ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 8, August 2017

# Synthetic Speech Spoofing Detection using MFCC and SVM

[1] Anagha Sonawane, [2] M.U.Inamdar
[1][2] Dept. of E&TC
[1][2] Siddhant College of Engineering, Sudumbare, Pune, India

***Abstract* - In the recent times synthetic voice is used to deceive a speaker recognition based biometric access systems. This paper presents synthetic speech detection in automatic speaker verification system (ASV) for spoof detection. Canonical Mel Frequency Cepstral Coefficients (MFCC) algorithm is used for feature extraction and Support Vector Machine (SVM) is used for classification of natural and synthetic voice. Several experiments are carried out on ASVspoof 2015 database, showing that nonlinear SVM performs better than linear SVM.**

**Index Terms: Synthetic speech detection; Spoof recognition; Automatic speaker verification; MFCC; SVM.**

## I.INTRODUCTION

The human voice is consisting of sounds generated by the opening and closing of the glottis by the vocal cords, which produces a periodic waveform with many harmonics. This basic sound is then filtered by the nose and throat (a complicated resonant piping system) to produce differences in harmonic content (formants) in a controlled way, creating the wide variety of sounds used in speech [10]. There is also other set of sounds, known as the unvoiced and plosive sounds, which are created or modified by the mouth in different fashions. There are two types of speech recognitions : text dependent automatic speaker verification (TDASV) and text independent automatic speaker verification (TIASV) [7]. While TDASV systems use fixed or randomly prompted utterances with known or same text content, TIASV works on arbitrary utterances, possibly spoken in different languages, modes, emotions, physical conditions [10]. Text-independent methods are best suited in the surveillance system implementation where speech signals are likely to originate from non-cooperative speakers [7][10]. In user authentication applications, text-dependent ASV with shorter speech utterances since better accuracy can then be achieved with shorter utterances [28]. Now a days more concentration is being provided on the text independent user authentication such as caller verification in telephone/mobile banking.

A speech signal has information in three parts :voice timbre, prosody and language content. Individual speaker can be mostly characterized by short-term spectral, prosodic[14]. Short-term spectral features are typically extracted from short frames of 20-30 milliseconds duration. They detail the short-term spectral envelope which is an acoustic correlate of voice timbre [18]. Principle Component Analysis(PCA) [21], Hidden Markov Model (HMM) [25], Mel Frequency Cepstral Coefficients (MFCC)[23], Linear Predictive Cepstral Coefficient (LPCC)[8], Perceptual Linear Prediction (PLP) [17] are all popular spectral features. Prosodic features such as pitch, energy and duration are extracted from longer segments such as syllables and word-like units to characterize speaking style and intonation. These features are less sensitive to channel effects but due to their sparsity, the extraction of prosodic features requires relatively large amounts of training data[2][3], and pitch extraction algorithms are generally unreliable in noisy environments [4].

Spoofing attack is a set of circumstances that has one program or user exactly pretence as another by falsifying data, thereby gaining an illegitimate advantage [24][28]. It is a direct attack to the sensor input of a biometric authentication system and the attacker does not need prior knowledge about the recognition algorithm [5]. Basically speech spoof recognition is applied for two types, namely isolated word recognition and isolated word recognition[29]. Isolated word speech spoof recognition performs better than isolated word recognition because of its invariance and shorter length of signal. Mainly spoofing in speech signals can be done in three ways: mimicking , replay, synthetic speech[. The most common attack is mimicking of prosodic and stylistic cues, it is perhaps considered more effective in fooling human listeners than today's state-of-the-art ASV systems[16][25].

Replay attack is a type of attack in which attacker makes use of previously-recorded speech from a genuine client in the form of continuous speech

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 8, August 2017

recordings, or sample speech resulting from the concatenation of shorter audio segments. Replay is a relatively low-technology and simple attack, because for this attacker need not to have the specialized knowledge in speech processing. Replay attack is effective as well as difficult to detect due to the availability of inexpensive, high-quality recording devices and digital audio editing software. Speech synthesis also known as text-to-speech (TTS), is a technique for generating intelligible, natural sounding artificial speech for any random text. VOCODER are used for generation of synthetic speech signal.

