ANALYSIS OF SUITABLE EXTRACTION METHODS AND CLASSIFIERS FOR SPEAKER IDENTIFICATION

R.ARUL JOTHI M.E

Applied Electronics, Dept. of ECE,
Anna University Regional Campus, Tirunelveli

---***---

Abstract - *Speaker Identification system plays an important role in the identification of human voice and disguised voices especially in the cases of telephoned bomb threat and demand of money in kidnapping cases etc. A suitable feature extraction technique which captures the unique characteristics of voice signal can be used to identify criminals. Here Purposed Mel Frequency Cepstrum Coefficient (MFCC) feature extraction method is used; Mel Frequency Cepstrum Coefficient can follow the frequency spectral properties.. Then an algorithm based on the extracted features are investigated with the Machine Learning techniques such as SVM(Support Vector Machine)Classifier is purposed to classifies the human voice and disguised voices. The experimental result indicates the proposed system, compared with MFCC with SVM classifier improves the performances of Speaker Identification system and accuracy rates up to 90%.*

Key Words: *Mel Frequency Cepstral Coefficient (MFCC), Support Vector Machine (SVM), Linear Predictive Cepstral Coefficient (LPCC), Fast* Fourier Transform (FFT), Discrete Cosine Transform (DCT).

1. INTRODUCTION

Speaker Identification system requires two recording voice samples i.e., an original voice of suspected person and a disguised or unknown voice sample. The disguised or unknown voice sample that can be recorded from the cellular phone calls, telephone calls or tape-recorder recording samples is usually taken from the police department.

The speech recording system of Andrey Baranov et al (2010) proposed system which, consists of several parameters to improve the recording speech sample they are overloading, reverberation and compression etc. The sorting of signal amplitude limitation is done by overloading, it occurs due to the dynamic range of recording chain sample which cannot match with the dynamic range of unknown recording voice signal. It causes the negative acceptance of frequency in the phonetic quality of voice signal.

Next lets to study about the voice disguising in voice forensic system, that is the process by which an utterer voice tone gets changed and helps in hiding his/her individuality. Voice disguise has two types they are Intentional voice and Unintentional voice. Intentional voice can also classify into two types they are Electronic disguise voice and Non Electronic disguise voice. Electronic disguised voice obtained by using electronic scrambling devices to modify frequency spectral properties such as the voice pitch and voice formants of an original voice. The original voice has been changed by using some software's such as AV Voice Changer and Skype Voice Changer etc. It can be used to change the perceived age and gender of a speaker. Non-electronic disguise is the method to alter the voice tone of an utterer by disturbing his human speech production system mechanically. Common non electronic voice disguising methods include pinching the nostrils, clenching the jaw, using a bite block, pulling the cheek, holding the tongue, speaking with an object in mouth, etc.

Kajarekar et al (2006) proposed an investigation in the effect of both electronic and non electronic voice modifications on the speaker recognition system. The identification includes data collection, where original and unknown voices are collected or recorded from conversation in the telephone. For comparison purposes, it also includes an investigation work similar to that for NIST database extended-data speaker recognition. The results show variation in both known/unknown voices and speaker recognition systems to suggest a potential for coactions between human identification and automatic recognition systems to address this phenomenon.

Then let's focus the voice disguise and its automatic detection of Perrot et al (2007) proposed the problem of voice alteration caused by channel distortion is not presented in automatic speaker identification. A large range of options are open to change the speaker voice and to trick a human ear identification system. A voice can be modified by changing the semitone of voice signal or more simply by tapping intra-speaker variance, modification of pitch, variations of the position of the vocal organ as lips or tongue which change the formant frequencies. A demerit of this method is slow learning of machine learning algorithm and text dependent.

Then another system that is proposed by the Mireia Farrús et al (2010) ,which describes measurement of voice imitation and conversion in the automatic speaker recognition. Voice can be deliberately changed by means of human caricature or voice conversion. In the proposed method first, analyzing some speech features extracted from voices of impersonators attempting to mimic a target voice and, second, using both intra gender and cross gender voice modification which uses the spectral-based features in the speaker recognition system. The results obtained in this method show that the identification in the fault error rate rises when testing with imitated voices, as well as when using changed voices, especially the cross gender conversions.

