Journal of Engineering and Applied Sciences 14 (21): 7922-7927, 2019
ISSN: 1816-949X

A Study on the Voice Authentication Security for Recorded Voice through Sound Color Marker Analysis

Bong-Young Kim and Myung-Jin Bae
Department of Information and Telecommunication Engineering, Soong-sil University,
Sang-do Ro, Dongjak-gu, 369 Seoul, Korea

Abstract: With the development of technology related to the 4th Industrial Revolution such as smart phone and internet of the things, more convenient financial services with biometric authentication are emerging one after another. Voice authentication is cheaper and easier to introduce than other biometrics authentication methods because it can provide the biometrics authentication function with only the microphone installed in the smart phone. Voice authentication is a very good authentication method because it is possible to apply the speaker presentation method such as sentence presentation in terms of security. Especially, since, it is very difficult to restore human voice to a common speaker, speech recognition is a highly secure biometric authentication method. In this study, we used sound color marker that express intuitively the voice components to see, if it is possible to distinguish between real voice and recorded voice. As a result of the experiment in Chapter 3, we could distinguish the real voice from the recorded voice by analyzing the similarity of the sound color marking and it was confirmed that the voice is very secure from the recorded voice and the processed voice.

Key words: Sound color marker, voice authentication, biometric authentication, security, recorded voice, authentication

INTRODUCTION

The Republic of Korea is a well-established IT powerhouse and many attempts, efforts and investments have been made in various fields to lead the 4th Industrial Revolution. In particular, the development of technologies related to the 4th Industrial Revolution such as smart phones and the internet of things, the emergence of new financial services such as fin tech, cloud funding and institutional support of the government are complex. This is followed by convenient non-face-to-face financial services that incorporate biometric authentication. A typical example is an internet-only bank where customers can use financial services without having to meet financial staff by replacing identity verification with video communication, biometrics, etc. Existing general banks are actively introducing financial services with biometric authentication. Some banks in the first financial sector have launched a service that allows them to conduct financial transactions such as account inquiry, sending money and exchanging money as a voice after iris authentication. And another common bank is preparing services for financial transactions through voice (Kim *et al*., 2015; Song and Kim, 2017).

In financial transactions through biometric authentication, security is an important factor that is comparable to convenience. Biometric authentication is superior in security and convenience compared with authentication methods such as password, token and security card. However, there is also a risk of biometrics replication or leakage potential. Biometrics can not be modified like a password and there is a great concern that it can still be exploited by a single leak. A representative example is that a German hacker group has replicated the iris of Russian President Vladimir Putin. With the development of video signal processing technology and 3D printing technology, biometrics hacking and duplication are becoming reality. Voice authentication is a safe and convenient way to lose, steal and steal. Even if the system is hacked, it is possible to secure enough security when introducing sentence independent speaker recognition and sentence presentation type speaker recognition. In addition, the voice is very secure because it is very difficult to duplicate it through a commercial speaker as a commercial microphone recognizes a real voice. Voice authentication is a biometric authentication method suitable for non-face-to-face financial services because it can introduce biometric authentication

Corresponding Author: Myung-Jin Bae, Department of Information and Telecommunication Engineering, Soong-sil University, Sang-do Ro, Dongjak-gu, 369 Seoul, Korea

functions at low cost by installing only microphones in ATM devices (Kim *et al*., 2015; Baek *et al*., 2010; Bae and Lee, 1998 and Lee, 2005).

In this study, we compare the voices of several people using sound color marker. "Real voice", "Recorded voice" and "voice output by amplifying the bass portion of recorded voice" were analyzed by voice data recognized by a microphone which is a biometric input device. Through this analysis, we want to see how voice authentication is suitable for non-face-to-face financial services by checking whether the recorded voice can be distinguished from real voice.

MATERIALS AND METHODS

Basic theory of voice generation

Voice generation principle: Each person has different voices, according to their priori and posteriori influences and is attracting attention as a biometric authentication means because it is convenient to store, safe to be lost and stolen. Opening and closing the vocal cords from the lungs creates the basic tone of the voice and through the throat through the mouth, teeth and nose, a complex ringing occurs and the voices are created by the complex combination of the sounds. These voices can be a feature that distinguishes a person by reflecting characteristics such as individual's physical structure as well as lifestyle (Lee, 2005; Park *et al*., 2016).

