Speaker Identity Tracing Using Fingerprint Data Hiding against Telecommunications Fraud

Hongxia Wang* and Jing Sang†
* College of Cybersecurity, Sichuan University, Chengdu, China
† School of Information Science & Technology, Southwest Jiaotong University, Chengdu, China
E-mail: hxwang@swjtu.edu.cn Tel: +86-13219049629

***Abstract*—One of the problems in telecommunication fraud investigation is speaker identity tracing and verification. It is difficult to trace identities due to the anonymity of speaker. Therefore the identity tracing of the anonymous user has been identified as a significant problem to assist the investigation. In this paper, a speaker identity tracing scheme using fingerprint data hiding is proposed for preventing telecommunication fraud. First, the fingerprint data of the speaker as the identity information is hidden into the recorded speech signal in the process of calling by smartphone, and then the OFDM of 4G wireless communication is used to modulate the speech signal containing the speaker identity information. Once telecommunication fraud occurs, the speaker can be successfully traced by extracting the fingerprint data from the speech signal. Experimental results show the proposed scheme can accurately trace the speaker identity in the remote speech communications, and prevent telecommunication fraud to some extent.**

I. INTRODUCTION

Recently, with the rapid development of digital speech communication and the extensive usefulness of smartphone, the transmission and application of digital speech are very common and popular. At present, the traffic caused by the digital speech holds a large proportion of the whole Internet and mobile network. At the same time, the creditability validation of digital speech has become an important issue [1]. On the one hand, it is easy to disguise or tamper the voice content by the audio processing software [2-3]. The audio editing software such as Audacity, CoolEdit, PARRT can provide disguising and tampering operation to conceal or forge a speaker's identity by changing his or her voice tone [4]. On the other hand, the reliability and creditability of the speech sources need to be verified in the remote speech communication, and the counterfeit speaker's identity and telecommunication fraud should be prevented. In recent years, the non-repudiation and accountability become more and more important to the e-banking transactions with the rise of telecommunications fraud. Therefore, it will be necessary to trace the identity of the speaker when the economic and legal disputes happen in the financial voice operations.

Speaker identity plays an important role in human-human and human-computer communication. The existing speaker identity verification approaches have achieved some better effects. In Ref. [5], a simultaneous cluster and naming (SCAN) algorithm was proposed, which can automatically learned speaker identity from the noisy audio and identity data derived from pervasive sensors. The speaker identity verification is usually released by combining with the acoustic features between the unknown speaker and the known speaker to verify the speaker identity [6-9]. Ref. [10] presented the varying degree of speaker identity information that is embedded across the MFCCs (Mel Frequency Cepstrum Coefficients) feature. However, many factors such as changing language types, emotions, health and environment usually affect the verification accuracy. Therefore, E. Argones- Ru´a et al. [11] addressed the advantage of adding quality information of the biometric signals into a multimedia-based (video and audio) identity verification system. Sani M. Abdullahi et al. [12] Concealed the fingerprint-biometric data into audio signals for identify authentication. In order to enhance the security of fingerprint data, a fingerprint identification algorithm based on minutiae and invariant moment was proposed in [13]. In our work, we use the binary image of speaker's biometric fingerprint as identity information for tracing and verification purposes in the incredible anonymous communication by the smartphone.

In this paper, we proposed a speaker identity tracing scheme using fingerprint data hiding for the secure anonymous communication. For the speech communication, the automatic recording function of smartphone is opening when unfamiliar call is received. Moreover, the biological fingerprint of the speaker is required to scan by the fingerprint scanner of smartphone. Then, we embed the fingerprint data into the recorded speech signal. When the speech signal containing the fingerprint data is transferred to the receiver side by a wireless channel, the speaker identity tracing can be realized by matching the extracted fingerprint data from the recorder speech and the one in the legal fingerprint database.

This paper is organized as follows. In Section II, we present the proposed speaker identity tracing scheme. The experimental result results are shown in Section III. Finally, the paper is concluded in Section IV.

