Hindawi
Security and Communication Networks
Volume 2022, Article ID 5755785, 10 pages
https://doi.org/10.1155/2022/5755785

Research Article

User Authentication Method via Speaker Recognition and Speech Synthesis Detection

Hyun Park and TaeGuen Kim

SoonChunHyang University, Asan, Republic of Korea

Correspondence should be addressed to TaeGuen Kim; tg.kim@sch.ac.kr

Received 28 October 2021; Accepted 17 December 2021; Published 10 January 2022

Academic Editor: Hao Peng

Copyright © 2022 Hyun Park and TaeGuen Kim. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

As the Internet has been developed, various online services such as social media services are introduced and widely used by many people. Traditionally, many online services utilize self-certification methods that are made using public certificates or resident registration numbers, but it is found that the existing methods pose the risk of recent personal information leakage accidents. The most popular authentication method to compensate for these problems is biometric authentication technology. The biometric authentication techniques are considered relatively safe from risks like personal information theft, forgery, etc. Among many biometric-based methods, we studied the speaker recognition method, which is considered suitable to be used as a user authentication method of the social media service usually accessed in the smartphone environment. In this paper, we first propose a speaker recognition-based authentication method that identifies and authenticates individual voice patterns, and we also present a synthesis speech detection method that is used to prevent a masquerading attack using synthetic voices.

1. Introduction

As online services that provide many functions to create a relationship between people, such as social media services, are widely used by people. Many users use the social media service not only to contact other people but also to get lots of information, and social media services can be easily accessed by any device that is connected to the Internet. The social media service is very common to the user who is familiar with information technology. Social media activities are just everyday things to them. The content on social media is mostly opened to the public, but it is considered as personal data that should be protected from unauthorized people or attackers.

Many attacks targeting the social media service have been increased dramatically along with the number of SNS users, and there are many threats [1, 2] on smart devices that might be used for social media. The cyber security survey [3] revealed that in 2019, 33 percent of organizations worldwide were targets of one to ten social media attacks. Even in this situation, many social media users do not carefully manage their accounts or security tokens such as a password. According to Thycotic's report [4], hacking social media accounts has never been easier. Attackers take advantage of poor password hygiene and usually hijack an account and hold it for ransom. It is difficult to get the real picture, but according to Facebook, accounts are hacked 600,000 times a day. And even worse, 80% of all cyber security attacks involve a weak or stolen password. In this regard, many security mechanisms such as malware detection and wireless network security [5, 6] have been researched, and among them, user authentication is considered as the fundamental concept for securing the user system.

Previously, many self-certification methods for user authentication were made using public certificates or resident registration numbers, but existing methods pose a high risk of recent personal information leakage accidents. Personal information leakage incidents are on the rise every year, and there is also the hassle of periodically updating the medium used in existing identity methods [7, 8]. One of the authentication methods used to compensate for these problems is biometric authentication technology [9].

Biometric authentication techniques [10–12] use biometric features to uniquely define individuals, and biometric authentication is considered to have a much lower risk of information theft, forgery compared to traditional security methods [9]. Typical biometric authentication technologies include face recognition, iris recognition, fingerprint recognition, and speaker recognition. Recently, biometric authentication technologies have begun to be applied to multiple devices such as smartphones and laptops, and the biometric market has expanded as the use of security-critical groups such as companies, government agencies, and financial institutions increases. It is also used in mobile applications and telecommunications companies are also using speaker recognition technology to ensure that users access their accounts. The representative equipment for acquiring voice is a microphone, which has the advantage of being less expensive than equipment for acquiring other biometric information such as face recognition, iris recognition, and fingerprint recognition.

In this research, we propose a speaker recognition method that identifies and authenticates individual voice patterns. Speaker recognition is a function of determining the owner of a voice by using features from an individual's voice. In the speaker recognition, the learning process for the registered voice data is performed by using a deep neural network, and whenever the voice-based authentication is conducted, the given individual voice patterns to be tested are discriminated based on the deep learning model learned by the registered user's void data. In addition to this, we conducted experiments for showing the risk of synthetic speech in speaker recognition techniques, and the method for preventing this problem is applied in the proposed method.

The main contribution of our research can be summarized as follows:

(i) The proposal of a deep learning method utilizing speech data for user authentication

(ii) The demonstration of a masquerading attack to avoid speech-based authentication

(iii) The proposal of a novel framework that consists of two main components: user authentication component and synthetic speech detection component

2. Related Work

Yang and Das in [13] propose two new features, ICQC (Inverted Constant-Q Coefficients) and ICQCC (Inverted Constant-Q Cepstral Coefficients), using DCT (discrete cosine transformations) in the inversion octave power spectra and inversion linear power spectra, respectively. In addition, the task is extended using DCT and redundant block transformations so that features extracted from the entire frequency band are not easily affected by noise from some specific frequency bands. Thus, two new features derived from inverted octave blocks and inverted linear blocks are called ICBC and ICLBC, respectively. Using the derived features, they detect spoofing attacks via synthetic speech, focusing on CQT (Constant-Q transform) based on high-frequency information investigation.

De Leon and Stewart in [14] detect the synthesized voice by requesting a sentence containing words that clearly distinguish between natural and synthetic speeches from the user. Preliminary work on synthetic speech detection is done by analyzing words that strongly distinguish between natural speech and mechanical synthesis speech in humans. It was based on an informal test in which a specific word is observed like a synthetic speech rather than another general voice, regardless of the synthesizer or vocoder. Such sound is most likely due to unnatural modeling of a particular phoneme but can be the basis for improving speech identification by analyzing common words. The target application is a text-dependent SV (speaker verification)-based authentication system that asks the user for specific phrases containing many words that distinguish natural and synthetic speeches.

