International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 2 (2019), pp. 157-161

A Study on the Validity of Speaker Identification Using Sound Color Marker

Bong-Young Kim[1], Sung-Tae Lee[2] and Myung-Jin Bae[3*]
[1]Soong-sil University, Department of Information and telecommunication Engineering, Seoul, 06978, Korea.
Orcid Id : 0000-0002-3553-039X
[2]Soong-sil University, Department of Information and telecommunication Engineering, Seoul, 06978, Korea.
[3] Soong-sil University, Department of Information and telecommunication Engineering, Seoul, 06978, Korea.
Orcid Id : 0000-0002-7585-0400
**Corresponding author*

Abstract:

The development of IT technology, such as smart devices and pin-tecs, is leading to convenient and highly secure biometric authentication. In biometric authentication, voice authentication is a very efficient biometric authentication method in terms of convenience, security, and economy. The voice is expressed through the complex energy distribution and change in the sense of personality and words. In this paper, we use the Sound Color Marker calculated from the time series average spectrum of voice signals to verify the validity of speaker identification. In experiments with adults 10 speech, extraction time for reference sound source is not less than 60 seconds, extraction time for the sound source to be discriminated if more than 10 seconds, the speaker identification rate using the Sound Color Marker was very high. Also, even if we do not know what it says, we found that the longer the extraction time for the reference sound source and the extraction time for the sound source to discriminate, the higher the speaker identification rate. These results show that context-independent type speaker identification is possible only by comparing the sound color markers, which are simply expressed by seven parameters.

Keyword: Biometric Authentication, Voice Signal, Sound Color Marker, Speaker Identification, Identification Rate

1. INTRODUCTION

In recent years, the development of IT technology such as smart devices, fin tech, big data, artificial intelligence, and block chain has made interest in biometric authentication more convenient and secure. Especially, the government of the Republic of Korea is trying to abolish the certificate system until the revision of the law, and in the related industries, there is an active movement to introduce biometric authentication as a means of replacing the certificate. Biometric authentication is a method of recognizing information using fingerprint, voice, face, iris, and the lines of the palm. There are many ways to improve accuracy and security by using several methods together. Among these, authentication method using voice is safe for theft, loss, duplication, and it is very effective authentication method because it can increase the security according to the operating method and the introduction cost is very low [1-5]. However, the voice is so complex and irregular that the information may appear very different depending on the contents and the state of the speaker, and it may be easily changed to the microphone characteristic of the bio-information collecting device and the background noise. Therefore, in case of authentication using voice, it is necessary to apply the method that is suitable for the object and purpose of authentication. We can improve the efficiency of authentication by using simple discrimination using parameters that are not influenced by various situations or peripheral characteristics, and by using fine discrimination using parameters of very complex characteristics [4-5].

Authentication of a speaker can be classified into context-independent type and context-independent type depending on the content of vocalization. The context-dependent type is a method of identifying a speaker by speaking a predetermined sentence and comparing it with the expected reference value. And the context-independent type is a method to identify the speaker by comparing feature vectors extracted from learning from various existing data and feature vectors extracted by uttering unexpected sentences. For context-independent types, much more training data should be collected than context-dependent types. Therefore, it is true that more research has been done on the Identificat method of context-dependent type. However, speaker identification that can not speak promised sentences is also very common, so speaker identification of context-independent type is also highly required [5-8].

In this paper, we applied Sound Color Marker as a very simplified parameter while minimizing the influence of various situations or peripheral characteristics. We also tried to verify the validity of context-independent type speaker identification using Sound Color Marker through experiments. Chapter 2 describes the characteristics of the time series average spectrum of the voice signal and the Sound Color Marker. Chapter 3 describes the experiment and the results. Chapter 4 concludes with conclusions.

2. THE CHARACTERISTICS OF TIME SERIES AVERAGE SPECTRUM OF VOICE SIGNAL AND SOUND COLOR MARKER

2.1 The characteristics of time series average spectrum of voice signal

A person's voice signal contains numerous personal characteristics and meanings of words in the distribution of a large number of frequency components that change in real time.

These complex personal characteristics and meanings of words can be confirmed by analysis of shot-time spectrum. The voice signal that changes in real time can be simplified by a statistical method. In this case, the personal characteristics or the meaning of words seen in real-time change are extremely lost. However, statistically simplifying the long-term voice signal can be said to approximate the most basic characteristics of the speaker's vocal organs. In Figure 1, person "A" reads the heading of the book "Principles and Application of Sound." The time series average spectrum is obtained by extracting arbitrarily 240 seconds, 60 seconds, 30 seconds, 10 seconds and 5 seconds of the entire range. For comparison with different persons, the spectrum of the person "B" extracted from the 60-second sound source is also displayed [9-14].

