Oo et al. EURASIP Journal on Audio, Speech, and Music Processing (2019) 2019:8
https://doi.org/10.1186/s13636-019-0151-2

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

# Replay attack detection with auditory filter-based relative phase features

Zeyan Oo[1†], Longbiao Wang[2*†], Khomdet Phapatanaburi[3†], Meng Liu[2], Seiichi Nakagawa[4], Masahiro Iwahashi[1] and Jianwu Dang[2]

**Abstract**

There are many studies on detecting human speech from artificially generated speech and automatic speaker verification (ASV) that aim to detect and identify whether the given speech belongs to a given speaker. Recent studies demonstrate the success of the relative phase (RP) feature in speaker recognition/verification and the detection of synthesized speech and converted speech. However, there are few studies that focus on the RP feature for replay attack detection. In this paper, we improve the discriminating ability of the RP feature by proposing two new auditory filter-based RP features for replay attack detection. The key idea is to integrate the advantage of RP-based features in signal representation with the advantage of two auditory filter-based RP features. For the first proposed feature, we apply a Mel-filter bank to convert the signal representation of conventional RP information from a linear scale to a Mel scale, where the modified representation is called the Mel-scale RP feature. For the other proposed feature, a gammatone filter bank is applied to scale the RP information, where the scaled RP feature is called the gammatone-scale RP feature. These two proposed phase-based features are implemented to achieve better performance than a conventional RP feature because of the scale resolution and. In addition to the use of individual Mel/gammatone-scale RP features, a combination of the scores of these proposed RP features and a standard magnitude-based feature, that is, the constant Q transform cepstral coefficient (CQCC), is also applied to further improve the reliable detection decision. The effectiveness of the proposed Mel-scale RP feature, gammatone-scale RP feature, and their combination are evaluated using the ASVspoof 2017 dataset. On the evaluation dataset, our proposed methods demonstrate significant improvement over the existing feature and baseline CQCC feature. The combination of the CQCC and gammatone-scale RP provides the best performance compared with an individual baseline feature and other combination methods.

**Keywords:** Relative phase information, ASVspoof 2017, Replay attack detection, Auditory filter, Score combination

## 1 Introduction

Automatic speaker verification (ASV) aims to determine whether the given speech belongs to a given speaker [1]. Recently, much progress has been made in the field of speaker recognition and verification [2–4]. There is great interest in the reliability and security of these ASV systems [5–7]. Many methods can deceive systems by imitating the property of speech, and most biometric security systems [8, 9] are sometimes prone to deception. Countering this deception has been one of the challenges in the field of speech processing. At present, increasing numbers of researchers are beginning to pay attention to the vulnerability of ASV systems. These above mentioned attacks which area called spoofing attacks can be divided into four types: impersonation, replay, text-to-speech (TTS) synthesis, and speech conversion [10]. In this paper, we focus on replay attack detection, which is a task that determines whether a speech sample contains genuine or replayed speech. The replayed speech is the speech recorded using the recording device and replayed using a loud speaker. Countermeasures on replayed speech have not been thoroughly researched because of the lack of publicly available databases and standardized benchmarks prior to ASVspoof 2017 challenge.

*Correspondence: longbiao_wang@tju.edu.cn
†Zeyan Oo, Longbiao Wang and Khomdet Phapatanaburi contributed equally to this work.
[2]Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China
Full list of author information is available at the end of the article

To detect replayed speech from genuine speech, there are several approaches. Some research has focused on tuning the classifier [11–14], whereas other research has focused on feature extraction [15, 16]. As an anti-spoofing task mainly focuses on the characteristics of the given speech, in this paper, we focus on a new feature extraction method that provides better discrimination between replayed speech and genuine speech. Most of the methods used in previous studies have focused on amplitude information. The Mel-frequency cepstral coefficient (MFCCs), inverse MFCC, and linear frequency cepstral coefficients (LFCC) were used in [10]. High-frequency cepstral coefficients, single frequency filtering cepstral coefficients (SFFCC), and constant Q cepstral coefficients (CQCC) were proposed in [17]. The novel variable length Teager energy separation-based instantaneous frequency feature (VESA-IFCC) was proposed in [18]. In addition to the aforementioned single features, fusion systems [19, 20] have also been proposed for combining scores from different feature/classifier-based replay attack detection. A combination of the voice source, instantaneous frequency, and cepstral features was implemented in [19]. The fused system of LFCC and rectangular filter cepstral coefficients (RFCC) was proposed in [20]. In all these works, the best performing system strongly depends on an individual amplitude information-based feature and the score combination, which fuses scores from different features/classifiers derived from amplitude information. However, there are few studies that have focused on phase information because of the phase wrapping problem. Phase wrapping occurs because phase points are constrained to the range $-180°$ to $+180°$, even when the actual phase is outside this range; hence, the phase value outside this range becomes unusable. In previous work [21, 22], the use of only the magnitude feature was sometimes not sufficient for the discrimination task. This is because phase information, which is half of the original speech, is ignored when discriminating between replay and genuine speech.

Phase-based features have also been successfully used for synthesized and converted speech detection [23, 24]. One of the most commonly used phase feature is the modified group delay (MGD)-based feature [11]. Typically, the feature is defined as the negative derivative of the phase information of the Fourier transform based on a signal. In fact, the MGD feature is not extracted using only phase information, but uses both phase and magnitude information. Thus, the MGD feature contains both phase and magnitude information, which may make the detection unable to discriminate between genuine speech and replayed speech. The MGD feature enhances important sections of the envelop of the short time speech spectrum which may lose representation in vocal source information of the given speech [25, 26]. In [12], the cosine phase feature was proposed and outperformed the MGD feature. Unlike the MGD feature, the phase information of the cosine phase feature is computed using only phase information. However, the cosine phase does not normalize the phase variation by cutting positions and the sine function is not applied to the unwrapped phase spectrum; thus, the cosine phase may lose some information, which can reduce the performance of detecting replayed speech. In our previous work, to overcome the problems of cosine phase feature extraction, we proposed a phase-based feature called the relative phase (RP) feature. A widely used merit associated with the RP feature is that it extracts precise phase information from speech because the phase variation is significantly reduced by cutting positions, and both the cosine function and sine function are applied to the normalized phase spectrum. RP information that is directly extracted from the Fourier transform of the speech wave has been proposed for speaker recognition/verification in various conditions [27, 28]. In [29, 30], the RP was also applied to synthesized and converted speech detection. The results indicate that the RP significantly outperformed the baseline feature set, such as the MFCC and MGD. Although the RP has been successfully applied in many speech applications, only a few studies have used the RP feature in replay attack detection. In fact, RP extraction is based on a linear scale and may not perform well when used to detect replayed speech. Recent works [1, 17] have shown that features with a nonlinear scale could provide better performance than features with a linear scale. Therefore, we expect that the extraction of the RP feature with the integration of a nonlinear scale may further improve the performance of the linear-scale-based RP feature.