According to Wu et al. in [15], the modulation function derived from the size/phase spectrum conveys long-term information of the voice, where it detects temporal artifacts due to frame-by-frame processing of speech signal synthesis.

According to Saratxaga et al. in [16], using that most of the speech processing techniques do not consider phase information, phase perturbation is detected to prevent synthetic impersonators from attacking the speaker verification system. Review systems based on Modified Group Delay and systems based on Relative Phase Shift.

Paul et al. in [17] use a new short-term spectrum function that is efficiently distinguished from each other in the characteristics of synthetic and natural voices.

Sanchez et al. in [18] detect synthetic speech based on different phase structures of natural speech and synthetic speech.

Yang et al. in [19] propose a new method called subband transformation. Subband transformation has been shown to capture artifacts more effectively in synthetic speech than overall band transformation. For constant-Q equal subband transformations (CQ-EST), constant-Q octave subband transformations (CQ-OST), and discrete Fourier-Mel subband transformations (DF-MST), they propose an iso-subband transform. Studies have demonstrated that functions based on subband transformation outperform those based on full-band transformation in clean and noisy conditions.

Hanilçi et al. in [20] compare and analyze existing commercial synthetic speech detectors in the state of abominable noise contamination, especially the front end, to view the performance of synthetic speech detection in a noise environment. Studies show that synthetic speech detection techniques in noisy and nonnoisy environments show significant differences. Wu et al. in [21] assume that the difference in distribution between natural and synthetic speeches is an important discriminatory feature and uses a method called functional unification that learns genuinization with CNN (convolutional neural network) using only the characteristics of natural voices.

Hassan and Javed in [22] propose an effective synthetic speech detector that uses the fusion of spectral characteristics. Specifically, a fusion feature vector consisting of MFCC (Mel-frequency Cepstral Coefficient), GTCC

(Gammatone Cepstral Coefficient), Spectral Flux, and Spectral Centroid is proposed. This can capture the voice variation attribute of the actual signal and the algorithm artifact of the synthesized signal.

You et al. in [23] propose an antispoofing system that uses One-class learning to detect unknown synthetic speech spoofing attacks, such as text speech conversion. The key idea is to compact the speech representation and inject an angular margin to separate the spoofing attacks in the embedding space. They outperformed all existing single systems by achieving an error rate of 2.19% on the evaluation set of automated speaker-verified spoofs for 2019 challenge logical access scenarios without the use of data scaling methods.

De Leon et al. in [24] propose a new function based on relative phase shift and suggests a method of improving the security of the speaker verification system using the corresponding classifier. Sanchez et al. in [25] present a synthetic speech detector that can be connected to the front end or back end of a standard speaker verification system. Proposed systems are binary classifiers based on Gaussian mixed models. Three state-of-the-art vocoders are selected and modeled using two sets of acoustic parameters: relative phase shift and standard MFCC.

3. Our Proposed Method

We developed the voice-based authentication model that learns and discriminates each user's voice data using a deep neural network. For training the model, the normal user voices are firstly collected, and their MFCC (Mel-frequency Cepstral Coefficient) feature vectors are extracted and used for tuning the model. In addition, we devised the synthetic speech detection module and it is added as a part of our proposed method. The overall processing flow of our method is described in Figure 1. The proposed method can be divided into two modules: MFCC based user authentication module and the Mel-Spectrogram-based synthetic speech detection module.

The voice-based authentication module and the synthetic speech detection module share the same processes; preemphasis process, framing & windowing process, FFT (Fast Furrier Transformation) process, Mel-Scale filter bank process. These common processes are conducted to extract the feature for two major modules.

3.1. Common Data Refinement Processes. The common data refinement processes are performed for the feature extraction, and the features retrieved from these processes are MFCC features and Mel-Spectrogram features. The MFCC algorithm focuses on the most important part of the signal by quantifying and reflecting the signal spectrum while simultaneously eliminating the microscopic part of the less important spectrum. MFCC features are extracted from Mel-Scale that does not analyze the entire voice data in batches but divides it into specific sizes to conduct spectral analysis for each interval. The Mel-Scale is defined as a unit of pitch such that equal distances in pitch sounded equally distant to the listener, and the Mel-Spectrogram feature, which is used in synthetic speech detection module, is a spectrogram where the frequencies are converted to the Mel-Scale. Each subprocesses are explained in the next sections.

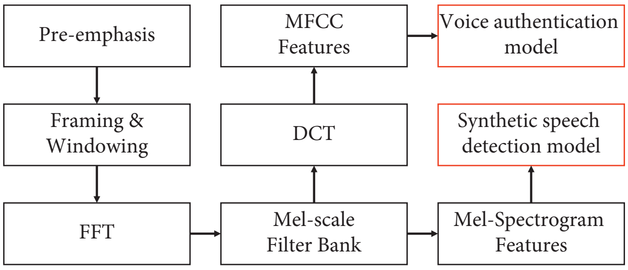


Figure 1: The overall processing flow of our proposed method.

3.1.1. Preemphasis. Preemphasis is a high-frequency amplification stage with high-bandwidth filters. High-frequency data are smaller in size compared to low-frequency data. Therefore, the spectrum is balanced by emphasizing the high-frequency component using filters. At the same time, denoising effects also occur, allowing better quality voice data to be obtained [26].

3.1.2. Framing and Windowing. Framing is the step of dividing voice data into a constant unit of time. Because the signal changes constantly, when analyzing the spectrum, it is necessary to transform Fourier by segmented frame rather than Fourier for the entire signal. Assuming that the signal is stationary for a short-segmented time, we frame it for spectroscopic analysis [26].