Speech spoofing detection can be used in customer verification for mobile banking at call centers, detection of intrusion in voice based password protected systems, automatic speaker verification etc. Sound speech recognition is two level system which consists of speech feature extraction and speech classification..

In this paper, synthetic speech is detected using MFCC feature extraction algorithm and SVM classifier as shown in Fig 1. MFCC algorithm is used because of its simple calculation, better ability of distinction and high robustness to noise[6]. Supervised binary support vector machine is trained using natural human speech and synthetic speech generated using vocoders.

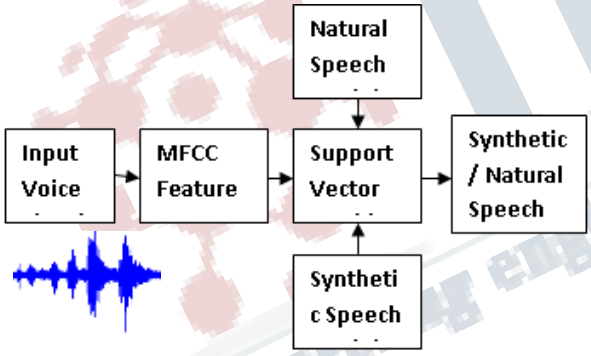



*Fig. 1 Synthetic speech detection system*

The paper is organized with six sections: The first section is an introduction including previous research on ASV and spoofing detection methods. The second section narrates the foundation of MFCC algorithm. The third section offers information about synthetic speech generation. Section four describes the SVM classifier. Section five provides experimental results of spoofing detection and performance analysis. Last section concludes the work.

## II. MFCC FEATURE EXTRACTION

MFCC feature extraction is based on human hearing perceptions which cannot perceive frequencies over 1KHz. Features which are obtained by MFCC algorithm are similar to known variation of the human cochlea's critical bandwidth with frequency [15][23]. The steps of MFCCs algorithm are shown in Fig. 2. The speech input is typically recorded at a sampling rate above 16000 Hz to minimize the effects of aliasing in the analog-to-digital conversion.

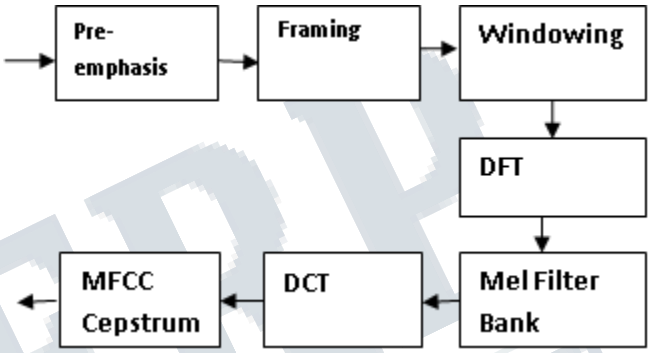



*Fig. 2 Generalized block diagram of MFCC feature Extraction*

### *A. Pre–emphasis*

This step processes the passing of signal through a filter which emphasizes higher frequencies as shown in Fig. 3(a-c). This process increases the energy of signal at higher frequency.

$$Y[n] = X[n] – 0.95 * X[n – 1] \quad (1)$$

It is Assumed that 95% of any one sample is originate from previous sample [1].

### *B. Frame Blocking*

The procedure of segmenting the speech samples that are obtained from analog to digital conversion (ADC) into a small frame with the length within the range of 20 to 40 msec. The voice signal is divided into frames of N samples with 50 % overlapping.

### *C. Windowing*

The next step in the processing is the application of hamming window to each individual frame to minimize the signal discontinuities at the beginning and end of each frame and to collect closer frequency components. The concept here is to make the spectral distortion minimum by using the window to taper the signal to zero at the beginning and end of

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 8, August 2017

each frame. Suppose if we define the window as , $w(n), 0 \le n \le N-1$ where N is the number of samples in each frame, then the result of windowing is the signal $y_l(n1)$ (2)

$$y_l(n) = x_l(n) * w(n), \quad 0 \le n \le N-1 \qquad (2)$$

Typically the Hamming window is given by (3):

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad 0 \le n \le N-1 \qquad (3)$$

***D. Fast Fourier Transform (FFT)***

Fast Fourier Transform will convert each frame of N samples from the time domain into the frequency domain. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of N samples {xn}, as shown in (4),

$$X_k = \sum_{n=0}^{N-1} x_n e^{-j2\pi kn/N}, \qquad k = 0,1,2,...,N-1 \quad (4)$$

In general Xk's are complex numbers and we only consider their absolute frequency magnitude values.