II. PROPOSED SPEAKER IDENTITY TRACING SCHEME

First, the scanned biological fingerprint is transformed to binary image, and then is embedded into the speech signal. Thus the speech and speaker identity are connected. The speech containing fingerprint data is called stego-speech. The stego-speech will be transmitted by 4G wireless communication channel with OFDM (Orthogonal Frequency Division Multiplexing) modulation. The flowchart of the

proposed speaker identity tracing scheme is shown in Fig. 1, which consists of two processes. One is the sender side of the stego-speech signal. The stego-speech signal is encoded to binary stream by PCM (Pulse Code Modulation), and then is modulated by QAM (Quadrature Amplitude Modulation)/ BPSK(Binary Phase Shift Keying)/QPSK(Quadrature Phase Shift Keying) and OFDM to increase the spectrum efficiency. The modulated stego-speech signal is transmitted by wireless channel to receiver side. So another process is receiver side of the stego-speech signal and speaker identity tracing. The stego-signal acquired from the wireless channel is demodulated by the same modulation method, and then the demodulated stego-speech signal is decoded (PCM decoding), after that, the fingerprint information is extracted by the decoded stego-speech. Finally, we can trace the speaker identity by matching the extracted fingerprint data with the one in the legal fingerprint database.

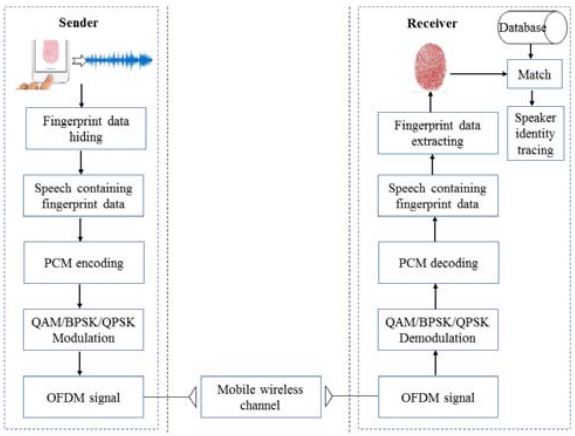


Fig. 1 Flowchart of the proposed speaker identity tracing scheme

A. *Fingerprint data hiding*

The proposed speaker identity tracing scheme is based on the fingerprint data hiding. The speech containing fingerprint data is transmitted by 4G wireless channel. So the fingerprint data hiding algorithm should be able to resist the quantification attack by PCM. In this paper, first, we transform the biological fingerprint scanned by the smartphone into the binary image, and then embed the binary fingerprint image into the recorded speech signal. Suppose the recorded speech signal is $A=\{a(i),1\le i\le L\}$, and $a(i)$ is the sample of speech and L is the number of samples. Let W represent the binary image of the fingerprint. The fingerprint data hiding process is as follows:

Step 1. Segmentation of speech frame. The recorded speech signal A is split into M non-overlapping frames, denoted as $A(p), p=1,2,\cdots,M$, and the length of each speech frame is N, so $M=L/N$.

Step 2. Binary fingerprint image scrambling. In order to resist cropping attack and extract visually identifiable fingerprint image under cropping attack, we use chaotic sequence to scramble the original binary fingerprint image $W=\{w_{ij}\mid w_{ij}\in\{0,1\}\}$ before being embedded into recorder speech, where $1\le i\le X, 1\le j\le Y$, and $X\times Y$ is the size of the fingerprint image. The detailed scrambling method is described in Ref.[14], and the scrambled binary fingerprint image is denoted by $W'=\{w'_{ij}\mid w'_{ij}\in\{0,1\}\}$.

