β-Order Energy-Weighting Modulation on Spectral Bins for Replay Speech Detection

Jichen Yang, *Senior Member, IEEE*, Chang Huai You, *Senior Member, IEEE*, Rohan Kumar Das, *Senior Member, IEEE*

***Abstract*—The usage of recording and playback devices for the generation of replay speech introduces some variation into replay speech. This kind of variation penetrates into the short-time details of the replay speech signal in a way of energy distribution. Together with the environmental noise during the process of recording, the disparity between replay speech and genuine speech is discernible. In this paper, in order to emphasize the disparity of the short-time energy distribution between replay speech and genuine speech, we propose a β-order energy-weighting (EW) method to modulate the spectral-energy into spectral bins in short time frame-wise duration. Specifically, we apply the β-order EW modulation in different spectral domains, i.e., octave spectral domain, Mel spectral domain and linear spectral domain. Subsequently, we propose three novel features: constant-Q EW octave coefficient (CEOC), Mel frequency EW cepstral coefficient (MFECC), and linear frequency EW cepstral coefficient (LFECC). Moreover, with the three proposed features, relevant replay speech detection (RSD) systems are developed with respective ResNet and DNN as backend classifiers. With experiments conducted on respective ASVspoof 2017 version 2.0, BTAS 2016 physical access, ASVspoof 2019 physical access and ASVspoof 2021 physical access corpora, the effectiveness of the three proposed features is evaluated in terms of the RSD performance. The evaluation results demonstrate that the β-order EW modulation exhibits superior discriminative capability in detecting replay speech. Furthermore, through the experiments, it can be observed that our proposed RSD systems outperform many state-of-the-art RSD systems.**

***Index Terms*—β-order energy-weighting, replay speech detection, CEOC, MFECC, LFECC**

I. Introduction

Spoofing detection is an essential aspect of ensuring the security and integrity of automatic speaker verification (ASV) systems, particularly in the context of data transmission and network communication. Over the past decades, there has been significant interest in ASV systems [1]–[4]. These systems have steadily matured and found their places in commercial product applications due to their exceptional convenience and touchless operability [5]–[7]. Most of these ASV systems are typically trained using genuine speech data, making them vulnerable to attacks in the shape of various spoofing speech techniques [8], [9].

Spoofing attacks against ASV systems mainly arise from replay speech [10], [11], text-to-speech (TTS) [12]–[14], voice conversion (VC) [15]–[17] and impersonation [18], [19]. Of these, impersonation presents a mild danger since it primarily revolves around mimicking the speaker's identity through specific intonations and behavioral signals, which are not commonly replicated. On the contrary, computer-generated speech, e.g. TTS generated speech or VC speech, is typically created using complex algorithms, which can be widespread replicated. Although machine learning methods can often distinguish between computer-generated and human-generated elements, high-quality computer-generated speech can still pose detection challenges, especially as spoofing technologies evolve. However, replay speech is obviously difficult to be detected with anti-spoofing algorithms because the replay signal just originates from the genuine speech itself; and also it can be replicated effortlessly by using recording and playback devices. By replaying the captured packets, the attacker aims to imitate one of the legitimate parties involved in the communication or gain unauthorized access to sensitive information. Moreover, due to the minimal distinction between authentic speech and replay speech, primarily resulting from the advancements in high-quality recording and playback devices, the task of detecting replay speech becomes increasingly challenging, and thus the development of high-performance countermeasures against replay speech detection (RSD) becomes imperative.

The most important technique for RSD concentrates at features that represent the information of spoofed components and genuine components. The features for RSD can be broadly categorized into two types: deep features [20]–[22] and handcrafted features [23]–[28]. On the one hand, deep feature lacks formal expressions and interpretability. The nature of deep feature is inherently shaped by the training data through training process driven by deep learning algorithms. Some prior work on deep feature extraction for RSD takes the advantage of convolutional neural network (CNN) or deep neural network (DNN) to learn deep features from the input of log power spectrum based on fast Fourier transform [20], [29]. Deep feature extractor can automatically learn highly discriminative representations from raw speech signals. The deep learning function makes deep feature effectively capture both low-level acoustic characteristics and high-level semantic information, and therefore enhances its ability to differentiate

This work was in part by key construction discipline project for research capability improvement of Guangdong Provincial Department of Education (2024ZDJS025), special projects in key areas of Guangdong Provincial Department of Education (2023ZDZX1006) and the Science and Technology Program (Key R&D Program) of Guangzhou, China (2023B01J0004)). (Both Jichen Yang and Chang Huai You are co-first authors.) (Corresponding author: Chang Huai You)

Jichen Yang is with the School of Cyber Security, Guangdong Polytechnic Normal University, Guangzhou, China. (e-mail: nisonyoung@163.com).

Chang Huai You is with the Institute for Infocomm Research, A∗STAR, Singapore. (e-mail: echyou@i2r.a-star.edu.sg).

Rohan Kumar Das is with Fortemedia Singapore, Singapore. (e-mail: ecerohan@gmail.com).

between genuine and spoofed speech. While deep feature has demonstrated superior performance, particularly when the training data is sufficiently representative, it is necessary to acknowledge that the effectiveness of deep feature heavily relies on the characteristics of the training data. Also, deep features are designed to capture intrinsic properties of the input signal, which can make them robust to variations between training and evaluation conditions. The complexity increases when conducting experiments with multiple deep feature extractors, as each one needs to be trained using different databases. In contrast, this complexity does not arise when utilizing handcrafted features. On the other hand, handcrafted feature can be explicitly represented by formulas [23]–[25], [30]. It can incorporate domain-specific knowledge and insights that might not be easily captured by deep learning models. In certain domains where expert knowledge is crucial, handcrafted feature can provide an interpretable and stable representation.

Due to the benefit of long-term constant-Q transform (CQT), the constant-Q cepstral coefficients (CQCC) [26], [27] has been the most popular feature for replay attack detection [31]–[34]. Especially, both ASVspoof 2017 and ASVspoof 2019 physical access challenge edition consider CQCC as baseline frontend. Compared with the representatives of short-term transform using discrete Fourier transform (DFT), such as linear filterbank slope [25] and power normalized cepstral coefficients (PNCC) [35], CQCC can capture more recording/playback information because characteristics of CQT emphasizes the higher frequency resolution at lower frequencies and higher temporal resolution at higher frequencies [26], [27]. The characteristics of CQT intensifies the difference of acoustic properties between the device informational component and human speech component, numerous features such as constant-Q statistics-plus-principal information coefficients [36] and extended CQCC (eCQCC) [37] have been derived based on the CQT.

Besides CQT related approaches, handcrafted feature extraction for RSD often adopt the discrete cosine transform (DCT) to extract the principal spectral information by consolidating signal energy into a selected set of coefficients while discarding residual information. Examples of such techniques include Mel frequency cepstral coefficients (MFCC) and linear frequency cepstral coefficients (LFCC). However, our previous investigation [37] demonstrates that DCT fails to effectively capture the information contained in the octave power spectrum. The reason is that some useful information in the residual part has been discarded. To tackle this issue, the concept of octave subband transform (OST) was introduced in [38]. The fundamental principle behind OST involves dividing the octave power spectrum into segments based on octaves, followed by the application of DCT to extract essential data from each individual octave. The results demonstrate the efficacy of OST in capturing octave power spectrum information. Consequently, we will further investigate the utilization of OST for octave power spectrum, DCT for Mel power spectrum, and DCT for linear power spectrum respectively in this paper.

Replaying device introduces the disparity between spoofed speech and genuine speech in a form of energy distributed over diverse spectral spaces. The disparity in energy information may serve as a crucial indicator for RSD [39]. This disparity has prompted investigations into the potential of energy as a reliable indicator of spoofing. One exemplified feature that illustrates energy utilization is the CQCC-E feature [34], which is derived by concatenating CQCC and energy.

Analysing the usage of energy information in CQCC-E [34], we provide some observations as follows: First, CQCC-E uses energy as additional information to form a concatenated feature vector. Second, although the energy makes the concatenated features (CQCC-E) more discriminative capability to detect replay speech than the primary CQCC [34], [40], the discriminative characteristics of the original frequency bins keeps unchange.

A noticeable difference in the energy patterns between genuine speech and the corresponding replay speech can be observed. Also, utilizing recording and playback devices in various environments adds some level of noise in the form of additive or convolutional way inevitably into the replay speech. Obviously, the variation brought to replay speech penetrates into the short-time details of the replay speech signal in a way of energy. We believe that the discriminative capability of the frequency bins can be improved if energy is used as the weight of the frequency bins. Therefore, to emphasize the variation in short-time energy between replayed and genuine speech, we propose a β-order energy weighting (EW) modulation on spectral bins for RSD, inspired by the concept of β-order minimum mean square error (MMSE) originally introduced in the context of speech enhancement [41]. Subsequently, we develop three novel features by applying β-order EW modulation in octave, Mel and linear spectral domains respectively. As a result, three spoofing detection systems with ResNet [42] as backend classifiers are developed as tools to rigorously evaluate the effectiveness of the β-order EW method. In order to verify the β-order EW, we developed another three spoofing detection systems with DNN [43] as backend classifier to further verify the validity of β-order EW method.

The remainder of the paper is organized as follows. Section II introduces the proposed β-order EW modulation and the three proposed features for RSD. The experimental evaluation are reported in Section III. We gives discussion on utility of energy in section IV. Finally, Section V concludes this work.

II. Proposed β-order energy-weighting modulation and novel features for RSD systems

There exists a discernible distinction in energy levels between genuine speech and corresponding replay speech to varying degrees. In this section, we propose a β-order EW modulation to discriminate the genuine speech and replay speech. Moreover, we explore three popular feature extraction techniques used for spoofing and replay detection: CQ-OST, MFCC, and LFCC. While CQ-OST has been proven effective in octave spectral domain, MFCC and LFCC are widely employed in the Mel spectral domain and linear spectral domain, respectively. On the basis of the three traditional features, we propose three novel features by introducing the β-order EW modulation.

