symmetry

MDPI

Article

New Acoustic Features for Synthetic and Replay Spoofing Attack Detection

Linqiang Wei [1], Yanhua Long [1,*], Haoran Wei [2] and Yijie Li [3]

[1] Key Innovation Group of Digital Humanities Resource and Research, Shanghai Normal University, Shanghai 200234, China; weilinqiang925@pingan.com.cn
[2] Department of ECE, University of Texas at Dallas, Richardson, TX 75080, USA; haoran.wei@utdallas.edu
[3] Unisound AI Technology Co., Ltd., Beijing 100096, China; liyijie@unisound.com
* Correspondence: yanhua@shnu.edu.cn

Abstract: With the rapid development of intelligent speech technologies, automatic speaker verification (ASV) has become one of the most natural and convenient biometric speaker recognition approaches. However, most state-of-the-art ASV systems are vulnerable to spoofing attack techniques, such as speech synthesis, voice conversion, and replay speech. Due to the symmetry distribution characteristic between the genuine (true) speech and spoof (fake) speech pair, the spoofing attack detection is challenging. Many recent research works have been focusing on the ASV anti-spoofing solutions. This work investigates two types of new acoustic features to improve the performance of spoofing attacks. The first features consist of two cepstral coefficients and one LogSpec feature, which are extracted from the linear prediction (LP) residual signals. The second feature is a harmonic and noise subband ratio feature, which can reflect the interaction movement difference of the vocal tract and glottal airflow of the genuine and spoofing speech. The significance of these new features has been investigated in both the t-stochastic neighborhood embedding space and the binary classification modeling space. Experiments on the ASVspoof 2019 database show that the proposed residual features can achieve from 7% to 51.7% relative equal error rate (EER) reduction on the development and evaluation set over the best single system baseline. Furthermore, more than 31.2% relative EER reduction on both the development and evaluation set shows that the proposed new features contain large information complementary to the source acoustic features.

Keywords: linear prediction residual signal; harmonic and noise ratio; anti-spoofing; system fusion

Citation: Wei, L.; Long, Y.; Wei, H.; Li, Y. New Acoustic Features for Synthetic and Replay Spoofing Attack Detection. *Symmetry* **2022**, *14*, 274. https://doi.org/10.3390/sym14020274

Academic Editor: Alexander Zaslavski

Received: 8 November 2021
Accepted: 11 January 2022
Published: 29 January 2022

1. Introduction

In recent years, the performances of automatic speaker verification (ASV) systems have been significantly improved. It has become one of the most natural and convenient biometric speaker recognition methods. The ASV technologies are now widely deployed in many diverse applications and services, such as call centers, intelligent personalized services, payment security, access authentication, etc. However, many studies in recent years have found that the state-of-the-art ASV systems fail to handle the speech spoofing attacks, especially for three major types of attacks: the replay [1], speech synthesis [2,3], and voice conversion. Therefore, great efforts are required to ensure adequate protection of ASV systems against spoofing. Due to the symmetry distribution characteristic between the genuine speech and spoof speech pair, the spoofing attack detection is challenging. As commonly used acoustic features between genuine and spoof speech share lots of similarities, specifically designed features for spoofing attack detection are in great demand.

Replay attacks refer to using pre-recorded speech samples collected from the genuine target speakers to attack the ASV systems. They are the most simple and accessible attacks, and pose huge threats to ASV, because of the availability of high quality and low-cost recording devices [1]. Unlike replay attacks, speech synthesis and voice conversion attacks are usually produced from sophisticated speech processing techniques. In these attacks,

the state-of-the-art text-to-speech synthesis (TTS) systems are utilized to generate the artificial speech signals, while the voice conversion (VC) systems are utilized to adapt a given natural speech to target speakers. Given sufficient training data, both the speech synthesis and voice conversion techniques can produce high-quality spoofing signals to mimic the ASV systems [2].

To help research works on anti-spoofing and provide common platforms for the assessment and comparison of spoofing countermeasures, the Automatic Speaker Verification Spoofing and Countermeasures (ASVSpoof) challenges were organized every two years in ASV society from 2015. The most recent ASVSpoof 2019 was the first to consider the replay, speech synthesis, and voice conversion attacks within a single challenge [2]. In ASVSpoof 2019, these attacks are divided into the logical access (LA, speech synthesis, and voice conversion attacks) and physical access (PA, replay attacks) scenarios according to different use case scenarios (https://www.asvspoof.org/(accessed on 1 November 2021)).

In this study, we focus on using the ASVSpoof 2019 database to explore automatic spoofing detection methods for both logical and physical access spoofing attacks. We investigate two types of new acoustic features to enhance the discriminative information between bonafide and spoofed speech utterances.

The first proposed features consist of two cepstral coefficients and one log spectrum feature, which are extracted from the linear prediction (LP) residual signals of speech utterances. As we know that the LP residual signal of each utterance implicity contains its excitation source information. Although the latest TTS and VC algorithms can even produce spoofed speech that is perceptually indistinguishable from bonafide speech under some well-controlled conditions, there is still a big difference in excitation source signals between these spoofed speech and the bonafide speech. Moreover, from the analysis of replayed speech, we find that most of the playback device distortions occur in excitation signals, they can not affect much in the speech spectrum trajectories. Therefore, compared with the typical acoustic features that are directly extracted from the bonafide speech, we speculate the cepstral features extracted from the residual signals can capture better discrimination between the bonafide and spoofed speech, either for the synthetic or voice converted attacks or for the replayed attacks.

The second new feature is the harmonic and noise subband ratio feature (HNSR). It is motivated by the conventional Harmonic plus Noise Model (HNM) [4] for high-quality speech analysis. Based on the HNM, the speech signal is decomposed into a deterministic component and a stochastic (noise or residual) component. According to the decomposition, we assume that the interaction movement difference of the vocal tract and glottal airflow of the bonafide and spoofed speech can be reflected and distinguishable in the HNSR features. The usefulness of the proposed features is examined in both the t-stochastic neighborhood embedding (t-SNE) space and the spoofing detection modeling space.

Besides that two new features, this paper also compared various single features on two major baseline systems, which are Gaussian mixture model (GMM) based system and Light Convolutional Neural Networks (LCNN)-based system. Then, to verify the complementarity between different features, the score-level fusion system is proposed.

2. Related Work

In the literature, the continuously organized ASVSpoof challenges [2,5,6] have resulted in a large number of spoofing detection countermeasures. These countermeasures can be divided into two main categories, binary classification modeling algorithms, and new discriminative features, either for the replayed, or the synthetic or voice converted speech detection.

2.1. Classifiers

At the classifier level, the algorithms or architectures used for replay speech and synthetic or converted speech detection are almost the same in recent ASVspoof challenges. They can be categorized into shallow models and deep models. In the literature, the Gaussian

mixture model (GMM) is the most widely used shallow model in most spoofing detection works [1,7–12]. Most of the deep models are based on the DNNs [13–15], recurrent neural networks [16,17], and convolution neural networks (CNN) [18,19]. It has been found that the deep models perform better than the shallow models given the same input features in the in-domain ASV tasks. However, our previous work in [20] found that the performance gains obtained from the development set was very difficult to generalize to the evaluation set, the shallow model GMMs showed better robustness to the unseen or new conditions in our works on ASVSpoof 2019 challenge.

2.2. Features

For replay speech detection, previous works mainly focus on extracting new features to reflect the acoustic level difference between original and replay speech. These differences have resulted from the recording or playback devices, and the recording environments. Ref. [21] explored the re-recording distortions by using the constant Q cepstral coefficients (CQCC) in the high-frequency sub-band (6–8 kHz). Ref. [22] proposed to extract phase information by incorporating the Constant Q Transform (CQT) with a Modified Group Delay (MGD) on ASVSpoof 2019. Ref. [23] extracted the mel-frequency cepstral coefficients (MFCC) from the linear prediction residual signal to capture the playback device distortions. Ref. [24] proposed a low-frequency frame-wise normalization in CQT domain to capture the artifacts in the playback speech. Though MFCC is a widely used feature for speech-related works, it is not utilized for the two ASVSpoof baseline systems conducted by this work. Besides these hand-crafted features, deep features using neural networks have also been investigated to detect playback speech. For instance, the convolutional neural network (CNN) was used to learn deep features from the group delay [25] and Siamese embedding the spectrogram [26]. In [27], the DNN-based frame-level and RNN-based sequence-level features were extracted to capture a better representation of playback distortions. In general, due to the strong feature modeling ability of deep neural networks, these deep features are more effective than hand-crafted features on the in-domain ASV tasks. However, these features may not easy to be generalized to out-of-domain ASV tasks, because their extraction is highly dependent on the DNN model training data.

