Modified Magnitude-Phase Spectrum Information for Spoofing Detection

Jichen Yang, *Senior Member, IEEE*, Hongji Wang, Rohan Kumar Das, *Senior Member, IEEE*, and Yanmin Qian, *Senior Member, IEEE*

***Abstract*—Most of the existing feature representations for spoofing countermeasures consider information either from the magnitude or phase spectrum. We hypothesize that both magnitude and phase spectra can be beneficial for spoofing detection (SD) when collectively used to capture the signal artifacts. In this work, we propose a novel feature referred to as modified magnitude-phase spectrum (MMPS) to capture both magnitude and phase information from the speech signal. The constant-Q transform is used to obtain the magnitude and phase information in terms of MMPS, which can be denoted as CQT-MMPS. We then use this information for the proposal of a handcrafted feature, namely, constant-Q modified octave coefficients (CQMOC). To evaluate the proposed CQT-MMPS and CQMOC features, three classic anti-spoofing models are adopted, including the Gaussian mixture model (GMM), the light CNN (LCNN) and the ResNet. Additionally, since there is usually no prior knowledge about the spoofing kind in real-world applications, two novel methods referred to as three-class classifiers with maximum spoofing-score (TCMS) and multi-task learning (MTL) are designed for unknown-kind SD (UKSD). The experimental results on ASVspoof 2019 corpus show that CQMOC outperforms most of the commonly-used handcrafted features, and the CQT-based MMPS performs better than the magnitude-phase spectrum and the commonly-used log power spectrum. Further, the MMPS-based systems can achieve comparable or even better performance when compared with the state-of-the-art systems. We find that the newly-designed TCMS and MTL methods outperform the combination-based method for UKSD and meanwhile, generalize much better than the respective-kind-based methods in cross-spoofing-kind evaluation scenarios.**

***Index Terms*—Constant-Q modified octave coefficients, modified magnitude-phase spectrum, unknown-kind spoofing detection.**

I. Introduction

AUTOMATIC speaker verification (ASV) aims to accept or reject an identity claim with reference to a person's voice samples [1]–[4]. Although the research on ASV has witnessed success for practical systems, they are vulnerable to various spoofing attacks [5]. There are four broad categories of spoofing attacks, which are replay [6]–[8], text-to-speech (TTS) [9]–[11], voice conversion (VC) [12]–[14], and impersonation [15]. Due to the lack of a standard database, impersonation attacks have received less attention for spoofing detection research. In this paper, we focus on the remaining three spoofing types.

Most of the spoofing detection (SD) frameworks have a front-end feature extraction module followed by a module of back-end classifier. Various works on SD either focus on investigating novel acoustic cues for front-end feature extractor [16]–[22] or emphasize on designing effective classifiers and neural-network-based systems [23]–[26]. Literature shows that most of the front-end modules for SD consider features derived from the power spectrum. Some of these include mel frequency cepstral coefficients (MFCC) [27]–[29], rectangular filter cepstral coefficients [30], inverted MFCC [31] (IMFCC), Gammatone filter bank cepstral coefficients (GFCC) and inverted GFCC (IGFCC) [32]. Additionally, mel-warped overlapped block transformation, inverted speech-based-signal overlapped block transformation, speech-signal frequency cepstral coefficients and inverted speech-signal frequency cepstral coefficients are also studied for SD [33]. Note that it has been shown that cepstral features based on inverted filter banks can perform better than the corresponding features based on filter banks in synthetic speech detection [29], [32], for example, IMFCC (IGFCC) outperforms MFCC (GFCC).

The features mentioned above are obtained using discrete Fourier transform (DFT), which transforms the signal from the time domain into the frequency domain. In contrast to this, a novel feature based on constant-Q transform (CQT) referred to as constant-Q cepstral coefficients (CQCC) is proposed in [19], [20]. The CQT benefits the CQCC feature to have a better time resolution in higher frequency regions, as well as a better frequency resolution in lower frequency regions. Further, CQCC

Manuscript received June 5, 2020; revised November 18, 2020 and February 16, 2021; accepted February 16, 2021. Date of publication February 22, 2021; date of current version March 10, 2021. This work was supported in part by the Programmatic Grant A1687b0033 through the Singapore Government's Research, Innovation, and Enterprise 2020 Plan (Advanced Manufacturing and Engineering domain), in part by the National Research Foundation Singapore through the AI Singapore Programme under Award AISG-100E-2018-006, in part by the Human-Robot Interaction Phase 1 under Grant 1922500054, and in part by the National Research Foundation, Prime Minister's Office, Singapore under the National Robotics Programme, and by the China NSFC Projects under Grants 62071288 and U1736202. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Zheng-Hua Tan. *(Jichen Yang and Hongji Wang are co-first authors.) (Corresponding author: Yanmin Qian.)*

Jichen Yang is with the School of Electronic and Information Engineering, South China University of Technology, Guangzhou 510641, P. R. China (e-mail: NisonYoung@163.com).

Hongji Wang and Yanmin Qian are with the SpeechLab, Department of Computer Science and Engineering and MoE Key Laboratory of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai 200240, P. R. China (e-mail: jijijiang77@sjtu.edu.cn; yanminqian@sjtu.edu.cn).

Rohan Kumar Das is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore 117583, Singapore (e-mail: rohankd@nus.edu.sg).

This article has supplementary downloadable material available at https://doi.org/10.1109/TASLP.2021.3060810, provided by the authors.

Digital Object Identifier 10.1109/TASLP.2021.3060810

outperforms many commonly-used power spectrum based features based on DFT [19], [20]. Inspired by the success of CQCC, two novel subband features from the octave and liner subband power spectrum are proposed in [34]. These studies showed that subband features could significantly improve the performance on SD.

Regarding the nature of information captured by the features, it can be seen that they are derived from the magnitude spectrum. In [18], two magnitude-spectrum-based features, namely log-magnitude spectrum and residual log-magnitude spectrum, are investigated for synthetic speech detection. As the power spectrum is the square of the magnitude spectrum, they can be regarded under the same category with magnitude information. Apart from such handcrafted features, many neural-network-based anti-spoofing systems also use features with magnitude information as the input. For instance, log power spectrum (LPS) is used as the input to the neural networks [35]–[41]. Additionally, the power spectrum is also considered as the input to deep Siamese networks previously [42].

Apart from the features derived from the magnitude spectrum, some features based on the phase information have also been used for SD. Group delay, modified group delay [43], modified group delay function (MODGDF) [44], instantaneous phase [45], [46], linear prediction residual phase features [47], instantaneous frequency derivative, baseband phase difference and pitch synchronous phase [18] are mentioned to be a few of them. Besides, cosine normalized phase (Cosphase) feature [48] and all-pole group delay function (APGDF) [49] were investigated for synthetic speech detection on the corpus of ASVspoof 2015 [50] in [29]. Among which, APGDF usually can give good performance because of its resemblance with spectral characteristics [49]. These works signify the scope of phase features for anti-spoofing.

Most of the works show that although the phase-based features may not be as well discriminative as the magnitude-based features, combining phase and magnitude information can help to improve the performance of spoofing detection. For example, in [40], the authors consider the phase and magnitude spectrum together as input to the anti-spoofing model, which could enhance performance for replay attack detection. They contribute this improvement to the complementary information contained in the magnitude and phase spectrum. The magnitude and phase spectrum are used together for many other speech processing tasks as well. However, most of the combinations are performed by developing separate systems for magnitude-based and phase-based features and then combine at the score level [45], [46].

To the best of our knowledge, there are very few attempts to extract features from both magnitude and phase spectra at the same time because it is difficult to group magnitude and phase information together. A few neural-network-based systems apply group delay gram [43] as the input of the networks [39], [51]. Though group delay gram has both magnitude and phase information, its phase information is obtained by time delay operation, which is unlike the magnitude information directly obtained from the transform such as DFT or CQT. As a result, group delay gram is not widely used like LPS in front-end feature extraction. This motivates us to find an effective way to collectively capture magnitude and phase information.

In this paper, we investigate CQT based features for SD. We believe that the magnitude and phase information derived from the CQT can be more useful and effective if captured collectively. Our previous work in [52] attempted to capture the magnitude and phase spectrum information collectively from CQT. However, it resulted in representations of all positive values due to consideration of magnitude over log magnitude-phase spectrum. We hypothesize that our previous representation of the magnitude-phase spectrum (MPS) can greatly benefit if we preserve the sign of the magnitude part.

Motivated by this, we propose a novel way of capturing both magnitude and phase information collectively that we refer to as modified magnitude-phase spectrum (MMPS), which extends our previous work on MPS [52]. The differences between MMPS and MPS are as follows:

- MMPS is obtained by modifying MPS, which preserves the sign of the magnitude part because magnitude information could play a more important role than phase information at most cases in SD.
- The values of MMPS can be either positive or negative, while that of MPS are always positive.

We then use the MMPS obtained from CQT to derive a novel handcrafted feature and apply them for SD. Specifically, this handcrafted feature is derived by combining CQT-MMPS and octave subband transform [34]. Therefore, we refer to it as constant-Q modified octave coefficients (CQMOC). We will formulate MMPS and CQMOC with detail in Section II.

Traditional SD is often modeled as a binary classification task. Further, logical access (LA) and physical access (PA) attack detection are regarded as two different tasks. The reason behind this is that there is much difference between LA and PA attacks. Specifically, LA attacks are derived logically using speech processing methods that involve various TTS or VC algorithms and vocoders, while PA attacks are just replayed versions of some genuine examples, which differ from the genuine speech due to device characteristics and recording environments. In this work, we regard LA and PA attacks as two different kinds of spoofing attacks. Since the spoofing kind is known in advance, traditional SD can be regarded as known-kind SD (KKSD).

However, we usually have no prior knowledge about the kind of spoofing attack in practice, which is different from the previous ASVspoof challenge series where we have known the spoofing kind in advance. In this paper, we term this task as unknown-kind SD (UKSD). In order to solve this problem, there is a requirement to assess the scope of generalized countermeasure for UKSD [53] and design spoofing detectors that can achieve promising performance for both spoofing attack kinds [54]. In [53], we proposed a generalized countermeasure for the UKSD by combining LA and PA as the spoofed class, which can be named as the combination-based method. However, we found that the performance of the combined-based method can not be satisfied because there is much difference between LA and PA attacks, as mentioned above. Thus, we believe that it would be better if anti-spoofing models can well

discriminate between LA and PA attacks rather than simply regarding them as a spoofed class.