A. β-order energy weighting modulation

Assume that X represents a sequence of consecutive power spectral vector of an utterance regardless of genuine or replay speech, and $\{X_i\}_{i=1}^{I}$ denotes the frames of X, and I is the total number of frames, $\{x_{ik}\}_{k=1}^{K}$ are the power of the k-th spectral bin of X_i for i-th frame, $i = 1, ..., I$, and K is the number of spectral bins. Thus, we have X be written as:

$$X = \begin{Bmatrix} X_1 \\ ... \\ X_i \\ ... \\ X_K \end{Bmatrix} = \begin{Bmatrix} x_{11},..., & x_{1K} \\ ... & \\ x_{i1},..., & x_{iK} \\ ... & \\ x_{I1},..., & x_{IK} \end{Bmatrix} \tag{1}$$

A β-order EW quantity, E_{X_i}, of the i-th frame is defined as follows:

$$E_{X_i} = \sum_{k=1}^{K} {x_{ik}}^{\beta} \tag{2}$$

where value of β is in the range $(-\infty, \infty)$. Modulating the β-order EW quantity into the power spectral bin in Eq. (1), we propose a β-order EW modulated sequence Y as follows:

$$Y = \begin{Bmatrix} x_{11}\sum_{k=1}^{K} {x_{1k}}^{\beta},..., & x_{1K}\sum_{k=1}^{K} {x_{1k}}^{\beta} \\ ... & \\ x_{i1}\sum_{k=1}^{K} {x_{ik}}^{\beta},..., & x_{iK}\sum_{k=1}^{K} {x_{ik}}^{\beta} \\ ... & \\ x_{I1}\sum_{k=1}^{K} {x_{Ik}}^{\beta},..., & x_{IK}\sum_{k=1}^{K} {x_{Ik}}^{\beta} \end{Bmatrix} \tag{3}$$

It has been known that the logarithmical compression in speech spectral component is to simulate the human perception of loudness. It converts the linear magnitude values into a logarithmic scale, resulting in a more compact representation of the spectral information. Applying logarithm into each spectral feature vector in Y, we have

$$\log\left\{Y\right\} = \log\begin{Bmatrix} x_{11},..., & x_{1K} \\ ... & \\ x_{i1},..., & x_{iK} \\ ... & \\ x_{I1},..., & x_{IK} \end{Bmatrix} + \log\begin{Bmatrix} E_{X_1} \\ ... \\ E_{X_i} \\ ... \\ E_{X_I} \end{Bmatrix} \tag{4}$$

Eq. (4) indicates that the logarithm of Y is actually composed of two terms, term of log power spectrum (LPS) and term of β-order EW in log-scale. In other words, power spectral vector modulated by frame-level β-order spectral energy can be analyzed into the original log power spectral vector itself and the additive log energy weight via logarithm operator.

We therefore introduce a new feature framework with the modulation of β-order EW for RSD as follows: First, speech signal is processed as a sequence of overlapped short-time frames through certain spectral transformation in a particular spectral domain such as octave, Mel or linear spectral domain. Second, the power spectral vector at frame-level is obtained, it is followed by applying β-order EW modulation; as a result, the β-order EW power spectral vector is realized. Third, the proposed β-order EW feature is completed by applying logarithm and the post-transformation. The flowchart of the β-order EW feature extraction framework is depicted in Fig. 1.

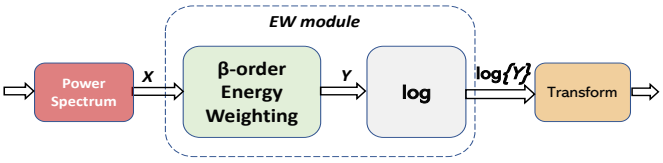

Fig. 1: The proposed framework of the β-order EW feature for RSD.

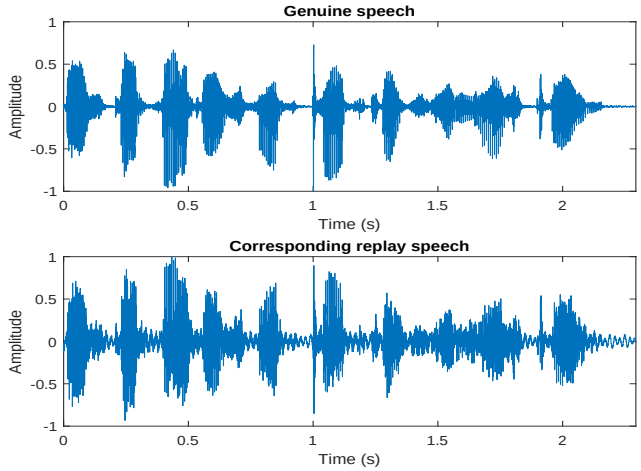

Fig. 2: Time-domain waveform comparison of genuine and replay speech. Differences in waveform characteristics, such as distortions and unnatural variations in replayed speech, are visually noticeable, highlighting the impact of recording and playback devices on the signal.

Replay attacks exploit genuine speech recordings to deceive speaker verification systems, presenting a substantial challenge because replayed speech retains many characteristics of authentic human speech. Unlike other spoofing techniques, such as TTS or VC speech, replayed speech contains real acoustic features that were captured during the initial recording. However, the processes of recording and playback inherently introduce subtle variations, especially in the short-time energy distribution of the speech. These variations arise due to factors like the quality of the recording device, environmental noise, and the playback medium. Although these energy distribution differences may be subtle, they can serve as crucial indicators for distinguishing between genuine and replayed speech.

Here, we provide some visual examples from the ASVspoof2017 physical access dataset to illustrate the difference between genuine and replayed utterances. Fig. 2 demon-

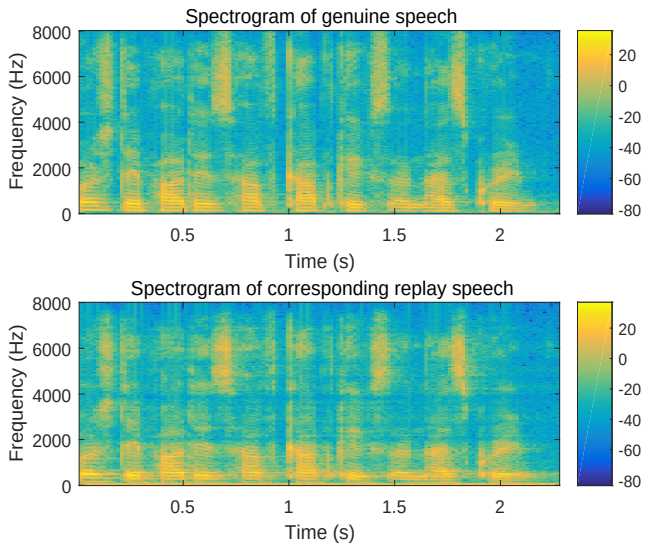

Fig. 3: Spectrogram comparison of genuine and replay Speech.

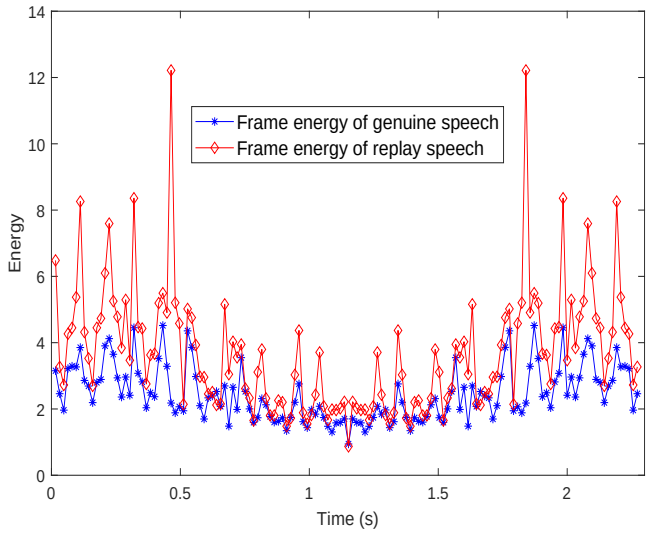

Fig. 4: Frame energy of genuine and replay speech over time. Replay speech exhibits abnormal energy distribution due to device-induced distortions, making it distinguishable from genuine speech.

strates what replay alters the natural smoothness and variation of the waveform; the replayed version exhibits a somewhat distorted waveform compared to the original genuine speech. This distortion arises because replay attacks introduce minor changes to the signal, such as added noise, device-induced alterations, and degradation from the recording and playback processes. Fig. 3 reveals differences in the frequency domain, where artifacts from replay attacks become apparent in specific frequency bands. It is observed that the signal of genuine speech changes naturally and unpredictably while two fixed marks at 4000 Hz and 2000 Hz obviously persist throughout the entire time domain in the replayed speech, reflecting the influence of the impulse response of replay device system. In fact, there are numerous such marks that go unnoticed by human listeners; however, advanced machine learning techniques are capable of detecting the subtle patterns introduced by devices.

The fundamental idea behind the β-order EW method is to enhance the short-time energy variations between replayed and genuine speech. Fig. 4 compares the frame energy of genuine and corresponding replayed speech over time, showing what replay attacks introduce subtle yet detectable fluctuations in the energy profile. The energy levels in replayed speech appear high dynamic, exhibiting unnatural peaks or troughs caused by the recording device and environment. These distortions in the energy distribution make replay attacks easier to be detected by identifying these energy anomalies.

The key benefits of β-order EW modulation include the following:

- **Enhanced Discriminative Capability**: By applying EW, the β-order modulation sharpens the system's sensitivity to the nuanced energy variations caused by replay attacks. Traditional features may overlook these subtle differences due to their focus on broader spectral characteristics. However, β-order energy-weighting accentuates specific variations in short-time energy, thereby making frequency bins more responsive to replay-induced anomalies. This heightened sensitivity improves the system's ability to accurately distinguish replayed speech from genuine speech, even when high-quality recording devices and quiet environments are used in the attack.
- **Adaptability and Robustness**: The adjustable β parameter offers a flexible mechanism to adapt the detection system to different replay attack scenarios. Replay attacks vary widely, influenced by the types of recording and playback devices used, the acoustic environment during recording, and the properties of the playback medium. The ability to fine-tune the β parameter means that the energy-weighting method can be optimized for specific scenarios, enhancing detection accuracy across a broad range of attack types. This adaptability of β against particular condition is valuable, as replay attacks can vary significantly depending on the recording and playback conditions, and β-order EW modulation provides an adaptable approach for detecting these variations.
- **Flexibility Across Spectral Domains**: By applying β-order EW in different spectral domains, e.g. octave, Mel, and linear domains, the method is capable to capture a wide range of spectral characteristics from different spectral domains. As a result, the system becomes more robust, capable of detecting replay attacks under various conditions by leveraging the strengths of each spectral domain.