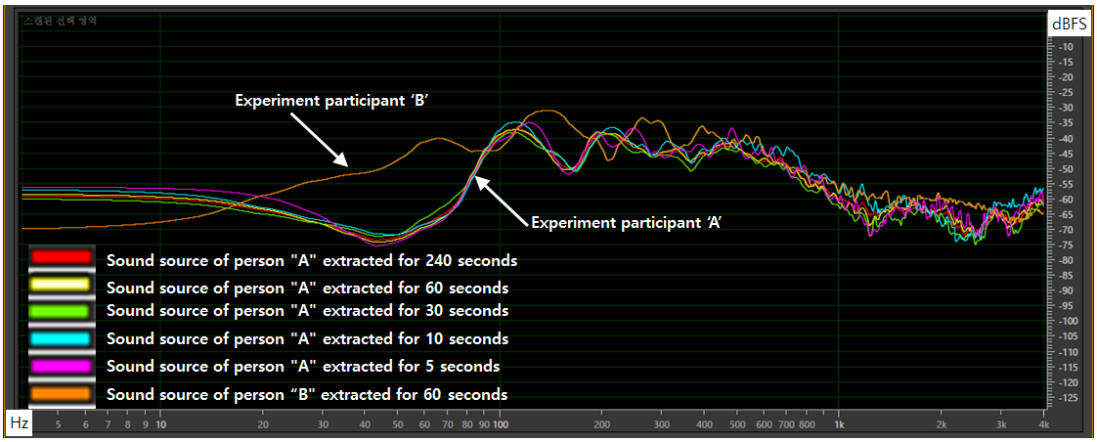

Fig 1. Time Series Average Spectrum

In Figure 1, the spectrums for "A" show that the longer the extraction time is, the more approximated the 240 second time series average spectrum, although the content is very different. Also, the extraction of the 5 second voice signal is very similar to the 240 second time series average spectrum. However, in case of "B", the time series average spectrum is very different from those of "A" regardless of extraction time.

2.2 Sound Color Marker

Sound Color Marker was developed to measure the sound of a human perceives a sound. It is used for the intuitive expression and analysis of sound by displaying the sense of the sound by the seven colors in allude to the visual sensation that the person recognizes the light. A person is recognized as a log scale when the frequency of sound increases. This range of recognition is divided into seven levels, and the magnitude of the sound is represented by the weight of each frequency step. This method of expression makes it possible for a person who does not have a knowledge of the volume scale to intuitively recognize what type of sound is distributed. Figure 2 shows the time series average spectrum extracted for 240 seconds in Figure 1 as the Sound Color Marker [15-17].

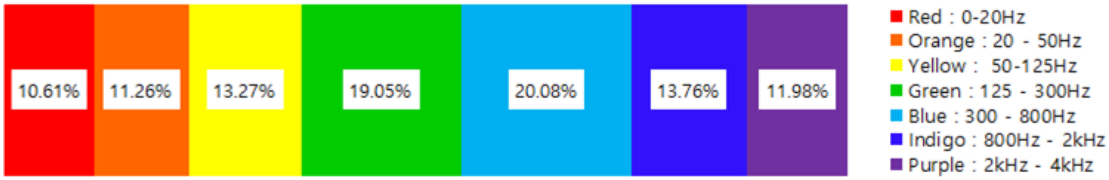

Fig 2. Example of a Sound Color Marker

3. EXPERIMENTS AND RESULTS

In this paper, we tried to verify whether the Sound Color Marker, which is simplified by 7 parameters, can be used as feature vectors of context-independent type through experiments. We also tried to verify whether the Sound Color Marker is useful for context-independent type speaker identification. In the experiment, 10 different men and women read the headings of "Principles and Application of Sound" published in "Cheong Moon Gak" as if they were always reading a book. The voice signal was collected with Galaxy Note 4, and the voice file was sampled at 8 kHz and quantized to 32 bits. Audition CC and Cool Edit Pro 2.1 were used for the analysis.