Windowing is one of the Finite Impulse Response (FIR) filtering methods and has the effect of accurately ensuring the frequency applied. To eliminate discontinuity between frames and maintain the original signal shape, the frame's frequency component is accurately displayed, and the frame's overlap makes it similar to the original signal.

3.1.3. FFT (Fast Fourier Transform). FFT is an algorithm that uses Fourier transformations to convert time domain data into frequency domains. FFT is an algorithm that applies the Discrete Fourier Transform (DFT) algorithm, which is designed for faster operation [26]. When there are n signal data of the time domain, the DFT algorithm requires n^2 operations, while the FFT algorithm requires nlogn operations.

3.1.4. Mel-Scale Filter Bank. Filter Bank is a set of triangular filters generated by Mel-Scale graphs. Filter Bank's triangular filters are generated in density at low frequencies, and the higher the frequency, the wider the gap, the wider the bandwidth of the filter. This is based on Mel-Scale's principle that human ears are sensitive to low frequencies and can hear better than high frequencies [26].

If this step is completed, Mel-Spectrogram data will be extracted. A Mel-Spectrogram is a spectrum in which the unit of frequency is changed to mel unit according to the following equation. m represents mel and f represents frequency.

$$m = 2595\log_{10}\left(1 + \frac{f}{700}\right). \tag{1}$$

3.1.5. Mel-Scale Generation. Mel-Scale is a scale transformation function that reflects the criteria for recognizing human ear tones. Human ears are sensitive to the lower the Hz of the sound and become insensitive to the higher frequency. With this reference, converting the frequency to Mel-Frequency and representing the result as a Mel-Scale graph allows us to accept it as linear at low frequencies below 1,000 Hz, and Log-Scale at high frequencies, like human ears [26].

3.1.6. DCT (Discrete Cosine Transform). DCT is a method that similarly models human eye sensitivity. The idea is derived from the fact that the human eye feels similar to its original form even if the data of high frequencies is reduced because it is insensitive to high frequencies compared to low frequencies [26]. The input signal is represented by N Cosine functions, and the DCT coefficients obtained from the transformation represent the frequency components of the data. These coefficients show that energy is concentrated in the low-frequency region and energy in the high-frequency region is reduced. Because of this feature, it is mainly used in algorithms that compare data.

In MFCC, data highlighted by certain frequencies are compressed via DCT to obtain a feature vector coefficient. MFCC is a coefficient obtained through Discrete Cosine Transform of equation (2) after grouping Mel-Scale Spectrum into a specific number of the frequency band.

$$X(k) = \sum_{n=0}^{N-1} x[n]\cos\left\{\frac{\pi}{N}\left(n + \frac{1}{2}\right)k\right\}, \tag{2}$$

where k represents a row, n represents a column, and N represents matrix size.

3.2. Voice-Based Authentication Process

3.2.1. Authentication Model Generation. After the MFCC based features are generated, the features are used to train the deep learning model. The deep learning algorithm we used is a feed-forward neural network algorithm. The feed-forward neural network is an artificial neural network wherein connections between the nodes do not form a cycle, and it consists of the input layer, the output layer, and the multiple hidden layers. The input layer adapts the MFCC feature of each user, and the output layer produces the classification result that describes how the input is classified. The multiple hidden layers are used to make the deep learning model more fine-grained for authentication. Figure 2 is an example of the feed-forward model for authentication. The hyperbolic tangent (tanh) function was used for the activation function of each hidden layer. Tanh function is a nonlinear activation function that makes the shape of weight and bias to have polymorphism [27]. The softmax function was used for the output function. The softmax function is an activation function used in a multiclass classification that is classified into three or more classes, and when there are n classes to be classified, it receives an n-dimensional vector to estimate the probability of belonging to each class [28].

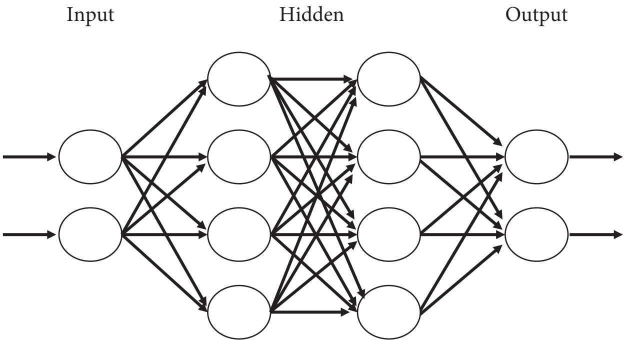


Figure 2: The example of the feed-forward neural network.

3.2.2. Voice-Based User Identification. After the authentication model is built with the normal user data, then it is utilized to identify which user is highly related to the given input voice data. The authentication model uses the softmax function as an output function, and it produces the possibility that the given input is classified to a specific registered user. The possibility can be interpreted as a user matching rate. If there are n users who are registered in the user authentication model, then n matching rates are produced by the model. The given input voice is classified to the user that has the maximum matching rate.

3.3. Synthetic Speech Detection Module. Synthetic speech is a speech that is produced by an electronic synthesizer activated by a keyboard or other electric devices. It is possible to imitate someone's speech using the previously collected voice data. With the synthetic speech data, the authentication method can be defeated because it is hard to distinguish the natural speech of a user and the synthetic speech artificially generated. We had some experiments about synthetic speech detection. The results are described in Section 4. As a result, it was found that no synthetic voice was detected by the voice-based authentication module that we developed assuming a normal situation that there is no masquerading attack. To make our proposed method more robust against the masquerading attack, we added a module capable of detecting synthetic voice.