Among the different methods and approaches the Speaker Identification system can use the various feature extraction methods and classifiers for identification of voice sample. The Feature extraction methods such as Mel Frequency Cepstrum Coefficient (MFCC) which extract the Statistical features of voice signal such as Mean and Correlation Coefficient of First Order Delta coefficient and Second Order Delta coefficient of Mel Frequency Cepstrum Coefficient (MFCC) this features can be used for the identification of the voice signal. Then the system uses the Classifiers as Machine Learning Techniques such as Support Vector Machine (SVM) classifiers for classification of human original voice and unknown or disguised voice, the classification is based on the features of voice samples.

2. METHODOLOGY

The Speaker Identification process can be used to identify the original voice and unknown/disguised voices. In the identification process it consists of two important stages, (I) feature extraction stage, which includes several feature extraction techniques to extract the unique features of the voice signal and (II) classification stage, which includes several classifier to classify the original and disguised voice. Let us discuss the two stages as detailed,

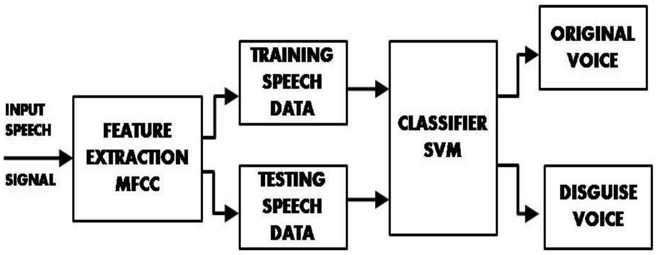


Fig-1: Block diagram of Speaker Identification System

The Block Diagram of MFCC with SVM classifier Fig: 2.1 describe the Operation of MFCC Feature Extraction method with the SVM classifier.

2.1 Feature Extraction

In the feature extraction process employs important method such as Mel Frequency Cepstral Coefficient (MFCC)

2.1.1 Mel Frequency Cepstral Coefficient (MFCC)

It takes the human perception sensitivity with respect to frequency properties of voice signal (VR). The word melody gives the short term word "Mel" to indicate the scale is based on pitch comparisons. Therefore voice tone with a real frequency *f*, measured in Hz, an imminent pitch is measured on a scale called the "Mel"scale. Mel scale follows a linear scaling of frequency less than 1 KHz and a logarithmic scaling of frequency above 1 KHz. MFCC is less sensible to additive noise than some other feature extraction techniques such as Linear Predictive Cepstral Coefficients (LPCC)(307). Delta and delta-delta coefficients of MFCC also called as the differential and acceleration coefficients. MFCC feature extraction steps includes six steps they are

2.1.2 Steps of MFCC

The detailed description of various steps involved in the MFCC feature extraction is explained below.

Pre-Emphasis

Pre-emphasis is the first step of MFCC, which refers to filtering voice signal to emphasize it to higher frequencies. The voice signal can recorded in a microphone or mobile phone from a long distance which has -6 dB/octave slope range approximately and few unwanted noises are add in the voice vocal cord signal(K.S.RAO). Therefore, pre-emphasis removes noises and unwanted sound in the signal and boost up the signal in high frequency. The pre emphasis filter is represent by the transfer function is,

$$Y(z) = 1 - bz^{-1}$$

, where the value of *b* controls the slope of the filter and is usually between 0.9 and 1.0.

Frame Blocking

The second step of MFCC is Frame Blocking. Here the voice signal is blocked in to frames of *M* samples, which is overlapped by N samples (N<M). The first frame consists of the first *M* samples. The second frame begins *N* samples after the first frame, and overlaps it by *M* – *N* samples. The speech signal is a slowly time-varying or quasi-stationary signal. For the static acoustic features, speech needs to be examined over a sufficiently short period of time.

Overlapping frames are not having much information loss and to maintain correlation between the adjacent frames.

Windowing

The next step of MFCC is to windowed the frames of each samples, to minimize the signal discontinuities at the starting and ending of the each frame .It is defined by the window function as *w (m), 0 ≤ m ≤ M -1*, where *M* is the number of samples in every frame, then the resulting windowing signal is

$$x(m) = y(m)w(m), 0 \leq m \leq M - 1$$

Typically the Hamming window is used, its equation is

$$w(m) = 0.54 - 0.46cos\left[\frac{2\pi m}{M - 1}\right], 0 \leq m \leq M - 1$$

Fast Fourier Transform

The next step is the Fast Fourier Transform, which converts each frame of *M* samples from the time domain into the frequency domain. FFT reduces the number of complex multiplications and it improves the speed.