***E. Mel-frequency Wrapping***

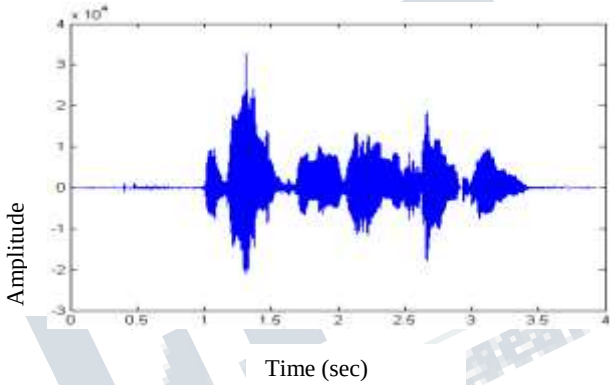
Human perception of the frequency contents of sounds for speech signals does not follow a linear scale [6][23]. So for each tone with the actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel' scale as given in (5). Typically the mel-frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz.

$$Mel(f) = 2595 \, X \, log\left(1 + \frac{f}{700}\right) \qquad (5)$$

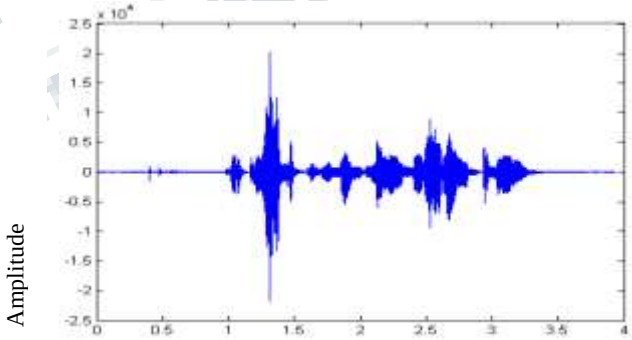
The Mel filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval as shown in Fig. 3(d). The number of Mel spectrum coefficients, K, is typically chosen as to be 24. Mel-wrapping filter bank is very useful to view each and every filter as a histogram bin (where bins have overlap) in the frequency domain.

***F. Cepstrum***

Finally log mel spectrum is converted back to time domain using Discrete Cosine Transform (DCT) which is called the mel frequency cepstrum coefficients (MFCC) refer Fig. 3(e-f). A very good representation of the local spectral properties of the speech signal for the given frame analysis is provided by the speech spectrum cepstral representation . Only the first two cepstral coefficients c0 and c1 have a meaningful interpretation. c0 is the power over all frequency bands and c1 is the balance between low and high frequency components within the signal frame. The other cepstral coefficients have no clear interpretation other than they contain the finer detail of the spectrum to discriminate the sounds [15][23].




(a)




(b)