Voice generation model and voice analysis: The linear model for voice generation was developed by Fant in the late 1950's. It is a linear prediction model that assumes the speech output as a signal through which the sound source passes through the frequency filter and regards each part of the sound source and the vocal track as independent. They used a quasi-periodic pulse for voiced sound, a white noise for voiceless sound and modeled the effect of vocal cords on the sound source. It models the voice as a vocal track frequency filter that the voice source and the voice source communicate with and enables independent mathematical analysis and problem solving by modeling each voice independently (Lee, 2005; Park *et al*., 2016) (Fig. 1).

Frequency spectrum: Most of the sounds in everyday space are composite sounds mixed with various elements. Among them, human voice is the most complex compound sound. Frequency analysis is the decomposition of what constitutes a complex mixture of voices or sounds. When a voice passes through a filter, it is decomposed into various frequency components. The frequency spectrum shows the sound intensity on the vertical axis and the frequency on the horizontal axis.

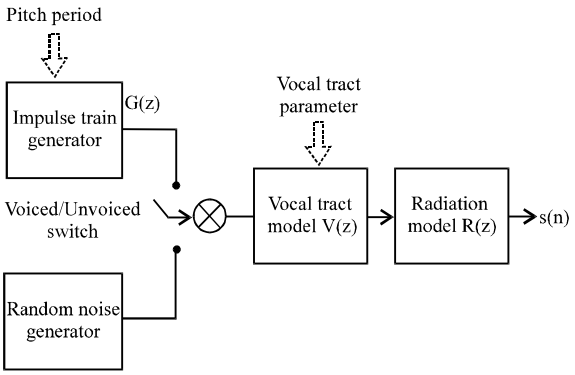

Fig. 1: Voice generator model (Lee, 2005)

Sound color marker: The sound color marker is expressed by assigning a rainbow color similar to the sense of light for each frequency band by dividing the frequency band by the log scale in the same way that the human auditory system responds to the mel scale. This is an indicator that intuitively recognizes the weight of energy of each element of sound through visual (Kim *et al*., 2018).

RESULTS AND DISCUSSION

Twenty different men and women were asked to read "There were mother pigs and three baby pigs in deep mountains" three times and they were collected as a sound source. Frequency analysis was performed for each case of the collected speech. Also, energy distribution by sound source was analyzed by using sound color marker. The sound source was collected with the SONY ICD-UX543F voice record and the voice file was sampled at 8000 Hz and 16-bit quantized. The software used was Audition CC and Cool Edit Pro 2.1.

The collected sound sources of women are called W1-W3 and male sound sources are called M1-M3. Voice data of 6 out of 20 people are divided into 4 types (A-D) for the voice of the same person. For the sound source of W1, W1-A is the first sound source. W1-B is the third acquisition source, W1-C is the rerecorded sound source for the first acquisition source and W1-D is the rerecorded source by amplifying the first acquisition source. For the analysis of the sound color marker for each sound source, the bandwidth was limited to 0~4 kHz