Synthetic speech detection consists of the following four steps:

(1) Extract Mel-Spectrogram data from natural and synthetic speeches

(2) By comparing the value of each element in Mel-Spectrogram with the threshold, the value of the

element that exceeds the threshold is marked as one, and in another case, the value is grounded to zero

(3) The detection model is trained using the reproduced Mel-Spectrogram data of the natural speech data and the synthetic speech data

(4) When the user authentication is performed, the detection model tests whether the given input speech data is synthetic or not

The synthetic speech detection model is also the feed-forward neural network, and its training strategy and testing method are also the same as the ones of the authentication model.

4. Experiments and Analysis

We had several experiments to show the performance of our proposed method. The datasets used for the experiments were collected in many ways. Firstly, we collected the public voice data of 5 people from [29] and the voice data of a speaker provided by [30]. In addition to this, we recorded two people's voices and used them for the experiments. To refer dataset clearly in the paper, the data from [29] and the data from [30] and our record data are denoted as users 1-5, user 6, users 7-8, respectively. The voice data of user1-6 is data recorded in a noise-free environment, and user 7 voice data is recorded in a weak, noisy environment and user 8 in a strong noisy environment. In detail, the voice data of users 1-6 were collected in an enclosed room, user 7 data were collected in a quiet outdoor area, and user 8 data were collected at the subway station entrance. Even we used only eight users' data, but each voice signal is segmented by time units, and in the case of user authentication, the segmented voice data are also converted into 54,549 MFCC features and 34,967 Mel-Spectrogram features. Therefore, it was enough to train the model properly. The detailed information about the dataset for the evaluation is described in Table 1.

4.1. The Effectiveness of the MFCCs from Users. We extracted the MFCCs of users 1–5 and expressed the MFCC data to heatmap to show how the MFCC values are varied. Heatmap is an image that outputs colorable information in a heat distribution form graphically. With the MFCC heatmaps, it is possible to confirm that the unique characteristic of each user is reflected in MFCC features, and it will be helpful to distinguish the users.

The MFCC heatmaps are depicted in Figure 3. MFCC heatmap for each speaker is varied differently, indicating that different speakers have different feature vectors extracted by MFCC. In each MFCC heatmaps of Figure 3, the x-axis represents the number of MFCCs, and the y-axis represents the length of voice data. Since we set the number of MFCCs per user voice data to 20, the length of the x-axis is 20. The length of the voice signal was set to 502, but we extracted only 50 values among them to draw the heatmap in a limited size. The librosa library [31] is used to calculate MFCCs, where normalized values are in between −1 and 1. The red color means the minimum coefficient value, a negative one. The green color means the maximum coefficient value one. As shown in Figure 3, each user's MFCC heatmap is considered visually different.

4.2. The Parameter Selection for the User Authentication Method

4.2.1. The Architecture of the Feed-Forward Neural Network. To improve the performance of the user authentication model, we experimented with measuring classification accuracies while varying the number of hidden layers of the model. The accuracy is the ratio of the data points that were correctly classified among the whole data points for the test. The accuracy is calculated using the following equation:

$$\text{acc} = \frac{TP + TN}{TP + TN + FP + FN}. \tag{3}$$

The experimental results are summarized in Table 2. According to this result, when there is no hidden layer in the model, the accuracy value was 59% which is the lowest. When the number of hidden layers was 8, the accuracy value was the highest, and when the number of hidden layers was 9, the accuracy was slightly decreased. Therefore, we decided to use 8 hidden layers for the user authentication model.

4.2.2. Learning Strategy: The Growth of the Cost Value in the Training Phase. It is necessary to set the appropriate number of epochs in the training phase to produce the best accuracy while reducing the overall training time. We measured the cost value for each epoch, and we found the moment when the cost value is saturated.