In this paper, we modify the RP-based spectrum using two auditory filters, that is, Mel- and gammatone-filter banks to capture phase information instead of linear scale in original RP. Mel-filter bank, which is a perceptual scale that helps to simulate the way that the human ear works, is applied to capture important information in the full dimension of the RP-based spectrum, and the RP feature with the decreasingly captured dimension is called the Mel-scale RP. Additionally, preliminary experiments have indicated that the Mel-scale RP provides better performance compared with the MGD cepstral coefficient (MGDCC) and MFCC, and CQCC. However, the detailed Mel-scale RP information extraction and analysis were not described in [31]. In the present study, we extend our previous paper [31] by performing more experiments and analysis. The contributions of this study are as follows: (1) According to [31], the detection model in the previous work, which used only a training subset of the ASVspoof 2017 database, might not lead to an impressive result on the evaluation of subset-based testing; hence, we incorporate a training subset and development subset to train

the model in our system. (2) In addition to the Mel-filter bank represented to scale the RP information in the previous work, we also apply a gammatone filter bank, thereby simulating the auditory system of humans to convert the important information of the full-band RP-based feature in the gammatone scale, where the RP feature with the reduced dimension is called the gammatone-scale RP. (3) To use the classifier-based complementary based on different features to further improve the reliable detection decision, a combination of the scores of the proposed gammatone-scale RP feature and the MFCC/standard CQCC is also applied in this paper.

The remainder of this paper is organized as follows: in Section 2, related work is described, which includes the details of the classifier and the baseline features for replay attack detection. The proposed auditory filter-based RP information extraction is introduced in Section 3. The experimental setup and results are reported in Section 4, and our conclusion is presented in Section 5.

## 2 Related work

Several methods have been proposed for replay attack detection in recent years. Techniques have been evaluated on the Automatic Speaker Verification Spoofing and Countermeasure Challenge dataset (ASVspoof 2017). The challenge acts as a common baseline by which researchers can compare their experiments and perform evaluations. In these works, a Gaussian mixture model (GMM) is used as a classifier, and two features, that is, the MFCC and CQCC, are used as baseline features. Moreover, the MGDCC is also considered in the present paper as a phase-based feature.

### 2.1 Classifier for replay attack detection

Several classifiers can be used in replayed speech detection methods, such as a GMM [32], support vector machine, and deep neural networks. The implementation of GMM [33, 34] is one of the easiest and demonstrates high performance for replayed speech detection. Moreover, the GMM classifier is the baseline classifier for the ASVspoof 2017 challenge. Hence, we use GMM as the replayed speech detector in our experiment. Figure 1 shows the process of typical replayed speech detection system. The decision of whether the given speech is human or replayed speech is based on the likelihood ratio:

$$\wedge_{\mathrm{GMM}}(O) = \log p(O|\lambda_{\mathrm{genuine}}) - \log p(O|\lambda_{\mathrm{replay}}), \quad (1)$$

where $O$ is the feature vector of the input speech, and $\lambda_{\mathrm{genuine}}$ and $\lambda_{\mathrm{replay}}$ are the GMMs for genuine and replayed speech, respectively. The MFCC, CQCC, MGDCC, conventional RP feature, and proposed RP feature described in Sections 2.2 and 3 are used as magnitude features and phase features.

From [30], we can see that phase and magnitude-based features may contain different characteristics which can be complementary to each others. Therefore in this paper, the likelihood ratios of features are also combined to produce a new score, $\wedge^n_{\mathrm{comb}}$, as follows:

$$\wedge^n_{\mathrm{comb}} = \alpha \wedge^n_{\mathrm{mag}} + (1-\alpha)\wedge^n_{\mathrm{phase}}, \quad (2)$$

$$\alpha = \frac{\wedge^n_{\mathrm{mag}}}{\wedge^n_{\mathrm{mag}} + \wedge^n_{\mathrm{phase}}},$$

where $\wedge^n_{\mathrm{mag}}$ denotes the likelihood ratios of the magnitude-based features, $\wedge^n_{\mathrm{mag}}$ denotes the likelihood ratios of the phase-based features, and $\alpha$ denotes the weighting coefficients.

### 2.2 Baseline features for replay attack detection

#### 2.2.1 MFCC

The MFCC is one of the most popular magnitude-based features in speech processing. It uses cepstral analysis on the log magnitude spectrum in the Mel scale. The MFCC contains vocal tract dynamics and its corresponding pulse train is related to glottal motor control. This makes the feature suitable for distinguishing converted speech from human speech. The MFCC is defined as

$$C_i = \sqrt{\frac{2}{N}} \sum_{j=1}^{N} \cos\left(\frac{\pi i}{N}(j-0.5)\right), \quad (3)$$

where $N$ is the number of Mel-frequency bins of log spectrum $L$ and $i$ is the number of cepstral coefficients [1]; we use 13 coefficients in this study.

#### 2.2.2 CQCC

In contrast to the MFCC feature, the CQCC feature [1] has variable spectrum resolution, and the time-frequency representation is very effective in replayed speech detection, which is more suitable for the ASVspoof task. The CQCC feature is used as a benchmark feature, and it has been proven to be a good feature in the ASVspoof system. Therefore, we use the CQCC as the baseline feature in our experiments for comparison.

The CQCC is an amplitude-based feature that uses the constant Q transform (CQT) in combination with traditional cepstral analysis. The frequency bins of $X^{cq}(k)$ are represented in geometric scale which is different from linear scale of typical DCT. Therefore, uniform sampling is applied to the constant Q power spectrum $\log\left(|X^{cq}(k)|^2\right)$ and the resulting $\log\left(|X^{cq}(i)|^2\right)$ can then be applied with DCT. The extraction of the CQCC features is illustrated in Fig. 2.

#### 2.2.3 MGDCC

The spectrum $X(\omega)$ of a signal is obtained by the discrete Fourier transform (DFT) of an input speech signal sequence $x_n$:

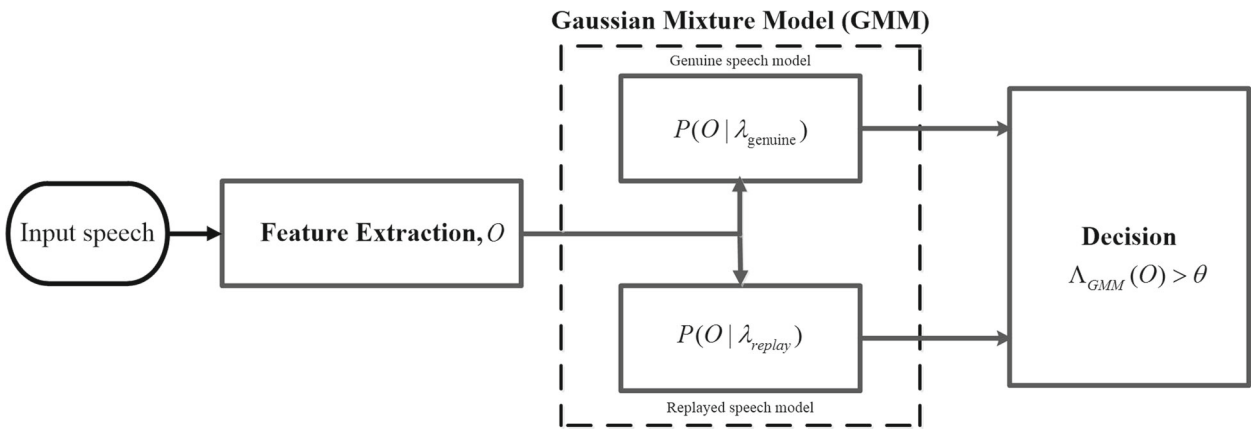


**Fig. 1** Flowchart of a GMM-based detection system

$$X(\omega,t) = |X(\omega,t)|e^{j\theta(\omega,t)}, \tag{4}$$

where $|X(\omega,t)|$ and $\theta(\omega,t)$ are the magnitude spectrum and phase spectrum at frequency $\omega$,time $t$, respectively.

The group delay [35] is defined as the negative derivative of the Fourier transform phase for the frequency:

$$\tau(\omega,t) = -\frac{d(\theta(\omega,t))}{d\omega}. \tag{5}$$

The group delay function can also be calculated directly from the speech signal using

$$\tau_x(\omega,t) = \frac{X_R(\omega,t)\,Y_R(\omega,t) + X_I(\omega,t)\,Y_I(\omega,t)}{|X(\omega,t)|^2}, \tag{6}$$

where the subscripts $R$ and $I$ denote the real and imaginary parts of the Fourier transform, respectively, and $X(\omega,t)$ and $Y(\omega,t)$ are the Fourier transforms of $x(n)$ and $nx(n)$ , respectively. Several studies have reported that the MGD is better than the original group delay [26, 36, 37]. The MGD function is defined as

$$\tau_m(\omega,t) = \frac{X_R(\omega,t)\,Y_R(\omega,t) + X_I(\omega,t)\,Y_I(\omega,t)}{|S(\omega,t)|^{2\gamma}}, \tag{7}$$

where $S(\omega,t)$ is the cepstrally smoothed spectrum of $X(\omega,t)$. To extract the cepstral coefficients from the MGD function, the DCT is applied as follows:

$$C_i = \sum_{m=0}^{M} \tau(m)\cos\left(\frac{\pi i}{M}(2m+1)\right), \tag{8}$$

where $M$ is the DCT order and $\tau$ (m) is the group delay function. In this case, the DCT acts as a linear decorrelator, which means that the diagonal covariance is available for the human/replayed speech modeling. The resulting $C$ is the MGDCC feature.

## 3 Auditory filter-based relative phase features

### 3.1 Original RP features

The phase varies depending on the framing position of speech at the same frequency [28]. The method to eliminate this variation is shown as follows. Let $s_1, s_2, \ldots, s_n, (s_{L_\omega+1} = s_1)$ be a sampling sequences for a unit circle function. The wave length in radian is $L_\omega = \frac{f_s}{f} = f_s\frac{2\pi}{\omega}$. However, the phase difference between sequences $s_1, s_2, ...., s_{L_\omega}$ and $s_2, ...., s_{L_\omega}, s_{L_\omega+1}$ is $\frac{2\pi}{L_\omega}$. To solve the problem of phase variation with respect to frame position, phase at a base frequency $\omega_b$ for all frames is kept constant, and the phases of other frequencies are calculated based on the set frequency. In this paper, the base frequency $\omega_b$ is set to 1000 $Hz$. This base frequency would

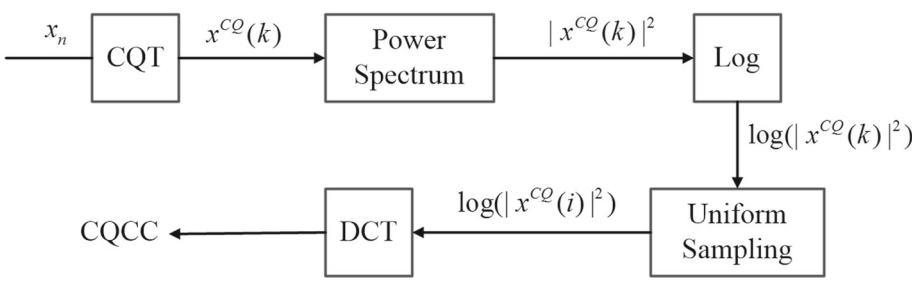


**Fig. 2** CQCC feature extraction process

not effect the performance as mentioned in [28]. If we set the base frequency $\omega$ to 0, we can achieve

$$X'(\omega,t) = |X(\omega,t)|e^{j\theta(\omega,t)} \times e^{-j(\theta(\omega,t))}. \quad (9)$$

The main difference between the typical expression of the phase in Eqs. (4) and (9) is $-\theta(\omega)$ but for other frequency which are $\omega' = 2\pi f'$ the difference is $(\omega'/\omega)(-\theta(\omega))$ ; therefore, we can obtain the following spectrum:

$$X'(\omega',t) = |X(\omega',t)|e^{j\theta(\omega',t)} \times e^{-j\frac{\omega'}{\omega}(\theta(\omega,t))}. \quad (10)$$

Thus, the phase can be further normalized as follows:

$$\tilde{\theta}(\omega',t) = \theta(\omega',t) + \frac{\omega'}{\omega}(-\theta(\omega,t)). \quad (11)$$

The base frequency is set to $2\pi \times 1000\ Hz$. Without performing any modification, the phase is limited by the phase-wrapping problem. Hence, we modify the phase into coordinates on a unit circle:

$$\tilde{\theta} \rightarrow \left\{\cos(\tilde{\theta}), \sin(\tilde{\theta})\right\} \quad , or \rightarrow \left\{\tilde{\theta}_{cos}(k), \tilde{\theta}_{sin}(k)\right\}. \quad (12)$$

Using the RP extraction method that normalizes the phase variation using the cutting position, we can reduce the phase variation problem. However, the normalization of the phase variation is still inefficient. The normalized phase depend on pitch,phonemes,channel etc., even if the speaker is fixed. Further information for addressing the variation is obtained using the statistical distribution model of the GMM.

If we split the utterance by each pitch cycle, the variation in phase information would be further obviated. Therefore, we use an extraction method that synchronizes the splitting section with the pseudo pitch cycle, as described in [29, 30]. To recombine the cutting section in the time domain, the method searches for the maximum amplitude at the center around the typical splitting section of an utterance waveform, and the peak of the utterance waveform is used as the center of the next corresponding window (pseudo pitch synchronization). Hence, the center of the frame has a similar maximum amplitude in all frames.