FFT is a fast algorithm is defined by the set of *M* samples *X (k)*, as follow:

$$X(k) = \sum_{m=0}^{M-1} x(m) e^{\frac{-i2\pi mk}{M}} \quad , 0 \le k \le M$$

Where *i* denote the imaginary unit, i.e. *i* = -1, *x (m)* is the complex number.

Mel-Frequency Warping

After the FFT the next step is Mel Frequency Warping it shown human perception of the frequency contents of speech signals, which does not follow the linear scale. For each voice tone with a real frequency, f, measured in Hz, and immanent pitch is measured on a scale called the 'Mel' scale. The Mel frequency scale is linear frequency spacing below 1 kHz and a logarithmic spacing above 1 kHz. Therefore the following formula can be used to calculate the Mel for a given frequency f_m in Hz is

$$mel(f_m) = 2595 log_{10}\left(1 + \frac{f_m}{700}\right)$$

Cepstrum

The next step is the cepstrum. Cepstrum is defined as the logarithm of the Mel frequency wrapping. It is the inverse Fourier transform of the logarithm of the power spectrum of a Mel frequency signal. It is useful for determining periodicities of the voice spectrum of signal. The Cepstrum is denoted by the mathematical equation is,

$$c(m) = ifft(log|fft(s(m))|)$$

Where *s (m)* is the sampled speech signal multiple by Mel filter, and *c (m)* is the Cepstral signal.

Discrete Cosine Transform

Final step is to convert the Cepstral coefficients to time domain using the Discrete Cosine Transform (DCT). The advantage of taking the DCT is that the resulting coefficients are real valued, which makes subsequent processing easier. The result is called the Mel frequency cepstrum coefficients (MFCC).

2.2 Classification

The classification process is used to classify the original and unknown voices by important technique such as Support Vector Machine (SVM)

2.2.1 Support Vector Machine (SVM)

Support Vector Machine (SVMs) is a useful technique for signal/speech classification. A classification task usually involves separating the available speech 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 speech data and is used to train SVM with linear kernel. Then the features are also extracted from the testing speech data. Based on the attributes the voice is classified to the two labels 'original' and 'Unknown'. SVM classifies speech data by finding the best hyper plane that separates all speech 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 (307).

3. RESULT

3.1 Database

The database of Speaker Identification analysis was collected from the students of Anna University Regional Campus, Tirunelveli. Database of about 176 voices are recorded from the 20 students and 3 Non teaching staff and 1 Assistant professor and 26 voices of various actors which are downloaded from internet. The voice recording was text and language independent. They were allowed to speak for more than 5s. The 30 voices are used for training the classifier. The 72 voices are used for testing the classifier.

3.2 Feature Extraction Result

3.2.1 Mel Frequency Cepstral Coefficient

The identification of original voice is done by using the Speaker Identification technique. The Mean and Correlation coefficient of MFCC and its delta and double delta coefficients are extracted. The values of delta and double delta coefficients also vary in original and unknown voices. The Result of Feature extraction of MFCC is shown in the tables.

Table no -1: Mean values of voice signal

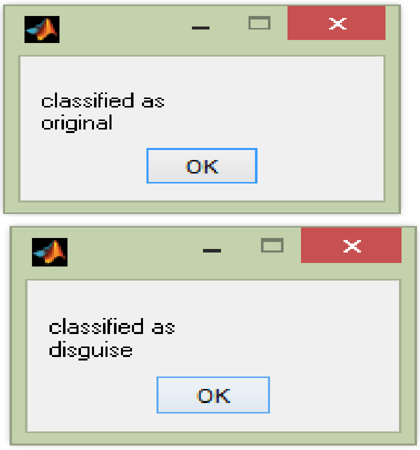
Features of voice signal	Original voice signal	Disguise voice signal
Mean of MFCC	2.394404e-01	1.302124e-01
Mean of Delta MFCC	4.083126e-01	3.657317e-01
Mean of Delta Delta MFCC	2.653543e-01	2.384403e-01

Table no -2: Correlation coefficient values of voice signal

Features of voice signal	Original voice signal	Disguise voice signal
Correlation Coefficient of MFCC	3.403892e-01	3.604562e-01
Correlation Coefficient of Delta MFCC	2.906409e-01	2.974974e-01
Correlation Coefficient of Delta Delta MFCC	6.142048e-01	6.482286e-01