To address the UKSD problem in real-world applications, two methods referred to as three-class classifiers with maximum spoofing-score (TCMS) and multi-task learning (MTL) for UKSD are proposed to improve neural-network-based anti-spoofing models in this work, which extend our previous work on the combination-based method [53]. Herein, the TCMS method is also inspired by the work of multi-class classifiers used for synthetic speech detection in [32]. The differences among the combination-based method, TCMS and MTL are as follows:

- The combination-based method regards LA and PA attacks as one class (spoofed class), while TCMS separates them into two different classes. In addition, MTL has another task to learn how to discriminate between them.
- TCMS applies a three-class output layer, including the bonafide, replay, and synthetic nodes. Also, the maximum spoofing-score strategy is adopted to compute the final score for a test utterance.
- MTL retains the two-class output layer (bonafide or spoofed) in the model, and further add a new branch with two-class outputs (LA and PA attacks). In other words, there are two two-class output layers in the MTL framework. One is to predict whether a test utterance is spoofed or not, and the other is to predict the spoofing kind for a spoofing attack.

To evaluate the proposed MMPS and CQMOC features for SD as well as the proposed TCMS and MTL methods for UKSD, we conduct all experiments on the recent ASVspoof 2019 corpus that includes both LA and PA attacks.

The contributions of this work can be summarized as below:

- Proposal of a novel feature, namely MMPS, to jointly capture magnitude and phase information for SD
- The use of CQT-MMPS to derive a novel handcrafted feature CQMOC that captures both magnitude and phase information
- Proposal of TCMS and MTL for UKSD

The remainder of this paper is organized as follows. Firstly, Section II introduces the proposed MMPS feature, based on which we propose a handcrafted feature CQMOC. Then Section III introduces the commonly-used anti-spoofing models for KKSD as well as our proposed TCMS and MTL methods for UKSD. Afterwards, Section IV and V evaluate the performance of KKSD and UKSD, respectively. Finally, this work is summarized and concluded in Section VI.

II. Modified Magnitude-Phase Spectrum

In this section, we present the details of extracting MMPS. Afterwards, we propose a novel handcrafted feature CQMOC based on CQT-MMPS.

A. Modified Magnitude-Phase Spectrum

For a given audio signal $x(n)$, its corresponding frequency domain signal can be obtained by using CQT or DFT, which can be written as:

$$X(\omega) = |X(\omega)|e^{j\phi(\omega)} \tag{1}$$

where $|X(\omega)|$ and $\phi(\omega)$ represent the magnitude spectrum and phase spectrum of $x(n)$, respectively. Herein, $\phi(\omega)$ can be obtained by computing the arctangent of imaginary part of $X(\omega)$ to real part of $X(\omega)$ ratio. The values of $\phi(\omega)$ are wrapped between -π and π, thus $\phi(\omega)$ can be regarded as a wrapped phase.

Then we can obtain Eq. (1) in log-scale with base-e:

$$\begin{aligned}\ln(X(\omega)) &= \ln(|X(\omega)|e^{j\phi(\omega)}) \\ &= \ln(|X(\omega)|) + j\phi(\omega)\ln(e) \\ &= \ln(|X(\omega)|) + j\phi(\omega)\end{aligned} \tag{2}$$

where $\ln|X(\omega)|$ denotes the log magnitude spectrum (LMS) with base-e.

The module of Eq. (2) is as follows:

$$\begin{aligned}|\ln(X(\omega))| &= |\ln|X(\omega)| + j\phi(\omega)| \\ &= \sqrt{(\ln(|X(\omega)|))^2 + \phi(\omega)^2}\end{aligned} \tag{3}$$

where $|\ln(X(\omega))|$ represents the MPS of $x(n)$, i.e., MPS$(x(n))$. We also can write it as:

$$\text{MPS}(x(n)) = \sqrt{(\ln(|X(\omega)|))^2 + \phi(\omega)^2} \tag{4}$$

From Eq. (4), it can be seen that MPS$(x(n))$ contains two parts: the magnitude part ($\ln(|X(\omega)|)$) and the phase part ($\phi(\omega)$). Considering the fact that magnitude information could play a more important role than phase information at most cases in SD, we modify MPS by preserving the sign of the magnitude part (i.e., $\ln(|X(\omega)|)$) to obtain MMPS. Further, for $x(n)$, its MMPS can be formulated as:

$$\text{MMPS}(x(n)) = \text{sgn}(\ln(|X(\omega)|))\sqrt{(\ln(|X(\omega)|))^2 + \phi(\omega)^2} \tag{5}$$

where sgn$(\cdot)$ represents the sign function.

B. Constant-Q Modified Octave Coefficients

We propose a novel handcrafted feature CQMOC using MMPS derived from CQT (i.e., CQT-MMPS). Fig. 1 illustrates the diagram of CQMOC extraction. As observed from Fig. 1, firstly, CQT is applied to the speech signal, and then the proposed MMPS is derived to capture the magnitude-phase information. As MMPS is obtained from CQT, its frequency bin characteristics resemble an octave scale. In other words, every frequency bin has different bandwidth, and the former octave bandwidth is one half of the latter octave bandwidth. Accordingly, we use the octave subband transform (OST) [34] that considers octave scale based subbanding followed by discrete cosine transform (DCT) on it. We note that our previous work on OST [34] showed improved results while considering such subband-based features over the full frequency band features for SD. Therefore, we apply the same strategy (OST) into the subbands of the proposed CQT-MMPS to derive more spoofing relevant information for SD.

In OST, the octave subbanding is used to segment the full band into subbands of octave 1, octave 2, . . . , octave V (V represents

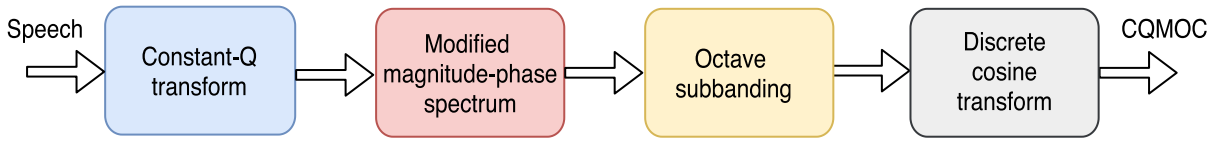


Fig. 1. Schematic diagram of constant-Q modified octave coefficients (CQMOC) extraction.

the number of octaves). DCT is then used to extract every octave subband spectral principal information. Afterwards, the top P coefficients of the DCT result are selected to form the final feature.

According the definition of CQT in [55], we can obtain the CQT for $x(n)$, denoted as $Y(k,n)$. Considering $MMPS_Y$ as the MMPS of $Y(k,n)$, we now explain the steps to derive the corresponding CQMOC feature $CQMOC_Y$ in detail.

First, $MMPS_Y$ is segmented into octave subbands according to the octave scale, which is as follows:

$$MMPS_Y = \left\{ MMPS_{Y_1}, MMPS_{Y_2}, \ldots, MMPS_{Y_V} \right\} \tag{6}$$

where $MMPS_{Y_1}, MMPS_{Y_2}, \ldots, MMPS_{Y_V}$ are the 1-st, 2-nd, $\ldots$, V-th octave subband of $MMPS_Y$, respectively.

Then, DCT is applied on $MMPS_{Y_1}$, $MMPS_{Y_2}$, $\ldots$, $MMPS_{Y_V}$, respectively, which is as follows:

$$F_1(p) = \sum_{k=0}^{B-1} MMPS_{Y_1} \cos\left\{ \frac{(k+\frac{1}{2})p\pi}{B} \right\} \tag{7}$$

$$F_2(p) = \sum_{k=B}^{2B-1} MMPS_{Y_2} \cos\left\{ \frac{(k+\frac{1}{2})p\pi}{B} \right\} \tag{8}$$

$$\cdots\cdots\cdots$$

$$F_V(p) = \sum_{k=(V-1)\times B}^{K-1} MMPS_{Y_V} \cos\left\{ \frac{(k+\frac{1}{2})p\pi}{B} \right\} \tag{9}$$

where $F_1(p)$, $F_2(p)$, $\ldots$, and $F_V(p)$ are the DCT results on the 1-st, 2-nd, $\ldots$, and V-th octave subband, respectively. In addition, p can be chosen from 0, 1, 2, $\ldots$, P-1.

Finally, as shown in Eq. (10), $CQMOC_Y$ can be obtained by concatenating $F_1(p), \ldots, F_V(p)$, and p is ranging from 0 to P-1.

$$CQMOC_Y = \left\{ F_1(0), \ldots, F_1(P\text{-}1), \ldots\ldots\ldots, F_V(0), \ldots, F_V(P\text{-}1) \right\} \tag{10}$$

Note that if the module of MMPS in Fig. 1 is replaced by LPS, the obtained feature is constant-Q transform octave subband transform (CQ-OST) [34].

III. Anti-Spoofing Models for KKSD and UKSD

In this section, we first introduce the back-end anti-spoofing models that are used to evaluate our proposed CQT-MMPS and CQMOC for KKSD. Herein, KKSD contains synthetic as well as replay speech detection. Afterwards, we will elaborate on considered neural-network-based models for UKSD. Specifically, we propose the TCMS and MTL methods for UKSD.

A. Anti-Spoofing Models for KKSD

In this work, we adopt three classic anti-spoofing models to evaluate our proposed methods. One is the frame-level Gaussian mixture model (GMM), the others are the utterance-level models: the light CNN (LCNN) and the ResNet.

1) GMM: GMM is widely used as the back-end classifier for SD [17], [19], [29], [56]. Meanwhile, the GMM-based systems are also the official baseline systems in the ASVspoof challenge series. Hence, we consider it for the study of the proposed handcrafted feature CQMOC in this work. We refer to GMM as a frame-level model because it is trained using speech frames instead of speech segments or utterances.