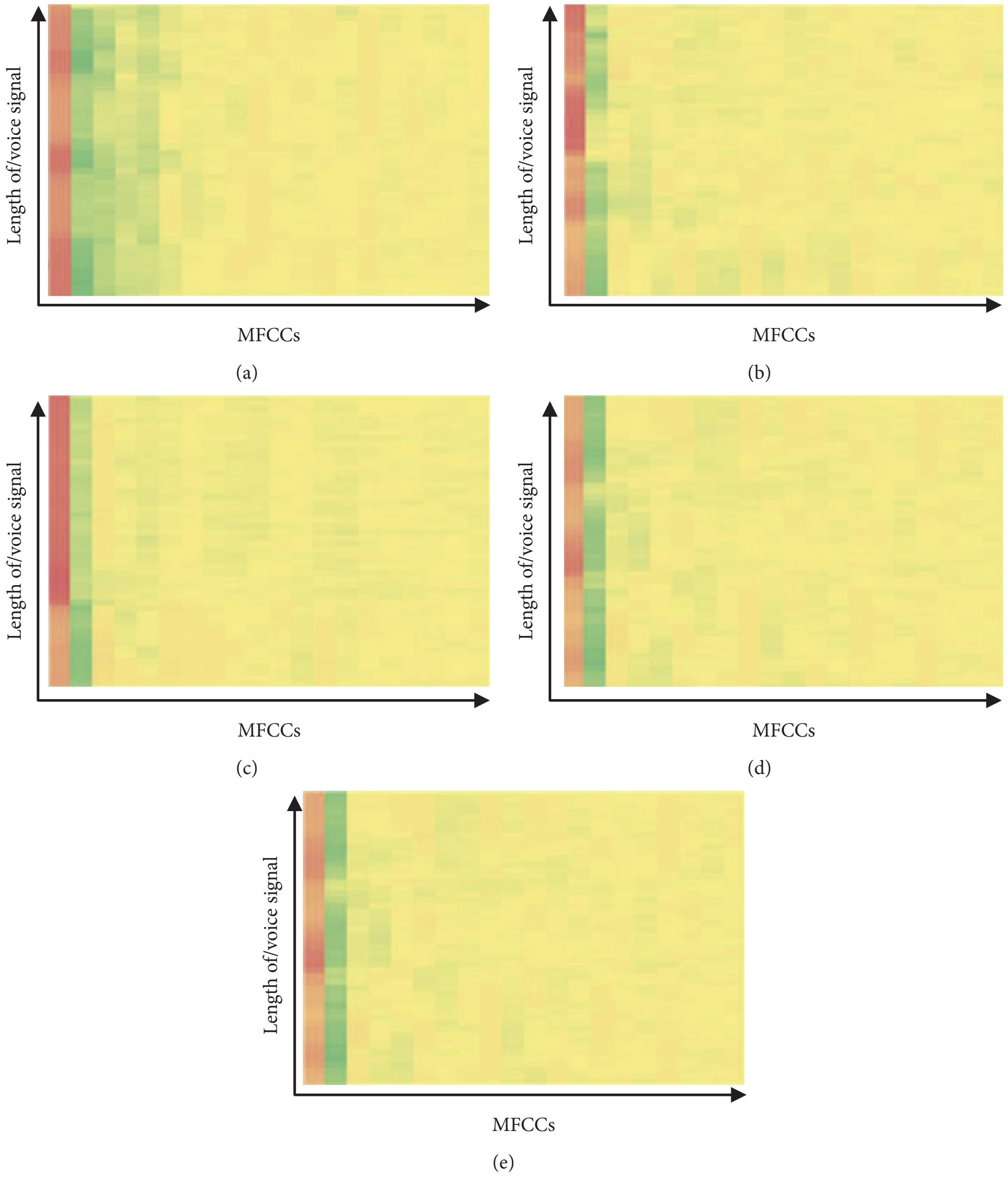
The cost function used in the training phase is the cross-entropy function which is a measure to calculate the difference between two probability distributions. We set the appropriate number of lessons by looking at the cost value according to the number of learning. The growth of the cost value is depicted in Figure 4. It was confirmed that the cost value converges to a specific value from the point when the number of lessons exceeded 750. Therefore, we trained the model by setting the number of epochs at least 750 times.

4.3. The Performance of the User Authentication: Speech Recognition. An experiment was conducted on which speaker the data is identified by inputting arbitrary voice data into the user authentication model.

Since the softmax function was used as an output function, the final result that the model produces is a set of possibilities. Each possibility (i.e., matching rate) implicates how well the given input is fitted to a specific user. For example, if there are n users, and n possibilities $\{p_1, p_2, p_3, \ldots p_n\}$ are output from the model, then the p_1 means a possibility that the input voice data is user 1's voice data. To check the user authentication accuracy, therefore, we measured the possibilities by using different user's test data repeatedly. The test for each different user was repeated 50 times, and the overall results are described in Table 3. The average of the possibility that the given input is classified as the correct user

Table 1: The dataset used in the evaluation.

Voice authentication module		Synthetic speech detection module	
Training data (MFCC)	Testing data (MFCC)	Training data (Mel-Spectrogram)	Testing data (Mel-Spectrogram)
USR 1 ~ 8	USR1 ~ 8	USR1 ~ 8 (Natural & Synthetic)	User1 ~ 5 (Natural), User6 ~ 8 (Synthetic)
The number of MFCC samples		The number of Mel-Spectrogram samples	
44,758	9,791	27,974	6,993



(a)

(b)

(c)

(d)

(e)

Figure 3: MFCC heatmaps of each speaker. (a) User 1. (b) User 2. (c) User 3. (d) User 4. (e) User 5.

is shown. The minimum average value was 0.8, which is the possibility when user 8's voice was input and classified to user 8. The reason is that the user 8 voice data is recorded in a relatively strong noisy environment. Even we assumed that user authentication might be performed in a noisy environment, it is possible to identify each user by setting the identification threshold to 0.8.

We also conducted an additional experiment to see if the user authentication model can identify an unregistered user that is not trained in the model. The matching threshold for the identification was set at 0.8. The test was also repeated 50 times. In every experiment, each user's voice data was excluded in the training phase one by one, and the matching rates with other users were measured.

Table 2: Accuracy by the number of hidden layers and the number of neurons.

# of hidden layers	The number of neurons										Accuracy
	1st layer	2nd layer	3rd layer	4th layer	5th layer	6th layer	7th layer	8th layer	9th layer	10th layer	
0	—	—	—	—	—	—	—	—	—	—	0.59
4	256	256	256	256	—	—	—	—	—	—	0.66
5	256	256	256	256	128	—	—	—	—	—	0.69
6	256	256	256	256	128	128	—	—	—	—	0.73
7	256	256	256	256	128	128	128	—	—	—	0.75
8	256	256	256	256	128	128	128	128	—	—	0.80
9	256	256	256	256	128	128	128	128	128	—	0.79
10	256	256	256	256	128	128	128	128	128	128	0.79

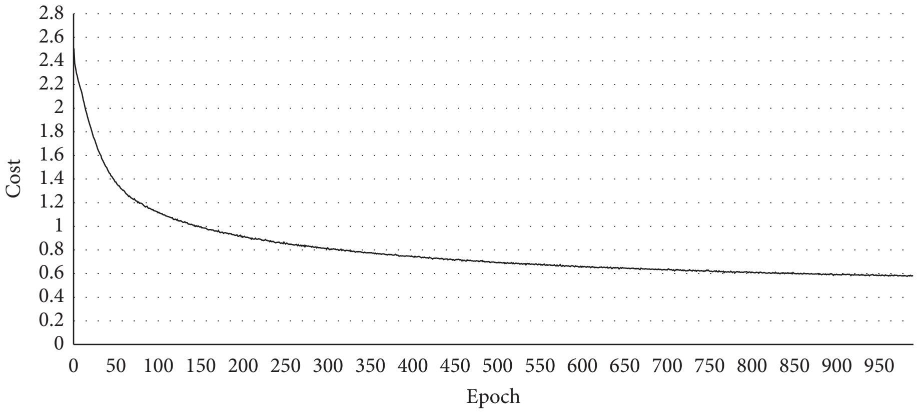


Figure 4: The growth of the cost value by epoch.