Step 3. Scrambled fingerprint data hiding process. Let $d_i(p)$ and $c_i(p)$ represent the percentile and thousandth digit of the ith sample in the pth frame $A(p)$, respectively. For the case of the following condition, $d_i(p)$ and $c_i(p)$ will be modified to embed the scrambled fingerprint data W' as follows:

$$d_i(p)=\begin{cases}3, & w'_{ij}=0 \,\&\, 5<d_i(p)\le 9\\ 7, & w'_{ij}=1 \,\&\, 1<d_i(p)<5\end{cases} \tag{1}$$

$$c_i(p)=\begin{cases}2, & w'_{ij}=0 \,\&\, d_i(p)=0 \,\&\, (c_i(p)>5 \text{ or } c_i(p)=0)\\ 6, & w'_{ij}=1 \,\&\, d_i(p)=0 \,\&\, c_i(p)<5\end{cases} \tag{2}$$

Otherwise, $d_i(p)$ and $c_i(p)$ will not be modified. Thus, the fingerprint data can be embedded into the recorded speech, and the stego-speech containing speaker identity information is obtained.

B. *Fingerprint data extracting and speaker identity tracing*

In real life, the anonymous phone calls usually result in telecommunications fraud, which makes personal property to suffer major losses. So we need trace the speaker identity once the telecommunications fraud occurs. Because the biological fingerprint is one of the important personal identity information, so the fingerprint registration system of the citizen can be made in the public security bureau. Moreover, the fingerprint information is also embedded into the chip of the second-generation ID card in China. In this paper, the fingerprint data are embedded into the recorder speech during the anonymous call, and then the stego-speech containing fingerprint information is transmitted by the 4G wireless channel with OFDM modulation. Once the telecommunications fraud occurs, we can extract the fingerprint information from the recorder speech to trace the anonymous speaker. The fingerprint data extracting and speaker identity tracing process are described as follows:

Step 1. Segmentation of stego-speech frame. Similarly to the fingerprint data hiding process, the stego-speech signal is divided into M frames, denoted as $A^*(p), p=1,2,\cdots,M$.

Step 2. Extraction of fingerprint data. Let $d_i^*(p)$ and $c_i^*(p)$ represent the percentile and thousandth digit of the ith sample in the pth stego-frame $A^*(p)$. The extracted scrambled fingerprint image $W^*=\{w_{ij}^*\mid w_{ij}^*\in\{0,1\}\}, 1\le i\le X, 1\le j\le Y$ can be obtained by

$$w_{ij}^*=\begin{cases}1, & \text{if } \ d_i^*(p)\ge 5 \text{ or } (d_i^*(p)=0 \text{ and } c_i^*(p)>5)\\ 0, & \text{else}\end{cases} \tag{3}$$

Step 3. Inversely scrambling the binary fingerprint image. We perform the inverse scrambling operation on the scrambled fingerprint image, and obtain the extracted binary fingerprint image for tracing speaker identity.

Step 4. Extracted fingerprint image denoising. Due to quantization and channel fading, the noise will be generated in the extracted fingerprint images. In order to make the fingerprint image purer, the denoise processing of the extracted binary fingerprint image is necessary. There are 5 types of denoise processing according to the different texture characteristics of the binary fingerprint image. We set the pixels that indicate the fingerprint texture to be white, and then the white pixel will be judged to be noise pixel under different cases in 5 types shown in Fig.2, so the white pixels shown in Fig.2 will be reset to black pixels to denoise. Thus, the denoised fingerprint image can be obtained. After denoise processing for the extracted fingerprint image, we can obtain the denoised binary fingerprint image $\tilde{W} = \{\tilde{w}_{ij} \mid \tilde{w}_{ij} \in \{0,1\}\}$, $1 \le i \le X, 1 \le j \le Y$, where $X \times Y$ is the size of the fingerprint image.