### 3.2 Mel-scale RP

For the original RP information, the phase information, which is mapped into coordinates on a unit circle by applying cosine and sine functions, is computed using 128 components of the sub-band spectrum before reducing them to the 19 lowest fixed components to minimize the feature parameters. In previous work, we observed that the RP feature using a linear scale could provide promising results in speaker recognition/verification tasks and spoofing attack detection focused on TTS (text-to-speech), synthesized speech, and voice conversion. However, the RP feature with a linear scale may not perform well when used to detect replayed speech, as shown in [1, 17].

In this paper, the RP-based feature and Mel-filter bank are exploited to propose a new feature extraction used for replay attack detection. The new proposed feature extraction is called the Mel-scale RP feature. The key idea is to integrate the advantage of the RP feature in phase-based information representation with the advantage of the Mel-filter in perceptual scaling, covering the range from 60 to 8000 Hz. The Mel-filter bank is a collection of triangular filters defined by center frequencies $f_c(m)$ and is given by

$$H_{\text{mel}}(k,m) = \begin{cases} 0 \text{ for } & f(k) < f_c(m-1) \\ \frac{f(k)-f_c(m-1)}{f_c(m)-f_c(m-1)} \text{ for } & f_c(m-1) \le f(k) < f_c(m) \\ \frac{f(k)-f_c(m+1)}{f_c(m)-f_c(m+1)} \text{ for } & f_c(m) < f(k) \le f_c(m+2) \\ 0 \text{ for } & f(k) \ge f_c(m+1). \end{cases} \quad (13)$$

Mel-filter bank $H_{\text{mel}}$ is an $F \times N$ matrix. It helps to capture the magnitude-based energy at each critical band and provides a rough approximation of the spectrum shape. In this paper, the Mel-filter bank is used to capture the phase information of the RP-based feature. The process of Mel-scale RP feature extraction is shown in Fig. 3b. After the phase information of the original RP feature is mapped into coordinates on a unit circle by applying cosine and sine functions, the normalized phase information, $\tilde{\theta}_{\text{con}}(k)$ and $\tilde{\theta}_{\text{sin}}(k)$, are scaled logarithmically using Mel-filter bank $H(k,m)$. Finally, based on the frame level, the scaled phase information, $[\tilde{\theta}_{\text{cos}}(k) * H_{\text{mel}}(k,m)]$, is augmented with $[\tilde{\theta}_{\text{sin}}(k) * H_{\text{mel}}(k,m)]$ as follows:

$$\tilde{\theta}_{\text{mel}} \rightarrow \left\{\left[\tilde{\theta}_{cos}(k) * H_{\text{mel}}(k,m)\right], \left[\tilde{\theta}_{\text{sin}}(k) * H_{\text{mel}}(k,m)\right]\right\}, \quad (14)$$

where $\tilde{\theta}_{\text{mel}}$ is the Mel-scale RP feature. By extracting the phase information using the Mel scale, the performance of replay attack detection is expected to improve compared with the linear scale RP-based spectrum.

### 3.3 Gammatone-scale RP

From the previous subsection, the Mel-filter bank has been used to capture the phase information from the linear-scale RP spectrum. However, a gammatone filter bank has not been applied to improve the performance of linear scale RP spectrum. The advantage of a gammatone filter that is based on the equivalent rectangular bandwidth (ERB) scale is the finer resolution at low frequencies than Mel scale, as observed in automatic speech recognition. Therefore we propose a new feature extraction method by integrating the gammatone filter bank into our RP feature. The impulse response of a gammatone filter bank centered at frequency $f$ is

$$H_{\text{gamma}}(f,t) = \begin{cases} t^{a-1}e^{2\pi bi}\cos(2\pi ft) , & t > 0 \\ 0 , & \text{otherwise,} \end{cases} \quad (15)$$

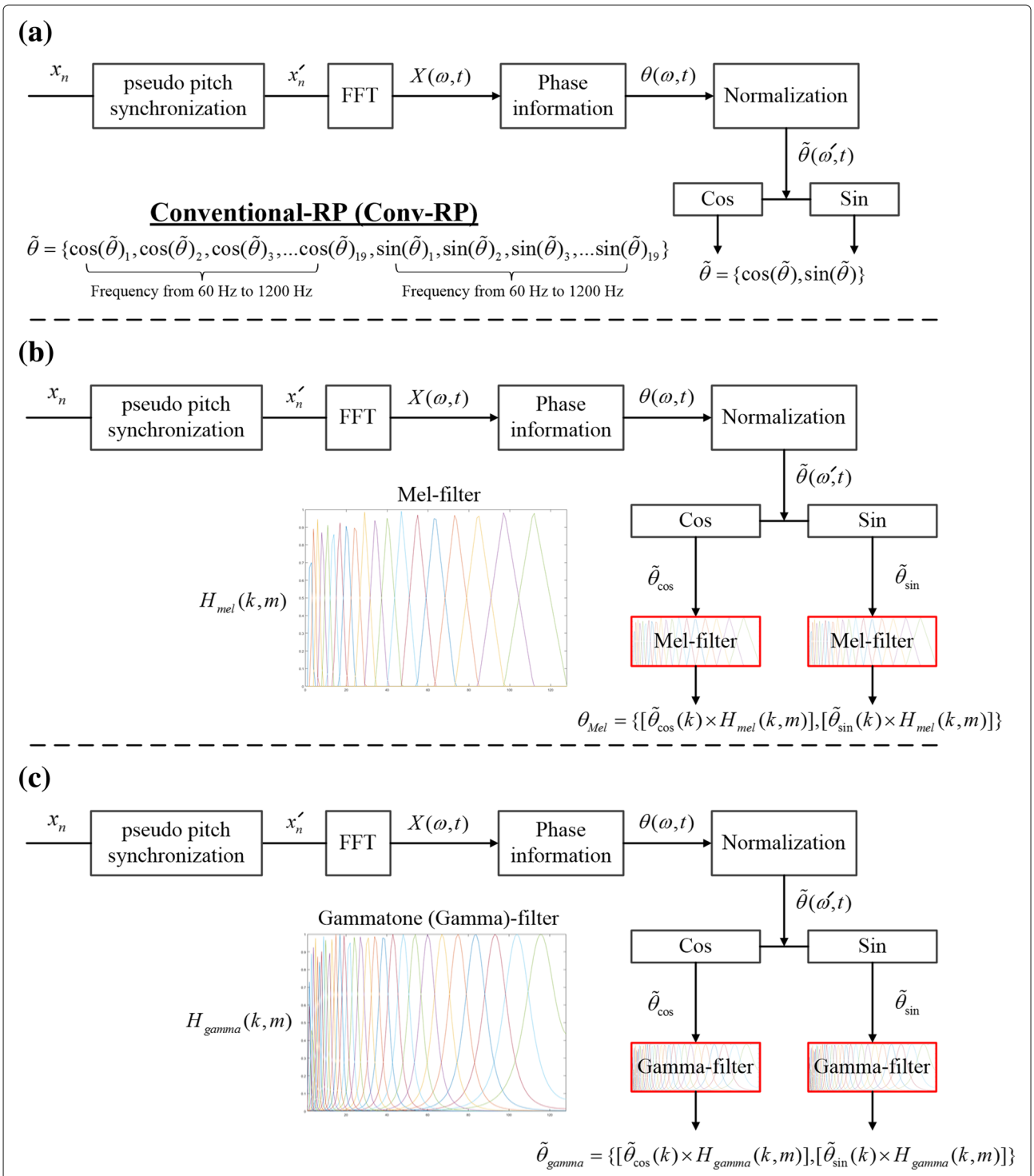


**Fig. 3** Flowchart of the conventional (original) RP, Mel-scale RP, and gammatone-scale RP feature extraction. **a** RP feature. **b** Mel- RP feature. **c** Gamma-RP feature

where $t$ refers to time, $a = 4$ is the order of the filter, and $b$ is the rectangular bandwidth, which increases with center frequency $f$. We use a bank of 128 filters whose center frequency ranges from 50 Hz to 8000 Hz. These center frequencies are equally distributed on the ERB scale [38], and the filters with higher center frequencies respond to wider frequency ranges.