Table 3: Matching rate with the voice of the same speaker.

	USR1	USR2	USR3	USR4	USR5	USR6	USR7	USR8
Average of matching rate	0.92	0.87	0.97	0.85	0.83	0.83	0.97	0.80

The average matching rates with the others are described in Table 4. It is confirmed that it is possible to identify unregistered user voice accurately when the identification threshold is set to 0.8.

Table 4: Matching rate with the voice of other speakers.

	USR1	USR2	USR3	USR4	USR5	USR6	USR7	USR8
USR 1	—	0.19	0.06	0.17	0.21	0.09	0.22	0.05
USR 2	0.11	—	0.01	0.15	0.61	0.01	0.00	0.10
USR 3	0.00	0.21	—	0.02	0.10	0.66	0.00	0.01
USR 4	0.02	0.49	0.03	—	0.27	0.13	0.01	0.05
USR 5	0.05	0.55	0.07	0.21	—	0.03	0.04	0.05
USR 6	0.11	0.39	0.31	0.05	0.12	—	0.01	0.00
USR 7	0.18	0.13	0.03	0.01	0.09	0.00	—	0.56
USR 8	0.01	0.13	0.02	0.06	0.22	0.01	0.55	—

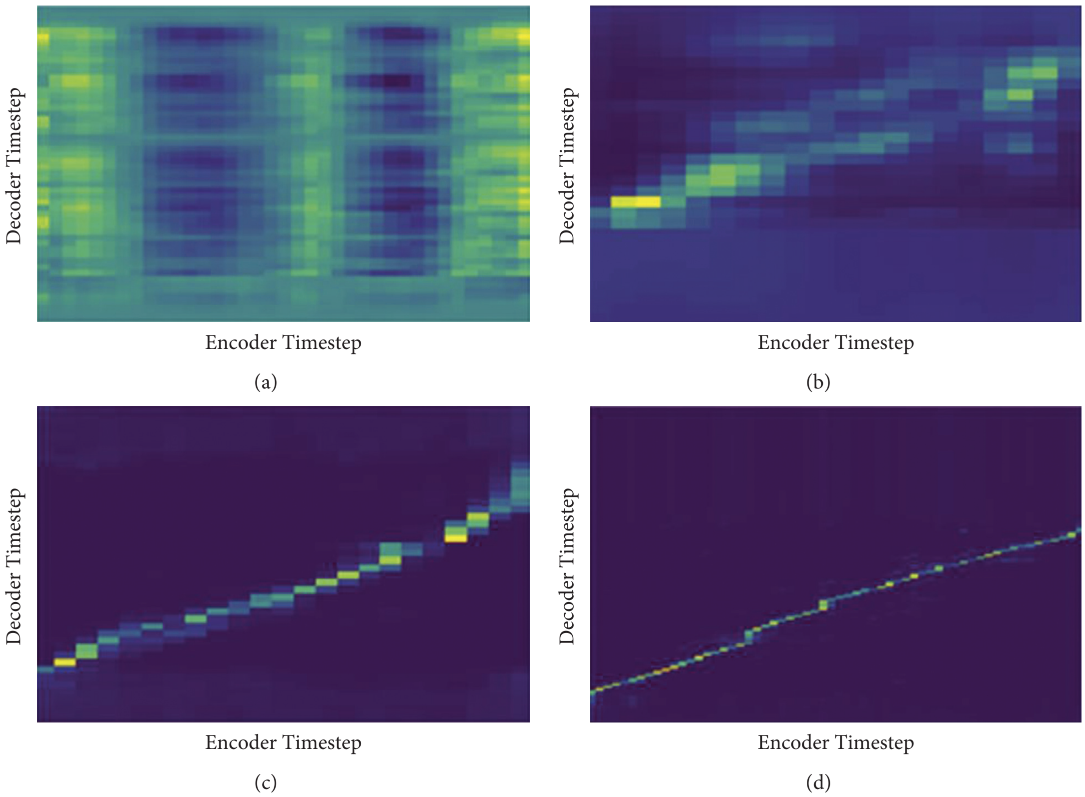
4.4. The Performance about the Synthetic Speech Detection

4.4.1. Masquerading Attack Simulation. The masquerading attack using the synthetic speech data is simulated. To generate the synthetic speech, Tacotron that is a deep learning-based voice synthesis model published by Google, was used [31]. The Tacotron model was trained using a user's natural voice data and generated the corresponding synthetic speech data.

The attention assignment graphs of the synthetic voice generated when creating synthetic speeches are shown in Figure 5. Each graph shows how well the input speech data and the synthetic speech are aligned. In the attention assignment graph, the x-axis represents the input feature vector and the y-axis represents the Mel-Spectrogram built on each input vector. Because the output comes out in the order of input, the graph is also produced in a straight line in parallel, and the color of each point in the graph represents the decibel value representing the intensity of the sound. When Tacotron's train step reached 100,000 steps, it was found that the synthetic speech data were well organized and aligned with the input data.