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 8, August 2017

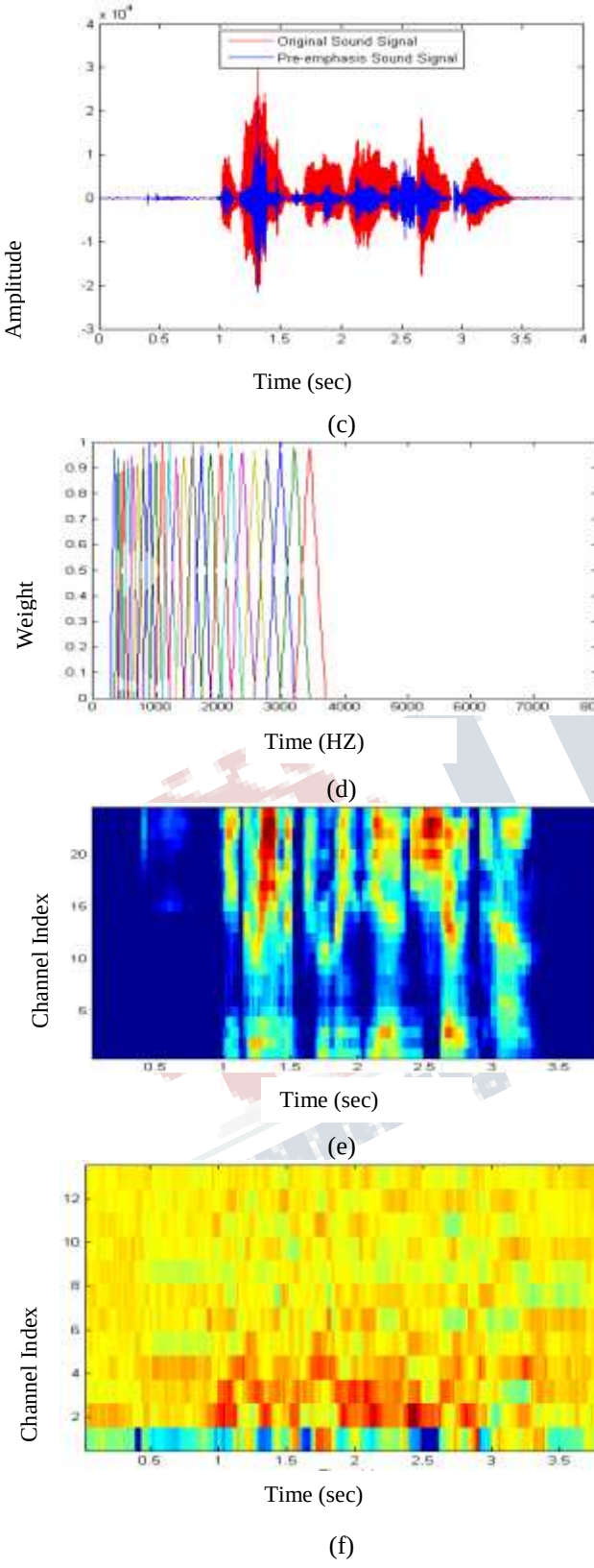



***Fig. 3 a) Original speech signal b) Pre-emphasis filtering output c) Effect of pre-emphasis d) Triangular filter bank response e) Log (Mel) filter bank energies f) Mel frequency Cepstrum***

## III. SYNTHETIC SPEECH GENERATION

Hidden Markov Model (HMM) and Harmonics plus noise model (HNM) based vocoders for statistical parametric speech synthesis are used for synthetic speech generation[3][9]. MLSA is the basic vocoder included in the HTS demo release [9]. During the analysis it estimates the fundamental frequency and performs Mel-cepstral analysis of order 24 for fs=16kHz. The waveform that is reconstructed is built by filtering a simple F0-dependent pulse/noise excitation through the so called MLSA filter which is related to the Mel-cepstral coefficients. The vocoder that is STRAIGHT based is available in the HTS demo release [3][4][9] which is high-quality speech analysis, manipulation and reconstruction tool that represents the speech signal by means of its fundamental frequency, a high-resolution spectral. AHOCODER, a recently proposed vocoder based on the harmonics plus noise model (HNM) which is applied to both speech synthesis and voice conversion. [3][12][13]. It parameterizes speech into three different streams namely, fundamental frequency, Mel-cepstral coefficients of order 39 for sampling frequency of 16kHz and maximum voiced frequency and uses HNM-related procedures for signal analysis and reconstruction.

## IV. SUPPORT VECTOR MACHINE

SVM normally used for classification of higher dimension data which can be separated by kernel function and for limited training data SVM gives higher classification performance[2][26][22]. SVM based binary classifier is trained using natural human speech and synthetic speech. While training SVM, for natural speech class is assigned as +1 and for synthetic speech class -1 as given in (6) and (7). (x,y) is set of training data, x is the MFCC feature set and y is class label. w is normal vector and bo is bias value.

$$\langle w.x \rangle + b_o \geq 1 \, , \forall y = 1 \qquad (6)$$

$$\langle w.x \rangle + b_o \geq -1, \forall y = -1 \qquad (7)$$

For the separation of training data, we used linear and Radial Basis Function (RBF) nonlinear kernels for separating hyper-planes for natural and synthetic spoofed data. Normally synthetic speech generated by VOCODER has close resemblance with natural

ISSN (Online) 2394-6849

International Journal of Engineering Research in Electronics and Communication Engineering (IJERECE)
Vol 4, Issue 8, August 2017

speech, therefore non-linear SVM perform betters than linear SVM. Linear kernel function and RBF kernel function with small positive number σ, are given in (8) and (9).

$$\text{Kernel } (x,y) = (x.y) \quad (8)$$

$$\text{Kernel } (x,y) = e^{\frac{-\|x-y\|^2}{2*\sigma^2}} \quad (9)$$

However, Since only fixed length data vectors are classified by the SVMs, this method cannot be readily applied to task length data has to be transformed to fixed length vectors before SVMs can be used[19].