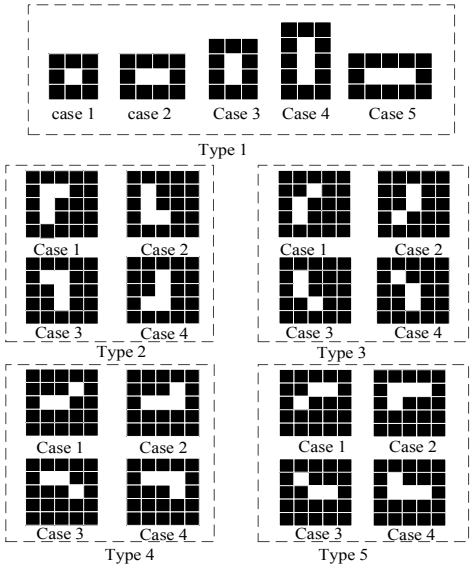


Fig.2 5 types of denoising

Step 5. Speaker identity tracing. The denoised binary fingerprint image is purer, so we can make better matching between the denoised binary fingerprint image and the registered original image in the fingerprint registration system. The accuracy rate ρ of binary fingerprint matching is defined by

$$\rho = 1 - \frac{\sum_{ij} \tilde{w}_{ij} \oplus w_{ij}}{X \times Y}, 1 \le i \le X, 1 \le j \le Y \qquad (4)$$

where w_{ij} is the pixel value of original binary fingerprint image in the registration system, and $\tilde{w}_{ij}$ is the pixel value of denoised binary fingerprint image extracted from the stego-speech, which is transmitted by the 4G wireless channel with OFDM modulation. The high accuracy rate of binary fingerprint matching implies the ideal reliability of speaker identity tracing.

III. EXPERIMENTAL RESULTS

A. *Inaudibility*

In our experiments, the test speech is from ITU-T audio databases [15], and the scanned fingerprint is obtained by our fingerprint recognition instrument with model specification XYZ-III. The stego-speech signal is transmitted by the 4G wireless channel with OFDM modulation and different SNR (Signal-to-Noise Ratio), the inaudibility of fingerprint data hiding will be different. If the SNR degrades, the speech waveform demodulated by OFDM will have a large distortion, and the inaudibility of fingerprint data hiding will become worse. Fig.3 shows the original speech signal sampled at 16 kHz. The Logistic chaotic map is used to scramble the binary fingerprint image, and the initial value of Logistic map is 3.97. Then, we embed the scrambled fingerprint into the recorded speech, and obtain the stego-speech which is transmitted by 4G wireless channel. As a sample, Fig.4 shows the stego-speech signal demodulated by OFDM in AWGN (Additive White Gaussian Noise) channel with different SNR. From Fig.4, we see the burrs of stego-speech waveform reduces as the SNR value increases. When SNR=10dB, the stego-speech waveform is more close to the original speech waveform. Therefore, the inaudibility of the proposed scheme is good under the higher SNR.

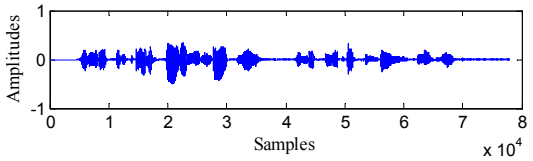
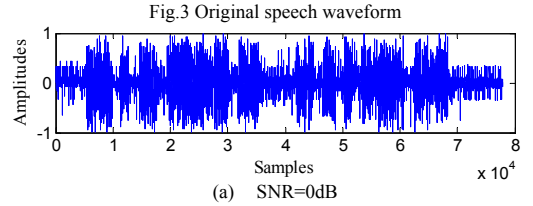
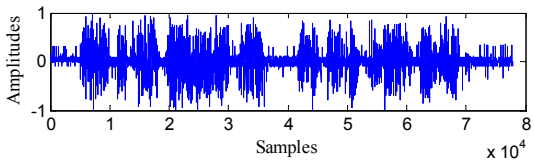