To capture the artifacts introduced during the TTS and VC speech manipulation, new features mainly focus on both the acoustic and prosodic level features. For instance, the modulation features from magnitude and phase spectrum proposed in [28] were used to detect temporal artifacts caused by the frame-by-frame speech synthesis processing. The best system [29] in ASVSpoof 2015 used a combination of standard MFCC and cochlear filter cepstral coefficients (CFCCs) with the change in instantaneous frequency (IF) to detect the discrimination between bonafide and spoofed speech. As the phase information is almost entirely lost in a spoofed speech in current synthesis/conversion techniques, a modified group delay-based feature and the frequency derivative of the phase spectrum have been explored in [7]. And in [8], the fundamental frequency variation features were also proposed to capture the prosodic difference between the bonafide and spoofed speech. Actually, in the ASVSpoof challenge, we find many acoustic level features are effective in both the logical and physical spoofing speech detection, the typical features are the CQCC [9] and the linear frequency cepstral coefficient (LFCC) [30] features that have been used to build the ASVSpoof 2019 baselines.

In this study, we also focus on exploring new acoustic features to capture the artifacts in logical and physical spoofing speech detection. The CQCC and LFCC features extracted from the LP residual signals, and the harmonic and noise subband ratio feature are first investigated together with the shallow GMM classifier. Then we validate the effectiveness of using deep CNNs to model the spectrum of natural speech and its residual signals. Details of all these new features are presented in the next sections.

3. New Cepstral Coefficients of Residual Signals

3.1. Linear Prediction Residual Modeling and Analysis

In the LP model of speech, the bonafide speech signal $S_t(n)$ is formulated as

$$S_t(n) = \hat{S}_t(n) + r_t(n) \tag{1}$$

$$\hat{S}_t(n) = \sum_{k=1}^{p} a_k S_t(n-k) \tag{2}$$

where $\hat{S}_t(n)$ models the vocal-tract component of the bonafide speech signal in terms of LP coefficients $a_k, k = 1, 2, \ldots, p$, and the error in the prediction $r_t(n)$, called as LP residual signal that models the excitation component.

As discussed in [23], the replayed speech $S_r(n)$ is the convolution of input speech $S_t(n)$ with impulse response $i(n)$ of the playback device as,

$$S_r(n) = S_t(n) * i(n) \tag{3}$$

According to Equations (1) and (2), $S_r(n)$ can be expanded to

$$S_r(n) = [\sum_{k=1}^{p} a_k S_t(n-k) + r_t(n)] * i(n) \tag{4}$$

Equation (4) can be expanded to Equation (5) as

$$S_r(n) = \sum_{k=1}^{p} a_k S_t(n-k) * i(n) + r_t(n) * i(n) \tag{5}$$

Equation (6) is a simplified version of Equation (5). Equation (7) is part of Equation (6), indicating the replay signal generated by linear prediction, the other part of Equation (6) is the residual signal.

$$S_r(n) = \hat{S}_r(n) + r_r(n) \tag{6}$$

$$\hat{S}_r(n) = \sum_{k=1}^{p} c_k S_r(n-k) \tag{7}$$

the $\hat{S}_r(n)$ models the vocal-tract component of replayed speech in terms of LP coefficients $c_k, k = 1, 2, \ldots, p$, and $r_r(n)$ corresponds to the excitation component. It is clear from Equation (5) that, both the vocal-tract and excitation source components of bonafide speech $S_t(n)$ are affected by the characteristics of playback device $i(n)$. Detail analysis in [23] showed that, compared with the vocal tract component, the source component is relatively more affected. It indicated that the $i(n)$ affected residual signal is more different from that of bonafide speech, so it has advantages for detecting whether the speech is bonafide or replayed.

For the synthesized speech, as most latest TTS and VC techniques are parametric-based approaches, not the traditional unit selection and waveform concatenation approaches, there is still a big difference in the excitation source signals between these synthesized spoofing and the bonafide speech.

Figures 1 and 2 demonstrate the time and spectral domain representation, and the corresponding LP residuals between a bonafide speech and its corresponding replayed and synthesized speech, respectively. It is clear to observe that, there is a big difference between time-domain representation comparisons under PA and LA scenarios. Compared with the large temporal differences observed in Figure 2, The temporal differences between the bonafide speech and replayed one, and their corresponding LP residuals are much smaller. This indicates that the excitation variations introduced by speech synthesis methods are much larger than the impulse response of the playback device.

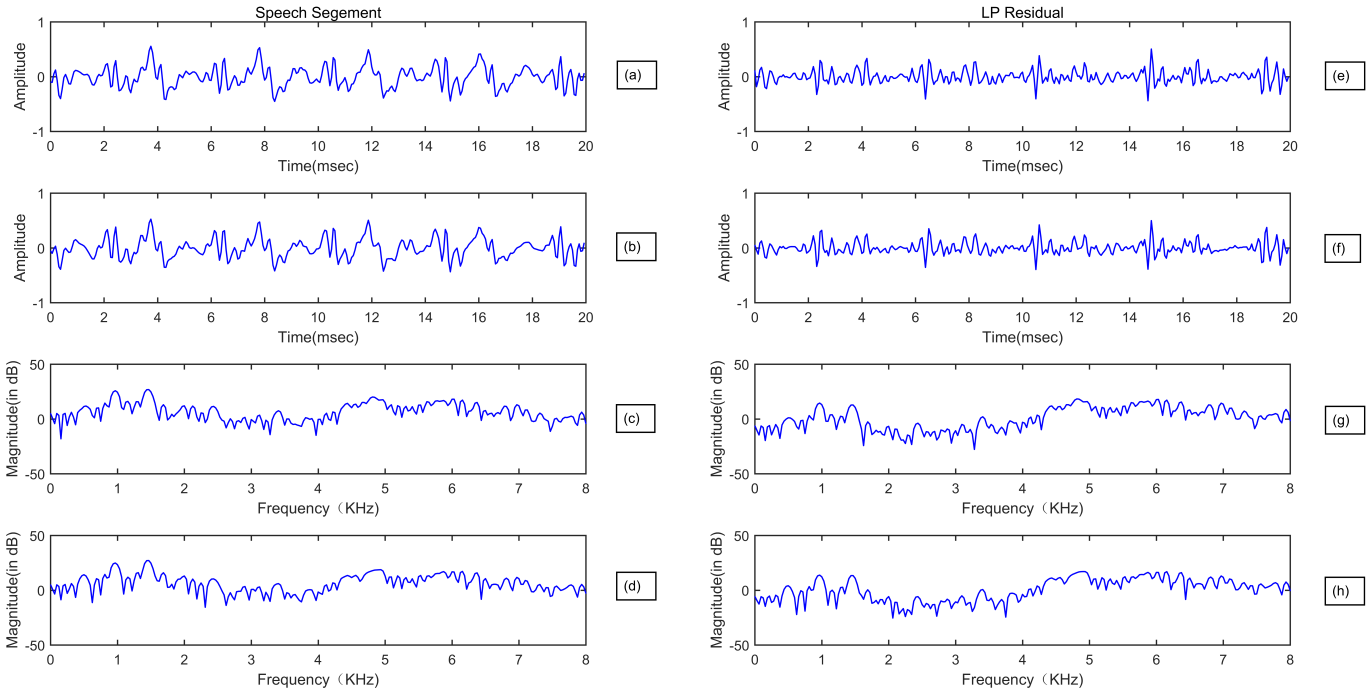

Figure 1. Time domain representation of a bonafide (**a**) and corresponding replay (**b**) speech segments, and their corresponding LP residuals (**e**,**f**) for a female speaker. Frequency domain representation of a bonafide (**c**) and corresponding replay (**d**) speech segments, and their corresponding LP residuals spectrum (**g**,**h**) for a female speaker. The LP order = 12.