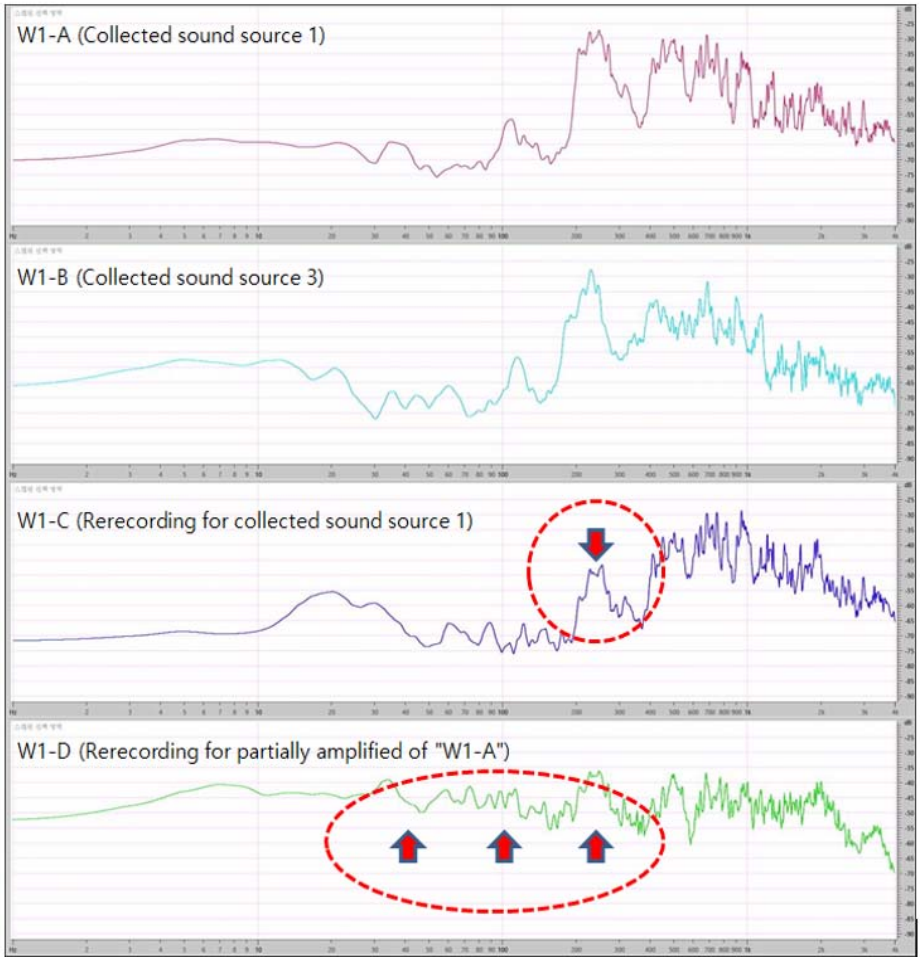

Fig. 2: Frequency spectrum of W1's sound sorce

for voice frequency analysis and the range was divided into 7 stages by mel scale. Red indicates a band of 0-20 Hz, orange 20-50 Hz, yellow 50-125 Hz, green 125-300 Hz, blue 300-800 Hz, indigo 800-2 kHz, purple 2-4 kHz it is the energy proportion of the band.

Figure 2 shows the frequency spectrum for the four sound sources of W1. In Fig. 2, we can see that 'W1-A' and 'W1-B' are very similar even though they are different sound sources. In the case of "W1-C" when the 'W1-A' sound source is rerecorded it can be seen that the range of 200~300 Hz is reduced to 20~25 dB. Also, in "W1-D", the range of 200~300 Hz of 'W1-A' sound source was amplified by 20~25 dB and rerecorded. However, compared with 'W1-A', the range of 200~ 300 Hz decreased (10~15 dB). Nevertheless, it can be seen that the interval of 200Hz or less of "W1-D" increases much which is much different from 'W1-A'. As a result, "W1-C" and "W1-D" were much different from 'W1-A' although, they were rerecorded to the same sound source as 'W1-A'. It can be seen that "W1-B" which is a different sound source is more similar to 'W1-A'.

Figure 3 shows the sound color marker for four cases of object 6. In Fig. 3, the sound sources A and B were not significantly different for each individual but the C sound sources rerecorded on the A sound source and A sound source differed greatly. It was confirmed that the sound source D which was rerecorded by amplifying 100~500 Hz of the sound source A is also significantly different from the sound source A. Figure 3 shows the sound color marker for four cases of object 6. In Fig. 3, the sound sources A and B were not significantly different for each individual but the C sound sources rerecorded on the A sound source and A sound source differed greatly. It was confirmed that the "D sound source" which was rerecorded by amplifying 100~500 Hz of the sound source A is also, significantly different from the sound source A.