3.3 Classification Result

Support Vector Machine Classification of voices into original and unknown. In SVM Classifier the data can be divided in to two classes. First class of data consists of +1 value which represent the voice is original. Then second class of data consists of -1 value which representing the voice is unknown voice.


classified as original

classified as disguise

Fig -2: Classification Result

4. CONCLUSION

An algorithm for identifying original and unknown or disguised voices is proposed. The identification is done by using the feature extraction techniques such as Mel Frequency Cepstral Coefficient which extracts twelve features such as the statistical moments of MFCC i.e., the mean values and correlation coefficients of MFCC vectors, Delta MFCC vectors and Delta Delta MFCC vectors are extracted as acoustic features. A statistical acoustic feature of MFCC indicates the feature components of original voices and unknown voices are altered. Thus these acoustic features can be used to separate unknown voices and original voices. The classification of identification system can be done by the SVM classifiers. The basic idea of the proposed algorithm is that it is possible to distinguish the original voice and unknown voices. The accuracy of the SVM classifier is 90%.

REFERENCES

[1] Andrey Barinov, 2010. Voice Samples Recording and Speech Quality Assessment for Forensic and Automatic Speaker Identification. Speech Technology Centre Ltd. Saint Petersburg. Russia.

[2] Audio forensics in www.en.wikipedia.org/wiki/Audio forensics for voice forensic.

[3] Darling, A.M., 1991. Properties and Implementation of the Gammatone Filter: A Tutorial.

[4] Guang-Bin Huang, 2013. Extreme Learning Machine. IEEE.

[5] Haojun Wu, Yong Wang, March 2014. Identification of electronic disguised voices. IEEE Trans. Information Forensics and Security.

[6] Jia-Ming Liu, Mingyu you et al, 2013. Cough Signal Recognition with Gammatone Cepstral Coefficients. IEEE.

[7] Kalamani, M., Valarmathy, S., 2015. Automatic Speech Recognition using ELM and KNN Classifiers. International Journal of Innovative Research in Computer and Communication Engineering.

[8] Koustav Chakraborty, Asmita Talele, November 2014. Voice Recognition Using MFCC Algorithm. International Journal of Innovative Research in Advanced Engineering (IJIRAE).

[9] Lindasalwa Muda, Mumtaj Begam, 2010.Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques. Journal of Computing.

[10] Lini T Lal, Avani Nath, N.J., 2015. Identification of disguised voices using feature extraction and classification. International Journal of Engineering Research and General Science.

[11] MFCC in www.practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs/ for mean and correlation coefficient.

[12] Prahlow, J., 2010. Forensic Pathology for Police. Death Investigators, Attorneys and Forensic Scientists. Springer Science Business Media. DOI 10.1007/978-1-59745-404-9_2.

[13] Pragnesh Parmar, Udhayabanu, R., Jan- March 2012. Voice Fingerprinting: A Very Important Tool against Crime. J Indian Acad Forensic Med. vol. 34, no. 1 ISSN 0971-0973.

[14] Sanghamitra Mohanty et al,2014. Speaker Identification from Oriya Voices Using SVM. International Journal of Innovations & Advancement in Computer Science (IJIACS). ISSN 2347 – 8616. Volume 3. Issue 8.

[15] Sonia Sunny, David Peter, S., 2013. Performance Analysis of Different Wavelet Families in Recognizing Speech. International Journal of Engineering Trends and Technology- Volume4Issue4- 2013.ISSN 2231-5381.

[16] Ta-Wen Kuan, An-Chao Tsai et al., 2016. A robust BFCC feature extraction for ASR system. Artificial Intelligence Research. Volume 5 Issue 2-1927-6974.

[17] Wang, X., Zhao, Shao, Y., Jul. 2012. CASA-Based robust speaker identification. IEEE Trans. Audio. Speech. Lang. Processing.

[18] Xavier Valero et al, 2012. Gammatone Cepstral Coefficients: Biologically Inspired Features for Non-Speech Audio Classification. IEEE Transactions on Multimedia.

[19] Xiaojia Zhao, DeLiang Wang, 2013. Analyzing Noise Robustness of MFCC and GFCC Features in Speaker Identification. IEEE.