In this paper, the gammatone filter bank, which is able to fine-tune the spectral resolution at a lower frequency band, is used to capture the phase information of the original RP spectrum, and the resulting feature is called the gammatone-scale RP feature. The process of gammatone-scale RP feature extraction is shown in Fig. 3c. After the phase information of the original RP feature is mapped into coordinates on a unit circle by applying cosine and sine functions, the two items of normalized phase information, $\tilde{\theta}_{\cos}(k)$ and $\tilde{\theta}_{\sin}(k)$, are scaled using gammatone filterbank $H(k,m)$. Finally, based on the frame level, the scaled phase information, $\left[\tilde{\theta}_{\cos}(k) * H_{\text{gamma}}(k,m)\right]$, is augmented with $\left[\tilde{\theta}_{\sin}(k) * H_{\text{gamma}}(k,m)\right]$ as follows:

$$\tilde{\theta}_{\text{gamma}} \rightarrow \left\{\left[\tilde{\theta}_{cos}(k) * H_{\text{gamma}}(k,m)\right], \left[\tilde{\theta}_{sin}(k) * H_{\text{gamma}}(k,m)\right]\right\}, \tag{16}$$

where $\tilde{\theta}_{\text{gamma}}$ is the gammatone-scale RP feature. By extracting the phase information using perceptual scaling, the performance of replay attack detection is expected to improve compared with the sub-band RP-based spectrum.

## 4 Experiments

### 4.1 Datasets

The ASVspoof database used in this paper is the part of ASVspoof 2017 challenge and contains three parts: the training set, development set, and evaluation set. These three sets are incoherent in terms of speakers and data collection locations. The speech in the training set was recorded at a single location and the development set was collected at two additional locations together with the location for the training set. Finally, the evaluation set was collected at two additional locations together with the locations of the training and development sets. Different recording/replay devices and acoustic conditions were used for the same location. The training set and development set contained 6 and 10 replay session, respectively, whereas the evaluation set contained the most replay sessions, at 161 sessions.

The database is mainly based on the recent text-dependent RedDots corpus [33]. It contains 10 common phrases that were recorded using different playback and recording devices as the dataset mainly focuses on replay attacks, which are the easiest and most common form of attack for an ASV system. The utterances were recorded using 16-bit precision and a 16 kHz sampling rate. The details of the dataset are further illustrated in Table 1.

According to [31], only training data were used to train the model. In this paper, for results in the development set, the model was trained using only the training set and the classifier was trained using a combination of the training and development sets when tested on the evaluation dataset.

**Table 1** Details of the ASVspoof 2017 database

| Subset | #Speakers | # Utterances | |
|---|---|---|---|
| | | Genuine | Replayed |
| Training (Train) | 10 | 1507 | 1507 |
| Development (Dev) | 8 | 760 | 950 |
| Evaluation (Eval) | 24 | 1298 | 12,008 |

### 4.2 Experimental setup

The MFCCs feature were extracted with 39 dimensions (13 MFCCs, 13 delta MFCCs, and 13 delta-delta MFCCs) and calculated using a 25-ms frame length and 10-ms frame shift. The CQCC had 96 bins-per-octave and 16 uniform samples in the first octave. The RP information was calculated using a 5-ms frame shift and 12.5-ms frame length. A spectrum of 128 components that consisted of the magnitude and phase were calculated using the DFT for every 256 samples. The MGDCC used a frame length of 25 ms and frame shift of 10 ms. We computed a spectrum of 128 components using the DFT for every 256 samples, and finally, 13 coefficients were computed using the DCT. Regarding the original RP, we computed 256 RP features that corresponded to 128 cos and 128 sin components. The range for searching the peak amplitude in the pseudo pitch synchronization was 2.5 ms, which is half the frame shift. Once pseudo pitch synchronization was complete, the Mel- and gammatone-filter banks were applied for the original RP with 128 cos and 128 sin of the RP to extract 38 dimensions of the Mel-scale RP and 64 dimensions of the gammatone-scale RP, respectively. The parameters used in the experiments follow our previous research and experiments [21, 22, 27–31]. From our experiments, the set parameter for baseline features such as MFCC,CQCC, and MGDCC has provided good results for spoofing attack detection. We did not use the same dimension as the Mel scale because the authors showed in [39] that decreasing the number of frequency bands in the gammatone filter bank decreases performance. Analysis conditions for features in this paper are described in Table 2.

Regarding classification, two GMMs for genuine and replayed speech models estimated by using maximum likelihood estimation had two 512-component models, which were trained using the expectation maximization algorithm with random initialization on genuine and replayed utterances, respectively. The score was computed as the log-likelihood ratio for the test utterance given both classifiers. We followed the baseline model that was provided by the organizers of the ASVspoof 2017 challenge.

**Table 2** Analysis conditions for all features

| | MFCC | CQCC | MGDCC | Original-RP | Mel-scale RP | Gammatone-scale RP |
|---|---|---|---|---|---|---|
| Frame length (ms) | 25 | – | 25 | 12.5 | 12.5 | 12.5 |
| Frame shift (ms) | 10 | – | 10 | 5 | 5 | 5 |
| FFT size (samples) | 512 | – | 256 | 256 | 256 | 256 |
| Dimensions | 39 | 90 | 39 | 38 | 38 | 64 |

### 4.3 Results and discussion

#### *4.3.1 Results for the development set*

In this subsection, we present the EERs (equal error rate) investigated on the development set. We compare our proposed Mel-scale RP and gammatone-scale RP features with three baseline features (MFCC, CQCC, and MGDCC) and the original RP feature. Moreover, we also refer to the results of four features used for recent replay attack countermeasures to compare them with the proposed feature. The results are reported in Table 3.

Table 3 shows that the Mel-scale RP and gammatone-scale RP feature had better EERs than the original RP. This is because the Mel-filter and gammatone-filter can capture the important information of the original RP, and the implementation of the auditory feature improved the robustness of the relative feature.

