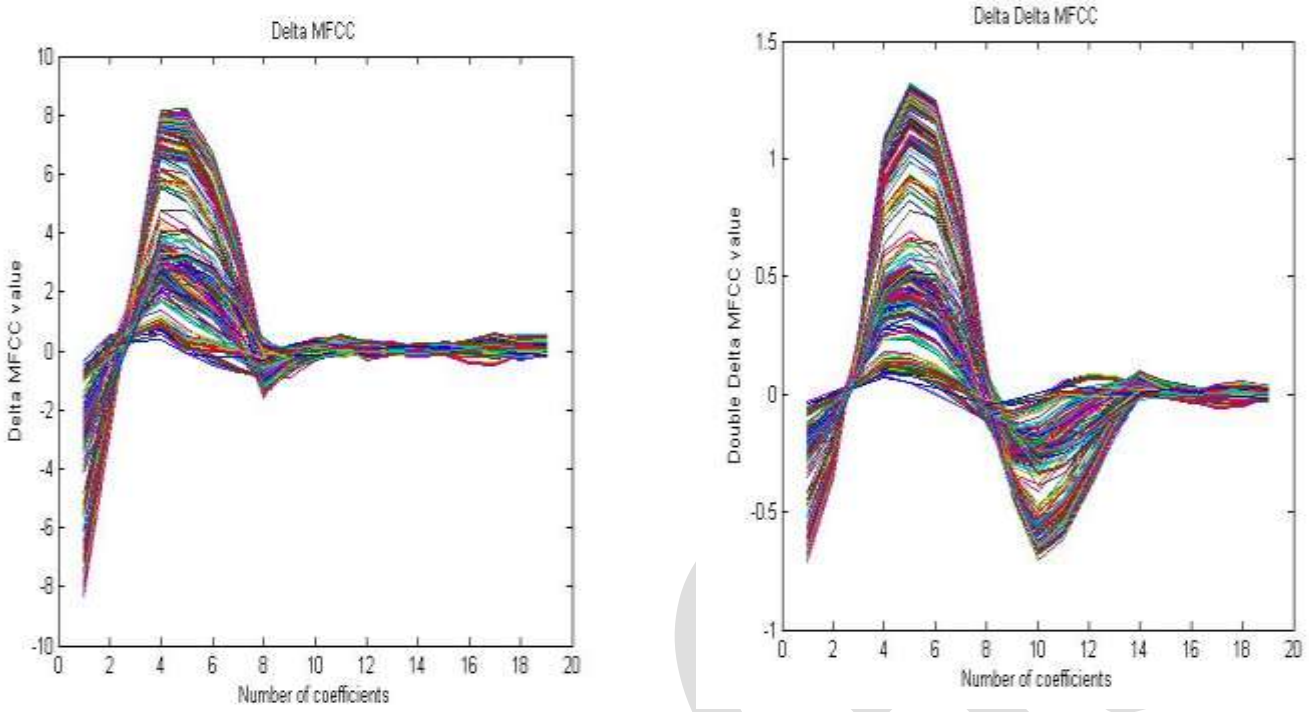


Fig..5. (a)Plot of double delta MFCC values of original signal (b) Plot of double delta MFCC values of disguised voices

Two groups or classes are available namely 'original' and 'disguised'. Each group contains five files of database. For each sound file 6 features are extracted. Mean values of MFCC, delta coefficient and double delta coefficients and correlation coefficients of MFCC, delta and double delta coefficients are extracted. The values obtained by training the SVM with original and disguised database is given in figure 6.

