**Discrimination between direct sound and its recorded playback sound**

**Kotaro SONODA1**,

**Corresponding author:** Kotaro SONODA, email: sonoda@cis.nagasaki-u.ac.jp (9 pt Cambria, Plain)

1 School of Information and Data Sciences, Nagasaki University, Bunkyo-machi 1-14, Nagasaki-City, Nagasaki, 852-8521, Japan

**Abstract** Automatic speaker verification system is a convenient means to communicate commands to a system without touch contact. However, there is a serious problem in that replayed speech recorded stealthily by attacker can be mistakenly verified as the speaker's own voice. This attack, called a replay attack, is very easy to implement, but is difficult to prevent. In this paper, we propose a method to discriminate between the direct sound and the replayed sound by cepstral analysis, focusing on the multi-recording nature of the replayed sound.

**Keywords:** Voice Spoofing, Speaker verification, 4-8 words/phrases which are representative for the paper – maximum 1 line (10 pt, plain)

**1. Introduction**

Biometric data is increasingly being used for identity authentication. In particular, fingerprint and face recognition have reached a practical level of use, such as when signing in mobile devices. At the same time, many countermeasures against spoofing are being researched, and vein images and facial depth images using infrared transmissive sensors are used to ensure that the input fingerprint or face is not a pre-captured image or video, but the one presented by the living person at the authentication site.

On the other hand, voice-based identity authentication technology is not used for sign-in and other applications due to its immature security, although it is used for controlling smart speakers and other applications. The reason for this is that countermeasures against voice spoofing are not yet developed.

Spoofing of the speaker's voice includes the following

* Mimicry of the speaker's voice by another person
* Mechanically synthesized fakes of the speaker's voice
* Recorded voice replay

Impersonation of a person's voice by another person affects the performance of speaker recognition.

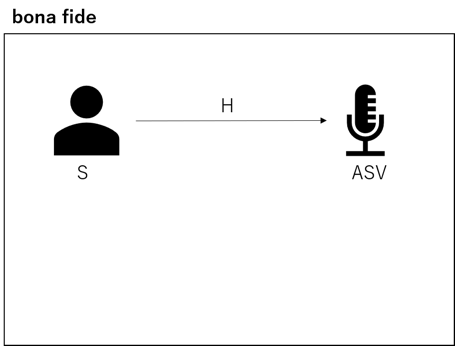
Mechanically synthesized voice mimics are created by synthesizing a voice based on a recording of the person's voice collected in advance and can be reproduced with high performance using recent DNN technology. In this case, however, it is still only a voice imitation, and is a problem related to the performance of speaker recognition.

The last, voice recording, is a playback of the recorded signal of the person's voice, but it is passthrough by normal speaker recognition.

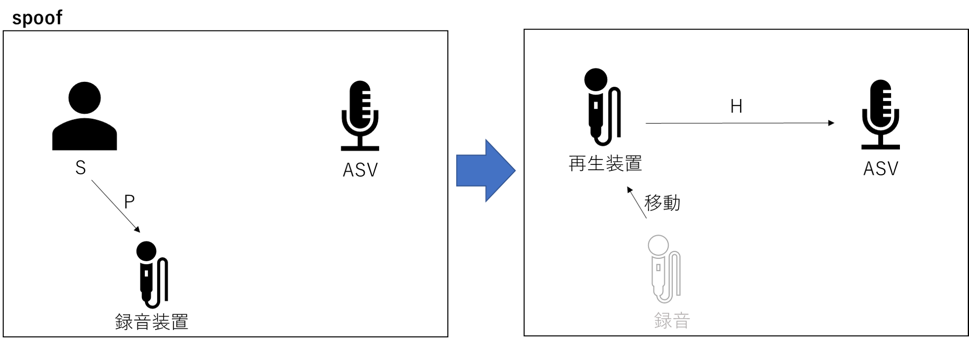
The presentation of the recorded voice to the person authentication is an attack method that can be realized very easily, yet it is vulnerable to speaker recognition.

T. Kinnunn et al. proposed the deep learning-based approach and marked the best performance in ASVspoof2017. Their method is highly accurate, achieving 6.7% EER (FRR-FAR equal error rate) by using LCNN, CNN and SVM. The second best is proposed by G. Lavrentyeva et al. and achieved 12.3% ERR using multiple features: CQCCs (Constant-Q Cepstral Coefficient), PLPs (Perceptual Linear Prediction), and MFCCs (Mel-Frequency Cepstral Coefficient) were all heavily used.