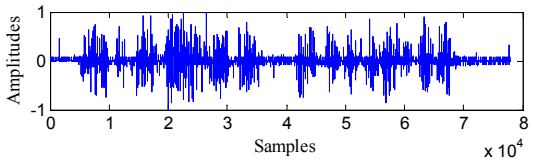
Fig.3 Original speech waveform



(a) SNR=0dB



(b) SNR=2dB



(c) SNR=4dB

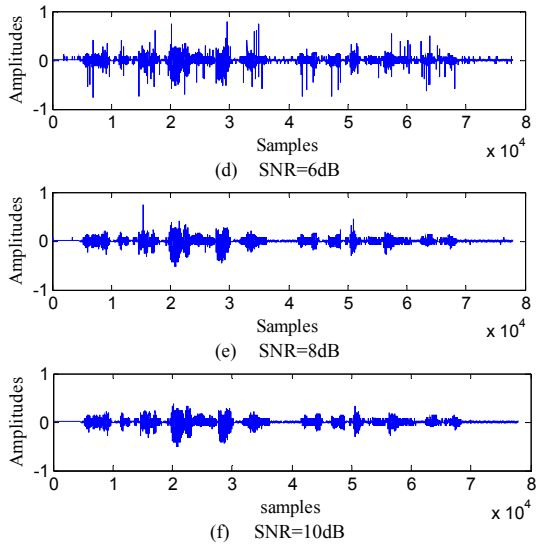


Fig.4 Stego-speech waveform modulated by OFDM in AWGN channel with different SNR

B. Identifiability of extracted fingerprint

In the proposed scheme, we can trace the speaker identity by matching the extracted fingerprint with the original one in the database. Fig.5 shows the original binary fingerprint image of size 256×304, and Fig.6 shows the denoised binary fingerprint image extracted from the stego-speech in AWGN channel with different SNR. From Fig.6, we see the fingerprint image is difficult to identify under SNR=0dB, while it is well be identified as the SNR value increases. The visual quality of the fingerprint image under SNR≥6dB is good in the AWGN channel. Similarly, Fig.7 shows the denoised binary fingerprint image extracted from the stego-speech in Rayleigh channel with different SNR. From Fig.7, we see the single-attenuating of the stego-speech in Rayleigh channel is stronger than AWGN channel. When SNR=8dB, the fingerprint image is still not clear. However, the visual quality of the fingerprint image under SNR≥16dB gets better in the Rayleigh channel.

Fig.5 Original fingerprint

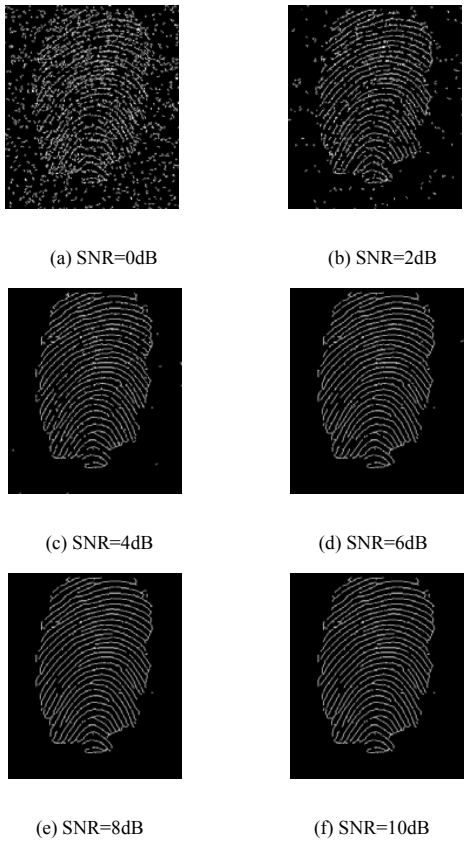
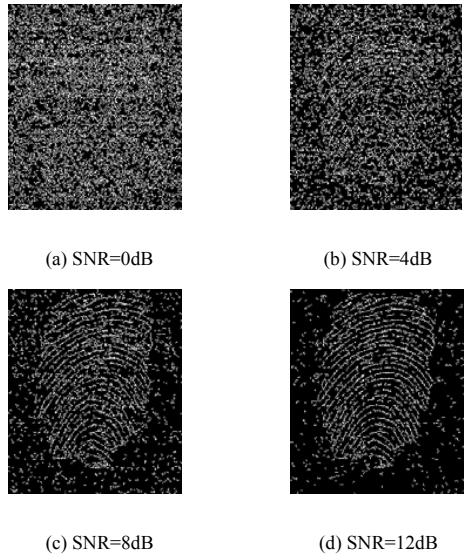