By comparing these result with the CQCC, we found that our proposed method did not perform well in development. This might be because the magnitude-based CQCC feature could provide superior discrimination power for the artificially generated speech in the development set, which had similar acoustic conditions to the training set.

**Table 3** Results for the development set

| Feature | EER (%) |
|---|---|
| MFCC | 13.78 |
| CQCC | 6.81 |
| MGDCC | 25.29 |
| Original-RP | 14.50 |
| Mel-scale RP | 9.57 |
| Gammatone-scale RP | 10.84 |
| MFCC + CQCC | 5.31 |
| CQCC + MGDCC | 12.08 |
| MFCC + Mel-scale RP | 6.05 |
| MFCC + gammatone-scale RP | 7.77 |
| CQCC + Mel-scale RP | 5.82 |
| CQCC + gammatone-scale RP | 5.33 |
| CQCC (6–8 kHz) (result in [17]) | 5.13 |
| VESA-IFCC (result in [18]) | 4.63 |
| Voice source + instantaneous frequency + cepstral features + CQCC (result in [20]) | 5.31 |
| RFCC + LFCC (result in [19]) | - |

The combination of scores of different features provides us with a promising result. First, we can see that there was a significant improvement in performance once a magnitude-based feature, such as the CQCC or MFCC, was combined with our phase-based feature (i.e., Mel-scale RP and gammatone-scale RP).

Second, from the results, we can also state that CQCC captured information that was more salient to the task of replayed speech detection compared with the MFCC, and thus, combining CQCC with our proposed RP features (Mel -scale and gammatonescale) provided better performance than combining it with MFCCs. The CQCC has already been one of the best performing magnitude-based features in ASV systems. Thus, by combining it with gammatone-scale RP, we achieved good performance. The above combination performed better than a combination of the CQCC and Mel-scale RP. This may be because gammatone-scale RP contains more filter bank than Mel-scale RP. The variable resolution of the CQCC may complement the more available frequency band in gammatone-scale RP.

#### *4.3.2 Results for the evaluation set*

In this subsection, we present the EERs investigated on the evaluation dataset. The difference between the development and evaluation sets is that the audio in the evaluation set was recorded in real-world conditions using a variety of recording devices. Hence, the performance on the evaluation set is much more significant than that on the development set. The results of evaluation dataset are shown in Table 4.

We experimented on three baseline features: the MFCC, CQCC, and MGDCC. The CQCC performed better than the MFCC as the CQCC had a variable resolution compared with the MFCC, which mainly focuses on low-frequency components. Second, the popular phase feature MGDCC did not provide good results. This may be because the MGDCC contained both magnitude and phase features, which may make the detection of replayed speech in-discriminating.

**Table 4** Results for the evaluation set

| Feature | EER(%) |
|---|---|
| MFCC | 29.11 |
| CQCC | 21.61 |
| MGDCC | 31.81 |
| Original-RP | 13.15 |
| Mel-scale RP | 10.98 |
| Gammatone-scale RP | 11.07 |
| MFCC + CQCC | 28.12 |
| CQCC + MGDCC | 22.48 |
| MFCC + Mel-scale RP | 15.34 |
| MFCC + Gammatone-scale RP | 14.51 |
| CQCC + Mel-scale RP | 10.74 |
| CQCC + Gammatone-scale RP | 9.48 |
| CQCC (6–8 kHz) (result in [17]) | 17.31 |
| VESA-IFCC (result in [18]) | 14.06 |
| Voice source + instantaneous frequency + cepstral features + CQCC (result in [20]) | 13.95 |
| RFCC + LFCC (result in [19]) | 10.52 |

The evaluation dataset was real-world recorded data, and therefore, may have contained noise or artifacts that distorted the speech. The relative phase feature normalizes the phase of other frequencies based on the frequency, so the phase values of the base frequency with different noises are the same. Therefore, the RP is more robust to noise [27]. Therefore, Mel-scale RP and gammatone-scale RP were robust to noise and provided an efficient representation. Moreover, our proposed methods performed significantly better than the CQCC baseline. Thus, we can conclude that our proposed feature contains efficient information to detect replayed speech, even in real-world conditions.

Finally, the score combination of Mel-scale RP/gammatone-scale RP and the MFCC/CQCC was applied to utilize the classifier-based complementary based on different features. Table 4 shows the results of the score combination. We can see that the replay attack detection based on the combined score of the CQCC/MFCC and gammatone-scale RP outperformed the systems that used the combined score of the CQCC/MFCC and Mel-scale RP, or used an individual feature. This is because the CQCC and gammatone-scale RP have a strong complementary nature, as described in the previous subsection. Moreover, our proposed combination of a magnitude and phase feature, that is, the CQCC and gammatone-scale RP, achieved a more significant relative error reduction than the combination of the baseline magnitude and phase feature, that is, the CQCC and MGDCC. We can also see that the combination of two magnitude features did not perform well compared with our proposed method. This proves that using only a magnitude feature may not be sufficient for real-world recorded data.

We also compared our results with other recent research publications and drew some conclusions. The CQCC at a high frequency in [17] performed better than the baseline CQCC. From this, we can conclude that for a magnitude feature, the high frequency contains discrimination information in replayed attacks; however, high-frequency components can be easily distorted by noise. By comparing our results with [17], we can see that the phase contains significant information for the replayed attacked detection task, and using only magnitude information could not provide good performance. The VESA-IFCC feature proposed in [18] achieved very good performance for a single feature, and Teager energy could be modified to achieve better robustness in the future as this method provided great performance on the development dataset. In [20], combining multiple types of features also improved performance. Although the performance of individual features did not exhibit good performance, multiple combinations of such features can increase the performance overall of replayed speech detection systems. In this paper, we have only investigated the combination of two features; therefore, further improvement could be achieved by combining more appropriate features.

## 5 Conclusion

In this paper, we proposed two novel features based on auditory filters to distinguish replayed speech from genuine speech. For the proposed features, the contribution of the Mel filter and gammatone filter were exploited to capture the important information of the conventional RP feature. We introduced the first implementation of the gammatone filter bank with the RP feature. Our proposed feature, Mel-scale RP, and gammatone-scale RP obtained significant performance improvement over the CQCC baseline and original RP feature. The results on ASVspoof 2017 showed that, for an individual feature with a GMM-based classifier, the Mel-scale RP and gammatone-scale RP performed relatively better than the baseline MFCC and CQCC features and other comparison feature (i.e., RFCC, CQCC (6–8 kHz), and VESA-IFCC) on the evaluation dataset. Moreover, the scores of the Mel-scale RP and gammatone-scale RP could be combined with the CQCC feature to obtain an additional improvement on the development and evaluation subsets. We confirmed that the proposed Mel-scale RP and gammatone-scale RP were very useful for replay attack detection.

In the future, we would like to explore the effectiveness of using the Teager energy operator phase information and other classifiers for replay attack detection.