B. Three novel features based on β-order energy weighting modulation

With the concept of β-order EW modulation, we propose three specific handcrafted features for spoofing detection as follows:

- CEOC: derived from octave domain by integrating CQT, β-order EW modulation and OST;

- MFECC: obtained from Mel domain by combining DFT, β-order EW modulation, Mel filter fbank, and DCT;
- LFECC: obtained from linear domain by combining DFT, β-order EW modulation, linear filterbank, and DCT.

The three proposed features originate from CQ-OST, MFCC and LFCC respectively. CQ-OST, MFCC and LFCC are the popular features in octave domain, Mel domain and linear domain, respectively. Apart from the introduction above that the three features belonging to different domains, we can find their difference as follows: (1) CQT is used in CQ-OST while DFT is used in MFCC and LFCC; because CQT is a long-term window transform while DFT is a short-term window transform, to some extent, we can say that CQ-OST is a feature based on long-term window transform while MFCC and LFCC are features based on short-term window transform. (2) CQ-OST is a subband feature while MFCC and LFCC are fullband features.

1) CEOC: First we consider the spectral energy in octave domain, from which CQ-OST feature is extracted. CQ-OST has emerged as a reliable feature for spoofing detection. Its effectiveness lies in capturing the unique properties of the original speech signal that are altered or lost in replay speech. CQ-OST is extracted from octave domain, wherein CQT is firstly used to transform the input time signal into frequency domain, then LPS is obtained by computing the power spectrum in log scale, finally OST is used to extract the principal information of octave.

Motivated by the CQ-OST feature, we propose a new feature CEOC for the purpose of RSD. CEOC feature is formed by using β-order EW module into a CQT spectral domain, and the logarithm of β-order EW is followed by octave domain transformation. In the case of CEOC, CQT spectral energy is modulated by β-order EW and then an OST transformation is applied. Conversely, with respect to CQ-OST, if the β-order EW modulation is excluded from CEOC, CQ-OST [38] can be obtained. Alternatively, we can state that the proposed CEOC feature is the β-order EW be applied to CQ-OST.

Fig. 5 shows how to extract CEOC. In this approach, the speech signal is transformed from the time domain to the constant-Q frequency domain using CQT. This is followed by the power spectrum module, which calculates the power value based on the CQT representation. To enhance the discriminative ability of detecting replay speech, the β-order EW modulation is applied. Additionally, logarithmic compression is employed to represent the values in a manner similar to the human auditory system. Finally, the OST module is used to extract the primary information for each octave.

2) MFECC: Besides CQ-OST that is proven to be effective feature for spoofing detection, MFCC, another prominent feature extraction technique, is also the popular feature related to Mel spectral domain for RSD respectively.

Inspired by the concept of MFCC, we propose the second novel feature: MFECC. MFECC is designed to capture essential characteristics of speech signals and is particularly expected for tasks of replay spoofing detection. The key innovation in MFECC lies in the modulation of the DFT spectral components by the β-order EW, which is then followed by a sequence of processing steps involving the Mel filter bank and the DCT. To provide a visual overview of this MFECC feature extraction process, Fig. 6 illustrates a block diagram that showcases the integration of DFT, β-order EW modulation, logarithmic scaling, Mel bank filtering, and DCT within the MFECC framework.

In the MFECC methodology, the β-order EW module assumes to play a pivotal role akin to CEOC feature extraction. This modular approach allows us to adapt and fine-tune MFECC to emphasize the disparity of spoofed speech from genuine speech. Benefiting from the inherent advantages of both the DFT and Mel filter bank for spectral analysis and the DCT for efficient feature dimensionality reduction, MFECC promises to be a powerful feature for RSD. It is necessary to emphasize that extracting MFCC can be achieved from Fig. 6 once the β-order EW component is eliminated.

3) LFECC: Despite the deviation from MFCC, LFCC retains a commonality with MFCC in its utilization of DCT as a critical step for feature dimensionality reduction. This cosine transformation not only helps to alleviate computational complexity but also plays a crucial role in retaining the most significant information within the feature vectors by capturing essential aspects of the spectral content while reducing the dimensionality.

Inspired by the advantageous characteristics of LFCC, we propose the third novel feature, i.e., LFECC. This innovative featuring approach leverages the power of β-order EW modulation within the framework of LFCC. In essence, LFECC takes the concept of LFCC a step further by incorporating the β-order EW into the computation of spectral energy. Specifically, the process begins with the extraction of the DFT spectral energy from the speech signal. This energy representation is then subjected to the β-order EW modulation, which enhances the variation level of non-speech distortion originated from recording/playback devices into the feature extraction process. Following this, a linear filter bank is applied to well preserve essential non-speech component caused by spoofing devices. Finally, the feature extraction process is completed with a DCT transformation, which provides a compact representation of the modulated energy values. The resulting LFECC feature holds promise spoofing signal analysis tasks, offering an enhanced capability for RSD.

Fig. 7 depicts a block diagram illustrating the LFECC feature extraction process, which involves the integration of DFT, β-order EW, linear filter bank, and DCT. In the context of LFECC, the β-order EW module fulfills a similar function to its role in MFECC feature extraction. It is important to mention that removing the β-order EW modulation from Fig. 7 yields LFCC.

C. RSD systems with the three β-order EW features

Backend classifier plays a critical role in the overall performance of a spoofing detection system. In order to evaluate the effectiveness of different features, it is necessary to select an appropriate backend classification method for detection of spoofing attacks. The frequently utilized classifiers in RSD include Gaussian mixture model (GMM), deep neural network (DNN), residual network (ResNet), and lightweight convolutional neural network (CNN). GMM is employed in various

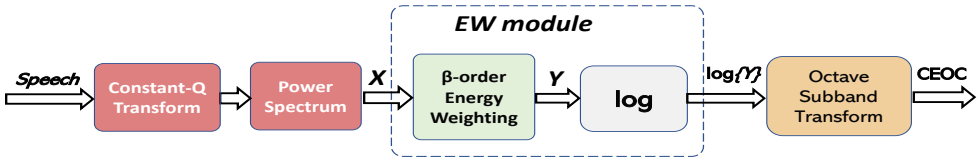

Fig. 5: Overview of the extraction methodology for Constant-Q Energy-Weighted Octave Coefficients (CEOC) used in replay speech detection.

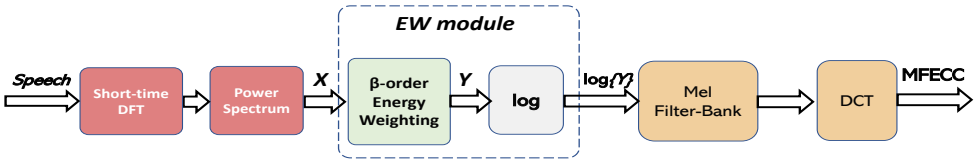

Fig. 6: Workflow diagram demonstrating the extraction of Mel Frequency Energy-Weighted Cepstral Coefficients (MFECC), highlighting spectral analysis in the Mel domain.

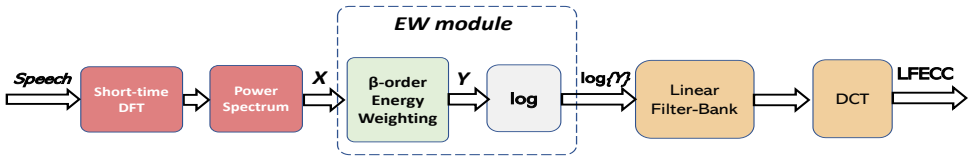

Fig. 7: Diagram illustrating the methodology for extracting Linear Frequency Energy-Weighted Cepstral Coefficients (LFECC), focusing on linear spectral domain processing for replay detection.

works for RSD [31], [44]–[46] [32], [47]. Similarly, DNN has been successfully utilized for RSD in [36], [37], [48], [49]. The utterance-based classifiers, unlike the others, take the utterance itself as the input and examine the artifacts at the utterance level. The work in [33], [50], [51] utilizes ResNet for RSD, while [20], [52], [53] employs light CNN.

1) ResNet: In this study, ResNet [42] is taken into account as the classifier due to its extensive usage in spoofing countermeasure research [33], [50], [51]. Employing ResNet can enhance the evaluation of spoofing detection features, thereby increasing their performance verification effectiveness. We utilized ResNet-based classifiers that adhere to the classic ResNet architecture established in the works of [42] and [54].

With the classic ResNet for backend classification, we develop three RSD systems: CEOC-ResNet, MFECC-ResNet and LFECC-ResNet.

2) DNN: In addition to ResNet, we utilize a DNN implemented in the Microsoft Cognitive Toolkit (CNTK) [43] as another backend classifier. Specifically, a series of four-layer DNN classifiers are trained for replay speech detection following our previous training methodology [40], [54]. The network consists of two hidden layers, each containing 512 nodes, and an output layer with two nodes. The input to the network is a feature vector constructed from an 11-frame context window, achieved by splicing five frames to the left and right of the central frame. Sigmoid activation functions are applied in the hidden layers to optimize performance, while cross-entropy with softmax is used as the training criterion. Mean and variance normalization is applied to the input data to ensure consistency. The DNN is trained using stochastic gradient descent (SGD) for optimization, with a specific training scheme following the default configuration of CNTK. The scheme includes 25 total epochs: the learning rate is set to 0.8 for the first epoch, increases to 3.2 for the next 14 epochs, and is reduced to 0.08 for the remaining 10 epochs. A minibatch size of 256 is used for the first epoch and 1024 for the remaining epochs, with a momentum value of 0.9 applied throughout. This setup, leveraging CNTK's training scheme, has been shown to yield optimal performance across different replay speech detection systems, including CEOC-DNN, MFECC-DNN, and LFECC-DNN.

III. Experiments And Evaluations

In this section, we evaluate the performance of three β-order EW modulated features, namely CEOC, MFECC, and LFECC, for the purpose of replay spoofing detection. To accomplish this, we conduct a series of experiments. The following subsections provide comprehensive details regarding the experimental conditions and the assessment results. We compare the performance of the proposed features against typical traditional features, thereby establishing a benchmark for assessment. By undertaking these evaluations, we aim to gain a deeper understanding of the effectiveness and potential advantages of the β-order EW modulated features for detection of replay attacks.