Their proposals used computationally expensive Deep Neural Network models and tons of features. In this paper, I would like to rebase the nature at the ASV systems’ input signal and find the difference between the direct live voice and the replayed voice.



**Fig.1.** bona fide path



**Fig.2.** spoof path

Figure 1 shows the voice signals’ path from the bona fide human utterance to ASV system. In this case, the ASV systems’ input signal y is consisted as follows:

y = h\*s,

where s is bona fide human utterance and h is the room transfer function between the utterance place to the received place of ASV system. In contrast, the replayed attack case, as shown in Figure 2, the attacker recorded the bona fide human utterance by his microphone at the different place and replays the recorded sound by his speaker. Therefore, the ASV systems’ input signal y’ in the spoofed case is formed as followed:

y’ = h\*(p\*s),

where h1 is the room transfer function and recording or replaying devices’ distortion.

The speech recognition and the speaker recognition require only the utterance signal s divided from y, but in generally the division is difficult problem and omitted in the recognition process. Moreover, bona fide utterance signal s is also included in the replayed signal. Therefore, to distinguish between the live and the replayed is the problem to distinguish the direct-path h and the relayed-path h\*p.

**2. lower coefficients of the minimum phase cepstrum**

The cepstrum analyzation is used to analyze the periodicity on the spectrogram, but not on the time-domain waveform. It can divide the envelope and detail from the spectrogram. If the signal is human voice, the envelope and the detail are 対応する with the formant components and the excitation components. The formant component is utilized in speech recognition and the excitation components are utilized in the speaker identification.

The observed signal in time-domain, y(t) is consisted of the convolution the excitation signal s(t) with the transfer functions h(t) as following equation (1):

y(t) = h(t) \* s(t).

In the Fourier transformation domain, the equation (1) is derived to

Y(omega) = H(omega) dot S(omega).

Their logarithmic spectrogram is consisted of the summation of the logarithmic spectrograms of the transfer functions and excitation signal. Therefore, the inverse Fourier transformation of the logarithmic form is derived to approximately the summation of the cepstrum of the transfer functions and excitation signal.

Y(q) = H(q) + S(q).

where Y (ω), S(ω), and H(ω) are the spectra of the observed signal, the source signal, and the reverberation characteristics at the angular frequency ω, respectively, and Y(q), S(q), and H(q) are the respective cepstral at the quefrency: q. Now, rcunwrap is performed. rcunwrap allows us to cancel the phase directly after the phase unwrapping process.

By doing rcunwrap, we can cancel the phase directly after the phase unwrapping process. Next, we perform cepstrum analysis. In the cepstrum analysis, we use the all-pass cepstrum analysis.

In cepstrum analysis, all-pass cepstrum analysis is used. In the cepstrum analysis of speech, the enveloping components of the spectrum are found in the low part of the quefrency. The higher part shows the fine structure of the spectrum. The low quefrency (the enveloping component of the spectrum) contains the source (minimum phase cepstrum) and the filter component, the room transfer function. By riffling the all-pass cepstrum and the low quefrency, the minimum phase cepstrum and the fine structure of the high quefrency in the low quefrency can be omitted, and only the room transfer function can be extracted.

In this section, we explain the all-pass cepstrum analysis. First, two types of cepstrum are constructed.

The real-valued cepstrum YA(q) and the complex-valued cepstrum Y(q) can be expressed as follows.

{

Minimally topologize for 1.

Then, the minimum topological cepstrum Cmin can be expressed as

Cmin = W - YA(q) Thus the all-pass cepstrum CAllpass can be expressed as

CAllpass = Y(q) - Cmin From the above, we can perform a cepstrum analysis of the room transfer function only.

In this paper, cepstrum is used to observe the recording pathes on the ASV input. It is expected that the spectrum of relayed-path is more complex than the direct-path. The complexity is 反映される　to the number of peaks on the cepstrum.

**3. Replay attack dataset**

To evaluate the complexity of the recording paths, the three persons’ voice samples on the direct sound and the replayed sounds in low/high/perfect quality from the public dataset of ASV spoof 2017 challenge.

In the room size 2-5 m^2 and the reverberation time 50-200 ms, the distance between the bona fide utterance to the ASV system is set to 10-50 cm. The target persons’ sample indexes in the ASV spoof 2017 challenge are No. 69 (low tone female), 70 (low tone male), and 74 (high tone female).

The qualities of replayed sound are depend on the fidelity of the recording device and the replaying device. Three types of quality are summarized as table 1.