Given two GMMs trained on bonafide and spoofed speech examples, respectively, denoted as λ_b and λ_s, we can obtain the score prediction of an input feature $CQMOC_Y$ based on their log-likelihood ratios:

$$\text{Score}(CQMOC_Y|\lambda_b,\lambda_s) = \log(CQMOC_Y|\lambda_b) - \log(CQMOC_Y|\lambda_s) \tag{11}$$

2) LCNN: The (9-layer) LCNN was the best system in ASVspoof 2017 [35], where a Max-Feature-Map (MFM) activation is used after each convolution (Conv) operation. It also performed well in ASVspoof 2019 [36], [37]. The MFM activation function is defined as:

$$\hat{y}_{ij}^{k} = \max(y_{ij}^{k}, y_{ij}^{k+\frac{F}{2}}) \tag{12}$$

where y is the input tensor of size $F \times H \times W$ and $\hat{y}_{ij}^{k}$ is the output tensor of size $\frac{F}{2} \times H \times W$. In addition, i and j represent the indices in time domain and frequency domain, respectively, and k is the filter index ranging from $[1, \ldots, \frac{F}{2}]$.

To further enhance modelling capability of LCNN, we adopt the 29-layer structure in this work. The details of the 29-layer LCNN architecture are described in Table I. Similar to our previous work [57], we apply global average pooling (GAP) in the time dimension after all convolution operations, making this model apply to various lengths of input features. Besides, since the ceiling mode is used in all max-pooling layers, it can be applied to very short utterances with less than 16 frames. All modules before MFM_FC1 are defined as the feature extractor that learns deep spoofing embeddings. In addition, the fully-connected (FC) FC2 and FC3 layers compose the spoofing detector that maps the embeddings into spoofing labels (bonafide or spoofed). To avoid over-fitting, we use dropout layers with a 0.5 ratio in both FC2 and FC3 layers.

After training the LCNN model, we can compute the score of a test utterance, which is the difference between the bonafide node and the spoofed node. For example, if the input feature is

TABLE I
THE ARCHITECTURE OF LCNN29 MODEL

Layer	Filter Size /Stride,Pad	Output Size	#Params
Conv1	$5 \times 5/1, 2$	$T \times 84 \times 64$	1.6K
MFM1	-	$T \times 84 \times 32$	-
MaxPool1	$2 \times 2/2, 0$	$T/2 \times 42 \times 32$	-
Conv2_x	$\begin{bmatrix} 3 \times 3/1, 1 \\ 3 \times 3/1, 1 \end{bmatrix} \times 1$	$T/2 \times 42 \times 32$	36.8K
Conv2a	$1 \times 1/1, 0$	$T/2 \times 42 \times 64$	2.0K
MFM2a	-	$T/2 \times 42 \times 32$	-
Conv2b	$3 \times 3/1, 1$	$T/2 \times 42 \times 128$	36.8K
MFM2b	-	$T/2 \times 42 \times 64$	-
MaxPool2	$2 \times 2/2, 0$	$T/4 \times 21 \times 64$	-
Conv3_x	$\begin{bmatrix} 3 \times 3/1, 1 \\ 3 \times 3/1, 1 \end{bmatrix} \times 2$	$T/4 \times 21 \times 64$	294.9K
Conv3a	$1 \times 1/1, 0$	$T/4 \times 21 \times 128$	8.1K
MFM3a	-	$T/4 \times 21 \times 64$	-
Conv3b	$3 \times 3/1, 1$	$T/4 \times 21 \times 256$	147.4K
MFM3b	-	$T/4 \times 21 \times 128$	-
MaxPool3	$2 \times 2/2, 0$	$T/8 \times 11 \times 128$	-
Conv4_x	$\begin{bmatrix} 3 \times 3/1, 1 \\ 3 \times 3/1, 1 \end{bmatrix} \times 3$	$T/8 \times 11 \times 128$	1769.4K
Conv4a	$1 \times 1/1, 0$	$T/8 \times 11 \times 256$	32.7K
MFM4a	-	$T/8 \times 11 \times 128$	-
Conv4b	$3 \times 3/1, 1$	$T/8 \times 11 \times 128$	147.4K
MFM4b	-	$T/8 \times 11 \times 64$	-
Conv5_x	$\begin{bmatrix} 3 \times 3/1, 1 \\ 3 \times 3/1, 1 \end{bmatrix} \times 4$	$T/8 \times 11 \times 64$	589.8K
Conv5a	$1 \times 1/1, 0$	$T/8 \times 11 \times 128$	8.1K
MFM5a	-	$T/8 \times 11 \times 64$	-
Conv5b	$3 \times 3/1, 1$	$T/8 \times 11 \times 128$	73.7K
MFM5b	-	$T/8 \times 11 \times 64$	-
MaxPool4	$2 \times 2/2, 0$	$T/16 \times 6 \times 64$	-
GAP	-	384	-
FC1	-	256	98.3K
MFM_FC1	-	128	-
FC2	-	128	16.3K
FC3	-	2	0.2K
Total	-	-	3263.5K

$MMPS_Y$, the formulation is as follows:

$$\begin{aligned} &\text{Score}(MMPS_Y|\text{LCNN}) \\ &= \log\left(\mathbf{ST_2}(\text{O}_1(MMPS_Y|\text{LCNN}))\right) \\ &\quad - \log\left(\mathbf{ST_2}(\text{O}_2(MMPS_Y|\text{LCNN}))\right) \end{aligned} \tag{13}$$

where $\text{O}_1(MMPS_Y|\text{LCNN})$ and $\text{O}_2(MMPS_Y|\text{LCNN})$ represent the first output (bonafide node) and the second output (spoofed node) from the trained LCNN for $MMPS_Y$, respectively. $\mathbf{ST}_2(\cdot)$ represents the softmax transform for two-class outputs.

3) ResNet: The ResNet variations used in ASVspoof 2019 achieved great performance in both PA and LA subtasks [37], [39], [58]–[60]. In this work, we implement the ResNet architecture following the standard one as depicted in [61]. In other words, the overall model structure of ResNet (e.g., the residual block) is almost the same as that in [61]. Specifically, we adopt the ResNet18 model consisting of 8 residual blocks $\{2, 2, 2, 2\}$, which is shown in Table II with details. Due to the use of an average pooling layer, the ResNet model can also apply to various lengths of input features. Here, we define all modules before the FC1 layer as the feature extractor. Similarly, the spoofing detector only consists of the FC2 and FC3 layers, where dropout layers are also used.

TABLE II
THE ARCHITECTURE OF RESNET18 MODEL. ALL FILTER SIZES ARE SET AS 3×3

Layer	Output Size	Blocks
Conv	$T \times 84 \times 16$	-
Res1	$T \times 84 \times 16$	2
Res2	$T/2 \times 42 \times 32$	2
Res3	$T/4 \times 21 \times 64$	2
Res4	$T/8 \times 11 \times 128$	2
Reshape & Average	128	-
FC1	128	-
FC2	128	-
FC3	2	-

Similarly, to score a test utterance using a trained ResNet model, we follow the same strategy as the LCNN model, as shown in Eq. (13).

B. TCMS and MTL for UKSD

Although anti-spoofing models are proposed for KKSD with promising performance, such as the LCNN and ResNet models mentioned above, they generalize poorly when it comes to cross-spoofing-kind SD. In real-world applications, however, we have no prior knowledge about the kind of spoofing attack for a test utterance. As a result, traditional anti-spoofing systems (with a binary output) for only replay speech detection or synthetic speech detection above-mentioned can not meet the requirements to detect real-world spoofing attacks with both kinds. To solve the UKSD problem, we investigate two methods to improve the neural-network-based anti-spoofing models, named as TCMS and MTL, respectively. Our ultimate goal is to make anti-spoofing models discriminative between PA and LA attacks instead of simply regarding them as a same spoofed class. We believe this would further improve the performance for UKSD, compared with the combination-based method in [53].

1) TCMS: A straightforward idea for UKSD is to adapt the model into a three-class fashion (i.e., bonafide, replay, or synthetic speech). By preparing bonafide speech examples as well as PA and LA attacks for training, we can expect the three-class model to discriminate well among them. We hypothesize that for a well-trained three-class anti-spoofing model, the probability with respect to replay speech is probably greater than that of synthetic speech for a replay attack. In the same way, the probability with respect to synthetic speech is probably greater than replay speech for a synthetic speech. Therefore, we can induce that the maximum between the probability of replay and synthetic speech is the spoofed probability.

On the basis of the analysis mentioned above, we propose a new method referred to as three-class classifiers with maximum spoofing-score (TCMS) for UKSD. Similarly, we take the $MMPS_Y$ feature and the LCNN model as an example to formulate the maximum spoofing-score method, which is as

follows:

$$\begin{aligned}&\text{Score}(MMPS_Y|\text{LCNN}_3)\\&= \log\left(\mathbf{ST_3}(\text{O}_1(MMPS_Y|\text{LCNN}_3))\right)\\&\quad - \log\left\{\mathbf{Max}\left(\mathbf{ST_3}(\text{O}_2(MMPS_Y|\text{LCNN}_3)),\right.\right.\\&\qquad \left.\left.\mathbf{ST_3}(\text{O}_3(MMPS_Y|\text{LCNN}_3))\right)\right\} \qquad (14)\end{aligned}$$

where $\text{O}_1(MMPS_Y|\text{LCNN}_3)$, $\text{O}_2(MMPS_Y|\text{LCNN}_3)$ and $\text{O}_3(MMPS_Y|\text{LCNN}_3)$ represent the first output (bonafide node), the second output (replay node) and the third output (synthetic node) of the trained three-class LCNN (denoted as LCNN_3) for $MMPS_Y$, respectively. In addition, $\mathbf{ST}_3(\cdot)$ represents the softmax transform for three-class outputs. $\mathbf{Max}(\cdot)$ denotes the maximum operation, which can be regarded as a score normalization operation.

Similar to the LCNN model, we can easily obtain the ResNet-based framework for TCMS.

2) MTL: Another approach we propose in this work is the multi-task learning (MTL) framework. Specifically, we retain the two-class output layer (bonafide or spoofed) in the model. In addition, another branch is connected after the feature extractor, which maps the embeddings into spoofing-kind labels (PA or LA).

In this work, we investigate the MTL framework for both LCNN and ResNet models. The new branch is implemented as a duplicate copy of the spoofing detector, which is composed of the FC2 and F3 layers, as shown in Table I and Table II. We term this new branch as the spoofing-kind discriminator in this paper. The training data for the MTL framework also consist of bonafide data, LA attacks and PA attacks. For LA and PA attacks, we train the whole network, including the feature extractor, the spoofing detector, and the spoofing-kind discriminator. However, for bonafide data, the spoofing-kind discriminator is fixed and we only train the remaining parts. In the testing stage, we only use the output of the spoofing detector to compute the final score, following the same strategy in KKSD, as shown in Eq. (13).