TABLE I: Dataset partitions of A17V, B16P and A19P, where # denotes the number of utterances wherein RRS and SRS stands for real replay speech and simulated replay speech, respectively.

DB	Set	#Genuine	#RRS	#SRS	#Total
A17V	Training	1,507	1,507	0	3,014
	Development	760	950	0	1,710
	Evaluation	1,298	12,008	0	13,306
B6P	Training	4,973	2,800	0	7,773
	Development	4,995	2,800	0	7,795
	Evaluation	5,576	4,800	0	10,376
A19P	Training	5,400	0	48,600	54,000
	Development	5,400	0	24,300	29,700
	Evaluation	18,089	0	134,630	152,719
A21P	Training	5,400	0	48,600	54,000
	Development	5,400	0	24,300	29,700
	Evaluation	816,480	126,630	0	943,110

A. Databases for Performance Evaluation

We consider ASVspoof 2017 version 2.0 (A17V) [11], [34], BTAS 2016 physical access (B16P) [55]–[57] (which is on the basis of AVspoof 2015 database [55]), ASVspoof 2019 physical access (A19P) [58] and ASVspoof 2021 physical access (A21P) [59] as the evaluation databases for replay speech detection.

A17V originates from the RedDots corpus [11] and has the most wide range of device information [34]. While the replay speech in B16P database is obtained by using four different scenarios, which are replay-phone1, replay-phone2, replay-laptop and replay-laptop-high quality [55]. Different from A17V and B6P corpora, the A19PA database is obtained by exploring replay attacks using a far more controlled evaluation setup in a form of simulated replay attacks and carefully controlled acoustic condition [60]. Additionally, A19P corpus is substantially larger than the A17V and B16P corpora. The bonafide examples of A19P are originated from VCTK corpus [61]; and the corpus comprises a total of 107 speakers, with 46 males and 61 females. Importantly, there is no speaker overlap among the three subsets, i.e. training, development, and evaluation datasets, of A19P. A19P and A21P have the same training set and development set. Since the replay speech in the training set and development set of A21P is simulated replay speech while that in the evaluation set of A21P is real replay speech, the mismatching between training data and evaluation data is apparent in A21P evaluation platform. Due to this mismatching, theoretically A21P model performs much worse than that of A19P.

Table I gives the data partitions containing the information on genuine utterances, replay utterances, as well as the data deployment for training, validation, and test in A17V, B16P, A19P and A21P, respectively.

The evaluation metric used for A17V and B16P is the equal error rate (EER), as stated in their respective challenge rules; whereas, for A19P and A21P, the tandem detection cost function (t-DCF) and EER serve as the primary and secondary evaluation metrics, respectively, as referenced by the work in [62].

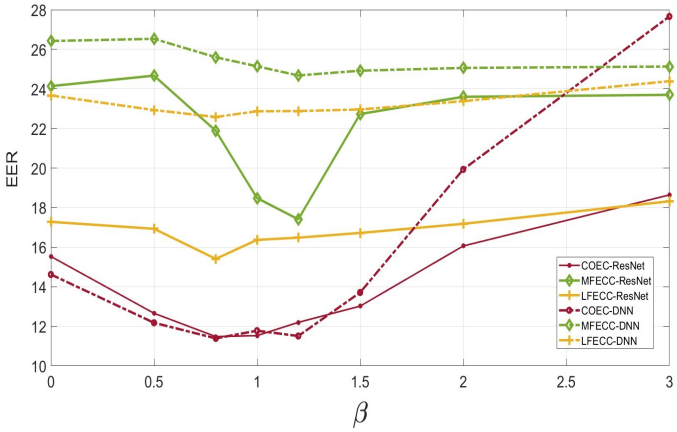

Fig. 8: Investigation of β value with the three proposed β-order EW modulation in terms of EER (%) of ResNet-based and DNN-based RSD.

B. Configuration of the experiments

In the experiments, we utilized identical parameters for CQT as those mentioned in [26], [27]. Specifically, for CEOC, our implementation includes 9 octaves, each containing 96 frequency bins in LPS. Like CQ-OST in [38], we choose the top 12 dimensions as the ultimate features after performing DCT on each octave. Consequently, the CEOC has a total of 108 dimensions (9×12). We configure the MFECC with 13 dimensions and the LFECC 20 dimensions.

C. Investigation on the value of β

First, we conducted a comprehensive analysis to investigate the effect of varying β values on the performance of our proposed EW features using the A17V evaluation set. Fig. 8 presents the results, showing the performance of both ResNet-based systems (CEOC-ResNet, MFECC-ResNet, and LFECC-ResNet) and DNN-based systems (CEOC-DNN, MFECC-DNN, and LFECC-DNN) in terms of EER across a range of β values for each of the three proposed features. The experimental results reveal that the choice of β significantly affects the system's ability to detect replay attacks. For each feature set, performance varies as β changes, highlighting the need for fine-tuning this parameter to achieve optimal results. Notably, different features exhibit different optimal β values. For example, the CEOC feature may perform best with a specific β value, while MFECC and LFECC might achieve superior results with different values of β. This indicates that β plays a critical role in modulating the energy-weighting process and that there is no single universally optimal β for all feature types. Overall, this investigation underscores the importance of selecting the right β value based on the specific feature set and model architecture, as it directly influences the discriminative power of the energy-weighted features in replay speech detection tasks.

Second, we further investigate the value of β with selected values such as 0.8, 1.0, and 1.2 across four major datasets: A17V, B16P, A19P, and A21P. Table II shows the performance

TABLE II: Experimental results in terms of EER (%) for the proposed EW-based features on A17V and B16P evaluation sets, EER and t-DCF for that on A19P and A21P evaluation sets. RR-EER denotes the relative reduction ratio of EER, herein, it indicates the EER reduction ratio of modulated system over its non-modulated (i.e. $\beta = 0$) system.

DB	Domain	Feature	β	ResNet			DNN		
				EER	RR-EER	t-DCF	EER	RR-EER	t-DCF
A17V	Octave	CQ-OST	0	15.53			14.62		
		COEC	**0.8**	**11.48**	26.07		**11.39**	22.09	
			1.0	11.55	25.63		11.78	19.43	
			1.2	12.20	21.44		11.51	21.27	
	Mel	MFCC	0	24.14			26.42		
		MFECC	0.8	21.88	9.36		25.60	3.10	
			1.0	18.48	23.45		25.14	4.84	
			1.2	**17.41**	27.87		**24.68**	6.59	
	Linear	LFCC	0	17.28			23.67		
		LFECC	**0.8**	**15.41**	10.82		**22.58**	4.60	
			1.0	16.37	5.27		22.87	3.38	
			1.2	16.48	4.63		22.88	3.34	
B16P	Octave	CQ-OST	0	0.60			2.99		
		COEC	**0.8**	**0.31**	48.30		**0.72**	75.92	
			1.0	0.34	43.33		0.96	67.89	
			1.2	0.47	21.67		1.19	60.20	
	Mel	MFCC	0	1.29			7.40		
		MFECC	0.8	0.74	42.64		7.01	5.27	
			1.0	0.89	31.01		6.81	7.97	
			1.2	**0.63**	51.61		**6.51**	12.03	
	Linear	LFCC	0	4.32			10.32		
		LFECC	**0.8**	**0.85**	80.32		**4.67**	54.75	
			1.0	0.88	79.63		4.73	54.17	
			1.2	0.96	77.78		4.93	52.23	
A19P	Octave	CQ-OST	0	3.75		0.095	16.02		0.357
		COEC	0.8	3.37	10.13	0.086	14.04	14.34	0.242
			1.0	**3.28**	12.53	**0.081**	**12.65**	21.04	**0.231**
			1.2	4.93	-31.47	0.124	14.05	12.30	0.264
	Mel	MFCC	0	2.45		0.066	19.47		0.467
		MFECC	0.8	1.18	57.55	0.032	20.10	-3.24	0.487
			1.0	**1.08**	55.92	**0.030**	19.89	-2.16	0.492
			1.2	1.38	43.67	0.039	**19.54**	-0.36	**0.468**
	Linear	LFCC	0	2.41		0.059	16.02		0.357
		LFECC	0.8	1.04	51.84	0.029	15.96	0.37	0.363
			1.0	**0.98**	59.34	**0.027**	16.12	-0.62	0.371
			1.2	1.10	54.36	0.032	**15.82**	1.25	**0.350**
A21P	Octave	CQ-OST	0	42.54		0.999	41.43		0.972
		COEC	-0.2	42.59	-0.12	0.999	49.61	-19.74	0.999
			-0.5	42.17	0.87	0.999	40.80	1.52	0.966
			-0.8	**40.89**	3.88	0.999	**39.93**	3.62	**0.950**
			0.8	44.39	-4.35	0.999	49.56	-19.62	0.999
			1.0	45.91	-7.92	0.999	49.61	-19.74	0.999
			1.2	45.89	-7.87	0.999	49.60	-19.72	0.999
	Mel	MFCC	0	40.80		0.999	35.33		0.927
		MFECC	**-0.2**	**35.92**	11.96	0.999	**35.90**	-1.61	**0.910**
			-0.5	41.84	-2.55	0.999	36.50	-3.31	0.918
			-0.8	43.18	-5.83	0.999	37.23	-5.38	0.937
			0.8	41.37	-1.40	0.999	**35.59**	-0.74	**0.930**
			1.0	42.92	-5.20	0.999	35.85	-1.47	0.938
			1.2	43.78	-7.30	0.999	35.98	-1.84	0.936
	Linear	LFCC	0	39.78		0.999	40.47		0.992
		LFECC	**0.8**	**42.09**	-5.81	**0.994**	**41.35**	-2.17	0.999
			1.0	42.80	-7.59	0.999	41.40	-2.30	0.999
			1.2	43.75	-9.98	0.999	41.74	-3.14	0.999

of the proposed β-order EW modulation for replay speech detection conducted on the four datasets. The comparison was carried out between the non-modulated features (CQ-OST, MFCC, LFCC) and their energy-weighted counterparts (CEOC, MFECC, LFECC). In terms of EER, the following observations can be seen:

- A17V Evaluation: On the A17V dataset, CEOC-ResNet demonstrated a 26.07% relative reduction ratio in EER compared to (CQ-OST)-ResNet, with an optimal β value of 0.8. MFECC-ResNet also showed a significant improvement, with a 27.87% reduction in EER compared to MFCC-ResNet. LFECC-ResNet provided a smaller, but still notable, improvement of 10.82%. These results indicate that the energy-weighting modulation is particularly effective in this dataset, especially in the Mel and octave domains. It is observed that the best performance is with β=0.8 where CEOC-DNN achieves 11.39% of EER and CEOC-ResNet 11.48%.