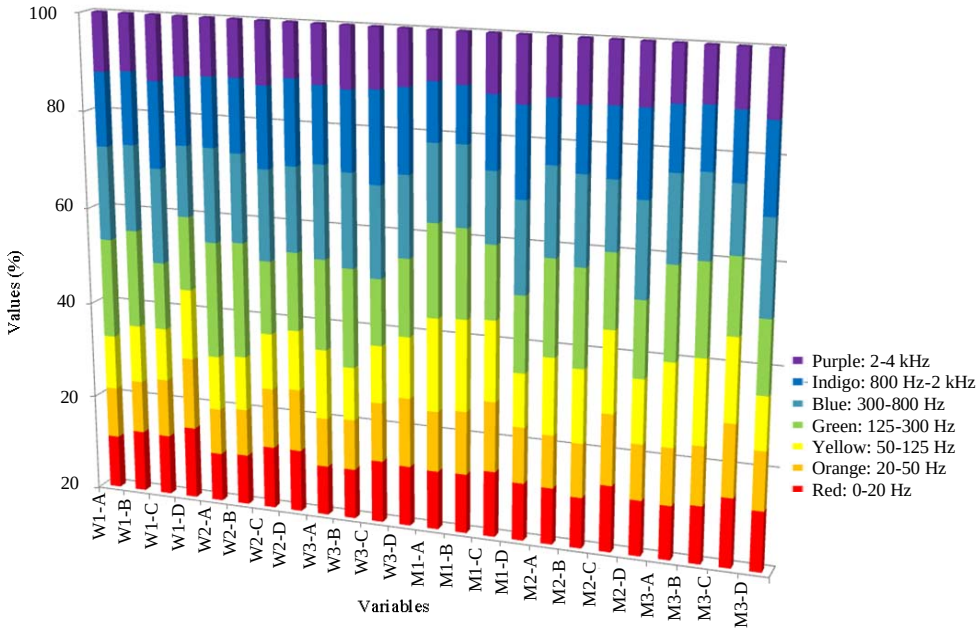

Fig. 3: Sound color markers for all sound sources

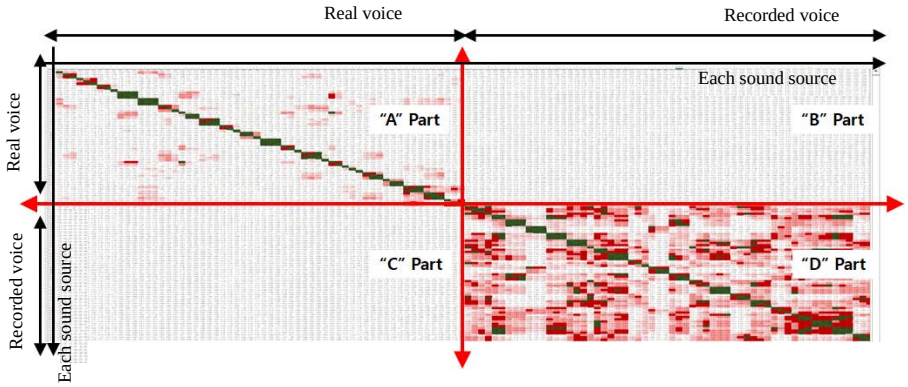

Fig. 4: Similarly analysis result for each sound source

Especially for women, the rerecord C sound source has a much lower energy ratio of the green section than the A source. However, in the case of the D source amplified by amplifying some of the bands, the energy ratio of the green region is lower than that of the A source and the energy ratio of the red region to the yellow region is much higher. In the case of the male, the energy ratio of the green section is much lower than that of the A source and the energy ratio of the green section is lower than that of the A source in the case of the rerecorded D source. The energy ratio of the high frequency band of the purple section is increased. For more precise comparison of the previous experiments, we compared the similarity of the sound sources of all 20 speakers. The comparative analysis was conducted on a total of 120 sound sources, including 60 sound sources collected from 3-20 persons and 60 sound sources outputting the same sound source to speakers. Similarity was calculated as the following Eq. 1:

$$\text{Similarity between sound sources}(\%) = 1 - \sum_{k=1}^{7} |x_k - y_k| \quad (1)$$

Where:

x_k, y_k = The ratio of energy in the k bands of the two sources to be compared (%)

k = Order of frequency band of sound color marker

Figure 4 shows the results of similarity analysis for each of 120 sound sources. When the similarity is over

Table 1: Voice identification rate calculation result

	Identification rate (Similarity threshold)		
	Over 94%	Over 95%	Over 96%
Identification of the same person	86	79	76
Identification of the different person	92	96	98
Recorded voice identification	100	100	100

94%, it is colored in the comparison result field and when the similarity is larger, it is displayed in dark color. The results of the similarity comparison shown in Fig. 4 show that the similarity between the voices of the same person is very high in the "A" part comparing the similarities between the actual voices. In the "D" part which compares the similarities between recorded voices, most sound sources are similar regardless of who the voice actually is. On the other hand, in the cases of "B" and "C" parts comparing real voice and recorded voice, it was confirmed that there is almost no similarity regardless of voice.