Fig.6 Denoised binary fingerprint image extracted from the stego-speech in AWGN channel with different SNR



(e) SNR=16dB (f) SNR=20dB

Fig.7 Denoised binary fingerprint image extracted from the stego-speech in Rayleigh channel with different SNR

C. *Accuracy of fingerprint matching*

OFDM system has two types of guard interval including CP (Cyclic Prefix) and ZP (Zero Padding). Table 1 ~ Table 4 list the accuracy rate of binary fingerprint matching under the AWGN channel and Rayleigh channel with CP and ZP guard interval, respectively. From Table 1~Table 4, we see the accuracy rates of binary fingerprint matching get higher as the SNR value increases. In addition, the accuracy rates under BPSK and QPSK demodulation are higher than 16-QAM and 64-QAM. As we known, the Rayleigh channel is more close to the real wireless channel than AWGN channel, so the Rayleigh channel fading is more serious. Therefore, the accuracy rates of binary fingerprint matching under the Rayleigh channel are lower than that of AWGN channel with both CP and ZP guard interval.

In the 4G communication standards, 256-QAM is also used, but the performance evaluations show that the introduction of 256-QAM does not significantly change the overall network performance in dense urban network deployments [16]. This is because the 256-QAM is only efficient when the serving cell's signal quality is good, which may not always be guaranteed due to the inter-cell interferences and channel characteristics. Therefore, in this paper, we will not perform the performance evaluations under the assumption of 256-QAM.

Table 1 Accuracy rate of binary fingerprint matching under the AWGN channel with CP guard interval

SNR/dB	0	2	4	6	8	10
BPSK	0.9230	0.9678	0.9881	0.9949	0.9960	0.9961
QPSK	0.9233	0.9678	0.9872	0.9944	0.9960	0.9961
16-QAM	0.7468	0.8151	0.8963	0.9547	0.9838	0.9936
64QAM	0.7340	0.7493	0.7743	0.7924	0.8169	0.8437

Table 2 Accuracy rate of binary fingerprint matching under the AWGN channel with ZP guard interval

SNR/dB	0	2	4	6	8	10
BPSK	0.9213	0.9670	0.9880	0.9943	0.9960	0.9961
QPSK	0.9227	0.9669	0.9884	0.9945	0.9961	0.9961
16-QAM	0.7355	0.8036	0.8849	0.9510	0.9830	0.9937
64-QAM	0.7341	0.7725	0.8141	0.8698	0.9249	0.9631

Table 3 Accuracy rate of binary fingerprint matching under the Rayleigh channel with CP guard interval

SNR/dB	0	4	8	12	16	20
BPSK	0.8832	0.9852	0.9857	0.9847	0.9928	0.9941
QPSK	0.7899	0.9718	0.9776	0.9854	0.9883	0.9945
16-QAM	0.6400	0.7985	0.8791	0.9679	0.9876	0.9942
64-QAM	0.7731	0.7458	0.7076	0.9952	0.9932	0.9940

Table 4 Accuracy rate of binary fingerprint matching under the Rayleigh channel with ZP guard interval

SNR/dB	0	4	8	12	16	20
BPSK	0.7198	0.9718	0.9862	0.9890	0.9943	0.9956
QPSK	0.8518	0.8690	0.9632	0.9863	0.9875	0.9888
16-QAM	0.6869	0.7992	0.9679	0.9741	0.9887	0.9953
64-QAM	0.7773	0.8317	0.9538	0.9748	0.9918	0.9947

IV. CONCLUSIONS

Speaker identity tracing is an effective solving method for the telecommunication fraud. In this paper, we propose a speaker identity tracing scheme against telecommunications fraud based on biological fingerprint data hiding. The speaker can be tracked by the extracted fingerprint information from the stego-speech via 4G wireless communication with OFDM technique. Therefore, the proposed scheme can be used to resolve the remote authentication of speaker identity, and also considered to apply in uncovering telecommunications fraud. Our future work will focus on the privacy protection of transmitting personal fingerprint information in the wireless channel.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (NSFC) under Grant No.U1536110

REFERENCES

[1] H. Wu, Y. Wang, and J. Huang, "Identification of electronic disguised voices," *IEEE Trans. on Information Forensics and Security*, vol. 9, No.3, pp. 489-500, 2014.