- B16P Evaluation: The B16P dataset demonstrated the strongest improvements, particularly with MFECC-ResNet at β = 1.2 achieving a 51.61% reduction in EER compared to MFCC-ResNet. CEOC-ResNet at β=0.8 showed a 48.30% reduction compared to CQ-OST, while LFECC-ResNet at β=0.8 achieves 80.32% reduction. The optimization of β was crucial in this dataset, as the value of 0.8 consistently delivered superior performance, especially in octave domain, where CEOC-ResNet achieves the best EER of 0.31%.
- A19P Evaluation: In the A19P dataset, which includes simulated replay speech, MFECC-ResNet and LFECC-ResNet provided substantial improvements, with MFECC-ResNet reducing EER by 55.92% and LFECC-ResNet by 59.34%. CEOC-ResNet showed a moderate reduction of 10.13%. The optimal β for MFECC and LFECC was 1.0, highlighting that energy-weighting modulation remains highly effective even in more controlled and simulated environments. The best EER is 0.98% with MFECC-ResNet. This is simulation dataset, it can be seen that the ResNet classifier consistently outperforms the DNN classifier in the three domains.
- A21P Evaluation: Based on the performance evaluation provided in Table II, the A21P dataset poses significant challenges due to the real-world nature of the replay attacks - the best EER is more than 35%; this is because that distinguishing between genuine and replayed speech remains difficult in this dataset due to mismatches between simulated training data (A19P training dataset) and real-world evaluation conditions (A21P evaluation dataset). However, the result demonstrates the potential of the DEOC and MFECC features in capturing replay artifacts with the mismatching conditions. It can be seen that the MFECC-ResNet achieved an EER of 35.92% with a β value of -0.2, demonstrating an improvement over the original MFCC feature, which recorded an EER of 40.80%. Similarly, the MFECC-DNN also reached an EER of 35.90% at β = -0.2, along with the best t-DCF score of 0.910. Additionally, the COEC feature at β = 0.8 showed notable performance, reducing the EER by 3.88% with the ResNet classifier and by 3.62% with the DNN classifier. These results highlight the effectiveness of the energy-weighted features in enhancing detection capabilities in the challenging A21P environment.
- The reason why negative beta only provided for A21P and not for other databases is that the replay attack types in A21P training set and that in A21P evaluation are different while those in other databases are the same. Furthermore, the replay attack types in A21P training set is simulated replay while that in A21P evaluation is real replay.

It can be noted that the t-DCF performance is almost consistent with EER. The experimental results demonstrate that β-order EW modulation significantly enhances the performance of traditional replay speech detection features (CQ-OST, MFCC, LFCC). Across the datasets, the proposed features CEOC, MFECC, and LFECC consistently outperformed their non-modulated counterparts. The optimal β value varied across datasets but was generally between 0.8 and 1.2, indicating that moderate energy modulation is effective in emphasizing short-time energy variations introduced by replay attacks. The improvements were particularly evident in real replay environments, such as B16P, where MFECC showed up to a 65.57% EER reduction. Overall, β-order EW modulation is a robust and adaptable technique for replay speech detection, offering substantial improvements in both real and simulated attack scenarios.

Based on the experimental results shown in Fig. 8 and Table II, we observed that for each feature type (CEOC, MFECC, or LFECC), the same β value led to optimal performance across both ResNet and DNN classifiers for a given evaluation dataset. This indicates that while the overall performance is influenced by the model architecture, the optimal energy-weighting modulation, determined by β, remains consistent for each feature type, regardless of whether ResNet or DNN is used as the backend classifier for a specific dataset (e.g., A17V, B16P, A19P, or parts of A21P). This suggests that optimal β in the energy modulation process is primarily determined by the characteristics of the feature representation itself, rather than the choice of classifier. As a result, once the optimal β value is determined for a specific feature, it can be applied to different classifier architectures with consistently positive results. For example, given a specifical database, if CEOC achieves optimal performance at a particular β value with ResNet, that same β value is likely to similarly benefit CEOC-DNN.

D. Comparison with some commonly used features

In this subsection, the proposed features with optimal β are compared with traditional features including instantaneous frequency cepstral coefficients (IFCC) [63], CQCC [26], [27], eCQCC [37], eCQCC concatenated log energy (eCQCC-E) , constant-Q statistics-plus-principal information coefficients (CQSPIC) [36] and PNCC. In this context, every feature is assessed using its own ResNet classifier, and the training configuration for the ResNet model remains consistent across all experiments for RSD.

Supposing the proposed β-order EW modulation is applied on CQCC, the obtained feature can be named as constant-Q energy-weighted cepstral coefficient (CQECC). Table III presents a comparative analysis of the performance of proposed features (CQECC, CEOC, MFECC, LFECC) against commonly used features (IFCC, CQCC, PNCC, eCQCC, CQSPIC) on multiple datasets (A17V, B16P, A19P, A21P).

- A17V: CEOC (β = 0.8) achieves an EER of 11.48%, outperforming all other features, including eCQCC, eCQCC-E and CQSPIC, which had EERs of 15.34%, 13.71% and 14.40%, respectively. MFECC and LFECC performed worse than eCQCC, eCQCC-E and CQSPIC, with EERs of 17.41% and 15.41%, respectively, indicating that the octave-based feature (CEOC) is more effective in this dataset, which may be due to the greater ability of CQT-based features to capture replay artifacts in real environments.

TABLE III: Performance comparison in EER (%) with some commonly used features under the classifier of ResNet on the evaluation sets of A17V, B16P, A19P and A21P, respectively.

DB	Type	Feature	EER	t-DCF
A17V	Commonly-used	IFCC	23.58	
		CQCC	18.42	
		PNCC	16.27	
		eCQCC	15.34	
		eCQCC-E	13.71	
		CQSPIC	14.40	
	Proposed	CQECC ($\beta = 0.8$)	16.94	
		CEOC ($\beta = 0.8$)	**11.48**	
		MFECC ($\beta = 1.2$)	17.41	
		LFECC ($\beta = 0.8$)	15.41	
B16P	Commonly-used	IFCC	1.48	
		CQCC	0.50	
		PNCC	5.13	
		eCQCC	0.46	
		eCQCC-E	0.36	
		CQSPIC	0.40	
	Proposed	CQECC ($\beta = 0.8$)	0.47	
		CEOC ($\beta = 0.8$)	**0.31**	
		MFECC ($\beta = 0.8$)	0.85	
		LFECC ($\beta = 1.2$)	0.63	
A19P	Commonly-used	IFCC	2.54	0.073
		CQCC	3.37	0.090
		PNCC	2.67	0.077
		eCQCC	1.85	0.047
		eCQCC-E	1.51	0.040
		CQSPIC	1.67	0.044
	Proposed	CQECC ($\beta = 1$)	3.03	0.087
		CEOC ($\beta = 1$)	3.28	0.081
		MFECC ($\beta = 1$)	**0.98**	**0.027**
		LFECC ($\beta = 1$)	1.08	0.030
A21P	Commonly-used	IFCC	47.30	0.999
		CQCC	46.30	0.999
		PNCC	41.41	0.999
		eCQCC	45.69	0.999
		eCQCC-E	41.41	0.999
		CQSPIC	42.72	0.999
	Proposed	CQECC ($\beta = -0.8$)	42.59	0.999
		CEOC ($\beta = -0.8$)	40.89	0.999
		MFECC ($\beta = -0.2$)	**35.92**	0.999
		LFECC ($\beta = 0.8$)	42.09	**0.994**

- B16P: CEOC (β = 0.8) demonstrates superior performance with an EER of 0.31%, outperforming CQSPIC, which had an EER of 0.40%. The relative reduction compared to IFCC (PNCC) is significant. MFECC and LFECC also perform well, with EERs of 0.85% and 0.63%, respectively. The results show that β-order energy-weighting modulation enhances feature performance in physical access replay conditions.
- A19P: MFECC (β = 1) achieves the best performance with an EER of 0.98%, outperforming all other features, including eCQCC (1.85%), eCQCC-E (1.51%), CQCC(3.37%), PNCC(2.67%) and CQSPIC (1.67%). LFECC (β = 1) follows closely with an EER of 1.08%, while CEOC (β = 1) performs less effectively with an EER of 3.28%. The results suggest that in simulated replay conditions, Mel and linear domain features benefit more from energy modulation.
- A21P: In the A21P dataset, the proposed energy-weighted features demonstrated notable performance, with CEOC achieving an EER of 40.89% at a beta value of -0.8. Meanwhile, MFECC and LFECC exhibited EERs of 35.92% and 42.09%, respectively. Although these results are competitive, they slightly surpass traditional features such as CQCC, which also presented EERs around 42% to 47%. Additionally, the LFECC feature recorded a t-DCF score of 0.994 at β=0.8, indicating a slight advantage in detection capability compared to other features under similar conditions.
- CQECC outperforms CQCC on the four databases. The reason is that the proposed β-order EW modulation is used in CQECC, which also confirms that the proposed β-order EW modulation.
- COEC performs better than CQECC on the four databases. The reason may be that CEOC only applies β-order EW modulation on the octave power spectrum while CQECC first applies β-order EW modulation on the octave power spectrum and then applies uniform resampling on octave power spectrum to obtain linear power spectrum.

The experimental results in Table III highlight that β-order EW modulation enhances the performance of traditional features across multiple datasets, particularly in real replay scenarios (A17V and B16P). CEOC (octave domain) is the most effective in physical replay conditions, while MFECC (Mel domain) excels in simulated replay attacks (A19P) and LFECC performs the best for simulated data (training) to real data (test) situation (A21P). The ability to optimize β for different environments provides flexibility and adaptability, making energy-weighted features a robust choice for replay speech detection.