Table1

|  |  |  |  |
| --- | --- | --- | --- |
| Quantity | Symbol | Unit | Value |
| Density (9 pt Cambria, Plain, Left Alignment) |  | kg/m3 | 8960 |
| Heat capacity |  | J/(kg∙K) | 3.84e+2 |
|  |  | m2/s | 1.16e-4 |

**4. Evaluation**

Figures (4.1), (4.3), and (4.5) show the time waveforms of ASV speech and the time variation of the cepstrum peaks of the room transfer function estimated by the person's own speech and re-recording (three levels of quality). The peak number of bona fide was found to be lower than the peak numbers of the three types of spoofs. No.70 (male with low voice) also showed a large change in peak number overall, but the peak number of bona fide was found to be constant compared to the peak numbers of the other spoofs. No.74 (female with high voice), the change in the peak number of spoofs was large, but the change in the peak number of bona fide was small. The number of peaks was mostly within the range of 180~195.

Figure (4.1), Figure (4.3), and Figure (4.5) show the mean values of the peak numbers in Figure (4.2), Figure (4.4), and Figure (4.6). For No. 69 (female with low voice), the peak number of bona fide was lower than that of spoof-low. However, spoof-perfect and spoof-high were about the same. No.70 (male with low voice), the peak number of bona fide was lower than any spoof. No.74 (female with high voice), the peak number of bona fide was lower than spoof-high. No.74 (female with high voice), the number of peaks of bona fide was lower than that of spoof-high. However, it was about the same as that of spoof-perfect and spoof-low.

|  |  |
| --- | --- |
| (a) | (b) |
| page13image16241120 | page13image16241328 |

図 4.1 No.69(低い声の女性) の音声の波形とケプストラムピーク数

図 4.2 No.69(低い声の女性) のピーク数の平均値

|  |  |
| --- | --- |
| (a) | (b) |
| page14image16213552 | page14image16213760 |

図 4.3 No.70(低い声の男性) の音声の波形とケプストラムピーク数

図 4.4 No.70(低い声の男性) のピーク数の平均値

|  |  |
| --- | --- |
| (a) | (b) |
| page15image16221616 | page15image16222864 |

図 4.5 No.74(高い声の女性) の音声の波形とケプストラムピーク数

図 4.6 No.74(高い声の女性) のピーク数の平均値

**5. Discussions**

As a result of the experiment, bona fide had the lowest number of peaks at 1.1 seconds in No.69 (female with low voice), 0.5 seconds, 0.8 seconds, 1.0 seconds in No.70 (male with low voice), and 1.2 seconds in No.74 (female with high voice), and the number of peaks increased as the recording quality deteriorated. It was thought that bona fide, which has the least multiple transfer function, would have the smallest number of peaks and spoof, which has the largest reverberation, would have the largest number of peaks, but this was not explicitly observed in the results. Since the room transfer function is supposed to remain unchanged unless the recording conditions change, let's look at the average value of the number of peaks for the entire time interval of the audio data. Figures (4.2), (4.4), and (4.6) show that bona fide has the smallest value, and we expected it to gradually increase as the quality of the spoof deteriorated. This suggests that the room transfer function may not be well separated from the low kerfuffle because there is no regularity in the amplitude and number of peaks depending on the voice data used.

**6. Conclusions**

In this study, we used the 2019 speech data of ASVspoof to examine the method of discriminating whether the voice is the person's own voice or a recorded voice. For the separation of the recording environment, only the room transfer function within the low quefrency was extracted and analyzed by riffling on an all-path cepstrum. However, Figure (4.1), Figure (4.3), and Figure (4.5), which show the number of peaks per second in the cepstral analysis of the person's voice and the recorded voice, did not show any difference in discrimination between the person's voice and the recorded voice. However, Figure (4.3) shows that the number of peaks for bona fide is lower than the quality of any spoof, although not in the expected way. The other figures also show that the quality of bona fide is lower than the quality of any one spoof, although it is not the quality of all spoofs. However, since we could not find a relationship between the amplitude of the temporal waveform and the quality of the spoof, we were not able to discriminate which voice was which just by looking at one data set, instead of comparing the two data sets of the person's voice and the recorded voice, which was the ultimate goal of this study. One of the future tasks is to make the cepstral analysis of speech data usable by taking measures to find the regularity. For this purpose, it is necessary to improve the algorithm to correctly separate the room transfer function from the speech.

**References**

**© 2021 by the Authors.** Licensee Poznan University of Technology (Poznan, Poland). This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).