Although SD is still modeled as a binary classification task in the MTL framework, it acquires the capability of distinguishing well between LA and PA attacks, benefitting from the spoofing-kind discriminator. We believe this would enhance performance for UKSD in real applications.

IV. KKSD Performance

In this section, we evaluate the proposed CQT-MMPS and CQMOC features on ASVspoof 2019 corpus for KKSD. Specifically, the ASVspoof 2019 LA portion is used for synthetic speech detection, while the ASVspoof 2019 PA portion is used for replay speech detection. We introduce the database and evaluation metric, as well as the experimental setup. Afterwards, the experimental results and analysis of synthetic speech detection and replay speech detection are presented, respectively.

TABLE III
SUMMARY OF THE ASVSPOOF 2019 CORPUS, WHICH INCLUDES THE ASVSPOOF 2019 LA PORTION AND THE ASVSPOOF 2019 PA PORTION

Portion	Subset	# Speakers		# Utterances	
		Male	Female	Bonafide	Spoofed
ASVspoof 2019 LA	Train	8	12	2,580	22,800
	Dev	8	12	2,548	22,296
	Eval	30	37	7,355	63,882
ASVspoof 2019 PA	Train	8	12	5,400	48,600
	Dev	8	12	5,400	24,300
	Eval	30	37	18,089	134,630

A. Database and Evaluation Metric

The ASVspoof 2019 database was released for the ASVspoof 2019 challenge [62], which is summarized in Table III. Both ASVspoof 2019 LA and ASVspoof 2019 PA portions contain three subsets: training (Train), development (Dev), and evaluation (Eval) set. They have the same number of speakers, respectively. However, it is observed that the data amount of the ASVspoof 2019 PA portion is larger than that of the ASVspoof 2019 LA portion. The spoofed data in the ASVspoof 2019 LA portion are generated using either text-to-speech synthesis or voice conversion algorithms, while the spoofed data in the ASVspoof 2019 PA portion are collected using a far more controlled simulation of replay spoofing attacks [62].

According to the ASVspoof 2019 evaluation plan, the tandem detection cost function (t-DCF) [63] and equal error rate (EER) are used as the primary and secondary evaluation metric, respectively. The EER is calculated using the scores from the countermeasure only, while t-DCF jointly considers the scores from the ASV system and the countermeasure to measure the final performance. Additionally, in the ASVspoof 2019 challenge, the ASV system is given and the ASV scores are fixed for fair comparison among all countermeasures from participators. In the same way, we use the ASV scores provided by the ASVspoof 2019 organizers to compute the t-DCF metrics in our experiments.

B. Experimental Setup

As mentioned in Section III-A, three classic anti-spoofing models are used in our systems: GMM, LCNN and ResNet. We will introduce how we set the hyper-parameters for the CQT-MMPS and CQMOC features, as well as the training procedure for these models.

1) GMM: For the GMM model, the parameters in CQT are set according to the work of [19]. For instance, the octave number V is set as 9, and the frequency bin number in each octave B is set as 96. As a result, the static dimension of CQT-MMPS is 863. In case of OST for CQMOC extraction, we follow the same parameters as our previous work [34], where P is set as 12. Thus the static dimension of CQMOC is $9 \times 12 = 108$. We further consider its delta and delta-delta coefficients, so the final feature dimension is $108 \times 3 = 324$.

In the experiments, we train two GMMs of 512 mixture components using bonafide and spoofed training samples, respectively, following the ASVspoof 2019 baseline system

TABLE IV
PERFORMANCE IN T-DCF AND EER (%) FOR THE PROPOSED CQT-MMPS AND CQMOC FEATURES ON ASVSPOOF 2019 LA CORPUS

Model	Feature	Development		Evaluation	
		t-DCF	EER	t-DCF	EER
GMM	CQMOC	0.161	4.87	0.199	7.10
LCNN	CQT-MMPS	0.064	2.00	0.176	5.99
	CQMOC	0.118	3.53	0.260	8.59
ResNet	CQT-MMPS	0.050	1.52	0.119	3.72
	CQMOC	0.084	2.51	0.197	6.49

specifications [51], [62]. Additionally, we find that GMMs with 512 mixtures can achieve considerable performance on both ASVspoof 2019 LA and PA development sets. In this work, voice activity detection (VAD) is not applied for data pre-processing because the nonspeech and boundary regions could contain discriminative features and distortions for SD, as illustrated in [36].

2) LCNN and ResNet: For the neural-network-based models, the parameters in CQT are set according to our previous work of [37]. For example, V and B are set as 7 and 12, respectively. As a result, the static dimension of CQT-MMPS is 84. Further, to extract the handcrafted CQMOC feature, we set P as 8. Thus the final dimension of CQMOC is $7 \times 8 = 56$.

Since utterance lengths differ, we pad all utterances to the maximum length by repeating their features within every batch, which enables them to be processed in parallel during the training process. Thus, the sizes of the input feature maps are $T \times 84$ for the 84-dimensional CQT-MMPS and $T \times 56$ for the 56-dimensional CQMOC, respective, where T denotes the varying feature lengths among batches. The batch size is set as 8 in the training stage. However, during the testing stage, we forward the input utterance one by one. In other words, the batch size is 1 and no padding is conducted in the testing stage. Therefore, regardless of the GPU memory limitation, both LCNN and ResNet models can be applied to test utterances with too large or short durations.

In this work, we implement both LCNN and ResNet models in PyTorch and initialize all network weights by Xavier method [64]. Further, cross-entropy loss is adopted as the loss criterion, and SGD optimizer with a momentum of 0.9 and a learning rate of 0.0001 is used during the training process. Similarly, VAD is not used here.

C. Synthetic Speech Detection

In this subsection, the proposed CQT-MMPS and CQMOC features are evaluated on ASVspoof 2019 LA portion for synthetic speech detection. We present the results and analysis as well as describe the related studies.

1) Results and Analysis: As discussed above, we use three models to evaluate the proposed features, including the frame-level model (GMM) and the utterance-level models (LCNN and ResNet). We note that GMM is not used for CQT-MMPS because it is of too high dimension (863). In this work, GMM is mainly used to evaluate the performance of handcrafted features. Table IV shows the experimental results on ASVspoof 2019 LA corpus. From Table IV, several conclusions can be obtained:

TABLE V
PERFORMANCE COMPARISON AMONG THE LPS-, MPS, AND MMPS-BASED FEATURES IN T-DCF AND EER (%) ON ASVSPOOF 2019 LA EVALUATION SET

Type	Model	Feature	t-DCF	EER
Frame-level	GMM	**CQ-OST**	**0.196**	**6.81**
		CQMOC	0.199	7.10
Utterance-level	LCNN	CQT-LPS	0.215	6.73
		CQT-MPS	0.215	7.18
		CQT-MMPS	**0.176**	**5.99**
		CQ-OST	0.270	8.78
		CQMOC	0.260	8.59
	ResNet	CQT-LPS	0.122	4.04
		CQT-MPS	0.123	4.03
		CQT-MMPS	**0.119**	**3.72**
		CQ-OST	0.198	6.64
		CQMOC	0.197	6.49

- These five systems achieve better performance on the development set than on the evaluation set consistently. The probable reason may be due to the fact that the same spoofing algorithms are shared in the training and development sets, while some variants and new spoofing algorithms are added into the evaluation set for ASVspoof 2019 LA portion.
- Considering three CQMOC-based systems, we observe that ResNet outperforms GMM on both development and evaluation sets, while LCNN performs worse than GMM on the evaluation set. This may be due to the large parameter size of the 29-layer LCNN model as well as the small training size of the ASVspoof 2019 LA corpus (25 380 samples in total).
- It can be seen that the (CQT-MMPS)-based systems significantly outperform the CQMOC-based systems, whatever LCNN or ResNet is used as the back-end model. The reason behind this is that the handcrafted CQMOC is extracted from CQT-MMPS, which could lose some spoofing-discriminative information and further cause performance degradation in SD.
- It can be found that the (CQT-MMPS)-ResNet system performs the best on both development and evaluation sets among all five systems. This reveals the high generalizability of the (CQT-MMPS)-ResNet system for synthetic speech detection.

2) Comparison With LPS and MPS: We now compare our proposed MMPS with LPS and MPS derived from CQT. The LPS is a commonly-used feature for SD, while MPS is proposed for replay spoofing detection in our previous work [52]. Herein, CQMOC and CQ-OST are derived from CQT-MMPS and CQT-LPS, respectively. Table V shows the experimental results on ASVspoof 2019 LA evaluation set. The utterance-level models, LCNN and ResNet, are used to evaluate the performance of the spectrum features (i.e., CQT-LPS, CQT-MPS, and CQT-MMPS). It is observed that CQT-MPS can obtain comparable performance with CQT-LPS, while CQT-MMPS outperforms both of them consistently in terms of t-DCF and EER. Firstly, this extends our previous work [52] and shows the effectiveness of MPS for synthetic spoofing detection. In addition, it verifies the effectiveness of MMPS by preserving the sign of the magnitude part based on MPS. By collectively capturing the magnitude

TABLE VI
PERFORMANCE COMPARISON IN T-DCF AND EER (%) WITH SOME COMMONLY-USED FEATURES ON ASVSPOOF 2019 LA EVALUATION SET

Feature	t-DCF	EER	Feature	t-DCF	EER
MFCC	0.238	8.71	CQSPIC	0.207	7.35
LFCC	0.227	8.32	CosPhase	0.541	23.98
IFCC	0.298	10.40	**APGDF**	**0.164**	**6.09**
CQCC	0.237	9.57	MODGDF	0.259	11.49
CQMOC	0.199	7.10			

TABLE VII
PERFORMANCE COMPARISON IN T-DCF AND EER (%) AMONG THE PROPOSED SYSTEMS AND SOME KNOWN SYSTEMS ON ASVSPOOF 2019 LA EVALUATION SET

Type	System	t-DCF	EER
GMM-based	CQCC-GMM [58]	0.237	9.57
	LFCC-GMM [58]	0.212	8.09
	ZTWCC-GMM [52]	0.141	6.13
LCNN-based	LFCC-LCNN [34]	0.100	5.06
	(LFCC-CMVN)-LCNN [34]	0.183	7.86
	FFT-LCNN [34]	0.103	4.53
ResNet-based	MFCC-ResNet [36]	0.204	9.33
	Spec-ResNet [36]	0.274	7.69
	CQCC-ResNet [36]	0.217	7.69
Proposed	**CQMOC-GMM**	**0.199**	**7.10**
	(CQT-MMPS)-LCNN	**0.176**	**5.99**
	CQMOC-LCNN	**0.260**	**8.59**
	(CQT-MMPS)-ResNet	**0.119**	**3.72**
	CQMOC-ResNet	**0.197**	**6.49**

and phase information, the MMPS-based systems can capture more artifacts and further perform better for synthetic speech detection.