## V. EXPERIMENTAL RESULTS

Experiments are performed on ASVspoof 2015 database [27]. This database consists of genuine speech of 106 speakers (45 male and 61 female) and with no significant channel or background noise effects. Spoofed synthetic signal generated by MLSA. STRAIGHT and AHOCODER model. Linear and RBF SVM is trained using natural human speech signal. Performance of algorithm is evaluated on the basis of percentage cross validation accuracy as shown in Fig. 4. RBF SVM performs better than linear SVM because of its ability of nonlinear separation of data and AHOCODER .

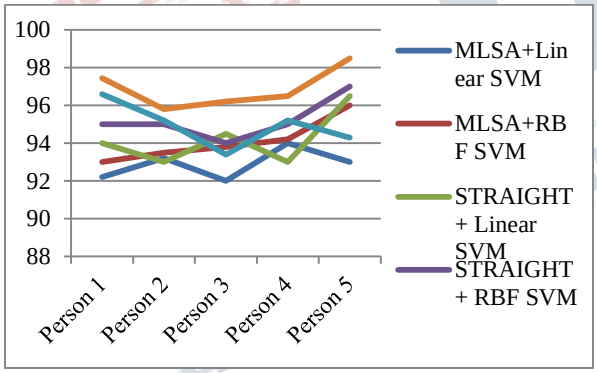



***Fig. 4 Cross validation accuracy of Linear and RBF SVM for MLSA, STRAIGHT, and AHOCODER***

## VI. CONCLUSION

In this paper, we have presented a synthetic speech detection system to prevent spoofing attacks to biometric speaker verification systems that use synthetic voice adaptation or conversion to generate the impostor signal. We have studied MFCC parameterization for synthetic signal feature extraction and SVM as binary classifier. Because of nonlinear and random nature of synthetic speech signal, RBF nonlinear SVM classifier outperforms linear SVM. The performance of algorithm is better for modified HNM AHOCODER synthetic speech generator because its high quality speech analysis. This algorithm faces challenges from variable length of input speech feature vector.

## REFERENCES

[1]. Adami, A., Mihaescu, R., Reynolds, D.A., and Godfrey, J.J., "Modeling prosodic dynamics for speaker recognition," Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, Hong Kong, 2003.

[2]. Ben-Hur, Asa, Horn, David, Siegelmann, Hava, and Vapnik, Vladimir; "Support vector clustering, "Journal of Machine Learning Research, pp.125–137, 2001.

[3]. D erro, I. Sainz E. Navas and I. Hernaez, "Imporoved HNM based vocoder for statistical synthesizers," in Porc. INTERSPEECH, 2011, pp. 1809-1812.

[4]. D. Erro, I. Sainz, E. Navas, and I. Hernáez, "Harmonics plus noise model based vocoder for statistical parametric speech synthesis," IEEE J. Sel. Topics Signal Process., vol. 8, no. 2, pp. 184–194, Apr. 2014.

[5]. F. Alegre, A. Amehraye, and N. Evans, "Spoofing countermeasures to protect automatic speaker verification from voice conversion," in Proc. IEEE ICASSP, May 2013, pp. 3068–3072.

[6]. Fang Zheng, Guoliang Zhang and Zhanjiang Song, "Comparison of Different Implementations of MFCC," J. Computer Science & Technology, vol. 16(6), pp. 582–589, 2001.

[7]. Gerhard "Pitch Extraction and Fundamental Frequency: History and Current Techniques" Technical Report TR-CS 2003-06, November, 2003.

[8]. H. Gupta and D. Gupta, "LPC and LPCC method of feature extraction in Speech Recognition System," 2016 6th International Conference - Cloud System and Big Data Engineering (Confluence), Noida, 2016, pp. 498-502.