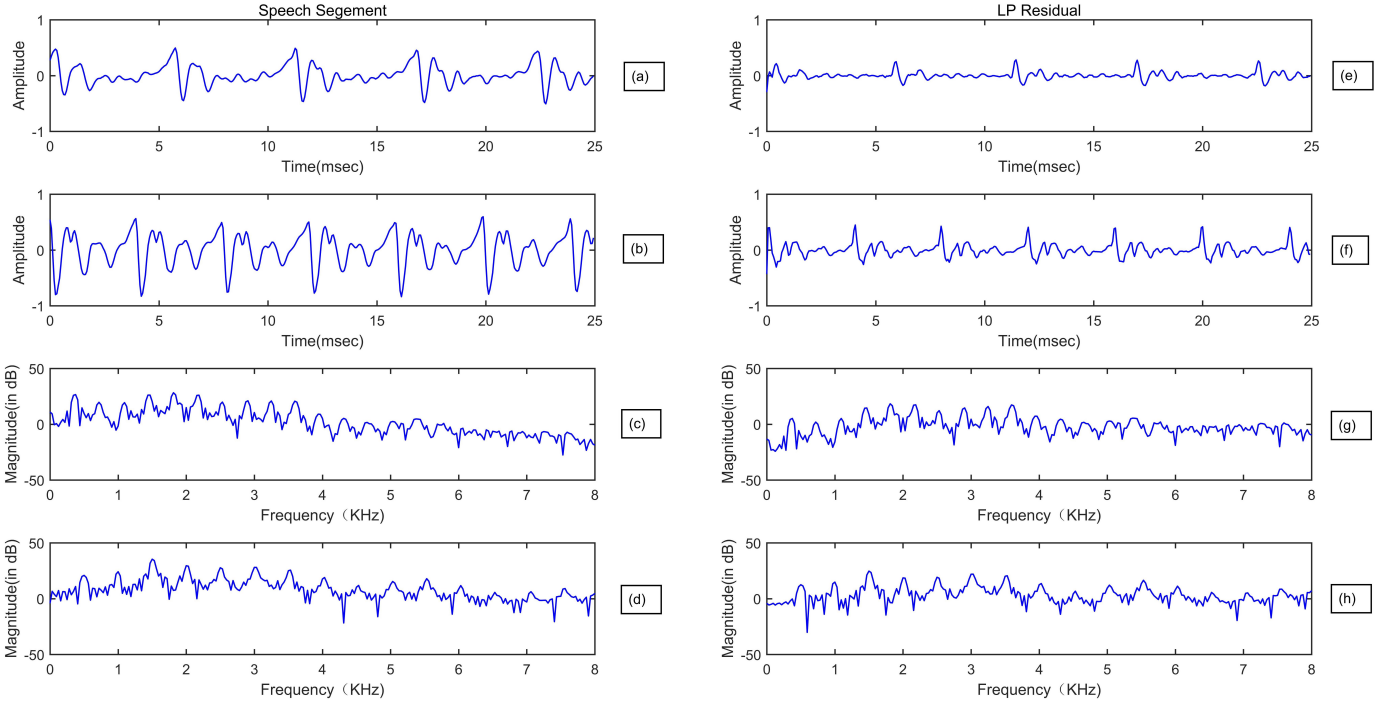

Figure 2. Time domain representation of a bonafide (**a**) and corresponding synthesized (**b**) speech segments, and their corresponding LP residuals (**e**,**f**) for a female speaker. Frequency domain representation of a bonafide (**c**) and corresponding synthesized (**d**) speech segments, and their corresponding LP residuals spectrum (**g**,**h**) for a female speaker The LP order = 12.

3.2. Conventional Cepstral Features

The constant Q cepstral coefficients (CQCCs) [9] and linear frequency cepstral coefficients (LFCCs) [30] are two types of effective conventional acoustic features used for anti-spoofing speech detection in ASV tasks. Both of them have been widely used in previous ASVSpoof Challenges, such as ASVSpoof 2015, 2017, etc. A brief description of these two features is presented in the next subsections.

3.2.1. CQCC Features

The extraction of CQCCs can be summarized in Figure 3. First, we calculate the constant-Q transform (CQT) $X^{CQ}(k,n)$ of the input discrete time-domain signal $x(n)$ as:

$$X^{CQ}(k,n) = \sum_{j=n-\frac{N_k}{2}}^{n+\frac{N_k}{2}} x(j)a_k^*(j-n+\frac{N_k}{2}) \tag{8}$$

where $k = 1, 2, \ldots, K$ is the frequency bin index, $a_k^*(n)$ is the complex conjugate of $a_k(n)$ and N_k are variable window lengths. $a_k(n)$ are cinolex-valued time-frquency atoms, the definition of $a_k(n)$ can be found from paper [9]. The notation $\lfloor . \rfloor$ infers rounding down towards the nearest integer. Unlike the regularly spaced frequency bins used in the standard Short-Time Fourier Transform (STFT), the CQT uses geometrically spaced frequency bins. This makes it offers a higher frequency resolution at lower frequencies and higher temporal resolution at higher frequencies.

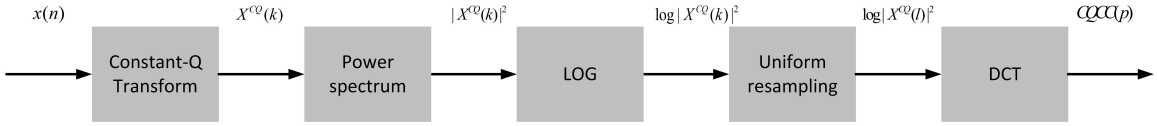

Figure 3. Block diagram of CQCC feature extraction.

Then, after the CQT, the log power spectrum of $X^{CQ}(k,n)$ is computed before performing a uniform re-sampling. As cepstral analysis cannot be applied directly to the CQT, due to the fact that the frequency bins are on a different scale to those of the basic functions of the discrete cosine transform (DCT). By applying the uniform re-sampling, the geometric space is then converted into the linear space for conventional cepstral analysis. Finally, in the linear space, the DCT is applied to get the CQCCs as

$$CQCC(p) = \sum_{l=1}^{L} \log|X^{CQ}(l)|^2 cos[\frac{p(l-\frac{1}{2})\pi}{L}] \tag{9}$$

where $p = 0, 1, \ldots, L-1$, and where l is the newly re-sampled frequency bins. More details of the CQCCs extraction, please find in [9].

3.2.2. LFCC

Like mel−frequency cepstral coefficients (MFCCs), Linear frequency cepstral coefficients (LFCCs) [30] are extracted the same way but the filters are triangular and spaced in linear scale as illustrated in Figure 4. The power spectrum is first integrated using overlapping band-pass filters and logarithmic compression followed by DCT is performed to produce the cepstral coefficients.

3.3. Residual Cepstral Coefficients

Motivated by the discussion on the potential effectiveness of residual signal for detecting the replayed and synthesized speech in Section 3.1, in this section, we investigate to extract the conventional acoustic features, CQCCs, and LFCCs from the LP residual signals instead of the raw audio. These new features are termed residual CQCC (RCQCC) and residual LFCC (RLFCC), respectively.

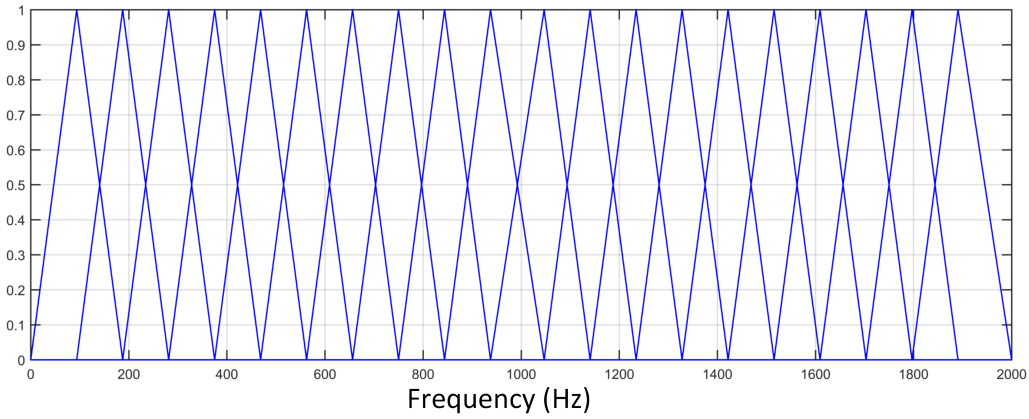

Figure 4. Filter bank used in the computation of LFCC.

Figure 5 shows the detailed block diagram of RCQCC feature extraction. Given a speech segment $s(n)$, it is the first overlap segmented into short-time frames using Hamming window. Then we perform the p-th order linear prediction analysis and inverse filtering to obtain the residual signal $r(n)$ frame-by-frame. These frames are then taken as the input signal to extract D-dimensional ($D = 30$) CQCC feature as shown in Figure 3, followed by a $\Delta + \Delta\Delta$ operation to achieve $2D$-dimensional dynamic features. These $3D$-dimensional features are concatenated together to form the final RCQCC acoustic features to train the spoofing detection classifiers.