### Abbreviations

ASV: Automatic speaker verification; CQCC: Constant Q transform cepstral coefficient; CQT: Constant Q transform; DCT: Discrete cosine transform; DFT: Discrete fourier transform; EER : Equal error rate; ERB: Equivalent rectangular bandwidth; GMM: Gaussian mixture model; LFCC: Linear frequency cepstral coefficients; ms: milliseconds; MFCC: Mel frequency cepstral coefficients; MGD: Modified group delay; MGDCC: Modified group delay cepstral coefficients; RFCC: Rectangular filter cepstral coefficients; RP: Relative phase; TTS: Text-to-speech; VESA-IFCC: Variable length Teager energy operator energy separation algorithm-instantaneous frequency cosine coefficients

### Acknowledgements

The research was supported partially by the National Natural Science Foundation of China (No. 61771333) and the Tianjin Municipal Science and Technology Project (No. 18ZXZNGX00330).

### Funding

The National Natural Science Foundation of China (No. 61771333) and the Tianjin Municipal Science and Technology Project (No. 18ZXZNGX00330).

### Availability of data and materials

Please contact author for data requests.

### Authors' contributions

ZO experimented on algorithm in Sections 2 and 3.1. KP designed the filter-bank in Section 3.3. LW and SN provided insight and recommendation on the experiments. ML helped us with the experiments on the existing method and baseline. MI and JD helped us with the analysis and provided support during the time of designing the algorithms and experiments. ZO and KP authors equally contributed to the research and this paper. All authors read and approved the final manuscript.

### Competing interests

The authors declare that they have no competing interests.

### Author details

[1]Nagaoka University of Technology, Nagaoka, Niigata, Japan. [2]Tianjin Key Laboratory of Cognitive Computing and Application, College of Intelligence and Computing, Tianjin University, Tianjin, China. [3]Department of Telecommunication Engineering, Faculty of Engineering and Architecture, Rajamangala University of Technology Isan Nakhonrachasrima, Nakhonrachasrima, Thailand. [4]Chubu University, Kasugai, Aichi, Japan.

Received: 5 February 2019 Accepted: 21 May 2019
Published online: 10 June 2019

### References

1. M. Todisco, H. Delgado, N. Evans, in *Speaker Odyssey Workshop, Bilbao, Spain*. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients, (2016), pp. 249–252. http://www.odyssey2016.org/papers/pdfs_stamped/59.pdf. Accessed 04 June 2019
2. W. Rao, M.-W. Mak, K.-A. Lee, in *2015 IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP)*. Normalization of total variability matrix for i-vector/plda speaker verification, (2015), pp. 4180–4184. https://ieeexplore.ieee.org/document/7178758. Accessed 04 June 2019
3. G. Heigold, I. Moreno, S. Bengio, S. Shazeer, in *2016 IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP)*. End-to-end text-dependent speaker verification, (2016), pp. 5115–5119. https://ieeexplore.ieee.org/abstract/document/7472652. Accessed 04 June 2019
4. N. W. Evans, T. Kinnunen, J. Yamagishi, in *Interspeech*. Spoofing and countermeasures for automatic speaker verification, (2013), pp. 925–929. http://cs.uef.fi/sipu/pub/IS131294.pdf. Accessed 04 June 2019
5. Z. Wu, N. Evans, T. Kinnunen, J. Yamagishi, F. Alegre, H. Li, Spoofing and countermeasures for speaker verification: A survey. Speech Comm. **66**, 130–153 (2015). https://www.sciencedirect.com/science/article/pii/S0167639314000788. https://doi.org/10.1016/j.specom.2014.10.005
6. Z.-F. Wabg, G. Wei, Q.-H. H, in *2011 International Conference On Machine Learning and Cybernetics (ICMLC), vol. 4*. Channel pattern noise based playback attack detection algorithm for speaker recognition, (2011), pp. 1708–1713. https://ieeexplore.ieee.org/document/6016982. Accessed 04 June 2019
7. D. A. Reynolds, R. C. Rose, Robust text-independent speaker identification using gaussian mixture speaker models. IEEE Trans. Speech Audio Process. **3**(1), 72–83 (1995). https://ieeexplore.ieee.org/document/365379
8. A. Ross, A. K. Jain, in *2004 12th EuropeanSignal Processing Conference*. Multimodal biometrics: An overview, (2004), pp. 1221–1224. https://ieeexplore.ieee.org/document/7080214. Accessed 04 June 2019
9. A. K. Jain, A. Ross, S. Prabhakar, An introduction to biometric recognition. IEEE Trans. Circ. Syst. Video Technol. **14**(1), 4–20 (2004). https://ieeexplore.ieee.org/document/1262027
10. Z. Wu, S. Gao, E. S. Cling, H. Li, in *2014 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. A study on replay attack and anti-spoofing for text-dependent speaker verification, (2014), pp. 1–5. https://ieeexplore.ieee.org/abstract/document/7041636. Accessed 04 June 2019
11. M. J. Alam, P. Kenny, T. Stafylakis, in *Sixteenth Annual Conference of the International Speech Communication Association,*. Combining amplitude and phase-based features for speaker verification with short duration utterances, (2015). https://pdfs.semanticscholar.org/6d22/330884f74d593afa3a672de39598b5f6ac11.pdf. Accessed 04 June 2019
12. Y. Liu, Y. Tian, L. He, J. Liu, M. T. Johnson, in *Sixteenth Annual Conference of the International Speech Communication Association*. Simultaneous utilization of spectral magnitude and phase information to extract supervectors for speaker verification anti-spoofing, (2015). https://www.isca-speech.org/archive/interspeech_2015/papers/i15_2082.pdf. Accessed 04 June 2019
13. H. Sailor, M. Kamble, H. Patil, in *Proc. Interspeech 2018*. Auditory filterbank learning for temporal modulation features in replay spoof speech detection, (2018), pp. 666–670. https://www.isca-speech.org/archive/Interspeech_2018/pdfs/1651.pdf. Accessed 04 June 2019
14. Z. Chen, W. Zhang, Z. Xie, X. Xu, D. Chen, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Recurrent neural networks for automatic replay spoofing attack detection, (2018), pp. 2052–2056. https://ieeexplore.ieee.org/abstract/document/8462644. Accessed 04 June 2019
15. S. Jelil, S. Kalita, S. M. Prasanna, R. Sinha, Exploration of compressed ilpr features for replay attack detection. Interspeech. **8**(760), 950 (2018). https://www.isca-speech.org/archive/Interspeech_2018/pdfs/1297.pdf
16. G. Suthokumar, V. Sethu, C. Wijenayake, E. Ambikairajah, in *Proc. Interspeech 2018*. Modulation dynamic features for the detection of replay attacks, (2018), pp. 691-695. https://www.isca-speech.org/archive/Interspeech_2018/pdfs/1846.pdf. Accessed 04 June 2019
17. M. Witkowski, S. Kacprzak, P. Zelasko, K. Kowalczyk, J. Gałka, in *18th Annual Conf. Int. Speech Communication Association (INTERSPEECH), Stockholm, Sweden*. Audio replay attack detection using high-frequency features, (2017), pp. 27–31. https://pdfs.semanticscholar.org/a2b4/c396dc1064fb90bb5455525733733c761a7f.pdf
18. H. A. Patil, M. R. Kamble, T. B. Patel, M. H. Soni, in *Proc. INTERSPEECH*. Novel variable length teager energy separation based instantaneous frequency features for replay detection, (2017), pp. 12–16. https://www.isca-speech.org/archive/Interspeech_2017/abstracts/1362.html
19. R. Font, J. Espn, M. J. Cano, in *Proc. INTERSPEECH*. Experimental analysis of features for replay attack detection results on the asvspoof 2017 challenge, (2017), pp. 7–11. https://www.isca-speech.org/archive/Interspeech_2017/abstracts/0450.html
20. S. Jelil, R. K. Das, S. M. Prasanna, R. Sinha, in *Proc. INTERSPEECH*. Spoof detection using source, instantaneous frequency and cepstral features, (2017), pp. 22–26. https://www.isca-speech.org/archive/Interspeech_2017/abstracts/0930.html
21. Z. Oo, Y. Kawakami, L. Wang, S. Nakagawa, X. Xiao, M. Iwahashi, in *Proc. INTERSPEECH*. DNN-based amplitude and phase feature enhancement for noise robust speaker identification, (2016), pp. 2204–2208. https://www.isca-speech.org/archive/Interspeech_2016/abstracts/0717.html. Accessed 04 June 2019