E. Comparison with state-of-the-art systems

Here, we would like to compare the proposed systems with state-of-the-art systems on the evaluation set of A17V, B16P, A19P and A21P. The experimental results comparison is given in Table IV, in which,

For the selected systems on A17V, the following approaches are compared:

- LCNN refers to a light convolutional neural network, FFT-LCNN represents light CNN based end-to-end system with fast Fourier transform (FFT) spectrograms as the input feature [56].
- CQCC-E is an energy-appended feature formed by concatenating CQCC with log energy [34].
- CMPOC stands for constant-Q magnitude-phase octave coefficients [64].

For the selected systems on B16P, DNN-, LSTM- and GRU-based systems are employed to compare their performance with the proposed ResNet-based RSD systems. Herein, MFCC and filterbank (Fbank) are used to feed these systems [57], respectively. In addition, FFT-LCNN is an end-to-end system as reported in [56].

For the selected systems on A19P, the following methods are used:

- ZTWCC stands for zero time windowing cepstral coefficients, and $\text{Dfea}_{\text{DFTLPS}}$) [29] represents deep feature with log power spectrum extracted from DFT.

TABLE IV: Performance comparison in EER (%) with the state-of-the-art single systems on the evaluation sets of A17V, B16P, A19P and A21P, respectively.

DB	Type	Systems	EER	t-DCF
A17V	LCNN	FFT-LCNN [56]	12.39	
	GMM	CQCC-GMM [34]	15.33	
		(CQCC-E)-GMM [34]	12.24	
	i-vector	CQCC-(i-vector) [34]	15.63	
		(CQCC-E)-(i-vector) [34]	12.93	
	DNN	CMPOC-DNN [64]	14.93	
		CMC-DNN [40]	14.45	
		CQSPIC-DNN [36]	**11.09**	
	Proposed	CEOC-ResNet($\beta = 0.8$)	11.48	
		MFECC-ResNet ($\beta = 1.2$)	17.41	
		LFECC-ResNet ($\beta = 0.8$)	15.41	
		COEC-DNN ($\beta = 0.8$)	**11.39**	
		MFECC-DNN ($\beta = 1.2$)	24.68	
		LFECC-DNN ($\beta = 0.8$)	22.58	
B16P	LCNN	FFT-LCNN [56]	8.44	
	DNN	MFCC-DNN [57]	2.06	
		Fbank-DNN [57]	2.00	
	LSTM	MFCC-LSTM [57]	2.15	
		Fbank-LSTM [57]	1.11	
	GRU	MFCC-GRU [57]	1.91	
		Fbank-GRU [57]	1.08	
	Proposed	**CEOC-ResNet(β=0.8)**	**0.31**	
		MFECC-ResNet (β=1.2)	0.63	
		LFECC-ResNet (β=0.8)	0.85	
		COEC-DNN (β=0.8)	0.72	
		MFECC-DNN (β=1.2)	6.51	
		LFECC-DNN (β=0.8)	4.67	
A19P	GMM	CQCC-GMM [58]	11.04	0.245
		LFCC-GMM [58]	13.54	0.302
		ZTWCC-GMM [65]	12.00	0.281
		IFCC-GMM [65]	15.59	0.357
		ICQCC-GMM [29]	10.94	0.250
		CMC-GMM [29]	11.22	0.250
	DNN	(CQ-OST)-DNN [38]	8.04	0.188
		$\text{Dfea}_{\text{DFTLPS}}$-DNN [29]	7.57	0.181
	LCNN	DCT-LCNN [66]	2.06	0.560
		LFCC-LCNN [66]	4.60	0.105
		CQT-LCNN [66]	1.23	0.030
	ResNet	CQCC-ResNet [51]	4.43	0.107
		Spec-ResNet [51]	3.81	0.099
		GDG-ResNet [50]	1.79	0.044
		JG-ResNet [50]	1.23	0.031
	PLDA	SIV-PLDA [67]	1.09	0.023
	ResNetGAP	GDG-DASP-ResNetGAP [60]	1.08	0.028
	ResNetW18	**MGCG-ResNetW18 [60]**	**0.52**	**0.014**
	Proposed	CEOC-ResNet(β=1.0)	3.28	0.081
		MFECC-ResNet (β=1.0)	1.08	0.030
		LFECC-ResNet (β=1.0)	**0.98**	**0.027**
		COEC-DNN (β=1.0)	12.65	0.231
		MFECC-DNN (β=1.2)	19.54	0.468
		LFECC-DNN (β=1.2)	15.82	0.350
A21P	GMM	CQCC-GMM [59]	38.07	0.943
		LFCC-GMM [59]	39.54	0.972
		World-GMM [68]	**24.88**	0.6853
	LCNN	LFCC-LCNN [59]	44.77	0.996
	CSGL	RawNet2 [59]	48.60	0.999
	Proposed	CEOC-ResNet(β=-0.8)	40.89	0.999
		MFECC-ResNet (β=-0.2)	35.92	0.999
		LFECC-ResNet (β=0.8)	42.09	0.994
		COEC-DNN (β=-0.8)	39.93	0.950
		LFECC-DNN (β=0.8)	41.35	0.999
		MFECC-DNN (β=0.8)	**35.59**	0.930
		MFECC-DNN (β=-0.2)	35.90	**0.910**

- In the ResNet-based systems, GDG and JG stand for group delay gram and jointly derived from GDG and short-term Fourier transform gram, respectively.
- In the PLDA-based system, SIV and PLDA stand for spoofing identity vector and probabilistic linear discriminant analysis, respectively. In which, SIV is actually extracted from gated recurrent convolutional neural networks (GRCNNs) with log magnitude spectrogram (LMS) and modified group delay (MGD) as well as their corresponding signal-to-noise mask (SNM) features as inputs. Similar with x-vector speaker verification, PLDA plays a role of classification in the SIV-PLDA system.
- The best and the second best single systems in ASVspoof 2019 PA challenge are also selected here for comparison. They are MGCG-ResNetW18 and GDG-DASP-ResNetGAP [60]. In the MGCG-ResNetW18 spoofing detection system, MGCG stands for Mel-gram concatenated with CQT-gram while ResNetW18 is a modified version of ResNet18 [60]. In the GDG-DASP-ResNetGAP system, GDG-DASP represents group delay gram (GDG) with data augmentation via speed pertubation (DASP) while ResNetGAP is based upon a ResNet34 architecture with a global average pooling (GAP) layer [60].

For the selected systems on A21P, the following baseline systems from the ASVspoof2021 challenge are compared:

- CQCC-GMM, LFCC-GMM, LFCC-LCNN and RawNet2 are baseline systems provided by the organizers of the ASVspoof 2021 challenge [59].
- World represents the World vocoder used to generate replay channel feature with log-spectrogram as input, and the system world-GMM gave the best performance for the single system in ASVspoof2021PA challenge, which applies a world-vocoder-based replay channel estimation feature with log-spectrogram as input while one-class and utterance-level GMM is used as the models.
- CSGL represents the classifier combining Sinc-filter, gate recurrent unit and linear-layer, more details of RawNet2 can be found in [69].

The performance of the proposed systems is compared with state-of-the-art approaches across four datasets: A17V, B16P, A19P, and A21P. As shown in Table IV, the experimental results illustrate that the proposed methods exhibit competitive performance against established models.

On the A17V dataset, the CEOC-DNN (β = 0.8) achieved an EER of 11.39%, which is marginally higher than CQSPIC-DNN's EER of 11.09%. However, the CEOC-DNN and CEOC-ResNet consistently outperform other methods, such as the (CQCC-E)-(i-vector) system and several GMM and DNN-based models. The slightly better performance of CQSPIC-DNN can be attributed to the use of a feature combination that integrates spectral, subband spectral, and statistical information, whereas CEOC relies solely on spectral features. Despite this, CEOC performs robustly with fewer features.

In the B16P dataset, the proposed systems demonstrated significant superiority over all known state-of-the-art methods. CEOC-ResNet (β = 0.8) delivered an outstanding performance, achieving the lowest EER of 0.31%. This result marked a substantial improvement over competitive systems such as Fbank-GRU and Fbank-LSTM, which had EERs of 1.08% and

1.11% respectively. The large margin by which CEOC-ResNet outperformed these methods highlights the exceptional capability of energy-weighted features in capturing replay-specific characteristics, particularly in physical replay conditions. Furthermore, MFECC-ResNet at β=1.2 and LFECC-ResNet at β=0.8 also demonstrated strong performance, with EERs of 0.63% and 0.85% respectively. These results clearly emphasize the robustness and reliability of the proposed energy-weighted modulation framework when applied to physical access scenarios, where subtle energy variations between replay and genuine speech are effectively captured and amplified, leading to more accurate detection outcomes.

For the A19P dataset, the LFECC-ResNet (β = 1) achieved the best performance, with an EER of 0.98% and a t-DCF score of 0.027, outperforming all other systems, including well-known models like GDG-DASP-ResNetGAP, which had an EER of 1.08%, and JG-ResNet, which reached an EER of 1.23%. This indicates the strong capability of linear-based energy-weighting in handling simulated replay attacks, where the proposed LFECC-ResNet method demonstrated superior robustness and accuracy. Additionally, the MFECC-ResNet (β = 1) also performed notably well, with an EER of 1.08%, coming very close to LFECC-ResNet, highlighting the effectiveness of Mel-based spectral analysis in detecting spoofed speech. Meanwhile, the CEOC-ResNet system showed a higher EER of 3.28%, indicating that the octave-based feature, though effective in other conditions, was less suited to the simulated replay environment of A19P compared to the Mel and linear domain features.

On the A21P dataset, the experimental results were mixed due to the difficulty of distinguishing between replayed and genuine speech in real-world scenarios. However, MFECC-DNN (β = 0.8) demonstrated a strong performance, achieving an EER of 35.59% and a t-DCF of 0.930. This result indicates that Mel frequency-based energy-weighted features, combined with the DNN backend, effectively capture replay artifacts in challenging conditions. MFECC-DNN outperformed several systems, including CQCC-GMM and LFCC-GMM, both of which recorded higher EERs of 38.07% and 39.54%, respectively. In addition, World-GMM, a system using log-spectrograms and replay channel estimation via a vocoder, performed particularly well with an EER of 24.88% and a t-DCF of 0.685. This system was among the best performers in the A21P challenge, showing that incorporating vocoder-based replay channel estimation can enhance the detection under the mismatching conditions. Despite the competitive results of World-GMM, the proposed MFECC-DNN still showcased its potential as a strong performer in terms of both flexibility and adaptability in detecting replay attacks under varying conditions.