[2] H. Wang and M. Fan, "Centroid-based semi-fragile audio watermarking in hybrid domain," *Science China Information Sciences*, vol. 53, No. 3, pp. 619-633, 2010.

[3] R. Singh, A. Jiménez, and A. Øland, "Voice disguise by mimicry: deriving statistical articulometric evidence to evaluate claimed impersonation," *IET Biometrics*, vol. 6, No. 4, pp. 282-289, 2017.

[4] W. Cao, H. Wang, H. Zhao, and Q. Qian, "Identification of electronic disguised voices in the noisy environment," *12th International Workshop on Digital-Forensics and Watermarking (IWDW 2016),* Beijing, China, *LNCS 10082,* pp. 75-87, 2017.

[5] C. Lu, H. Wen, and S. Wang, "SCAN: learning speaker identity from noisy sensor data," *16th ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN 2017)*, Pittsburgh, PA, USA, pp.67-78, 2017.

[6] Z. Li, X. Hu, and N. Guo, "Judicial expertise of speaker identity based on improved pitch algorithm," *13th International Conference on Electronic Measurement & Instruments (ICEMI 2017)*, Yangzhou, China, pp. 401-405, 2017.

[7] A. Chadha and J. Nirmal, "A full band adaptive harmonic model based speaker identity transformation using radial basis function," *11th International Conference on Intelligent Systems and Control (ISC 2017*), Calgary, Canada, pp. 217-223, 2017.

[8] H. Kashani, A. Sayadiyan, and H. Sheikhzadeh, "Vowel detection using a perceptually-enhanced spectrum matching conditioned to phonetic context and speaker identity," *Speech Communication*, vol. 91, pp. 28-48, 2017.

[9] V. Vestman, D. Gowda, M. Sahidullah, P. Alku, and T. Kinnunen, "Speaker recognition from whispered speech: A tutorial survey and an application of time-varying linear prediction," *Speech Communication*, vol. 99, pp. 62-79, 2018.

[10] S. Barlaskar, M. Laskar, N. Shome, and R. Laskar, "Study on the varying degree of speaker identity information reflected across the different MFCCs," *International Conference on Inventive Computation Technologies (ICICT, 2016),* Coimbatore, India, pp. 1-6, 2016.

[11] E. Argones-Ru´a, J. Alba-Castro, and C. Garcı´a-Mateo, "On the use of quality measures in face and speaker identity verification based on video and audio streams," *IET Signal Processing*, vol. 3, No. 4, pp. 301-309, 2009.

[12] S. Abdullahi, H. Wang, and Q. Qian, "Concealing fingerprint-biometric data into audio signals for identify authentication," *12th International Workshop on Digital-Forensics and Watermarking (IWDW 2016),* Beijing, China, LNCS 10082, pp. 129-144, 2017.

[13] J. Sang, H. Wang, Q. Qian, H. Wu, and Y. Chen, "An efficient fingerprint identification algorithm based on minutiae and invariant moment," *Personal and Ubiquitous Computing*, vol. 22, No. 10, pp. 71-80, 2018.

[14] H. Wang and M. Fan, "NDFT-based image steganographic scheme with discrimination of tampers," *KSII Trans. on Internet and Information Systems*, vol. 5, No. 12, pp. 2340-2354, 2011.

[15] ITU-T G-series: Voice and audio database, "ITU-T test signals for telecommunication systems," http://www.itu.int/net/itu-t/sigdb/menu.aspx, 2015.

[16] I. Kim, J. Um, and S. Park, "Implementation and performance evaluation of 256-QAM in Vienna system level simulator," *20th International Conference on Advanced Communications Technology (ICACT 2018)*, chuncheon, Korea, pp. 556-559, 2018.