22. Z. Oo, L. Wang, K. Phapatanaburi, M. Iwahashi, S. Nakagawa, J. Dang, Phase and reverberation aware dnn for distant-talking speech enhancement. Multimed. Tools Appl. **77**(14), 18865–18880 (2018). https://link.springer.com/article/10.1007/s11042-018-5686-1
23. J. Sanchez, I. Saratxaga, I. Hernaez, E. Navas, D. Erro, T. Raitio, Toward a universal synthetic speech spoofing detection using phase information. IEEE Trans. Inf. Forensic Secur. **10**(4), 810–820 (2015). https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=7029029
24. Z. Wu, E. S. Chng, H. Li, in *Thirteenth Annual Conference of the International Speech Communication Association*. Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition, (2012). https://pdfs.semanticscholar.org/617d/f2f1be497d98c0e255d66eb690af5a97b259.pdf. Accessed 04 June 2019
25. F. Itakura, T. Umezaki, in *1987 IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP)*. Distance measure for speech recognition based on the smoothed group delay spectrum, (1987), pp. 1257–1260. https://ieeexplore.ieee.org/abstract/document/1169476
26. R. M. Hegde, H. A. Murthy, G. R. Rao, in *2004 IEEE International Conference On Acoustics, Speech, and Signal Processing (ICASSP), vol. 1*. Application of the modified group delay function to speaker identification and discrimination, (2004), p. 517. https://ieeexplore.ieee.org/document/1326036
27. L. Wang, K. Minami, K. Yamamoto, S. Nakagawa, in *2010 IEEE International Conference On Acoustics Speech and Signal Processing (ICASSP)*. Speaker identification by combining mfcc and phase information in noisy environments, (2010), pp. 4502–4505. https://ieeexplore.ieee.org/document/5495586. Accessed 04 June 2019
28. S. Nakagawa, L. Wang, S. Ohtsuka, Speaker identification and verification by combining mfcc and phase information. IEEE Trans. Audio Speech Lang. Process. **20**(4), 1085–1095 (2012). https://ieeexplore.ieee.org/document/6047571
29. L. Wang, Y. Yoshida, Y. Kawakami, S. Nakagawa, in *Sixteenth Annual Conference of the International Speech Communication Association*. Relative phase information for detecting human speech and spoofed speech, (2015). http://www.asvspoof.org/asvspoof2015/longbiao.pdf. Accessed 04 June 2019
30. L. Wang, S. Nakagawa, Z. Zhang, Y. Yoshida, Y. Kawakami, Spoofing speech detection using modified relative phase information. IEEE J. Sel. Top. Sign. Process. **11**(4), 660–670 (2017). http://www.slp.cs.tut.ac.jp/nakagawa/pdfs/wang.ieee.2017.pdf
31. D. Li, L. Wang, J. Dang, M. Liu, Z. Oo, S. Nakagawa, H. Guan, X. Li, in *Proc. Interspeech*. Multiple phase information combination for replay attacks detection, (2018), pp. 656–660. https://www.isca-speech.org/archive/Interspeech_2018/pdfs/2001.pdf. Accessed 04 June 2019
32. G. S. Kumar, K. P. Raju, M. R. CPVNJ, P. Satheesh, Speaker recognition using gmm. Int. J. Eng. Sci. Technol. **2**(6), 2428–2436 (2010). https://pdfs.semanticscholar.org/3593/eba26ee4aac7dca4f4bd75f79cdce46b4894.pdf
33. T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamäki, D. Thomsen, A. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco, in *2017 IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP)*. Reddots replayed: A new replay spoofing attack corpus for text-dependent speaker verification research, (2017), pp. 5395–5399. https://ieeexplore.ieee.org/document/7953187. Accessed 04 June 2019
34. C. Hanilçi, T. Kinnunen, M. Sahidullah, A. Sizov, Spoofing detection goes noisy: An analysis of synthetic speech detection in the presence of additive noise. Speech Comm. **85**, 83–97 (2016). https://www.sciencedirect.com/science/article/pii/S0167639316300681
35. R. M. Hegde, H. A. Murthy, V. R. R. Gadde, Significance of the modified group delay feature in speech recognition. IEEE Trans. Audio Speech Lang. Process. **15**(1), 190–202 (2007). https://ieeexplore.ieee.org/document/4032772
36. R. Padmanabhan, S. H. Parthasarathi, H. A. Murthy, in *Proc. INTERSPEECH*. Robustness of phase based features for speaker recognition, (2009), pp. 2535–2358. https://www.isca-speech.org/archive/archive_papers/interspeech_2009/papers/i09_2355.pdf. Accessed 04 June 2019
37. J. Kua, J. Eppsi, E. Ambikairajah, E. Choi, in *Proc. INTERSPEECH*. LS Regularization of Group Delay Features for Speaker Recognition, (2009), pp. 2887–2890. https://www.isca-speech.org/archive/archive_papers/interspeech_2009/papers/i09_2887.pdf. Accessed 04 June 2019
38. Y. Shao, Z. Jin, D. Wang, S. Srinivasan, in *2009 IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP)*. An auditory-based feature for robust speech recognition, (2009), pp. 4625–4628. https://ieeexplore.ieee.org/document/4960661. Accessed 04 June 2019
39. X. Zhao, D. Wang, in *2013 IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP)*. Analyzing noise robustness of mfcc and gfcc features in speaker identification, (2013), pp. 7204–7208. https://ieeexplore.ieee.org/document/6639061. Accessed 04 June 2019


## Publisher's Note