The voices of 10 participants were divided into A to J according to the order. Then, nine random sources were extracted from arbitrary sections of each experimental participant sound source. The first sound source was obtained by extracting the sound source with 240 second interval in each sound source. And the extraction of the two sound sources in the 60 second interval is defined as the second and third, respectively. In the

same way, the two sound sources extracted for 30 second interval are defined as 4th and 5th, respectively, and the two sound sources extracted for 10 second interval are defined as 6th and 7th, respectively. Finally, the two sound sources extracted in the 5 second interval were set as 8th and 9th, respectively. The fifth sound source of the fourth participant is represented by 'D-5'. 'I-8' represents the 8th extraction sound source of the 9th participant. Each extracted sound source was converted into a Sound Color Marker, and the similarity between the sound sources was compared and analyzed. The similarity between the sound sources is calculated by the following equation (1).

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

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 3 shows the similarity between each sound source extracted from experiment participants A and B. When the similarity is more than 94%, the similarity mark is displayed in color, and the more similarity is, the darker the color is.

Reference sound source	Discrimination target sound source																	
	A-1	A-2	A-3	A-4	A-5	A-6	A-7	A-8	A-9	B-1	B-2	B-3	B-4	B-5	B-6	B-7	B-8	B-9
A-1		0.9888	0.9844	0.9874	0.9878	0.9932	0.9852	0.9828	0.9899	0.8590	0.8639	0.8578	0.8610	0.8610	0.8662	0.8468	0.8476	0.8610
A-2			0.9887	0.9814	0.9847	0.9890	0.9928	0.9864	0.9870	0.8545	0.8657	0.8536	0.8564	0.8564	0.8621	0.8423	0.8454	0.8637
A-3				0.9723	0.9888	0.9853	0.9836	0.9751	0.9780	0.8470	0.8578	0.8461	0.8490	0.8490	0.8547	0.8348	0.8379	0.8540
A-4					0.9817	0.9850	0.9809	0.9839	0.9902	0.8645	0.8681	0.8614	0.8658	0.8658	0.8698	0.8556	0.8512	0.8647
A-5						0.9838	0.9816	0.9735	0.9776	0.8485	0.8526	0.8455	0.8499	0.8499	0.8539	0.8378	0.8353	0.8488
A-6							0.9838	0.9799	0.9882	0.8617	0.8663	0.8599	0.8636	0.8636	0.8683	0.8494	0.8497	0.8632
A-7								0.9861	0.9832	0.8513	0.8625	0.8504	0.8532	0.8532	0.8589	0.8391	0.8422	0.8614
A-8									0.9873	0.8632	0.8744	0.8623	0.8652	0.8652	0.8708	0.8510	0.8541	0.8753
A-9										0.8669	0.8734	0.8660	0.8689	0.8689	0.8746	0.8547	0.8577	0.8711
B-1											0.9866	0.9879	0.9970	0.9970	0.9886	0.9820	0.9868	0.9841
B-2												0.9840	0.9881	0.9881	0.9849	0.9727	0.9789	0.9911
B-3													0.9865	0.9865	0.9914	0.9758	0.9846	0.9858
B-4															0.9898	0.9829	0.9855	0.9848
B-5															0.9898	0.9829	0.9855	0.9848
B-6																0.9798	0.9814	0.9860
B-7																	0.9821	0.9697
B-8																		0.9764
B-9																		

Fig 3. Comparison Result of Similarity between A and B Sound Sources

Figure 3 shows that the similarity between the sound sources 'A-1' and 'A-9' of the same person with different utterances and extraction time is over 94% in all cases. The similarity between the different sound sources 'B-1' ~ 'B-9' of the same person was also very high, over 94% in all cases. However, similarity between 'A-1 ~ A-9' and 'B-1 ~ B-9', which are the sound sources of different people, is very low. The similarity between 'A-1' extracted 240 seconds and 'A-9' randomly extracted for 5 seconds is 98.99%. The similarity between 'A-4' extracted for 30 seconds and 'A-9' extracted for 5 seconds is 99.02%. Through these comparisons, it can be seen that if the speaker is the same, the extraction information for 240 seconds, the information extracted for 30 seconds, and the information extracted for 5 seconds are very similar. However, if the speakers are different, the similarity is very low.