From the Table IV, the significant differences in EER across the datasets particularly the much lower EER in B16P and A19P, and the notably higher EER in A21P can be attributed to several key factors related to the nature of the datasets, attack types, and evaluation environments. The much lower EER values in B16P and A19P are largely due to the controlled and simulated nature of the replay attacks, which makes detecting artifacts easier. On the other hand, A21P poses a much more difficult challenge because it contains real-world replay attacks with more unpredictable acoustic conditions and higher variability, leading to significantly higher EER values. This reflects the current difficulty of detecting replay attacks in real-world, uncontrolled environments, where the line between genuine and spoofed speech becomes increasingly blurred. The proposed CEOC, MFECC, and LFECC systems provide competitive performance, and in many cases, surpass state-of-the-art methods, particularly in physical access replay detection scenarios.

The experimental results across multiple datasets, i.e. A17V, B16P, A19P, and A21P, clearly show that the β-order energy-weighting modulation leads to significant improvements in both EER and t-DCF metrics compared to non-modulated features. Its flexibility and adaptability across various datasets and conditions emphasize its effectiveness as a powerful and reliable approach for enhancing anti-spoofing systems.

TABLE V: Performance comparison in terms of EER (%) between the proposed energy-weight features (PEF) (β=1) and the corresponding energy-appended features (EF) under the classifiers of ResNet on the evaluation set of A17V, B16P and A19P, wherein RR (%) stands for relative reduction of EER, respectively.

DB	Domains	EF	EER	PEF	EER	RR
A17V	Octave	CQ-OST-E	13.95	CEOC	11.55	17.20
	Mel	MFCC-E	20.10	MFECC	18.48	8.06
	Linear	LFCC-E	16.76	LFECC	16.37	2.33
B16P	Octave	CQ-OST-E	0.54	CEOC	0.34	37.04
	Mel	MFCC-E	1.09	MFECC	0.89	18.35
	Linear	LFCC-E	1.12	LFECC	0.88	21.43
A19P	Octave	CQ-OST-E	3.48	CEOC	3.28	5.75
	Mel	MFCC-E	1.27	MFECC	1.08	14.96
	Linear	LFCC-E	1.29	LFECC	0.98	24.03

TABLE VI: Performance comparison in terms of t-DCF between the proposed energy-weight features (PEF) (β=1) and the corresponding energy-appended features (EF) under the classifiers of ResNet on the evaluation set of ASVspoof2019PA, wherein RR (%) stands for relative reduction of t-DCF.

Domains	EF	t-DCF	PEF	t-DCF	RR
Octave	CQ-OST-E	0.086	CEOC	0.081	5.81
Mel	MFCC-E	0.037	MFECC	0.030	18.92
Linear	LFCC-E	0.039	LFECC	0.027	30.77

IV. Discussion

A. Comparison with energy-appended features

In this subsection, we are to discuss the application of energy information in the feature extraction for RSD by comparing our proposed EW modulation features with traditional energy-appended features [34]. The similarity and discrepancy between appending energy information and EW modulation can be explicitly observed as follows. Both features originate from the same original feature, and they integrate the energy component into individual discriminative features in different ways. Traditional energy-appended feature improves the discriminative nature by concatenating energy while β-order

TABLE VII: Some details about ASVspoof2019LA corpus, # stands for number.

Subset	#Male	#Female	#Bonafide	#Spoofed
Training	8	12	2,580	22,800
Development	4	6	2,548	22,296
Evaluation	21	27	7,355	63,882

EW modulation feature enhances discriminative nature by using energy modulation. The energy information in energy-appended feature serves as an additional bin with which the original feature concatenates, while the energy information in EW modulation is utilized to affect energy distribution representing the disparity of spoofed components over the short-term sequence through original feature. The typical representatives of the traditional energy-appended features are CQ-OST-E, MFCC-E and LFCC-E.

We observed that the recording and replaying devices changes much on the short-time speech energy in spectral domain. This short-time energy can be used to differentiate the distortion from the genuine speech. Based on the observation, we proposed the β-order EW modulation framework. The comparison results between energy-appended features and the proposed EW-based features with standard $\beta = 1$ on A17V, B16P and A19P evaluation sets are presented in terms of EER in Table V, respectively. Table VI shows the performance comparison on A19P evaluation sets in terms of t-DCF. From Tables V and VI, it can be found that the proposed EW modulated features outperform the corresponding energy-appended features on the three evaluation sets in terms of EER and t-DCF.

B. Evaluation on synthesized speech on logical access database

In this subsection, we evaluate the performance of the proposed EW features on synthetic speech using the ASVspoof 2019 logical access (LA) database. The LA database comprises TTS and VC attacks, where no replay distortions are involved. Table VII provides an overview of the ASVspoof2019LA dataset, while Table VIII presents the performance results in terms of EER and t-DCF for the proposed EW features on this corpus.

Table VIII summarizes the performance of various EW-modulated features, comparing their results on the ResNet and DNN backends. The results demonstrate that while EW modulation performs poorly for synthetic speech attacks. This is evident from the negative RR-EER values in many cases, which suggest that the EW-modulated features degrade detection performance compared to their non-modulated counterparts. For instance, the β-order LFCC-EW feature exhibits an RR-EER of -107.73%, indicating a significant reduction in EER performance for detecting TTS and VC attacks.

This decline in performance occurs because EW modulation is primarily designed to capture energy variations introduced by playback devices in replay attacks. In synthetic speech, where no playback or recording device is involved, EW modulation cannot effectively differentiate between bonafide and spoofed utterances. Synthetic speech often lacks the energy distribution characteristics present in replay attacks, thus making EW-based detection less effective for this type of spoofing. Consequently, while EW modulation improves the detection of replay attacks, it is not suitable for synthetic spoofing detection.

V. CONCLUSION

In this paper, we have proposed a β-order Energy-Weighted (EW) modulation for replay feature enhancement. The β-order EW method addresses the challenge of capturing energy variations induced by different recording and playback devices and environments, offering a more adaptive mechanism for feature extraction. This modulation resulted in the development of three new features: CEOC, MFECC, and LFECC, each operating in distinct spectral domains. These features, work together with both ResNet and DNN systems, were extensively evaluated the effectiveness of the proposed modulation across multiple benchmark datasets, including ASVspoof2017V2, BTAS2016PA, ASVspoof2019PA, and ASVspoof2021PA.

The β-order EW modulation played a pivotal role in enhancing the performance of both the ResNet and DNN-based detection systems. Our results show that this modulation technique consistently improved the robustness of replay speech detection by allowing the system to dynamically adapt the value of β to various types of spoofing attacks. Moreover, while the classifier type influences overall performance, we observed that the optimal β value for EW modulation is more dependent on the feature type than the classifier. This insight simplifies the system optimization process, as the same β value can be applied for both ResNet and DNN classifiers once determined for a specific feature.

Performance comparisons across the datasets revealed that CEOC-ResNet and CEOC-DNN excelled on real-world replay scenarios, particularly on ASVspoof2017V2 and BTAS2016PA datasets. On the other hand, MFECC and LFECC, derived from Mel and linear spectral domains, respectively, were more effective in simulated replay conditions, as observed on ASVspoof2019PA. This finding underscores the critical impact of spectral domain choice when designing features for replay attack detection.

REFERENCES

[1] C. H. You, K. A. Lee, and H. Li, "An svm kernel with gmm-supervector based on the bhattacharyya distance for speaker recognition," *IEEE Signal Process. Lett.*, vol. 16, pp. 49–52, 2009.

[2] T. Stafylakis, M. J. Alam, and P. Kenny, "Text-dependent speaker recognition with random digit strings," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, pp. 1194–1203, 2016.

[3] R. Tong, B. Ma, K. A. Lee, C. H. You, D. Zhu, T. Kinnunen, H. Sun, M. Dong, E. S. Chng, and H. Li, "Fusion of acoustic and tokenization features for speaker recognition," in *Chinese Spoken Language Processing: 5th International Symposium (ISCSLP)*, 2006, pp. 566–577.

[4] J. Thienpondt, N. Madhu, and K. Demuynck, "Margin-mixup: A method for robust speaker verification in multi-speaker audio," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2022, pp. 1–5.

[5] K. A. Lee, A. Larcher, H. Thai, B. Ma., and H.Li, "Joint application of speech and speaker recognition for automation and security in smart home," in *12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2011, pp. 3317–3318.

TABLE VIII: Performance in EER(%) and t-DCF for proposed EW-based features on ASVspoof2019LA corpus.

Domains	Features	β	ResNet			DNN		
			EER	RR-EER	t-DCF	EER	RR-EER	t-DCF
Octave	CQ-OST	0.0	7.70		0.231	8.13		0.188
	COEC	0.8	10.14	-31.69	0.292	9.36	-15.13	0.206
		1.0	8.72	-13.25	0.254	8.48	-4.31	0.198
		1.2	7.77	-0.91	0.224	8.30	-2.09	0.200
Mel	MFCC	0.0	6.64		0.154	8.02		0.207
	MFECC	0.8	9.75	-46.84	0.199	7.72	3.74	0.200
		1.0	9.54	-43.67	0.195	7.64	4.74	0.200
		1.2	9.82	-47.89	0.208	7.69	4.11	0.200
Linear	LFCC	0.0	2.33		0.070	10.30		0.252
	LFECC	0.8	4.92	-111.16	0.138	10.41	-1.07	0.285
		1.0	4.84	-107.73	0.131	10.52	-2.14	0.283
		1.2	5.46	-134.33	0.151	10.65	-3.40	0.290

[6] R. K. Das, S. Jelil, and S. R. M. Prasanna, "Development of multi-level speech based person authentication system," *Journal of Signal Processing Systems*, vol. 88, no. 3, pp. 259–271, Sep 2017.

[7] S. Jelil, A. Shrivastava, R. K. Das, S. R. M. Prasanna, and R. Sinha, "SpeechMarker: A voice based multi-level attendance application," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria, 2019, pp. 3665–3666.

[8] Z. Wu, N. Evans, and J.Yamagishi T.Kinnunen, F.Alegre, and H. Li, "Spoofing and countermeasure for speaker verification: a survey," *Speech communication,*, vol. 66, pp. 130–153, 2015.