Considering the handcrafted features, CQ-OST and CQMOC, we observe that they can achieve comparable performance on synthetic speech detection. Specifically, CQ-OST slightly outperforms CQMOC on GMM, while CQMOC outperforms CQ-OST marginally on both LCNN and ResNet. These results reveal that for these well-handcrafted features, the back-end models could play a less important role in detecting synthetic attacks.

3) Comparison With Other Commonly-Used Features: Table VI shows the comparison among our proposed CQMOC feature and some other commonly-used handcrafted features, such as MFCC, linear frequency cepstral coefficient (LFCC), instantaneous frequency cepstral coefficients (IFCC) [65], CQCC, constant-Q statistics-plus-principal information coefficients (CQSPIC) [66], CosPhase, APGDF, and MODGDF. GMM is adopted as the back-end model to evaluate the performance of these handcrafted features. From Table VI, it can be seen that our proposed feature CQMOC outperforms most of the commonly-used features except APGDF. This shows the effectiveness of the MMPS-based CQMOC feature to capture both magnitude and phase information for synthetic speech detection.

4) Comparison With Some Known Systems: Table VII compares our proposed systems with some known systems on ASVspoof 2019 LA evaluation set. The ZTWCC-GMM stands for zero time windowing cepstral coefficients with a GMM classifier [56], while FFT-LCNN represents the LCNN-based neural network system with fast Fourier transform (FFT) as the input [36]. Accordingly, LFCC-LCNN and (LFCC-CMVN)-LCNN represent the LCNN-based neural network systems using LFCC and LFCC with the cepstral mean and variance normalization (CMVN) as the input, respectively. Similarly, MFCC-ResNet, Spec-ResNet, and CQCC-ResNet represent the ResNet-based neural network systems with MFCC, DFT-based LPS, and CQCC as the inputs, respectively [38]. The results of these various existing systems are cited from the respective published works.

It is observed that our systems achieve comparable or better performance with the previous works. These show that the magnitude-phase information captured by our proposed MMPS and CQMOC features can help anti-spoofing systems to achieve better results. It should be noted that FFT-LCNN and LFCC-LCNN perform the best and the second best among all single systems on ASVSpoof 2019 LA sub-challenge. Although our proposed (CQT-MMPS)-ResNet performs slightly worse than FFT-LCNN and LFCC-LCNN in terms of t-DCF, it performs better in terms of EER. The reason behind this may be that the ASV scores are computed by the x-vector ASV model using magnitude-based features such as MFCCs and filterbanks [62], [67]. In other words, the front-end feature could be more similar and compatible in FFT-LCNN and LFCC-LCNN than in (CQT-MMPS)-ResNet, compared with that in the x-vector ASV model. If SD is viewed as a stand-alone task, the (CQT-MMPS)-ResNet system is shown to perform the best (3.72% in EER), with significant improvements compared with the other systems in Table VII. Therefore, considering both t-DCF and EER metrics, we can say that the proposed (CQT-MMPS)-ResNet system achieves comparable results with state-of-the-art systems on ASVspoof 2019 LA evaluation set.

TABLE VIII
PERFORMANCE IN T-DCF AND EER (%) FOR THE PROPOSED CQT-MMPS AND CQMOC FEATURES ON ASVSPOOF 2019 PA CORPUS

Model	Feature	Development		Evaluation	
		t-DCF	EER	t-DCF	EER
GMM	CQMOC	0.053	2.60	0.135	6.95
LCNN	CQT-MMPS	0.010	0.34	0.024	0.90
	CQMOC	0.019	0.64	0.036	1.32
ResNet	CQT-MMPS	0.014	0.50	0.031	1.08
	CQMOC	0.015	0.55	0.038	1.36

D. Replay Speech Detection

We now focus on the studies on replay speech detection using ASVspoof 2019 PA portion. Experimental results and analysis of our proposed systems are given along with their comparison to some existing systems.

1) Results and Analysis: Table VIII presents the results on ASVspoof 2019 PA corpus. The GMM, LCNN, and ResNet models are used here to evaluate our proposed features. We note that the GMM is not used for the high-dimensional CQT-MMPS feature. From Table VIII, some observations can be found:

- The performance on the development set is better than that on the evaluation set for all five systems. The reason behind this is that both bonafide and replay data in both training

and development sets are generated using the same set of randomly-selected acoustic and replay configurations. In contrast, the evaluation set has different or unknown acoustic and replay configurations that are far more diverse and challenging.

- Considering the CQMOC-based systems, we can observe that both LCNN and ResNet significantly outperform GMM, which is different from the phenomenon on synthetic speech detection, as discussed in Section IV-C1. Our explanation behind this is that the larger training size of the ASVspoof 2019 PA corpus (54 000 samples in total) can help to train the large-size LCNN and ResNet models better and avoid over-fitting, thus improving the generalization performance.
- Similarly, the (CQT-MMPS)-based systems outperform the CQMOC-based systems consistently, whatever LCNN or ResNet is used. This is also owing to the information lost in the process of extracting CQMOC from CQT-MMPS.
- It can be found that the (CQT-MMPS)-LCNN system shows the best performance on ASVspoof 2019 PA evaluation set (0.024 in t-DCF and 0.90% in EER) among all five systems, and (CQT-MMPS)-ResNet performs the second-best and comparable result (0.031 in t-DCF and 1.08% in EER). This reveals the strong modelling capability of the neural-network-based models (LCNN and ResNet) to capture the artifacts for replay speech detection.

Comparing the results between synthetic speech detection (Table IV) and replay speech detection (Table VIII), it can be found the performance of replay speech detection on ASVspoof 2019 PA corpus is much better than that of synthetic speech detection on ASVspoof 2019 LA corpus, which is consistent with the results in the ASVspoof 2019 challenge. One reason is that the ASVspoof 2019 PA corpus is a simulated replay dataset, while the ASVspoof 2019 LA corpus is a synthetic speech dataset. The variance in simulated data is less than that in synthetic data generated by various algorithms. Another reason is that the training size of the ASVspoof 2019 LA corpus (25 380 samples in total) is much smaller than that of the ASVspoof 2019 PA corpus (54 000 samples in total). The data-driven large-size neural-network-based models (LCNN and ResNet) can benefit a lot from a larger training set in the PA sub-challenge.

2) Comparison With LPS and MPS: Table IX shows the performance comparison among the LPS-, MPS-, and MMPS-based features derived from CQT on ASVspoof 2019 PA evaluation set. We observe that the MMPS-based systems can perform better than the corresponding MPS-based systems on ASVspoof 2019 PA evaluation set, and both of them outperform the corresponding LPS-based systems in terms of t-DCF and EER. This again confirms our idea of the proposal of MMPS based on MPS. By capturing both magnitude and phase information, they can outperform LPS with only magnitude information in replay speech detection.

In addition, it can be seen that CQMOC slightly outperforms CQ-OST for GMM, LCNN and ResNet classifiers. Although the improvements are small for these handcrafted features, they also reveal the effectiveness of our proposed CQT-MMPS and CQMOC features for replay speech detection.

TABLE IX
PERFORMANCE COMPARISON AMONG THE LPS-, MPS, AND MMPS-BASED FEATURES IN T-DCF AND EER (%) ON ASVSPOOF 2019 PA EVALUATION SET

Type	Model	Feature	t-DCF	EER
Frame-level	GMM	CQ-OST	0.149	7.28
		CQMOC	**0.135**	**6.95**
Utterance-level	LCNN	CQT-LPS	0.040	1.39
		CQT-MPS	0.033	1.22
		CQT-MMPS	**0.024**	**0.90**
		CQ-OST	0.037	1.31
		CQMOC	0.036	1.32
	ResNet	CQT-LPS	0.038	1.30
		CQT-MPS	0.035	1.19
		CQT-MMPS	**0.031**	**1.08**
		CQ-OST	0.040	1.38
		CQMOC	0.038	1.36

TABLE X
PERFORMANCE COMPARISON IN T-DCF AND EER (%) WITH SOME COMMONLY-USED FEATURES ON ASVSPOOF 2019 PA EVALUATION SET

Feature	t-DCF	EER	Feature	t-DCF	EER
MFCC	0.360	14.21	CQSPIC	0.164	7.73
LFCC	0.302	13.54	CosPhase	0.484	19.71
IFCC	0.357	15.85	APGDF	0.328	13.51
CQCC	0.245	11.04	MODGDF	0.242	11.24
CQMOC	**0.135**	**6.95**			

TABLE XI
PERFORMANCE COMPARISON IN T-DCF AND EER (%) AMONG THE PROPOSED SYSTEMS AND SOME KNOWN SYSTEMS ON ASVSPOOF 2019 PA EVALUATION SET

Type	System	t-DCF	EER
GMM-based	CQCC-GMM [58]	0.245	11.04
	LFCC-GMM [58]	0.302	13.54
	ZTWCC-GMM [52]	0.281	12.20
LCNN-based	DCT-LCNN [34]	0.560	2.06
	LFCC-LCNN [34]	0.105	4.60
	CQT-LCNN [34]	0.030	1.23
ResNet-based	CQCC-ResNet [36]	0.107	4.43
	Spec-ResNet [36]	0.099	3.81
	(GD gram)-ResNet [22]	0.044	1.79
	(GD gram)-ResNet-DA [22]	0.028	1.08
Proposed	**CQMOC-GMM**	**0.135**	**6.95**
	(CQT-MMPS)-LCNN	**0.024**	**0.90**
	CQMOC-LCNN	**0.036**	**1.32**
	(CQT-MMPS)-ResNet	**0.031**	**1.08**
	CQMOC-ResNet	**0.038**	**1.36**

3) Comparison With Other Commonly-Used Features: As shown in Table X, our proposed CQMOC feature is compared with some other commonly-used features. Similarly, we consider MFCC, LFCC, IFCC, CQCC CQSPIC, CosPhase, APGDF, and MODGDF with GMM as the back-end model. We find that the proposed CQMOC feature can perform much better than the other features, which reveals the effectiveness and robustness of CQMOC for replay attack detection. Besides, it also confirms that the proposed idea of the modified magnitude-phase spectrum is correct.