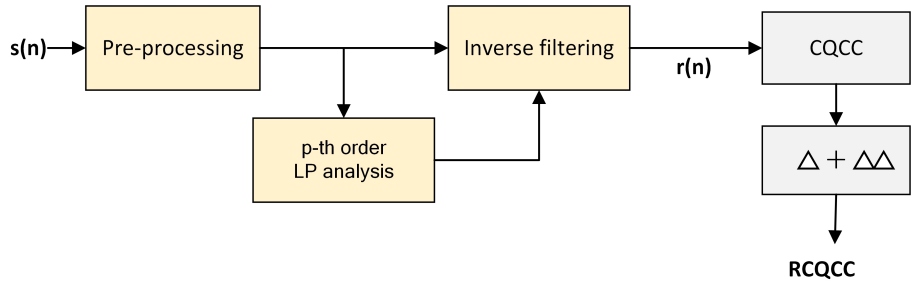

Figure 5. Block diagram of RCQCC feature extraction.

Figure 6 demonstrates the extraction diagram of RLFCC features. The frame-by-frame residual signal $r(n)$ is obtained in the same way as in RCQCCs. Then the DFT is used to transform the time-domain residual signal to the spectrum domain. After performing the linear-frequency scale uniform triangular band-pass filter banks on the power spectrum, the discrete cosine transform (DCT) is applied on the logarithm of the subband energies obtained from the triangular-filter banks to get the LFCC features. And as the RCQCCs, we also extract the $\Delta + \Delta\Delta$ dynamic features to form the final RLFCC features.

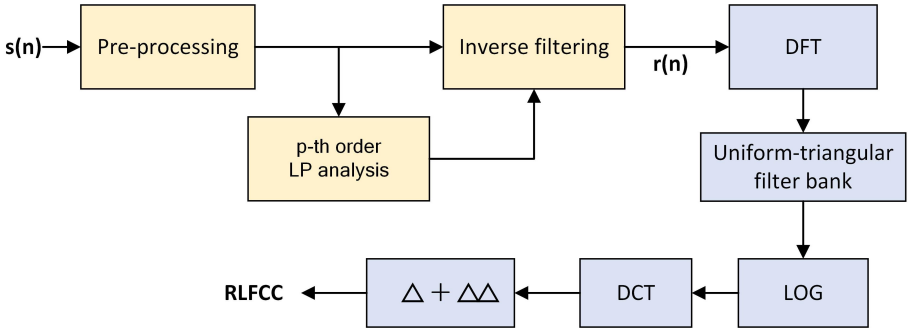

Figure 6. Block diagram of RLFCC feature extraction.

3.4. t-SNE Feature Visualization

Before using the above proposed residual features to build ASVSpoof systems, in this section, the discrimination ability between the genuine and spoof speech segments are investigated, using the t-stochastic neighborhood embedding (t-SNE) [31] visualization.

Figures 7 and 8 demonstrate the effectiveness of the proposed residual acoustic features RCQCCs and RLFCCs over the standard CQCCs and LFCCs, under the logical access and physical access scenarios, respectively. In both t-SNE figures, the two-dimensional visualization of LFCCs and RLFCCs are transformed from the 60-dimensional raw feature (including $\Delta + \Delta\Delta$) as described in Sections 3.2.2 and 3.3; while the two-dimensional visualization of CQCCs and RCQCCs are transformed from the 90-dimensional raw features (including $\Delta + \Delta\Delta$).

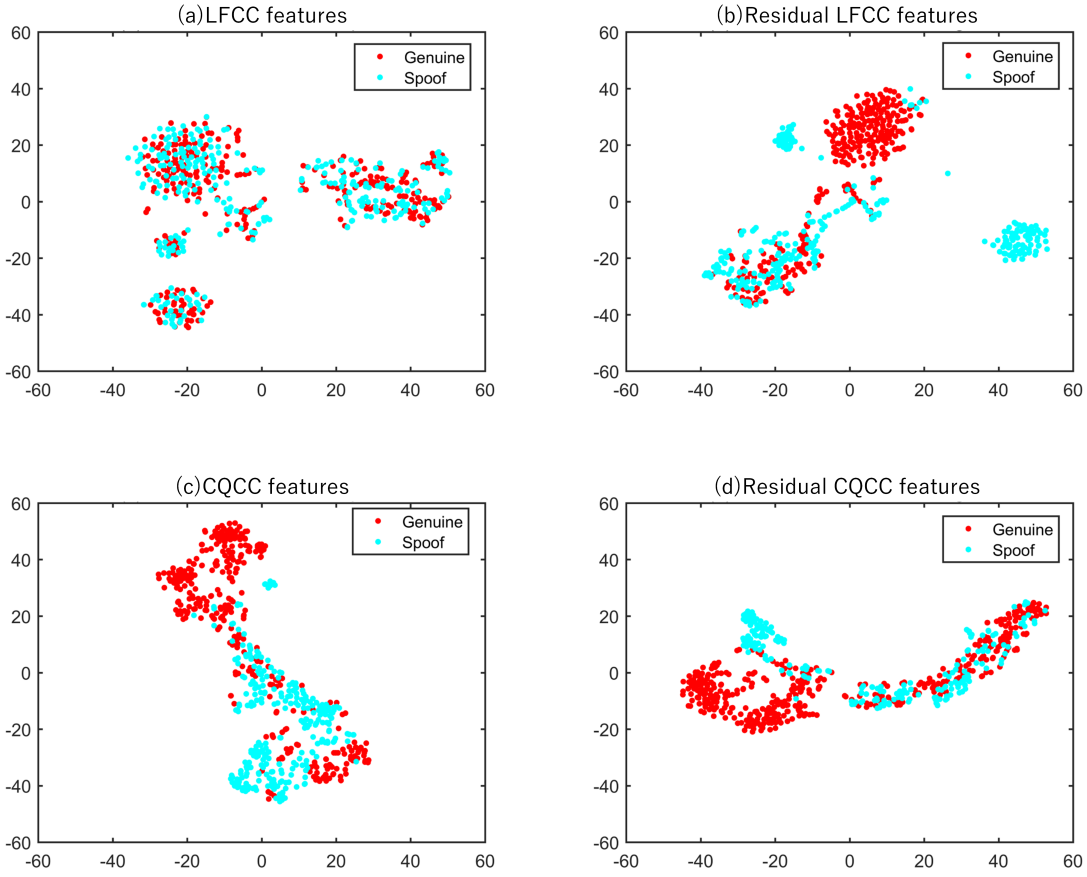

Figure 7. t-SNE visualization of (**a**) LFCCs (**b**) RLFCCs and (**c**) CQCCs (**d**) RCQCCs of the genuine and spoof speech segments, under the logical access scenario.

From both Figures 7 and 8, we see that there is a significant overlap between the original genuine and spoof speech segments, either in the CQCC (subfigure (c)) or the LFCC (subfigure (a)) features spaces, and whatever is under the LA or the PA scenarios. However, from sub-figure (b) and (d), the acoustic features of residual signals of genuine and spoof speech are clearly separated, especially for the RCQCCs and RLFCCs of examples under the PA scenario. It tells us that the acoustic feature discrimination between the residual signal of bonafide and spoof speech is much larger than that in the original/raw speech signals. Therefore, using RLFCCs and RCQCCs can build better ASVSpoof systems.