[9] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," in *21st Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2020, pp. 4213–4217.

[10] T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamaki, D. Thomsen, A. Sarkar, Z.-H. Tan, H. Delgado, M. Todisco, N. Evans, V. Hautamaki, and K. A. Lee, "RedDots replayed: a new replay spoofing attack corpus for text-dependent speaker verification research," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2017, pp. 5395–5399.

[11] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection," in *Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 2–6.

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

[13] H. Zen, M. J. F. Gales, Y. Nankaku, and K. Tokuda, "Product of experts for statistical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 794–805, 2012.

[14] M. Shannon, H. Zen, and W. Byrne, "Autoregressive models for statical parametric speech synthesis," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 587–597, 2013.

[15] Daniel Erro, Asuncion Moreno, and Antonio Bonafonte, "Voice conversation based on weighted frequency warping," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 5, pp. 922–931, 2010.

[16] D. Erro, E. Navas, and I. Hernaez, "Parametric voice conversation based on bilinear frequency warping plus amplitude scaling," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 21, no. 3, pp. 556–566, 2013.

[17] X. Tian, S. Lee, Z. Wu, E. S. Chng, and H. Li, "An example-based approach to frequency warping for voice conversation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 10, pp. 1863–1875, 2017.

[18] K. A. Lee, A. Larcher, H. Thai, B. Ma., and H. Li, "Joint application of speech and speaker recognition for automation and security in smart home," in *12th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Lyon, France, 2011, pp. 3317–3320.

[19] R. G. Hautamaki, T. Kinnunen, V. Hautamaki, and A.-M. Laukkanen, "Automatic versus human speaker verification:the case of voice mimicrys," *Speech Communication,*, vol. 72, pp. 13–31, 2015.

[20] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudasher, and V. Shchemelinin, "Audio replay attack detection with deep learning framework," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 82–86.

[21] F. Tom, M. Jain, and Prasenjit Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 681–685.

[22] K. Sriskandaraja, V. Sethu, and E. Ambikairajah, "Deep siamese architecture based replay detection for secure voice biometric," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 671–675.

[23] G. Suthokumar, V. Sethu, C. Wijenayake, and E. Ambikairajah, "Modulation dynamic features for the detection of replay attacks," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 691–695.

[24] B. Wickramasinghe, S. Irtza, E. Ambikairajah, and J. Epps, "Frequency domain linear prediction features for replay spoofing attack detection," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 661–665.

[25] M. S. Saranya and H. Murthy, "Decision-level feature switching as a paradigm for replay attack detection," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 686–690.

[26] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients," in *Speaker and Language Recognition Workshop (ODYSSEY)*, 2016, pp. 283–290.

[27] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic Speaker verification," *Computer Speech & Language,*, vol. 45, pp. 516–535, 2017.

[28] Y. Lou, S. Pu, J. Zhou, X. Qi, Q. Dong, and H. Zhou, "A deep one-class learning method for replay attack detection," in *23rd Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2022, pp. 4765–4769.

[29] R. K. Das, J. Yang, and H. Li, "Long range acoustic and deep features perspective on ASVspoof 2019," in *Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2019, pp. 1018–1025.

[30] T. Gunendradasan, B. Wickramasinghe, N. P. Le, E. Ambikairajah, and J. Epps, "Detection of replay-spoofing attacks using frequency modulation features," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 636–640.

[31] M. Kamble, H. Tak, and H. A. Patil, "Effectiveness of speech demodulation-based features for replay detection," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 641–645.

[32] H. B. Sailor, M. R. Kamble, and H. A. Patil, "Auditory filterbank learning for temporal modulation features in replay spoof speech detection," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 666–670.

[33] Z. Chen, Z. Xie, W. Zhang, and X. Xu, "ResNet and model fusion for automatic spoofing detection," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 102–106.

[34] H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. Lee, and J. Yamagishi, "ASVspoof 2017 version 2.0: meta-data analysis and baseline enhancements," in *Speaker and Language Recognition Workshop (ODYSSEY)*, 2018, pp. 296–303.

[35] C. Kim and R. M. Stern, "Power-normalized cepstral coefficients (PNCC) for robust speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, pp. 1315–1329, 2016.

[36] J. Yang, C. You, and Q. He, "Feature with complementarity of statistics and principal information for spoofing detection," in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2018, pp. 651–655.

[37] J. Yang and R. K. Das, "Improving anti-spoofing with octave spectrum and short-term spectral statistics information," *Applied Acoustic*, vol. 157, 2020.

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

[39] Prasad A. Tapkir, Madhu R. Kamble, Hemant A. Patil, and Maulik Madhavi, "Replay spoof detection using power function based features," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018.

[40] J. Yang, R. K. Das, and N. Zhou, "Extraction of octave spectra information for spoofing attack detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, pp. 2373–2384, 2019.

[41] C.H. You, S.N. Koh, and S. Rahardja, "β-Order MMSE Spectral Amplitude Estimation for Speech Enhancement," *IEEE Trans. Speech and Audio Processing*, Vol. 13, No. 4, pp. 475-486, Jul. 2005.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[43] Frank Seide and Amit Agarwal, "CNTK:Microsoft's open-source deep learning toolkit," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 2135–2135.

[44] H. A. Patil, M. R. Kamble, T. B. Patel, and M. Soni, "Novel variable length teager engery separation based on instantaneous frequency features for replay detection," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 12–16.

[45] Anupama Paul, Rohan Kumar Das, Rohit Sinha, and S.R. Mahadeva Prasanna, "Countermeasure to handle replay attacks in practical speaker verification systems," in *International Conference on Signal Processing and Communications*, 2016, pp. 12–15.

[46] S. Jelil, R. K. Das, S. R. M. Prasanna, and R. Sinha, "Spoof detection using source, instantaneous frequecny and cepstral features," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 22–26.

[47] R. K. Das and H. Li, "Instantaneous phase and excitation source features for detection of replay attacks," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1030–1037.

[48] R. K. Das, J. Yang, and H. Li, "Long range acoustic features for spoofed speech detection," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 1058–1062.

[49] J. Yang and R. K. Das, "Low frequency frame-wise normalization over constant-Q transform for playback speech detection," *Digital Signal Processing*, vol. 80, pp. 30–39, 2019.

[50] W. Cai, H. Wu, D. Cai, and M. Li, "The DKU replay detection system for the ASVspoof 2019 challenge on data augmentation, featurerepresentation, clssifier, and fusion," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 1023–1027.

[51] M. Alzanto, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 1078–1082.

[52] P. Nagarsheth, E. Khoury, K. Patil, and M. Garlandi, "Replay attack detection using DNN for channel discrimination," in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2017, pp. 97–101.

[53] Y. Yang, H. Wang, H. Dinkel, Z. Chen, S. Wang, Y. Qian, and K. Yu, "The SJTU robust anti-spoofing system for the ASVspoof 2019 challenge," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 1038–1042.

[54] L. Xu, J. Yang, C. You, X. Qian, and D. Huang, "Device features based on linear transformation with parallel training data for replay speech detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 31, pp. 1574–1586, 2023.

[55] S. K. Ergunay, E. K., A. Lazaridis, and S. Marcel, "On the vulnerability of speaker verification to realistic voice spoofing," in *Proceedings of the seventh IEEE International Conference on Biometric Theory, Application ans Systems (BTAS)*, 2015, pp. 1–6.

[56] H. Wang, H. Dinkel, S. Wang, Y. Qian, and K. Yu, "Cross-domain replay spoofing attack detection using domain adversarial training," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, Graz, Austria, 2019, pp. 2938–2942.

[57] Z. Chen, W. Zhang, Z. Xie, X. Xu, and D. Chen, "Recurrent neural networks for automatic replay spoofing attack detection," in *IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP)*, 2018, pp. 2052–2056.

[58] M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evas, T. Kinnunen, and K. A. Lee, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 1008–1012.

[59] J. Yamagishi, X. Wang, M. Todisco, and etc., "Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection," in *ISCA 2021 edition of ASVspoof*, 2021, pp. 47–53.

[60] A. Nautsch, X. Wang, N. Evans, T. H. Kinnunen, V. Vestman, M. Todisco, H. Delgado, M. Sahidullah, J. Yamagishi, and K. A. Lee, "ASVspoof 2019: Spoofing countermeasures for the detection of synthesized, converted and replayed speech," *IEEE Transactions on Biometrics, Behavior and Idemtity Science*, vol. 3, pp. 252–265, 2021.

[61] "Superseded - cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit," *http://dx.doi.org/10.7488/ds/1994*.

[62] T. Kinnunen, K. A. Lee, H. Delgado, N. Evans, M. Todisco, M. Sahidullah, Junichi Yamagishi, and Douglas. A. Reynolds, "t-DCF: a detection cost function for tandem assessment of spoofing countermeasures and automatic speaker verification," in *Speaker and language recognition workshop (ODYSSEY)*, 2018, pp. 312–319.

[63] K. Vijayan, P. Raghavendra Reddy, and K. S. R. Murty, "Significance of analysis phase of speech signals in speaker verification," *Speech Communication*, vol. 81, pp. 54–71, 2016.

[64] J. Yang and L. Liu, "Playback speech detection based on magnitude-phase spectrum," *Electronics Letters*, vol. 54, no. 14, pp. 901–903, 2018.

[65] K N R K R. Alluri and A. K. Vupala, "IIIT-H spoofing countermeasures for automatic speaker verification spoofing and countermesures challenge 2019," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 1043–1047.

[66] G. Lavrentyva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlos, "STC antispoofing systems for the ASVspoof2019 challenge," in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2019, pp. 1033–1037.

[67] A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, and A. M. Gomez, "A gated recurrent convolutional neural network for robust spoofing detection," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 27, pp. 1985–1999, 2019.

[68] X. Wang, X. Qin, T. Zhu, C. Wang, S. Zhang, and M. Li, "The DKU-CMRI system for the asvspoof 2021 challenge: vocoder based replay channel response estimation," in *ISCA 2021 edition of ASVspoof*, 2021, pp. 16–21.

[69] H. Tak, J. Patino, M. Todisco, A. Nautsch, N. Evans, and A. Larcher, "End-to-end anti-spoofing with RawNet2," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 6369–6373.