4) Comparison With Some Known Systems: Table XI shows the performance comparison among our proposed systems and some known systems on ASVspoof 2019 PA evaluation set. Most of the systems considered here are similar to those in Section IV-C4. In [39], the group delay gram (GD gram) is used as the input

to the ResNet models, denoted as (GD gram)-ResNet. The (GD gram)-ResNet-DA refers to the model where data augmentation (DA) is additionally applied.

Considering the GMM-based systems, we observe that our proposed CQMOC feature can perform much better than the CQCC, LFCC, and ZTWCC features. Significant improvements are also obtained if the utterance-based models are employed, such as LCNN and ResNet. As shown in Table XI, the (CQT-MMPS)-ResNet system consistently outperforms the other published ResNet-based systems without DA. Moreover, it can achieve a comparable result with the (GD gram)-ResNet-DA system. Among all the LCNN-based systems, the (CQT-MMPS)-LCNN system is shown to perform the best in terms of both t-DCF (0.024) and EER (0.90%) metrics. These results reveal that our proposed CQT-MMPS and CQMOC features are more effective for replay speech detection for GMM as well as neural-network-based LCNN and ResNet models.

V. UKSD Performance

In this section, the proposed TCMS and MTL methods for UKSD are evaluated on ASVspoof 2019 corpus. Our previous work, the combination-based method (CM) for UKSD [53], is used as the baseline here. In addition, the CQT-MMPS proposed in this work is considered as the input feature.

A. Training Data

As shown in Table III, the ASVspoof 2019 PA training set contains bonafide and replay speech examples, whereas the ASVspoof 2019 LA training set consists of bonafide and synthetic speech examples. In order to train the TCMS or MTL frameworks, we combine the ASVspoof 2019 LA and PA training set together and obtain a new training set that contains three classes, which are bonafide, replay, and synthetic speech examples. Specifically, there are 7980, 48 600, and 22 800 utterances for bonafide, replay, and synthetic speech examples in the new training set, respectively. Therefore, there are 71 400 spoofed samples if replay and synthetic speeches are jointly classified as the spoofed class.

B. Experimental Setup

From the experimental results on KKSD, it can be found that MMPS-based systems consistently outperform the corresponding CQMOC-based systems when LCNN or ResNet is considered as the classifier. As a result, here we only use MMPS-based systems to evaluate the performance of UKSD. In other words, only (CQT-MMPS)-LCNN and (CQT-MMPS)-ResNet systems are constructed for our proposed TCMS and MTL methods in this work. Besides, the parameters for extracting features such as CQT and MMPS are the same as mentioned above. Additionally, the training strategies of (CQT-MMPS)-LCNN and (CQT-MMPS)-ResNet are also reserved here. For the TCMS method, the only difference is that three nodes are used in the output layer. In addition, the t-DCF metric for TCMS is measured using the same approach as that of the two-class countermeasure. For the MTL approach, a new branch is additionally constructed, serving as the spoofing-kind discriminator, as illustrated in Section III-B2.

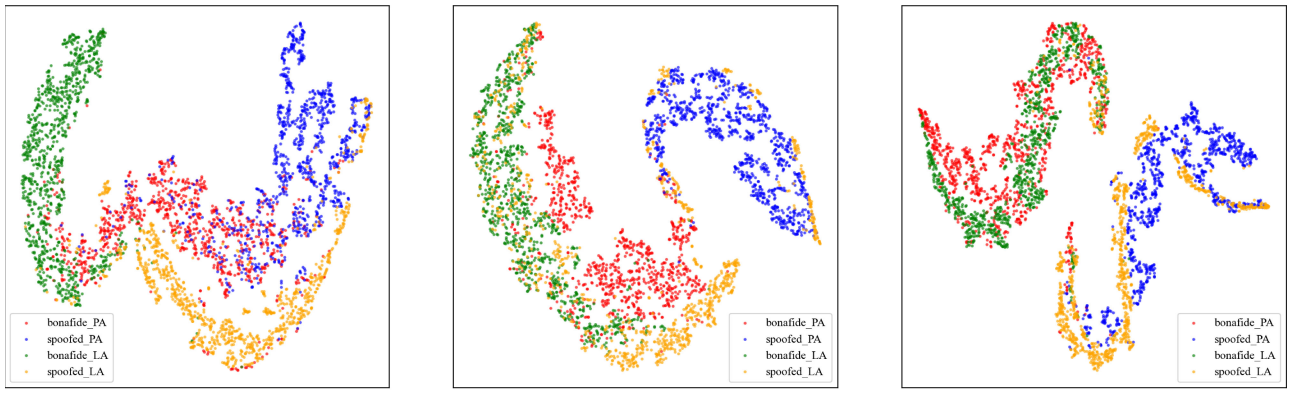
C. Results and Analysis

Table XII shows the performance comparison among the LCNN-based and ResNet-based models using the CQT-MMPS feature as input in terms of t-DCF and EER (%). Herein, LCNN_2-LA and LCNN_2-PA denote the traditional two-class LCNN models trained on ASVspoof 2019 LA and PA training sets, respectively, and LCNN_2-CM represents LCNN_2 is trained using the combination-based method (CM). The LCNN_3-TCMS refers to the three-class LCNN model with TCMS method, while LCNN_2-MTL represents the adapted MTL framework based on LCNN_2. In the same way, the corresponding definitions of the ResNet-based models are similar to that of the LCNN-based models. From Table XII, some observations can be found:

- Though the respective-kind-based models can obtain promising results for same-spoofing-kind attack detection, significant performance degradation can be observed when it comes to cross-spoofing-kind attack detection scenarios. Considering LCNN_2-PA as an example, it achieves the best result on ASVspoof 2019 PA evaluation set with 0.024 in t-DCF and 0.90% in EER, while generalizes poorly on ASVspoof 2019 LA evaluation set with only 0.347 in t-DCF and 18.24% in EER. These results imply that traditional KKSD methods cannot work well for UKSD. It also reveals the necessity of exploring new methods to solve the problem of UKSD.
- In case of UKSD, our proposed TCMS and MTL methods consistently outperform the combination-based method on both ASVspoof 2019 LA and PA evaluation sets for LCNN-based as well as ResNet-based models. The reason behind this is probably that both TCMS and MTL methods regard LA and PA attacks as two different spoofing kinds and are trained to distinguish between them, while the combination-based method simply considers LA and PA attacks together as a joint spoofed class. This reveals that the fine-grained discriminability in spoofing kinds (PA and LA) could further enhance performance for UKSD. In addition, it also confirms that the proposed idea of TCMS and MTL is correct. The MTL outperforms TCMS for LCNN-based models slightly, while TCMS outperforms MTL marginally for ResNet-based models. Overall, these two methods are both effective for UKSD with comparable performance.
- Comparing all systems in Table XII, we can observe that the best result on each evaluation set is achieved by KKSD methods. Specifically, the ResNet_2-LA model achieves the best result on ASVspoof 2019 LA evaluation set, while the LCNN_2-PA model performs the best on ASVspoof 2019 PA evaluation set. However, the UKSD methods, including the CM baseline as well as the proposed TCMS and MTL methods, can obtain comparable results with the best systems on both evaluation sets. Furthermore, if we only compare the results based on the same model

TABLE XII
PERFORMANCE COMPARISON IN T-DCF AND EER (%) AMONG LCNN-BASED AND RESNET-BASED MODELS USING THE CQT-MMPS FEATURE AS INPUT. THE NUMBERS IN BOLD FONT ARE THE BEST RESULTS ON EACH EVALUATION SET FOR THE LCNN-BASED OR RESNET-BASED MODELS

Type	Model	Training Set	Testing Set			
			ASVspoof 2019 LA Eval		ASVspoof 2019 PA Eval	
			t-DCF	EER	t-DCF	EER
LCNN-based	$LCNN_2$-LA	ASVspoof 2019 LA Train	0.176	5.99	0.416	16.10
	$LCNN_2$-PA	ASVspoof 2019 PA Train	0.347	18.24	**0.024**	**0.90**
	$LCNN_2$-CM	ASVspoof 2019 LA+PA Train	0.198	7.27	0.050	1.80
	$LCNN_3$-TCMS	ASVspoof 2019 LA+PA Train	0.196	6.32	0.034	1.21
	$LCNN_2$-MTL	ASVspoof 2019 LA+PA Train	**0.155**	**5.07**	0.031	1.06
ResNet-based	$ResNet_2$-LA	ASVspoof 2019 LA Train	**0.119**	**3.72**	0.431	16.89
	$ResNet_2$-PA	ASVspoof 2019 PA Train	0.469	24.17	0.031	1.08
	$ResNet_2$-CM	ASVspoof 2019 LA+PA Train	0.159	5.16	0.038	1.32
	$ResNet_3$-TCMS	ASVspoof 2019 LA+PA Train	0.119	3.84	**0.026**	**0.91**
	$ResNet_2$-MTL	ASVspoof 2019 LA+PA Train	0.133	4.32	0.028	0.96



(a) $ResNet_2$-LA (b) $ResNet_2$-PA (c) $ResNet_2$-CM

(d) $ResNet_3$-TCMS (e) $ResNet_2$-MTL

Fig. 2. The t-SNE visualization of evaluation data embeddings that are extracted by $ResNet_2$-LA, $ResNet_2$-PA, $ResNet_2$-CM, $ResNet_3$-TCMS, and $ResNet_2$-MTL, respectively. "bonafide_PA" (red) and "spoofed_PA" (blue) mean bonafide and spoofed data in the ASVspoof 2019 PA evaluation set, respectively, while "bonafide_LA" (green) and "spoofed_LA" (orange) refer to bonafide and spoofed data in the ASVspoof 2019 LA evaluation set, respectively. 1000 samples are randomly chosen for each label type (color).

architecture (LCNN or ResNet), we observe that some UKSD methods can outperform the corresponding KKSD method. For example, $LCNN_2$-MTL outperforms $LCNN_2$-LA on ASVspoof 2019 LA evaluation set, and $ResNet_3$-TCMS outperforms $ResNet_2$-PA on ASVspoof 2019 PA evaluation set. This indicates that by using the proposed TCMS or MTL methods, the synthetic training samples and the replay training samples might benefit each other in some way like data augmentation or model regularization, which further enhances performance for UKSD.