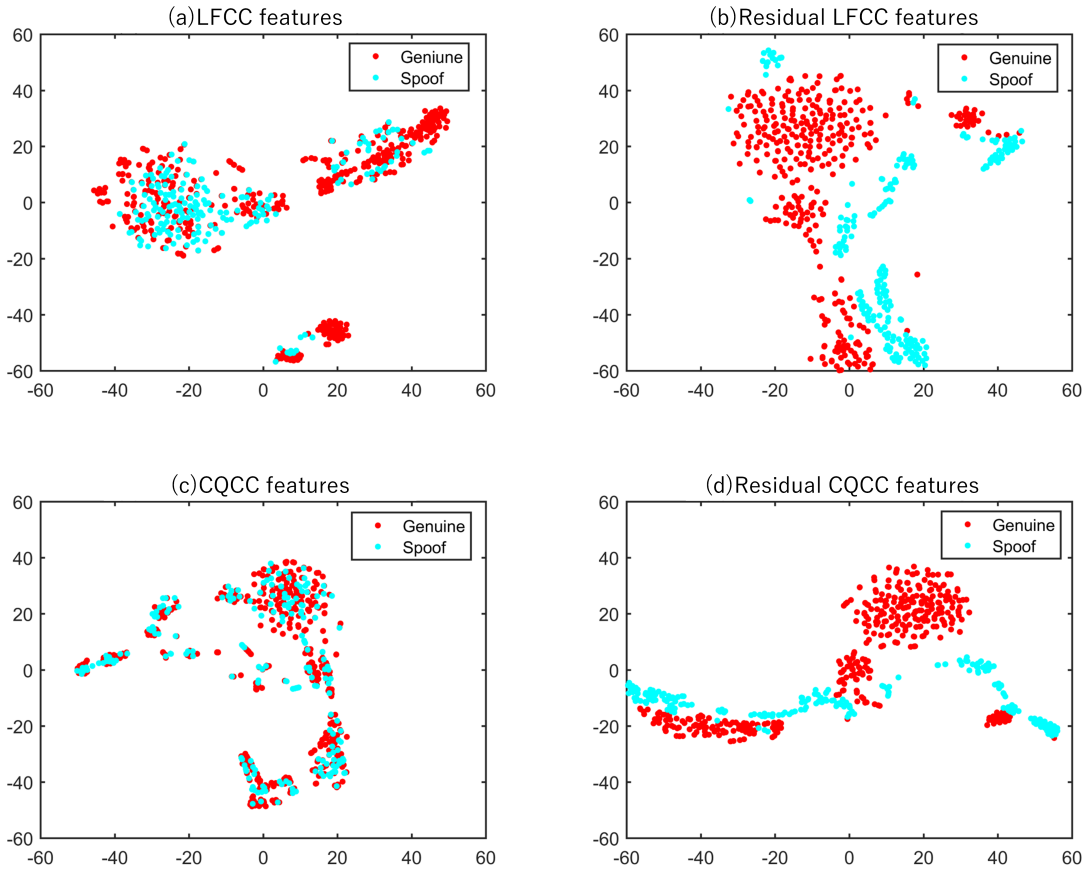

Figure 8. t-SNE visualization of (**a**) LFCCs (**b**) RLFCCs and (**c**) CQCCs (**d**) RCQCCs of the genuine and spoof speech segments, under the physical access scenario.

4. Harmonic and Noise Interaction Features

From the speech production mechanism and the success of Harmonic plus Noise Model (HNM) [4] for high-quality speech analysis and synthesis, we know that the generation of speech can be regarded as the interaction movement of the vocal tract and glottal airflow. Our previous work [32] has proposed a new feature called the spectral subband energy ratios (SSER) to reflect the interaction property, and experimental results have proved its effectiveness to characterize the speaker identity. Therefore, we think that the natural interaction property in the synthesized or replayed spoofing speech may be distorted by the speech synthesis algorithms or the speech recording devices, etc. These distortions may result in very different interaction features between the bonafide and spoofed speech. Therefore, in this section, we try to explore the effectiveness of the interactive features for the ASVSpoof task.

As proposed in our previous work of [32], here, we also use the spectral subband energy ratios as the interactive features but extract them differently as in [32]. So, in order to distinguish these features from SSER, we call them "Harmonic and Noise Subband Ratio (HNSR)". The principle of the HNSR is shown in Figure 9.

Given a speech segment $s(t)$, we first estimate the fundamental frequency (F0) and make the unvoiced/voiced decision using the Normalized Cross-Correlation Function (NCCF). Both F0 and NCCF are estimated using the Kaldi toolkit [33]. The pitch-markers are computed as in [34,35]. All the unvoiced frames are discarded, only voiced frames are used to extract the HNSRs. The pitch-synchronous frames are then extracted by using Hamming windows of two pitch period, and centered at each pitch marker. Finally, the harmonic component $h(t)$ is then estimated with the HNM analysis [4]. Instead of using the high-pass filtering method to get the stochastic noise part as in [4] for speech

synthesis, in this study, we choose to directly subtract the $h(t)$ by the original windowed speech signal $s(t)$ to achieve the $n(t)$.

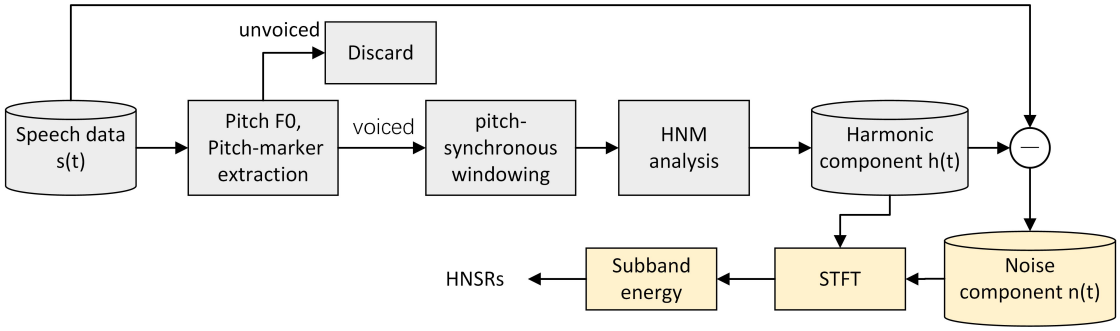

Figure 9. HNSR feature extraction.

Once the $h(t)$ and $n(t)$ are available, we then transform them into the spectrum domain and compute the frequency subband energies as

$$E_h^b = \sum_{k=B_s}^{B_e} |STFT\{h(t)\}|_k^2 \tag{10}$$

$$E_n^b = \sum_{k=B_s}^{B_e} |STFT\{n(t)\}|_k^2 \tag{11}$$

where $STFT$ is the short-time Fourier transform, Bs and Be represent the starting and ending frequencies of the spectral subband. For each frequency subband, we set the bandwidth $Bw = B_e - B_s$ to the averaged $F0$ (225 Hz) value of all the training datasets to have as many spectral subbands as possible. The maximum voiced frequency used in HNM analysis is fixed to 8 kHz. So, we can obtain a 35-dimensional HNSR feature vector for each voiced frame as

$$HNSR_{feature} = 10 * \ln(\frac{E_h^b}{E_n^b}) \tag{12}$$

As the RCQCC and RLFCC visualization, we also apply the t-SNE to see the distribution of HNSR features as shown in Figure 10. Compared with Figures 7 and 8, the discrimination of HNSRs is much poor. It indicates that the HNSR features can not well distinguish the genuine and spoof speech, but in this study, we can verify whether there is complementarity information between HNSR features and other acoustic features.

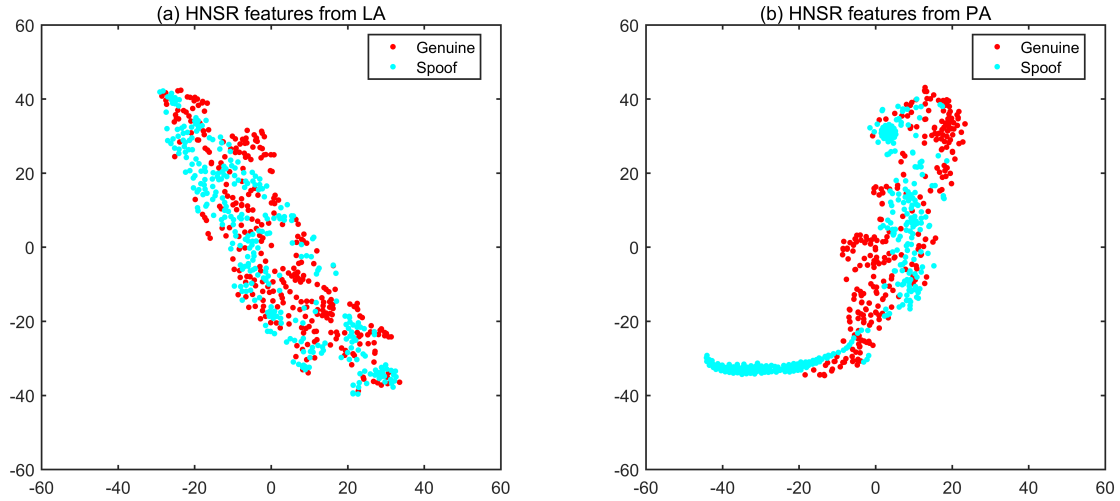

Figure 10. HNSR t-SNE plots under logical access (**a**) and physical access (**b**) scenarios.

5. Experiments and Results

5.1. Dataset-ASVSpoof 2019

All of our experiments are performed on 2019 Automatic Speaker Verification Spoofing and Countermeasures (ASVSpoof) challenge [2]. Different from previous ASVSpoof challenges, the 2019 year challenge is the first to focus on countermeasures for all three major attack types, namely those stemming from TTS, VC, and replay spoofing attacks. It contains two sub-challenges, namely the logical access (LA) task, and physical access (PA) task. Brief descriptions [20] of these two tasks are as follows.