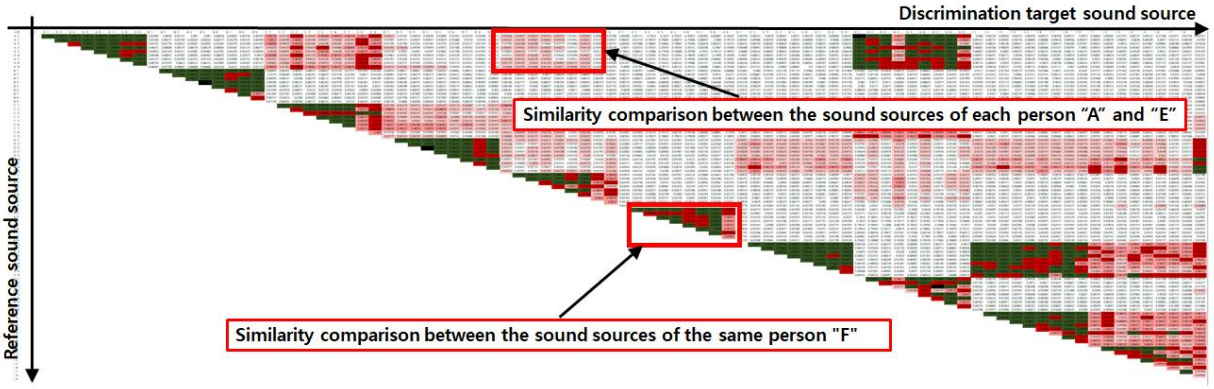

Fig 4. Comparison Result of Similarity between All Sound Sources

Figure 4 shows a cross-sectional comparison of the similarities among all the sound sources extracted nine times from ten people. Figure 4 shows that although the extraction time of the voice signal is different, if the speaker is the same person, the similarity is generally high. However, the similarity between the experimental participant 'A' and the experimental participant 'H' is somewhat higher, and the similarity between the experimental participant 'G' and the experimental participant 'I' is somewhat higher.

Table 1 shows two probabilities from the results of Figure 4. One is the probability of discriminating the target sound source with its own voice, and the other is the probability of discriminating the target sound source with the other. The method of discrimination is discrimination by checking whether the discrimination target sound source is most similar to any person of the reference sound sources.

Table 1. The Result of Calculating Speaker Identification Rate

Extraction time for reference sound source (sec)	Probability of speaker recognition				Probability that both sound sources fail the speaker recognition			
	Extraction time for the sound source to be discriminated (sec)							
	60	30	10	5	60	30	10	5
240	100%	95%	95%	90%	0%	0%	0%	0%
60		90%	95%	75%		10%	0%	0%
30			80%	75%			10%	10%
10				60%				20%

In the results of Table 1, when discrimination target sound sources, which are extracted randomly for 60 seconds regardless of the content of speech, are compared with extracted reference sound sources for 240 seconds, the comparison result was successful in discriminating the actual owner of the voice. On the other hand, when comparing discrimination target sound sources extracted randomly for 5 seconds with reference sound sources extracted for 10 seconds, only 60% discriminated the actual owners of the voices. As a result, it was confirmed that the longer the extraction time of the reference sound sources and the discrimination target sound sources, the higher the discrimination rate.

Compared with the same person's two sound sources (eg, 'C-2', 'C-3') extracted at the same time and the reference sound sources with an extraction time of 240 seconds, the probability of failing speaker discrimination is zero percent. In addition, the probability of failing speaker discrimination was 20% when comparing two sound sources of the same person with an extraction time of 5 seconds to reference sound sources with an extraction time of 10 seconds. As a result, it is confirmed that the probability of failure of both discrimina- tions in both data becomes lower as the extraction time of the reference sound source becomes longer, and the rate of failure to discrimination becomes lower as the extraction time of the discrimination target sound source becomes longer as well.

4. CONCLUSION

The voice authentication method is a very efficient biometric authentication method that can secure the theft, loss, duplication, etc. and apply various operation methods. Among these authentication methods, there is a context-independent type speaker identification that extracts feature vectors from learning data from a variety of existing data and identifies speakers by comparing unfamiliar sentences with feature vectors. The human voice expresses the personality and the meaning of the language with the composition and the change of the very complicated sound component. Thus, in human voices, there can be a wide variety of ways to extract feature vectors.