Table 1 shows the result of calculating the identification rate using the sound color marker based on the result of Fig. 4. As shown in Table 1 above, it can be confirmed that the identification of the voice and recorded voice is possible with only a very high probability by the similarity analysis of the sound color marker. Also, the ability to distinguish the voices of others in comparison with others is 92~98% depending on the threshold. However, in the comparison of the voices of the same person, 76~86% of the thresholds were found to be somewhat lower than those of the other voices. These results show that the sound color marker can be used to distinguish the real voice from the recorded voice easily. In addition, the sound color marker can be used very usefully for individual voice identification.

CONCLUSION

Due to the development of technology related to the 4th Industrial Revolution, biometric authentication which can replace public authentication certificates, passwords and security cards which are the main authentication methods in financial services is continuously being introduced. Biometric authentication is superior to other authentication methods in terms of security and convenience. However, higher security is needed because biometrics can be continuously exploited if duplicated or leaked. Voice authentication is very secure because it is very safe to lose and steal and it is very difficult to record or play the same as a real voice.

In this study, we have confirmed whether real voice and recorded voice can be distinguished through frequency analysis and sound color marker analysis of voice and whether rerecord voice amplified from 100~ 500 Hz band can be distinguished from real voice. In addition, we confirmed "Identification of the same person", "Identification of the different person" and "Recorded voice Identification" by comparing the similarity between 120 real voice and 120 recorded voice sources. Through this analysis, we have confirmed how voice authentication using voice is appropriate for financial services.

Experimental results show that the similarity of the first and third sound sources collected from the same person is very high. However, even if the same sound source is used, the rerecord sound source has a very low similarity with the original sound source. Also, the similarity of the rerecord sound source with the original sound source is very low by amplifying a part of the original sound source. In addition, it was possible to easily distinguish the recorded sound source from the processed sound source by only frequency analysis and sound color marker analysis. It can be seen that it is very difficult to reproduce the voice the same as an individual's real voice due to the frequency response characteristics of the speaker outputting the sound source and the microphone collecting it. Therefore, it is almost impossible for a third party other than myself to reproduce the recorded voice with unjustifiable intention to replace the legitimate rights holder. This conclusion suggests that voice authentication is a very good biometric authentication method and is very suitable for financial services. As such, we expect that voice authentication which has excellent security will be expanded to provide more convenient and secure financial services.

REFERENCES

Bae, M.J. and S.H. Lee, 1998. Digital Speech Analysis. Dong Young Diamond Industrial Co. Ltd., South Korea..

Baek, G.R., J.J. Yun and M.J. Bae, 2010. [A study on a similarity of a voice in the same lineage (In Korean)]. Proc. Acoust. Soc. Korea, 29: 447-448.

Kim, B.Y., E.Y. Yi and M.J. Bae, 2018. A study on sound a color about distinguishing voices characteristic. Asia Pac. J. Multimedia Serv. Convergent Art Humanities Sociology, 8: 13-21.

Kim, S.H., Y.S. Cho and D.S. Chi, 2015. FinTech Era: Needs for the innovation of user authentication technologies. Commun. Korean Inst. Inf. Sci. Eng., 33: 17-22.

Lee, Y.J., 2005. A study on robust mixture model with an optimal number of mixtures for speaker recognition. Ph.D Thesis, Soongsil University, Seoul, South Korea.

Park, H.W., S.H. Jee and M.J. Bae, 2016. Study on the confidence-parameter estimation through speech signal. Asia Pac. J. Multimedia Serv. Convergent Art Humanities Sociology, 6: 101-108.

Song, J.H. and I.S. Kim, 2017. A study on the utilization of biometric authentication for digital signature in electronic financial transactions: Technological and legal aspect. J. Soc. E. Bus. Stud., 21: 42-53.