Logical access task: Compared with the ten spoofing types in ASVSpoof 2015 [5], the spoofing utterances of LA sub-challenge in the ASVSpoof 2019 were synthetic speech using the most recent technology [10]. The quality of the synthetic speech from the ASVSpoof 2019 has improved a lot, which poses substantial threats to ASV. This sub-challenge contained training, development, and evaluation partitions. The genuine speech was collected from 107 speakers with no significant channel or background noise. The spoofed speech was generated from the genuine data using various spoofing approaches. No speaker overlap among the three subsets.

Physical access task: The speech of replay attacks in the ASVSpoof2019 was based upon simulated and carefully controlled environments. The training and development data of the PA sub-challenge was created according to 27 different acoustic environments, consisting of 3 room sizes, 3 levels of reverberation, and 3 microphone distances. There were 9 different replay configurations generated by 3 categories of recording distances, and 3 audio qualities. Detailed information about the training, development, and evaluation sets of ASVSpoof 2019 is illustrated in Table 1.

Table 1. Number of non-overlapping target speakers and the number of utterances in training, development, and evaluation sets of the ASVSpoof 2019 database.

Subset	Speakers		Utterances			
	Male	Female	Logical Access		Physical Access	
			Bonafide	Spoof	Bonafide	Spoof
Training	8	12	2580	22,800	5400	48,600
Development	4	6	2548	22,296	5400	24,300
Evaluation	21	27	7355	63,882	18,090	116,640

5.2. Experimental Configurations

5.2.1. Classifiers

Official baselines: Two official baseline countermeasure systems were made available to ASVSpoof 2019 participants. Both use a common Gaussian mixture model (GMM) back-end classifier with either constant-Q cepstral coefficient (CQCC) features [9] (B01) or linear frequency cepstral coefficient (LFCC) features [30] (B02). The GMMs for both B01 and B02 are with 512 components. The bonafide and spoofing GMM models are trained separately. Baselines are trained separately for LA and PA scenarios.

LCNN classifier: Besides the official GMM baselines, we also investigate using the Light Convolutional Neural Networks (LCNN) [18] as the classifier, which was the best system submitted to the ASVSpoof 2017 Challenge. As in [18], the same normalized log power magnitude spectrum (logspec) obtained by FFT is utilized as input of LCNN. To get a unified time-frequency (T-F), the shape of input features. the normalized FFT spectrograms along the time axis with the size of $864 \times 400 \times 1$ is truncated as the input of LCNN. Short files are extended by repeating their contents to keep the same in length. The LCNN classifier used in this paper is a reduced CNN architecture with Max-Feature Map activation (MFM), which provides feature selection function for LCNN. Detail about LCNN is the same as LCNN-9 that was used in work [18], with only a minor change of the outputs of FC6 layer, here, we use 256×2 instead of the 32×2.

5.2.2. Features

CQCC-based baselines use a constant-Q transform (CQT), which is applied with a maximum frequency of $f_{nyq} = fs/2$ (fs = 16 kHz); The minimum frequency is set to nine octaves below the maximum frequency $fmin = fmax/2^9 \cong 15$ Hz. The number of bins per octave is set to 96. The resulting geometrically-scaled CQT spectrogram is re-sampled to a linear scale using a sampling period of 16. The 29 + 0th order DCT static coefficients plus its $\Delta + \Delta\Delta$ dynamic coefficients that are computed using two adjacent frames are taken as our final CQCC features.

LFCC-based baselines use a short-term Fourier transform. Each frame is 20ms with Hamming window and 10 ms shift. The power magnitude spectrum of each frame is calculated using a 512-point FFT. A triangular, linearly spaced filter-bank of 20 channels is applied. Different from the CQCCs, only 19 + 0th order DCT static coefficients are extracted, plus its $\Delta + \Delta\Delta$, the final LFCCs are 60-dimensional acoustic feature vectors.

Table 2 presents all other detail parameters that are not mentioned in the baseline features or in Section 3. These parameters can generate the best performance for this experiment. "FL" and "FS" refer to frame-length and frame-shift in milliseconds(ms), respectively, "LA/PA" means the parameters for LA and PA condition, LP-order is the LP order to obtain the residual signal, "GMM(LA/PA)" means the GMM components for LA and PA conditions, respectively. All these parameters are tuned on the development datasets of the ASVSpoof 2019 challenge.

Table 2. Configurations of the proposed features.

Features	FL (LA/PA)	FS (LA/PA)	LP-Order	GMM (LA/PA)
RLFCC	30/25	15/12.5	12	512/256
RCQCC	30/30	15/15	12	256/128
HNSR	25/25	12.5/12.5	-	512/256
RLogSpec	25/20	10/10	12	-

5.3. Evaluation Metrics

Besides the equal error rate (EER) that was used for evaluating previous ASVSpoof challenge systems, in ASVSpoof 2019, a new ASV-centric metric referred to as the tandem detection cost function (t-DCF) [36] was first proposed as the primary evaluation metric. Use of the t-DCF means that the ASVSpoof 2019 database is designed not for the standalone assessment of spoofing countermeasures but their impact on the reliability of an ASV system when subjected to spoofing attacks. In this study, we use both the EER and t-DCF to evaluate the effectiveness of our proposed methods.

5.4. Results on LA Task

As presented in Section 5.2.1, two classifiers are used to build ASVSpoof detection systems, one is the GMM, and the other is the LCNN, either for the LA task or for the PA task. We first evaluate our proposed features for the LA task as follows.

5.4.1. Detection Results of Single Systems

Results for the LA task of ASVSpoof 2019 are shown in Table 3. `L1` and `L3` are the official GMM-based baselines, and `L6` is our LCNN baseline. By comparing `L2` with `L1`, the proposed residual CQCCs achieve similar results with the CQCCs on the development set, but with around absolute 2.14% worse EER than the CQCC baseline on the evaluation set. However, when we compare `L4` with its baseline `L3`, significant performance improvements are obtained, either in EER or in t-DCF for both the development and evaluation sets, such as on the development set, the EER is reduced from 2.72% to 1.16%, and t-DCF is reduced from 0.0663 to 0.0358, and on the evaluation set, consistent performance gains are achieved. Relative equal error rate (EER) improvements of 57.2%, 24.8% are achieved for development and evaluation set over the official baseline, respectively. It tells us that the RLFCC features extracted from the residual signal are much effective than the LFCCs

extracted from the original source signal. Moreover, from L5, it is clear that the proposed HNSR features are much worse than other features with GMM classifier, however, we hope that it may provide some complementary information to other acoustic features during the system fusion, because the HNSRs, RCQCCs, RLFCCs are extracted in a totally different way, they may capture the spoofing acoustic characteristics of speech synthesis, voice conversion in different aspects.

Table 3. The t-DCF and EER (%) results for the LA task on the development and evaluation sets. The lower the values of t-DCT and EER, the better is the performance.

ID	Classifier	Features	Dev.		Eval.	
			EER	**t-DCF**	**EER**	**t-DCF**
L1	GMM	CQCC	0.43	0.0123	9.57	0.2359
L2		**RCQCC**	0.40	0.0108	11.71	0.2817
L3		LFCC	2.71	0.0663	9.09	0.2115
L4		**RLFCC**	1.17	0.0358	6.84	0.1855
L5		**HNSR**	13.34	0.3064	26.41	0.6403
L6	LCNN	LogSpec	0.20	0.0059	20.83	0.2971
L7		**RLogSpec**	0.86	0.0286	10.07	0.1925

Furthermore, we also investigate extracting LogSpec features from residual signals to see its effectiveness for detecting spoofing speech. Comparing system L7 with its baseline L6, it is interesting to find that the EER and t-DCF on the evaluation set are significantly reduced, but the gains on the development set are exactly the opposite. We also hope the proposed RLogSpec can provide some complementary information for the LCNN-based countermeasures. This experiment shows the efficiency of residual signals compared to their respective baseline features for the ASVSpoof LA task. However, the performance of RCQCCs and RLFCCs for the LCNN model and RLogSpecs for the GMM model are missing. The efficiency of the proposed features for other ASVSpoof LA tasks needs further evaluation in the future.

5.4.2. System Fusion

To verify whether the different acoustic features are complementary, we perform system fusion on the score level instead of the feature level. The bosaris toolkit [37] is used. This toolkit provides a logistic regression solution, which can perform the fusion by learning the weights from scores of the development set and evaluation set. Results for the LA task of the ASVSpoof 2019 Challenge are shown in Table 4. In this table, we investigate many system fusion strategies to exploit complementary information of the acoustic features. By comparing these results with those ones in Table 3, it is clear to see that, score-level fusion can result in significant performance gains than single systems, either for the CQCCs, LFCCs, or the RCQCCs and RLFCCs. Considering both the performances on development and evaluation sets, we see the LF1 + RLFCC achieves the best fusion results on the GMM-based systems.