D. t-SNE Visualization

To better understand the mechanism of the proposed TCMS and MTL methods, we use t-SNE projection [68] to visualize

embedding distributions of the models. An example of the ResNet-based models is considered, which is shown in Fig. 2.

Considering the KKSD methods, ResNet_2-LA and ResNet_2-PA, they distinguish well between bonafide data and the known-kind spoofing attacks, while discriminate the unknown-kind spoofing attacks much more poorly. Considering ResNet_2-LA as an example, although it distinguishes well between "bonafide_LA" (green) and "spoofed_LA" (orange), it fails to separate "bonafide_PA" (red) and "spoofed_PA" (blue).

For the UKSD methods, ResNet_2-CM, ResNet_3-TCMS, and ResNet_2-MTL can separate the bonafide data from the spoofed data better in comparison with the KKSD methods discussed above. Comparing the CM baseline with our proposed TCMS and MTL methods, we can observe that ResNet_2-CM mixes up all spoofed samples, while both ResNet_3-TCMS and ResNet_2-MTL can well distinguish "spoofed_LA" from "spoofed_PA". Moreover, more samples are misclassified by ResNet_2-CM, which reveals the effectiveness of our newly-proposed TCMS and MTL methods for UKSD. Interestingly, ResNet_3-TCMS almost mixes up all bonafide data, while ResNet_2-MTL could still discriminate between "bonafide_PA" and "bonafide_LA". From our point of view, the result for ResNet_2-MTL could be a little counterintuitive. Our explanation is that ResNet_2-MTL may learn to distinguish "spoofed_LA" from "spoofed_PA" by some common characteristics in the dataset level, such as speaker traits and text. In other words, the bonafide data and spoofed data in two subsets are coupled by some means. Thus "bonafide_PA" and "bonafide_LA" could be separated at the same time.

VI. Conclusion and Future Work

This work attempts to utilize the magnitude and phase information collectively to improve the performance of SD. The MMPS is proposed as a novel feature to capture both magnitude and phase information. On the basis of MMPS obtained from the CQT (i.e., CQT-MMPS), a handcrafted feature CQMOC is further proposed. Three classic anti-spoofing models are considered to evaluate our proposed CQT-MMPS and CQMOC features, including the frame-level model (GMM) and the utterance-level models (LCNN and ResNet). In addition, the TCMS and MTL methods are proposed for UKSD in real-world applications because there is usually no prior knowledge about the kinds of spoofing attacks.

The experimental results show that the newly-proposed MMPS can outperform both LPS and MPS derived from CQT for both synthetic and replay speech detection in our implementations. In addition, CQT-MMPS can achieve better or comparable performance in comparison with the state-of-the-art systems. We also find that the proposed handcrafted CQMOC outperforms most of the handcrafted features on ASVspoof 2019 corpus. The strong modelling capabilities of the neural-network-based models (LCNN and ResNet) are also validated from their promising performance on both synthetic and replay speech detection. Moreover, it is shown that the proposed TCMS and MTL methods can outperform the combination-based method when we have no prior information about the spoofing kind. In addition, compared with the respective-kind-based methods, the TCMS and MTL achieve comparable results for the same-spoofing-kind attack detection, while they show much better performance in cross-spoofing-kind evaluation scenarios.

As mentioned above, we conducted cross-corpora experiments on the ASVspoof 2019 LA and PA databases to evaluate our proposed UKSD methods. Considering ASVspoof 2019 PA is a stimulated replay database that has a significant difference with a real replay speech database, we will investigate our proposed methods on some other realistic replay speech databases such as BTAS 2016 and ASVspoof 2017 in the future.

VI. Acknowledgment

The authors would like to acknowledge the handling editor, anonymous reviewers, Professor Haizhou Li from National University of Singapore, and Professor Kai Yu from Shanghai Jiao Tong University for their advice and suggestions to improve the quality of this paper. In addition, Jichen Yang and Hongji Wang have the equal contributions for this work.

References

[1] T. Kinnunen and H. Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech Commun.*, vol. 52, no. 1, pp. 12–40, 2008.

[2] L. Xu, B. Ren, G. Zhang, and J. Yang, "Linear transformation on x-vector for text-independent speaker verification," *Electron. Lett.*, vol. 55, no. 15, pp. 864–866, 2019.

[3] S. Wang, Z. Huang, Y. Qian, and K. Yu, "Discriminative neural embedding learning for short-duration text-independent speaker verification," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 27, pp. 1686–1696, Nov. 2019.

[4] S. Wang, Y. Yang, Z. Wu, Y. Qian, and K. Yu, "Data augmentation using deep generative models for embedding based speaker recognition," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 28, pp. 2598–2609, 2020, doi: 10.1109/TASLP.2020.3016498.

[5] R. K. Das, X. Tian, T. Kinnunen, and H. Li, "The attacker's perspective on automatic speaker verification: An overview," in *Proc. 21st Annu. Conf. Int. Speech Commun. Assoc.*, Shanghai, China, 2020, pp. 4213–4217.

[6] T. Kinnunen *et al.*, "The ASVspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Stockholm, Sweden, 2017, pp. 5395–5399.

[7] C. H. You and J. Yang, "Device feature extraction based on parallel neural network training for replay spoofing detection," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 28, pp. 2308–2319, 2020, doi: 10.1109/TASLP.2020.3011320.

[8] J. Yang, L. Xu, and B. Ren, "Constant-q deep coefficients for playback attack detection," *IEICE Trans. Inf. Syst.*, vol. E103-D, no. 02, pp. 464–468, 2020.

[9] H. Zen, M. J. F. Gales, Y. Nankaku, and K. Tokuda, "Product of experts for statistical parametric speech synthesis," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 20, no. 3, pp. 794–805, Mar. 2012.

[10] X. Wang, S. Takaki, J. Yamagishi, S. King, and K. Tokuda, "A vector quantized variational autoencoder (VQ-VAE) autoregressive neural f0 model for statistical parametric speech synthesis," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 28, pp. 157–170, 2020, doi: 10.1109/TASLP.2019.2950099.

[11] R. Liu, B. Sisman, F. Bao, J. Yang, G. Gao, and H. Li, "Exploiting morphological and phonological features to improve prosodic phrasing for mongolian speech synthesis," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 29, pp. 274–285, 2021, doi: 10.1109/TASLP.2020.3040523.

[12] X. Tian, S. Lee, Z. Wu, E. S. Chng, and H. Li, "An example-based approach to frequency warping for voice conversion," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1863–1875, Jul. 2017.

[13] J.-X. Zhang, Z.-H. Ling, L.-J. Liu, Y. Jiang, and L.-R Dai, "Sequence-to-sequence acoustic modeling for voice conversion," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 27, no. 3, pp. 631–644, Mar. 2019.

[14] B. Sisman, M. Zhang, and H. Li, "Group sparse representation with wavenet vocoder adaptation for spectrum and prosody conversion," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 27, no. 6, pp. 1085–1097, Jun. 2019.

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

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

[17] T. B. Patel and H. A. Patil, "Combining evidences from mel cepstral, cochlear filter cepstral and instantaneous frequency features for detection of natural vs. spoofed speech," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Dresden, Germany, 2015, pp. 2062–2066.

[18] X. Xiao, X. Tian, S. Du, H. Xu, E. S. Chng, and H. Li, "Spoofing speech detection using high dimensional magnitude and phase features: The NTU approach for ASVspoof 2015 challenge," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Dresden, Germany, 2015, pp. 2052–2056.

[19] M. Todisco, H. Delgado, and N. Evans, "A new feature for automatic speaker verification anti-spoofing: Constant Q cepstral coefficients," in *Proc. Speaker Lang. Recognit. Workshop*, Bilbao, Spain, 2016, pp. 283–290.

[20] M. Todisco, H. Delgado, and N. Evans, "Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification," *Comput. Speech Lang.*, vol. 45, pp. 516–535, 2017.

[21] R. K. Das, J. Yang, and H. Li, "Long range acoustic features for spoofed speech detection," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria, 2019, pp. 1058–1062.

[22] J. Yang, R. K. Das, and N. Zhou, "Extraction of octave spectra information for spoofing attack detection," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 27, no. 12, pp. 2373–2384, Dec. 2019.

[23] Y. Qian, N. Chen, and K. Yu, "Deep features for automatic spoofing detection," *Speech Commun.*, vol. 85, pp. 43–52, 2016.

[24] H. Dinkel, Y. Qian, and K. Yu, "Investigating raw wave deep neural networks for end-to-end speaker spoofing detection," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 26, no. 11, pp. 2002–2014, Nov. 2018.

[25] R. K. Das, J. Yang, and H. Li, "Long range acoustic and deep features perspective on ASVspoof2019," in *Proc. Autom. Speech Recognit. Understanding Workshop*, Sentosa, Singapore, 2019, pp. 1018–1025.

[26] Y. Qian, N. Chen, H. Dinkel, and Z. Wu, "Deep feature engineering for noise robust spoofing detection," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1942–1955, Oct. 2017.

[27] S. Davis and P. Mermelstein, "Comparison of parametric representation for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[28] T. Kinnunen, Z. Wu, K. A. Lee, F. Sedlak, E. S. Chng, and H. Li, "Vulnerability of speaker verification systems against voice conversion attacks: The case of telephone speech," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Kyoto, Japan, 2012, pp. 4402–4404.

[29] M. Sahidullah, T. Kinnunen, and C. Hanilci, "A comparison features for synthetic speech detection," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, Dresden, Germany, 2015, pp. 2087–2091.

[30] T. Hasen, S. O. Sadjadi, G. Liu, N. Shokouhi, H. Boril, and J. H. L. Hansen, "CRSS systems for 2012 NIST speaker recognition evaluations," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, Vancouver, Canada, 2013, pp. 6783–6787.

[31] S. Chakroborty, A. Roy, and G. Saha, "Improved closed set text-independent speaker identification by combining MFCC with evidence from flipped filter banks," *Int. J. Signal Process.*, vol. 4, no. 2, pp. 114–122, 2007.