[9]. HMM-Based Speech Synthesis System (HTS). [Online]. Available:

[10]. J. P. Campbell, "Speaker recognition: A tutorial," Proc. IEEE, vol. 85, no. 9, pp. 1437–1462, Sep. 1997.

[11]. J. Sanchez, I. Saratxaga, I. Hernáez, E. Navas, and D. Erro, "A crossvocoder study of speaker independent synthetic speech detection using phase information," in Proc. INTERSPEECH, 2014, pp. 1663–1667.

[12]. J. Yamagishi et al., "Robust speaker-adaptive HMM-based text-to-speech synthesis," IEEE Trans. Audio, Speech, Language Process., vol. 17, no. 6, pp. 1208–1230, Aug. 2009.

[13]. J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," IEEE Trans. Audio, Speech, Language Process., vol. 17, no. 1, pp. 66–83, Jan. 2009.

[14]. Kajarekar, S., Ferrer, L., Venkataraman, A., Sonmez, K., Shriberg, E., Stolcke, A., Bratt, H., Gadde, V.R.R., "Speaker recognition using prosodic and lexical features," IEEE Speech Recognition and Understanding Workshop, St. Thomas, U.S. Virgin Islands, pp. 19–24.

[15]. L. Muda, M.Begam and I. Elamvazuthi, "Voice Recognition Algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) Techniques", Journal of Computing, Vol. 2, No. 3, March 2010, pp. 138-143.

[16]. N. Evans, T. Kinnunen, and J. Yamagishi, "Spoofing and countermea-sures for automatic speaker verification," in INTERSPEECH 2013,

[17]. Peng Yuan, Mu Lin, Kong Xiangli, Lin Zhengqing, Wang Lei, "A study on echo feature extraction based on the modified relative spectra (RASTA) and perception linear prediction (PLP) auditory model", Intelligent Computing and Intelligent Systems (ICIS) 2010 IEEE International Conference on, vol. 2, pp. 657-661, 2010.

[18]. R. W. M. Ng, T. Lee, C. C. Leung, B. Ma and H. Li, "Spoken Language Recognition With Prosodic Features," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 21, no. 9, pp. 1841-1853, Sept. 2013.

[19]. R.E. Fan; K.W. Chang; C.J. Hsieh; X.-R. Wang; C.J. Lin "LIBLINEAR: A library for large linear classification". Journal of Machine Learning Research. vol. 9, pp. 1871–1874, 2008.

[20]. S. Imai, "Cepstral analysis synthesis on the mel frequency scale," in Proc. IEEE ICASSP, Apr. 1983, pp. 93–96.

[21]. S. Shabani and Y. Norouzi, "Speech recognition using Principal Components Analysis and Neural Networks," 2016 IEEE 8th International Conference on Intelligent Systems (IS), Sofia, 2016, pp. 90-95.

[22]. Smola, Alex J.; Schölkopf, Bernhard "A tutorial on support vector regression," Statistics and Computing. vol. 14 ,pp. 199–222, 2004.

[23]. T. Ganchev, N. Fakotakis, and G. Kokkinakis, "Comparative evaluation of various MFCC implementations on the speaker verification task," in 10th International Conference on Speech and Computer (SPECOM 2005), Vol. 1, pp. 191–194, 2005.

[24]. T. Kinnunen, Z.-Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li,"Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech," in Proc. IEEE ICASSP,Mar. 2012, pp. 4401–4404.

[25]. T. Satoh, T. Masuko, T. Kobayashi, and K. Tokuda, "A robust speaker verification system against imposture using an HMM-based speech synthesis system," in Proc. INTERSPEECH, 2001, pp. 759–762.

[26]. Vapnik, V "Support-vector networks". Machine Learning, vol. 20, pp. 273–297, 1995.

[27]. Wu Zhizhen, Kinnunen Tomi, Evans Nokolas, Yamagishi Junichi, "Automatic Speaker Verification Spoofing and Countermeasures Challenge (ASVspoof 2015) Database, 2015

[28]. Z. Kons and H. Aronowitz, "Voice transformation-based spoofing of text-dependent speaker verification systems," in Proc. INTERSPEECH, 2013, pp. 945–949.