The synthetic speech data of user 6 was generated for the experiment, and it was applied to our proposed user authentication model. The matching rates with all the users are described in Table 5.

As shown in Table 5, the matching rate with user 6 was 0.88, which exceeds the 0.8 thresholds. It means that the synthetic speech of user 6 was classified as user 6. The user authentication model was failed to identify that it is fake data for masquerading.



(a)

(b)

(c)

(d)

FIGURE 5: Attention alignment graph according to the training step. (a) # of epochs: 1000; (b) # of epochs: 3000; (c) # of epochs: 5000; (d) # of epochs: 10000.

TABLE 5: The evaluation result with the synthetic speech.

Voice data	USR1	USR2	USR3	USR4	USR5	USR6	USR7	USR8
Matching rate	0.03	0.02	0.06	0.00	0.00	0.88	0.00	0.00

TABLE 6: The evaluation result with the synthetic speech (modeling with the synthetic speech data).

Voice data	USR1	USR2	USR3	USR4	USR5	USR6	USR7	USR8	USR6 (synthetic)
Matching rate	0.00	0.01	0.03	0.00	0.00	0.14	0.00	0.00	0.82

After training the model with the synthetic speech data of user 6, we measured the matching rate again. The experimental result is included in Table 6.

Naturally, the user authentication model was able to identify the synthetic speech. However, since generating and learning synthetic speeches for all registered users is highly inefficient and practically impossible, it is not suitable to use this approach as a method for preventing the masquerading attack.

4.4.2. The Effectiveness of Mel-Spectrogram in Synthetic Speech Detection Module. Although it is possible to perform synthetic voice detection using MFCC data, our synthetic speech detection module uses only Mel-spectrogram data of each user. Regarding this, we had additional experiments to show the effectiveness of the model-spectrogram data of users compared with MFCC in synthetic voice detection. We performed detection using the MFCC data and the Mel-spectrogram 10 times, and the matching ratios between synthetic voice data were measured. In the experiment, we used user 7's synthetic voice data. The experimental result in Table 7 shows that the matching ratio of the Mel-spectrogram is higher than the result when MFCC data is being used. The percentage of the difference between the Mel-spectrogram and the MFCC data was in the range of 7% to 12%.

4.4.3. The Robustness of the Proposed Synthetic Speech Detection. It was found that the user authentication model is not enough to deal with the synthetic speech data, so we designed the synthetic speech detection using Mel-spectrogram. The evaluation result of the proposed detection model is described in Table 8. The synthetic speech of user 6 was used in this experiment again. Our detection method uses a grounding threshold to transform each value of the Mel-spectrogram to zero or one. If an element value equals

Table 7: MFCC vs. Mel in Synthetic Speech Detection Module.

	Matching ratio									
# of executions	1	2	3	4	5	6	7	8	9	10
MFCC	0.55	0.56	0.56	0.56	0.56	0.54	0.56	0.57	0.55	0.55
Mel-Spectrogram	0.65	0.63	0.65	0.64	0.64	0.64	0.66	0.64	0.65	0.63

Table 8: Experiment to set a threshold.

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
The avg of matching rate	0.04	0.30	0.25	0.15	0.20	0.25	0.17	0.26	0.30
Threshold	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0
The avg of matching rate	0.30	0.32	0.40	0.47	0.47	0.49	0.52	0.49	0.51
Threshold	10.0	11.0	12.0	13.0	14.0	15.0	-	-	-
The avg of matching rate	0.53	0.54	0.58	0.58	0.58	0.58	-		-

or exceeds the threshold, then the value is changed to one. In another case, the value will be zero.

As shown in Table 8, when the grounding threshold was set to 12.0, the matching rate was the highest value, 0.58. The highest matching rate value, 0.58, can be the detection threshold to identify the synthetic speech data finally.

5. Conclusions and Future Work

In our research, we propose a user authentication method using the deep learning method. The user authentication model uses the MFCC feature. We conducted the experiments with the user voice data recorded in the different environments, and it was found that the user authentication model can accurately distinguish each registered user. In addition to this, the synthetic speech detection method was also proposed together to examine the masquerading attack. According to our masquerading attack simulation, it was possible to pass the user authentication by using the synthetic speech data of a registered user. If the synthetic speech data of all users are used to train the user authentication model, the synthetic speech data can be detected very easily, but this simple method is not suitable to be used in the real world considering the time overhead for the synthetic speech generation. In this regard, we added a detection method that uses Mel-spectrogram data, and according to our evaluation, it was found that it can be used to prevent the masquerading attack.

In the future, we have the plan to develop a novel method using the convolutional neural network. As we have seen in this research, the MFCCs can be expressed as an image-like heatmap. MFCC data are considered applicable to be used as input of the convolutional neural network. The coefficient values in the similar area seem to have similar properties, so it is expected that the CNN algorithm will be suitable to process the MFCC heatmap data for user identification.

Data Availability

Some or all data, models, or code generated or used during the study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the Soonchunhyang University Research Fund (no. 20210441) and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (NRF-2021R1A4A2001810).

References

[1] S. K. Wong and S. M. Yiu, “Location spoofing attack detection with pre-installed sensors in mobile devices,” *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl*, vol. 11, no. 4, pp. 16–30, 2020.

[2] A. S. Kitana, T. Issa, and W. G. Isaac, “Towards an epidemic SMS-based cellular botnet,” *J. Internet Serv. Inf. Secur*, vol. 10, no. 4, pp. 38–58, 2020.

[3] “Global Social media Attacks Rate Among Businesses ,” 2019.