In this paper, we have experimentally confirmed whether speakers can be appropriately identified in context-independent type speaker identification using Sound Color Marker, which expresses speech characteristics intuitively with only 7 parameters from time series average spectrum. As a result, if the extraction time of the reference sound source is more than 60 seconds and the extraction time of the discrimination target sound source is more than 10 seconds, even if a sound source does not have the same content, the probability of identification by the speaker is more than 90%. Also, if the extraction time of the reference sound source is more than 30 seconds, for two sound sources of the same person with an extraction time of 5 seconds, the probability of failing speaker identification was less than 10%. These results show that although the Sound Color Marker is represented by seven parameters using only statistical techniques without any learning, it plays a proper role as a feature parameter of context-independent type speaker identification. It is a high level.

Although the speaker identification rate may be very different depending on the number of subjects to be discriminated or the criterion of discrimination, in the case of identification of non-cooperative speaker identifiers or extraction of similar groups in too many identifiers, Sound Color Marker Is very effective. In addition, if the speaker is identificated by combining the

sound color marker with another feature parameter, it is expected that it will be sufficiently applicable even if the number of objects to be discriminated increases.

REFERENCE

[1] Jae-Hun Song and In-Seok Kim. "A Study on the Utilization of Biometric Authentication for Digital Signature in Electronic Financial Transactions - Technological and Legal Aspect -", *The Journal of Society for e-Business Studies*, Vol.21, No.4, November 2016, pp.41-53.

[2] S. H. Kim, Y. S. Joo, D. S. Chi, "FinTech Era: Needs for the innovation of user authentication technologies", *Communications of the Korean Institute of Information Scientists and Engineers*, Vol.33, No.5, May 2015, pp.17-22.

[3] S. M. Jin, S. W. Cho, K. S. Lee, S. M. Jung and Y. W. Jung, "Speaker recognition based smart card security technology", *Proceedings of KIIS Spring Conference* Vol.25, No.1, 2015, pp.25-26.

[4] Y. J. Lee. A Study on Robust Mixture Model with an Optimal Number of Mixtures for Speaker Recognition, Ph.D paper, SoongSil Univ., 2005.

[5] H. Y. Chung, "Speaker Recognition Technology", *Journal of Communications of KIISE*, Vol.19, No.7, July 2001, pp.32-44.

[6] D. H. Kim, W. K. Seong and H. K. Kim, "Performance Comparison of Deep Feature Based Speaker Verification Systems", *Journal of the Korean Society of Speech Sciences, ISSN 2005-8063*, Vol.7, No.4, December 2015, pp.9-16.

[7] [http://terms.naver.com] Speaker Verification (Telecommunications Technology Association, Korea)

[8] [http://terms.naver.com] Speaker Verification (Maeil Business News Korea)

[9] S. T. Lee. Principles and Application of Sound, Cheong Moon Gak, Korea, 2004.

[10] M. J. Bae and S. H. Lee. *Digital Speech Analysis*, DongYoung, Korea, 1998.

[11] H. W. Park, S. H. Jee and M. J. Bae. "Study on The Confidence-Parameter Estimation Through Speech Signal", *Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, Vol.6, No.7, 2016, pp.101-108.

[12] S. G. Bae, B. M. Lim and M. J. Bae, "A study on Kim Jung-un's Obesity through Speech Signal Correlation Analysis", *Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, Vol.6, No.7, 2016, pp.109-116.

[13] Doo-Heon Kyon and Myung-Jin Bae, "A Voice Similarity Study about Vocal Organ Proportion", *Conference Proceedings of Voice Communication and Signal Processing, The Acoustical Society of korea*, Vol, 28, No.1, 2011, pp.103-104.

[14] W. H. Lee, G. R. Baek, M. J. Bae, Editors. "Valid-frame Distance Deviation of Drunk and non-Drunk Speech", *The Korean Institute of Communication and Information Sciences (ISSN 2287-2639)*, Vol.53, 2014, pp.876-877.

[15] Bong-Young Kim and Myung-Jin Bae, "A Study on Identification Rate Difference of Sound Color Marker According to Bandwidth-limited of Voice Signal", *International Journal of Engineering Research and Technology*, Vol.11, No.9, 2018, pp. 1463-1470.

[16] Bong-Young Kim, Eun-Young Yi, and Myung-Jin Bae, "A Study on Sound A Color about Distinguishing Voices Characteristic", *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology*, Vol.8, No.2, February 2018, pp.13-21.

[17] Eun-Young Yi, Uk-Jin Song and Myung-Jin Bae, "A Study on Noise Characteristics in Subway Using Sound A Color", *Conference Proceedings of Convergence Research Letter, Convergent with Art, Humanities, and Sociology,* January 2017, Vol.3, No.1, pp.1053-1056.