Furthermore, by comparing LF7 with L6 in Table 3, we see that adding residual LogSpec features can provide significant complementary information to the original LogSpec features, e.g., the EERs are reduced from 0.2%, 20.83% to 0.12%, 9.67% on the development and evaluation set, respectively. From LF8, we see further improvements by combining the best GMM-based system (LF5 + HNSR) with the LCNN-based system (LF7), the EER and t-DCF on the development set are reduced to zeros on the development set, and these numbers are also significantly reduced to 2.57% and 0.073931 on the evaluation test set. All these performance gains can prove that our provided residual acoustic features are useful and effective. From the big performance difference between the development

and evaluation set, it is clear that the generalization ability of current GMM-based and LCNN-based systems is limited.

Table 4. The score-level system fusion results (EER% and t-DCF) for the LA task on the development and evaluation sets.

ID	Classifier	Features	Dev.		Eval.	
			EER	t-DCF	EER	t-DCF
LF1	GMM	LFCC + CQCC	**0.04**	0.000449	5.08	0.130134
LF1-1		LF1 + **RLFCC**	**0.04**	0.000493	**4.42**	**0.118015**
LF1-2		LF1 + **RCQCC**	**0.04**	**0.000404**	5.11	0.130379
LF1-3		LF1 + **HNSR**	**0.04**	**0.000404**	5.05	0.132326
LF2		**RLFCC** + **RCQCC**	0.09	0.002972	6.46	0.168841
LF3		LFCC + **RLFCC**	1.01	0.032072	5.3	0.153161
LF4		CQCC + **RCQCC**	0.16	0.003633	5.92	0.15229
LF5		LF1 + LF2	**0.04**	0.000628	4.57	0.118691
LF5-1		LF5 + **HNSR**	**0.04**	0.000493	4.49	0.118943
LF6		LF3 + LF4	0.08	0.000897	4.53	0.119633
LF6-1		LF6 + **HNSR**	0.07	0.001121	4.5	0.120528
LF7	LCNN	LogSpec + RLogSpec	0.12	0.002854	9.67	0.180001
LF8	-	bestGMM + LF7	0	0	2.57	0.073931

5.5. Results on PA Task

As the above experimental investigations for the LA task, results for the PA task are as follows. The same acoustic features, system fusion strategies are validated.

5.5.1. Detection Results of Single Systems

Results for the PA task of ASVSpoof 2019 are shown in Table 5. As in Table 3, here, `P1` and `P3` are the official GMM-based baselines, and `P6` is the LCNN baseline. Results on the LA and PA tasks in Tables 3 and 5, show consistent performance on the behavior of residual acoustic features, such as, the RLFCCs achieve better results than LFCCs, while RCQCCs are not better or even worse than baselines, a single GMM system using HNSRs is still very bad, and the RLogSpec features are also very effective for the PA spoofing detection.

Table 5. The t-DCF and EER% results for the PA task on the development and evaluation sets. The lower the values of t-DCT and EER, the better is the performance.

ID	Classifier	Features	Dev.		Eval.	
			EER	t-DCF	EER	t-DCF
P1	GMM	CQCC	10.57	0.2103	12.64	0.2916
P2		**RCQCC**	16.96	0.3601	19.34	0.4463
P3		LFCC	9.97	0.2089	11.24	0.2494
P4		**RLFCC**	8.54	0.1921	10.45	0.2567
P5		**HNSR**	23.72	0.5839	29.54	0.6869
P6	LCNN	LogSpec	5.44	0.1795	6.59	0.1908
P7		RLogSpec	2.96	0.0913	4.25	0.1345

This experiment shows the efficiency of residual signals compared to their respective baseline features for the ASVSpoof PA task. But the performance of RCQCCs and RLFCCs for the LCNN model and RLogSpecs for the GMM model are missing. The efficiency of the proposed features for other ASVSpoof PA tasks needs further evaluation in the future.

5.5.2. System Fusion

Table 6 presents the score-level system fusion results for the PA task of the ASVspoof 2019 Challenge. As in Table 4, the same fusion strategies are examined, and not as LA task, the best fusion results are achieved from system PF6 + HNSR, whose EER and t-DCF are much lower than the official baseline system fusion (PF1) results, by introducing the complementary information of the proposed residual acoustic features RCQCCs and RLFCCs.

Table 6. The score-level system fusion results (EER% and t-DCF) for the PA task on the development and evaluation sets.

ID	Classifier	Features	Dev.		Eval.	
			EER	t-DCF	EER	t-DCF
PF1	GMM	LFCC + CQCC	9.22	0.180252	10.17	**0.234929**
PF1-1		PF1 + **RLFCC**	8.22	0.175117	**9.51**	0.239068
PF1-2		PF1 + **RCQCC**	9.26	0.181334	10.18	0.236621
PF1-3		PF1 + **HNSR**	9.02	0.176851	10.11	0.235087
PF2		**RLFCC** + **RCQCC**	8.29	0.185531	10.38	0.267975
PF3		LFCC + **RLFCC**	8.17	0.191823	9.96	0.256397
PF4		CQCC + **RCQCC**	10.56	0.218662	11.56	0.286208
PF5		PF1 + PF2	8.7	0.175911	9.55	0.238543
PF5-1		PF5 + **HNSR**	**8.63**	0.17278	9.54	0.238543
PF6		PF3 + PF4	8.73	0.172954	**9.51**	0.23673
PF6-1		PF6 + **HNSR**	7.97	**0.16983**	**9.51**	0.236967
PF7	LCNN	LogSpec + RLogSpec	2.69	0.084703	3.93	0.13176
PF8	-	bestGMM + PF7	1.85	0.058915	2.47	0.073727

For the LCNN systems, the fused system PF7 further reduces the EER from 5.44%, 6.59% to 2.69%, 3.39%, the t-DCF from 0.1795, 0.1908 to 0.104703, 0.13176 on the development and evaluation set, respectively. Moreover, system fusion between the best GMM-based system (PF6 + HNSR) and PF7 further improves the final system performances, e.g., PF8 achieves relative 31.2% and 37.1% EER reductions over PF7 on the development and evaluation sets, respectively. All these improvements are consistent on both the LA and PA tasks. And from the performance gap between development and evaluation sets in Tables 5 and 6, we also see that system fusion can reduce the performance gap between different test conditions, and improve the system robustness under PA task.

6. Conclusions

In this paper, we investigate two types of new acoustic features to improve the detection of synthetic and replay speech spoofing attacks. One is the residual CQCCs, LFCC, and LogSpecs, the other is the HNSR interaction feature. From the experimental results, we find that the acoustic features extracted from residual signals behave much better than those extracted from source speech signals. Although the single system built from HNSR features get bad results, they still show a little bit of complementary information at the score-level system fusion. Furthermore, we investigate different score-level fusion strategies and find that all of the proposed residual features can provide significantly complementary information to official baselines, and the GMM-based and LCNN-based systems have further complementary information (more than 30% relative EER reduction) to improve the final system performances.

Though the experiment results show the efficiency of proposed features, these features may not outperform all other acoustic features for the ASVSpoof task. Future works will focus on evaluating other acoustic features to conduct a comprehensive feature level comparison for the ASVSpoof task, and improving the generalization ability of anti-spoofing countermeasures under different conditions.