[4] “5 Shocking Insights into the Social Network Habits of Security Professionals,” 2017, https://thycotic.com/company/blog/2017/05/30/5-shocking-insights-into-the-social-network-habits-of-securityprofessionals-and-infographic/.

[5] G. S. Kasturi, A. Jain, and J. D. Singh, “Detection and classification of radio frequency jamming attacks using machine learning,” *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl*, vol. 11, no. 4, pp. 49–62, 2020.

[6] A. L. Marra, F. Martinelli, F. Mercaldo, A. Saracino, and M. Sheikhalishahi, “A distributed framework for collaborative and dynamic analysis of android malware,” *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl*, vol. 11, no. 3, pp. 1–28, 2020.

[7] D. Berbecaru, A. Lioy, and C. Cameroni, “Supporting Authorize-then-Authenticate for Wi-Fi access based on an electronic identity infrastructure,” *J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl*, vol. 11, no. 2, pp. 34–54, 2020.

[8] S. H. K. Wong and S. M. Yiu, “Identification of device motion status via Bluetooth discovery,” *J. Internet Serv. Inf. Secur*, vol. 10, no. 4, pp. 59–69, 2020.

[9] J. A. Unar, W. C. H. Seng, and A. Abbasi, “A review of biometric technology along with trends and prospects,” *Pattern Recognition*, vol. 47, no. 8, pp. 2673–2688, 2014.

[10] Y. Lu, L. Li, H. Peng, and Y. Yang, “An enhanced biometric-based authentication scheme for telecare medicine

information systems using elliptic curve cryptosystem," *Journal of Medical Systems*, vol. 39, no. 3, pp. 32–38, 2015.

[11] S. Chatterjee, S. Roy, A. K. Das, S. Chattopadhyay, N. Kumar, and A. V. Vasilakos, "Secure biometric-based authentication scheme using Chebyshev chaotic map for the multi-server environment," *IEEE Transactions on Dependable and Secure Computing*, vol. 15, no. 5, pp. 824–839, 2018.

[12] P. Padma and S. Srinivasan, "A survey on biometric-based authentication in cloud computing," in *Proceedings of the International Conference on Inventive Computation Technologies*, pp. 1–5, ICICT), Coimbatore, India, August 2016.

[13] J. Yang and R. K. Das, "Long-term high-frequency features for synthetic speech detection," *Digital Signal Processing*, vol. 97, Article ID 102622, 2020.

[14] P. L. De Leon and B. Stewart, "Synthetic speech detection based on selected word discriminators," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3004–3008, Vancouver, British Columbia, Canada, October 2013.

[15] Z. Wu, X. Xiao, E. S. Chng, and H. Li, "Synthetic speech detection using temporal modulation feature," in *Proceedings of the IEEE International Conference On Acoustics, Speech And Signal Processing*, pp. 7234–7238, British Columbia, Canada, March 2013.

[16] I. Saratxaga, J. Sanchez, Z. Wu, I. Hernaez, and E. Navas, "Synthetic speech detection using phase information," *Speech Communication*, vol. 81, pp. 30–41, 2016.

[17] D. Paul, M. Pal, and G. Saha, "Spectral features for synthetic speech detection," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 4, pp. 605–617, 2017.

[18] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, and D. Erro, "A cross-vocoder study of speaker-independent synthetic speech detection using phase information," in *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association*, Singapore, September 2014.

[19] J. Yang, R. K. Das, and H. Li, "Significance of subband features for synthetic speech detection," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 2160–2170, 2020.

[20] C. Hanilçi, T. Kinnunen, M. Sahidullah, and A. Sizov, "Spoofing detection goes noisy: an analysis of synthetic speech detection in the presence of additive noise," *Speech Communication*, vol. 85, pp. 83–97, 2016.

[21] Z. H. Wu, R. K. Das, J. Yang, and H. Li, "Light Convolutional Neural Network with Feature Genuinization for Detection of Synthetic Speech Attacks," in *Proceedings of the Interspeech 2020*, Shanghai, China, October 2020.

[22] F. Hassan and A. Javed, "Voice spoofing countermeasure for synthetic speech detection," in *Proceedings of the 2021 International Conference on Artificial Intelligence*, pp. 209–212, ICAI), April 2021.

[23] Z. You, F. Jiang, and Z. Duan, "One-class learning towards synthetic voice spoofing detection," *IEEE Signal Processing Letters*, vol. 28, pp. 937–941, 2021.

[24] P. L. De Leon, M. Pucher, J. Yamagishi, I. Hernaez, and I. Saratxaga, "Evaluation of speaker verification security and detection of HMM-based synthetic speech," *IEEE Transactions on Audio Speech and Language Processing*, vol. 20, no. 8, pp. 2280–2290, 2012.

[25] J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, and T. Raitio, "Toward a universal synthetic speech spoofing detection using phase information," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 810–820, 2015.

[26] 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, 2010.

[27] F. Samaa, M. Alshraideh, and T. Mahafza, "A medical decision support system for ent disease diagnosis using artificial neural networks," *Proceedings of International Journal of Artificial Intelligence and Mechatronics*, vol. 4, no. 2, pp. 45–54, 2015.

[28] G. Bolin and L. Pavel, "On the Properties of the Softmax Function with Application in Game Theory and Reinforcement Learning," April 2017, https://arxiv.org/abs/1704.00805.

[29] "Public Voice Data," https://github.com/dydtjr1128/Speaker-Recognition-using-NN.

[30] "Korean Single Speaker voice dataset," https://www.kaggle.com/bryanpark/korean-single-speaker-speech-dataset.

[31] "Librosa library," https://librosa.org/doc/latest/index.html.