Input	MFCC_mean	ΔMFCC_mean	ΔΔMFCC_mean	MFCC_cor	ΔMFCC_cor	ΔΔMFCC_cor
Original_1	-0.854513861	0.43766744	0.037328908	0.713409656	0.806615573	0.85267349
Original_2	-0.852312305	0.232190341	0.021636408	0.746558737	0.82483205	0.883283102
Original_3	-0.366831535	0.408029645	0.035184401	0.855394937	0.943177259	0.969329013
Original_4	-1.588524462	0.611154582	0.052898724	0.835482474	0.921329618	0.945658435
Original_5	-0.515569469	0.420110008	0.035373873	0.749316113	0.87930297	0.91942945
Disguised_1	-0.849806189	0.412126409	0.035192529	0.684584729	0.732634794	0.777914411
Disguised_2	-1.460441037	0.228004574	0.023757484	0.734898562	0.768952403	0.850083591
Disguised_3	-0.406159968	0.365971249	0.031963578	0.83519446	0.922125002	0.959562725
Disguised_4	-2.023699569	0.599729079	0.052397196	0.831394262	0.910726224	0.939287983
Disguised_5	-0.536636052	0.40241669	0.033599232	0.731551888	0.856941227	0.901733759

Fig. 6. Feature values

CONCLUSION

This work focuses on the identification of disguised voices. Disguised voices can cheat human ears and Automatic Speaker Recognition Systems (ASRS). Voice disguise is being widely used for illegal purposes. An offender can disguise his voice and create fake audio evidences. Thus it will negatively influence the authenticity of evidences. So the identification of disguised voices is inevitable in the field of audio forensics. The identification of disguised voices can be used as preliminary step in speaker recognition tasks to know whether the testing voice is disguised or not. Mel Frequency Cepstral Coefficients (MFCC) based features are used here for separating disguised voices from original voices. The idea used here is that the MFCC statistical moment values vary when a voice gets disguised. So the mean value and correlation coefficients of the MFCC features and its derivative coefficients are calculated. Based on the acoustic feature vector obtained, classification of a given speech database as 'original' or 'disguised' is done using SVM classifier.

REFERENCES:

[1] Haojun Wu,Yong Wang,Jiwu Huang "Identification of electronic disguised voices" IEEE Trans. Information Forensics and Security., vol.9, no.3, pp. 489-500, March. 2014.
[2] R. Rodmann "Speaker recognition of disguised voices: A program for research" in Proc. Consortium Speech Technol. Conjunct. Conf. Speaker Recognit. Man Mach.,Direct Forensic Appl., pp. 9-22, 1998.
[3] P. Perrot, G. Aversano, G. Chellot "Voice disguise and automatic detection: Review and perspectives" in Progress in Nonlinear Speech Processing, NY. USA: Springer-Verlag, pp. 101-117, 2007.
[4] P. Perrot, G. Chellot "The question of disguised voices" J. Acoust. Soc. Amer., vol 123, no.5, pp. 3878-1-3878-5, June 2008.
[5] H. J Kunzel, J. Gonazalez-Rodriguez, J.Ortega Gracia "Effect of voice disguise on the performance of a forensic automatic speaker recognition system" in Proc. IEEE Int. Workshop Speaker Lang. Recognit.,pp.1-4, June 2004.
[6] Haojun Wu, Yong Wang, Jiwu Huang "Blind detection of electronic disguised voices" in Proc, IEEE ICASSP, vol.1, no.3, pp. 3013-3017, February 2013.
[7] Haojun Wu, Yong Wang, Jiwu Huang, Y.Deng "Blind detection of electronic voice transformation with natural disguise" in Proc, Int. Workshop on Digital Forensics Watermarking, LNCS 7809, pp. 336-343, 2012.
[8] S.S Kajarekar, H.Bratt, E.Shriberg, R.de Leon "A study of intentional voice modifications for evading automatic speaker recognition" in Proc, IEEE Int. Workshop Speaker Lang. Recognit., pp. 1-6, June 2006.
[9] T. Tan "The effect of voice disguise on automatic speaker recognition" in Proc, IEEE Int. CISP, vol 8.pp. 3538-3541, October 2010.
[10] Cuiling Zhang "Acoustic analysis of disguised voices with raised and lowered pitch" in Proc. ISCSLP} pp. 353-357, 2012.

[11] H. Hollien "The acoustics of crime: The new science of Forensic Phonetics", NewYork: Plenum press ,1990.
[12] Chih Wei hsu, Chih Chuang Chang, Chih-Jen Lin "A practical Guide to Support Vector Classification" Department of Computer science, National Taiwan University,April 15, 2010