Author Contributions: Conceptualization, L.W. and Y.L. (Yanhua Long); methodology, L.W., Y.L. (Yanhua Long) and Y.L. (Yijie Li); software, L.W.; validation, Y.L. (Yanhua Long), H.W. and Y.L. (Yijie Li); formal analysis, L.W. and Y.L. (Yanhua Long); investigation, Y.L. (Yanhua Long), H.W. and Y.L. (Yijie Li); resources, Y.L. (Yanhua Long) and Y.L. (Yijie Li); data curation, L.W. and Y.L. (Yanhua Long); writing—original draft preparation, L.W. and Y.L. (Yanhua Long); writing—review and editing, Y.L. (Yanhua Long), H.W. and Y.L. (Yijie Li); visualization, L.W. and Y.L. (Yanhua Long); supervision, Y.L. (Yanhua Long) and Y.L. (Yijie Li); project administration, Y.L. (Yanhua Long) and Y.L. (Yijie Li); funding acquisition, Y.L. (Yanhua Long). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China grant No. 62071302 and 61701306.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Wu, Z.; Evans, N.; Kinnunen, T.; Yamagishi, J.; Alegre, F.; Li, H. Spoofing and countermeasures for speaker verification. *Speech Commun.* **2015**, *66*, 130–153. [CrossRef]
2. Wang, X.; Yamagishi, J.; Todisco, M.; Delgado, H.; Nautsch, A.; Evans, N.; Sahidullah, M.; Vestman, V.; Kinnunen, T.; Lee, K.A.; et al. ASVspoof 2019: A large-scale public database of synthetized, converted and replayed speech. *Comput. Speech Lang.* **2020**, *64*, 101114. [CrossRef]
3. Valizada, A.; Jafarova, S.; Sultanov, E.; Rustamov, S. Development and Evaluation of Speech Synthesis System Based on Deep Learning Models. *Symmetry* **2021**, *13*, 819. [CrossRef]
4. Stylianou, Y. Harmonic plus Noise Models for Speech, Combined with Statistical Methods, for Speech and Speaker Modification. Ph.D. Thesis, Ecole Nationale Superieure des Telecommunications, Palaiseau, France, 1996.
5. Wu, Z.; Kinnunen, T.; Evans, N.; Yamagishi, J.; Hanilçi, C.; Sahidullah, M.; Sizov, A. ASVspoof 2015: The first automatic speaker verification spoofing and countermeasures challenge. In Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech), Dresden, Germany, 6–10 September 2015; pp. 2037–2041.
6. Kinnunen, T.; Sahidullah, M.; Delgado, H.; Todisco, M.; Evans, N.; Yamagishi, J.; Lee, K.A. The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection. In Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech), Stockholm, Sweden, 20–24 August 2017; pp. 2–6.
7. Wang, L.; Yoshida, Y.; Kawakami, Y.; Nakagawa, S. Relative phase information for detecting human speech and spoofed speech. In Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech), Dresden, Germany, 6–10 September 2015; pp. 2092–2096.
8. Pal, M.; Paul, D.; Saha, G. Synthetic speech detection using fundamental frequency variation and spectral features. *Comput. Speech Lang.* **2018**, *48*, 31–50. [CrossRef]
9. Todisco, M.; Delgado, H.; Evans, N. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Comput. Speech Lang.* **2017**, *45*, 516–535. [CrossRef]
10. Todisco, M.; Wang, X.; Vestman, V.; Sahidullah, M.; Delgado, H.; Nautsch, A.; Yamagishi, J.; Evans, N.; Kinnunen, T.; Lee, K.A. ASVspoof 2019: Future Horizons in Spoofed and Fake Audio Detection. In Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech), Graz, Austria, 15–19 September 2019; pp. 1008–1012.
11. Alam, J. On the use of fisher vector encoding for voice spoofing detection. *Proceedings* **2019**, *31*, 37. [CrossRef]
12. Jelil, S.; Das, R.K.; Prasanna, S.M.; Sinha, R. Spoof detection using source, instantaneous frequency and cepstral features. In Proceedings of the Interspeech 2017, Stockholm, Sweden, 20–24 August 2017; pp. 22–26.
13. Villalba, J.; Miguel, A.; Ortega, A.; Lleida, E. Spoofing detection with DNN and one-class SVM for the ASVspoof 2015 challenge. In Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech), Dresden, Germany, 6–10 September 2015; pp. 2067–2071.
14. Nagarsheth, P.; Khoury, E.; Patil, K.; Garland, M. Replay attack detection using DNN for channel discrimination. In Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech), Stockholm, Sweden, 20–24 August 2017; pp. 97–101.
15. Yu, H.; Tan, Z.H.; Ma, Z.; Martin, R.; Guo, J. Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features. *IEEE Trans. Neural Networks Learn. Syst.* **2017**, *29*, 4633–4644. [CrossRef] [PubMed]
16. Zhang, C.; Yu, C.; Hansen, J. An investigation of deep-learning frameworks for speaker verification antispoofing. *IEEE J. Sel. Top. Signal Process.* **2017**, *11*, 684–694. [CrossRef]

17. Chen, Z.; Zhang, W.; Xie, Z.; Xu, X.; Chen, D. Recurrent neural networks for automatic replay spoofing attack detection. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 15–20 April 2018; pp. 2052–2056.
18. Lavrentyeva, G.; Novoselov, S.; Malykh, E.; Kozlov, A.; Kudashev, O.; Shchemelinin, V. Audio replay attack detection with deep learning frameworks. In Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech), Stockholm, Sweden, 20–24 August 2017; pp. 82–86.
19. Tak, H.; Patil, H.A. Novel linear frequency residual cepstral features for replay attack detection. In Proceedings of the Interspeech 2018, Hyderabad, India, 2–6 September 2018; pp. 726–730.
20. Feng, Z.; Tong, Q.; Long, Y.; Wei, S.; Yang, C.; Zhang, Q. SHNU Anti-spoofing Systems for ASVspoof 2019 Challenge. In Proceedings of the IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 548–552.
21. Witkowski, M.; Kacprzak, S.; Zelasko, P.; Kowalczyk, K.; Gałka, J. Audio replay attack detection using high-frequency features. In Proceedings of the 18th Annual Conference of the International Speech Communication Association (Interspeech), Stockholm, Sweden, 20–24 August 2017; pp. 27–31.
22. Cheng, X.; Xu, M.; Zheng, T. Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspoof 2019. In Proceedings of the IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 18–21 November 2019; pp. 540–545.
23. Singh, M.; Pati, D. Usefulness of linear prediction residual for replay attack detection. *Int. J. Electron. Commun.* **2019**, *110*, 152837. [CrossRef]
24. Yang, J.; Das, R. Low frequency frame-wise normalization over constant-Q transform for playback speech detetion. *Digit. Signal Process.* **2019**, *89*, 30–39. [CrossRef]
25. Tom, M.F.; Dey, P. End-to-end audio replay attack detection using deep convolutional networks with attention. In Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech), Graz, Austria, 15–19 September 2018; pp. 681–685.
26. Sriskandaraja, K.; Sethu, V.; Ambikairajah, E. Deep Siamese architecture based replay detection for secure voice biometric. In Proceedings of the 19th Annual Conference of the International Speech Communication Association (Interspeech), Graz, Austria, 15–19 September 2018; pp. 671–675.
27. Qian, Y.; Chen, N.; Yu, K. Deep features for automatic spoofing detection. *Speech Commun.* **2016**, *85*, 43–52. [CrossRef]
28. Wu, Z.; Xiao, X.; Chng, E.S.; Li, H. Synthetic speech detection using temporal modulation feature. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vancouver, BC, Canada, 26–31 May 2013; pp. 7234–7238.
29. Patel, T.; Patil, H. Cochlear filter and instantaneous frequency based features for spoofed speech detection. *IEEE J. Sel. Top. Signal Process.* **2016**, *11*, 618–631. [CrossRef]
30. Sahidullah, M.; Kinnunen, T.; Hanilci, C. A comparison of features for synthetic speech detection. In Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech), Dresden, Germany, 6–10 September 2015; pp. 2087–2091.
31. Maaten, L.v.d.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.
32. Long, Y.; Yan, Z.J.; Soong, F.K.; Dai, L.; Guo, W. Speaker characterization using spectral subband energy ratio based on harmonic plus noise model. In Proceedings of the 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Prague, Czech Republic, 22–27 May 2011; pp. 4520–4523.
33. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the 2011 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Waikoloa, HI, USA, 11–15 December 2011.
34. Rabiner, L.; Cheng, M.; Rosenberg, A.; McGonegal, C. A comparative performance study of several pitch detection algorithms. *IEEE Trans. Acoust. Speech Signal Process.* **2003**, *24*, 399–418. [CrossRef]
35. Reinier, W.; Kortekaas, L.; Kohlrausch, A. Psychoacoustical evaluation of the pitch-synchronous overlap-and-add speech-waveform manipulation technique using single-format stimuli. *J. Acoust. Soc. Am.* **1997**, *101*, 2202–2213.
36. Kinnunen, T.; Lee, K.A.; Delgado, H.; Evans, N.; Todisco, M.; Sahidullah, M.; Yamagishi, J.; Reynolds, D.A. t-DCF: A detection cost function for the tandem assessment of spoofing countermeasures and automatic speaker verification. *arXiv* **2018**, arXiv:1804.09618.
37. Brümmer, N.; De Villiers, E. The bosaris toolkit: Theory, algorithms and code for surviving the new dcf. *arXiv* **2013**, arXiv:1304.2865.