[32] H. Yu, Z.-H. Tan, Z. Ma, R. Martin, and J. Guo, "Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 10, pp. 4633–4644, Oct. 2018.

[33] D. Paul, M. Pal, and G. Saha, "Spectral features for synthetic speech detection," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 4, pp. 605–617, Jun. 2017.

[34] J. Yang, R. K. Das, and H. Li, "Significance of subband features for synthetic speech decetion," *IEEE Trans. Inf. Forensics Security*, vol. 15, pp. 2160–2170, 2020, doi: 10.1109/TIFS.2019.2956589.

[35] G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudasher, and V. Shchemelinin, "Audio replay attack detection with deep learning framework," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc.*, Stockholm, Sweden, 2017, pp. 82–86.

[36] G. Lavrentyva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, and A. Kozlos, "STC antispoofing systems for the ASVspoof2019 challenge," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria, 2019, pp. 1033–1037.

[37] Y. Yang *et al.*, "The SJTU robust anti-spoofing system for the ASVspoof 2019 challenge," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria, 2019, pp. 1038–1042.

[38] M. Alzanto, Z. Wang, and M. B. Srivastava, "Deep residual neural networks for audio spoofing detection," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria, 2019, pp. 1078–1082.

[39] W. Cai, H. Wu, D. Cai, and M. Li, "The DKU replay detection system for the ASVspoof 2019 challenge on data augmentation, feature representation, clssifier and fusion," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria, 2019, pp. 1023–1027.

[40] J.-W. Jung, H.-J. Shim, H.-S. Heo, and H.-J. Yu, "Replay attack detection with complementary high-resolution information using end-to-end for the ASVspoof 2019 challenge," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria, 2019, pp. 1083–1087.

[41] H. Wang, H. Dinkel, S. Wang, Y. Qian, and K. Yu, "Dual-adversarial domain adaptation for generalized replay attack detection," in *Proc. Interspeech*, 2020, pp. 1086–1090.

[42] K. Sriskandaraja, V. Sethu, and E. Ambikairajah, "Deep siamese architecture based replay detection for secure voice biometric," in *Proc. 19th Annu. Conf. Int. Speech Commun. Assoc.*, Hyderabad, India, 2018, pp. 671–675.

[43] R. M. Hegde and H. A. Murthy, "Significance of the modified group delay features in speech recognition," *IEEE Trans. Audio, Speech, Language Process.*, vol. 15, no. 1, pp. 190–202, Jan. 2007.

[44] H. A. Murthy and V. Gudde, "The modified group delay function and its application to phone recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Hognkong, China, 2003, pp. *I-68-I-71*.

[45] S. Jelil, R. K. Das, S. R. M. Prasanna, and R. Sinha, "Spoof detection using source, instantaneous frequecny and cepstral features," in *Proc. 18th Annu. Conf. Int. Speech Commun. Assoc.*, Stockholm, Sweden, 2017, pp. 22–26.

[46] R. K. Das and H. Li, "Instantaneous phase and excitation source features for detection of replay attacks," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Honolulu, Hawaii, 2018, pp. 1030–1037.

[47] K. Srinivas, R. K. Das, and H. A. Patil, "Combining phase-based features for replay spoof detection system," in *Proc. Int. Symp. Chin. Spoken Lang. Process.*, Taipei, Taiwan, 2018, pp. 151–155.

[48] Z. Wu, E. S. Chng, and H. Li, "Detecting converted speech and natural speech for anti-spoofing attack in speaker recognition," in *Proc. 13th Annu. Conf. Int. Speech Commun. Assoc.*, Portland, USA, 2013, pp. 1700–1703.

[49] P. Rajan, T. Kinnunen, C. Hanilci, J. Pohjalainen, and P. Alk, "Using group delay functions from all-pole models for speaker recognition," in *Proc. 14th Annu. Conf. Int. Speech Commun. Assoc.*, Lyon, France, 2013, pp. 2489–2493.

[50] Z. Wu *et al.*, "Anti-spoofing for text-independent speaker verification: An initial database, comparison of countermeasures, and human performance," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 4, pp. 768–783, Apr. 2016.

[51] F. Tom, M. Jain, and P. Dey, "End-to-end audio replay attack detection using deep convolutional networks with attention," in *Proc. 19th Annu. Conf. Int. Speech Commun. Assoc.*, Hyderabad, India, 2018, pp. 681–685.

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

[53] R. K. Das, J. Yang, and H. Li, "Assessing the scope of generalized countermeasures for anti-spoofing," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Barcelona, Spain, 2020, pp. 6589–6593.

[54] J. Monteiro, J. Alam, and T. H. Falk, "An ensemble based approcah for generalized detection of spoofing attack to automatic speaker recognizers," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Barcelona, Spain, 2020, pp. 6599–6604.

[55] J. C. Brown, "Calculation of a constant q spectral transform," *J. Acoust. Soc. Amer.*, vol. 89, pp. 425–434, 1991.

[56] K. N. R. K. R. Alluri and A. K. Vupala, "IIIT-H spoofing countermeasures for automatic speaker verification spoofing and countermesures challenge 2019," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria, 2019, pp. 1043–1047.

[57] H. Wang, H. Dinkel, S. Wang, Y. Qian, and K. Yu, "Cross-domain replay spoofing attack detection using domain adversarial training," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria, 2019, pp. 2938–2942.

[58] X. Cheng, M. Xu, and T. F. Zheng, "Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspoof2019," in *Proc. Asia Pacific Signal Inf. Process. Assoc. Annu. Summit Conf.*, Lanzhou, China, 2019, pp. 540–545.
[59] C.-I Lai, N. Chen, J. Villaba, and N. Dehak, "ASSERT: Anti-spoofing with squeeze-excitation and residual networks," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria, 2019, pp. 1013–1017.
[60] J. Monteiro and J. Alam, "Development of voice spoofing detection systems for 2019 edition of automatic speaker verification and countermeasures challenge," in *Proc. IEEE Autom. Speech Recognit. Understanding Workshop*, Sentosa, Singapore, 2019, pp. 1003–1010.
[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[62] M. Todisco *et al.*, "ASVspoof 2019: Future horizons in spoofed and fake audio detection," in *Proc. 20th Annu. Conf. Int. Speech Commun. Assoc.*, Graz, Austria, 2019, pp. 1008–1012.
[63] T. Kinnunen *et al.*, "t-DCF: A detection cost function for tandem assessment of spoofing countermeasures and automatic speaker verification," in *Proc. Speaker Lang. Recognit. Workshop*, Bilbao, Spain, 2018, pp. 312–319.
[64] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, Sardinia, Italy, 2010, pp. 249–256.
[65] K. Vijayan, P. R. Reddy, and K. S. R. Murty, "Significance of analysis phase of speech signals in speaker verification," *Speech Commun.*, vol. 81, pp. 54–71, 2016.
[66] J. Yang, C. You, and Q. He, "Feature with complementarity of statistics and principal information for spoofing detection," in *Proc. 19th Annu. Conf. Int. Speech Commun. Assoc.*, Hyderabad, India, 2018, pp. 651–655.
[67] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 5329–5333.
[68] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. Nov, pp. 2579–2605, 2008.

Jichen Yang (Senior Member, IEEE) received the Ph.D degree from the South China University of Technology (SCUT), Guangzhou, China, in 2010. From 2011 to 2015, he was a Postdoctor with SCUT. From 2016 to 2020, he was a Research Fellow with the Department of Human Language Technology, Institute for Infocomm Research (I^{2}R), A*STAR, Singapore, and then with the Human Language Technology Lab, Department of Electrical and Computer Engineering, National University of Singapore, Singapore. His research interests mainly include speaker recognition, voice conversion, and speech synthesis.

Hongji Wang received the B.S degree from the Department of Computer Science and Technology, Nanjing University, Nanjing, China, in 2018. He is currently working toward the M.E. degree with Shanghai Jiao Tong University (SJTU), Shanghai, China. He is a Member with SpeechLab, SJTU, under the supervision of Professor Kai Yu. His research interests mainly include antispoofing, speaker recognition, speaker diarization, and keyword spotting.

Rohan Kumar Das (Senior Member, IEEE) received the B. Tech degree in electronics and communication engineering from North-Eastern Hill University (NEHU), Shillong, India, in 2010 and the Ph.D degree from the Indian Institute of Technology Guwahati, Guwahati, India, in 2017. His Ph.D. work was focused on speaker verification using short utterances from the perspective of practical application oriented systems. Prior to his research in the field of speech processing, he has been a Project Scientist with Assam Science Technology and Environment Councilfrom 2010 to 2011. In 2017, after completing doctoral studies, he was a Data Scientist with a multinational company, Kovid Research Labs, and was involved in speech analytics based application services. He is currently a Postdoctoral Research Fellow with the National University of Singapore, Singapore, and continuing postdoctoral research work. He has authored or coauthored more than 80 research papers in peer-reviewed journals and conferences. His research interests include speech signal processing, speaker verification, antispoofing, machine learning, and pattern recognition. He is one of the organizers for special sessions on The Attacker's Perspective on Automatic Speaker Verification, Far-Field Speaker Verification Challenge 2020 in Interspeech 2020, and the Voice Conversion Challenge 2020. He was the Publication Chair of the IEEE Automatic Speech Recognition Understanding (ASRU) Workshop 2019 and one of the Chairs of Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020. He was the recipient of several travel fellowships from organizations, such as the IEEE Signal Processing Society, International Speech Communication Association, Microsoft Research India, Xerox Research Centre India, and Science and Engineering Research Board (SERB), and Government of India to present research works in top tier conferences, such as ICASSP and Interspeech.

Yanmin Qian (Senior Member, IEEE) received the B.S. degree from the Department of Electronic and Information Engineering, Huazhong University of Science and Technology, Wuhan, China, in 2007 and the Ph.D. degree from the Department of Electronic Engineering, Tsinghua University, Beijing, China, in 2012. In 2013, he joined the Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China, where he is currently an Associate Professor. From 2015 to 2016, he was also an Associate Researcher with the Speech Group, Department of Engineering, University of Cambridge, Cambridge, U.K. His current research interests include the acoustic and language modeling in speech recognition, speaker and language recognition, key word spotting, and multimedia signal processing.