Yang *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2020) 2020:6
https://doi.org/10.1186/s13636-020-00173-5

EURASIP Journal on Audio, Speech, and Music Processing

RESEARCH Open Access

Discriminative features based on modified log magnitude spectrum for playback speech detection

Jichen Yang[1†], Longting Xu[2*†], Bo Ren[3] and Yunyun Ji[4]

Abstract

In order to improve the performance of hand-crafted features to detect playback speech, two discriminative features, constant-Q variance-based octave coefficients and constant-Q mean-based octave coefficients, are proposed for playback speech detection in this work. They rely on our findings that variance-based modified log magnitude spectrum and mean-based modified log magnitude spectrum can enhance the discriminative power between genuine speech and playback speech. Then constant-Q variance-based octave coefficients (constant-Q mean-based octave coefficients) can be obtained by combining variance-based modified log magnitude spectrum (mean-based modified log magnitude spectrum), octave segmentation, and discrete cosine transform. Finally, constant-Q variance-based octave coefficients and constant-Q mean-based octave coefficients are evaluated on ASVspoof 2017 corpus version 2.0 and ASVspoof 2019 physical access, respectively. Experimental results show that variance-based modified log magnitude spectrum and mean-based modified log magnitude spectrum can produce discriminative features toward playback speech. Further results on the two databases show that constant-Q variance-based octave coefficients and constant-Q mean-based octave coefficients can perform better than some common features, such as mel frequency cepstral coefficients and constant-Q cepstral coefficients.

Keywords: Discriminative feature, Playback attack detection, Modified log magnitude spectrum, Constant-Q variance-based octave coefficients, Constant-Q mean-based octave coefficients

1 Introduction

Replay attacks present serious threat to automatic speaker verification (ASV) system. In which the source recordings of playback are from the legitimate clients [1, 2]. Thus, replay attacks can pose the threat to ASV system. This motivates our focus on playback speech detection.

Since the ASVspoof 2017 challenge [1, 2], more and more researchers begin to focus on playback speech detection [3–10]. Similar to many speech signal processing systems, most of all playback speech detection systems usually consist of front-end feature and back-end classifier [11–18]. For the end-to-end systems such as [19–23], the output layer of the neural network can be seen as a virtual classifier and the rest of the neural network can be seen a deep feature extractor. In this paper, we mainly focus on how to extract discriminative feature for playback speech detection.

1.1 Related works

Before 2017, several studies about playback speech detection have been reported. The earlier ones [24–26] were based on small-scale databases, where only a small number playback and recording conditions were taken into account. For example, in [24, 27], three playback and recording devices were used to collect the database; in [25, 28], one recording device and one playback device were used to create the database, which is named as authentic and playback speech database (APSD); in [29],

*Correspondence: xlt@dhu.edu.cn
†Jichen Yang and Longting Xu contributed equally to this work.
[2]College of Information Science and Technology, Donghua University, Shanghai China
Full list of author information is available at the end of the article

the database was built by four smartphones; and in [26], four devices were used to create the playback utterances in the database, which is named as (audio-visual spoofing 2015) AVspoof 2015.

Different from the above databases, the launch of the ASVspoof 2017 corpus provided a large common database, obtained using 26 playback devices, 25 recording devices, and 26 environments [1, 2, 30]. So, ASVspoof 2017 corpus can be used to evaluate a playback speech detection algorithm justly because it has more channel and acoustic conditions than any previous databases [24–26]. In addition, the recent released ASVspoof 2019 physical access also can be used in this study. Hence, ASVspoof 2017 and ASVspoof 2019 physical access not only can support researchers to develop countermeasures, but also can protect ASV system to avoid replay attack [1].

After ASVspoof 2017 and ASVspoof 2019 physical access corpus were released in 2017 and 2019, respectively, some effective methods were proposed to detect playback speech based on the two databases. According to how the used features generate, these methods can be categorized into two types: hand-crafted design features and deep features. In general, hand-crafted features are the features which are obtained by using math formula to design while deep features are the features which are obtained by learning from neural networks for the input.

Deep feature extraction usually contains two steps: the fist is to train a classifier using neural network and the input, and the second is to remove the output layer of the classifier. Because the end-to-end system can be seen a virtual classifier (the output layer) and a deep feature extractor (the rest part), in other words, the end-to-end system can be seen a deep feature extractor if the output layer is removed. So in this study, we also regard the deep features that can be obtained from the end-to-end systems as special type of deep features. According to the neural networks used, there are several types deep features. For example, light convolutional neural network was used to learn deep feature for the input of log power spectrum of constant-Q transform (CQT) and fast Fourier transform in [14, 20, 21], deep Siamese that is formed two convolutional neural networks with the input of spectrogram are used to learned to obtain Siamese embedding features in [31], residual network (ResNet) was used to learn deep feature from the input of group delay gram in [19, 32, 33].

For the hand-crafted design features, which mainly include the following categories:

- **CQT based features:** which include constant-Q cepstral coefficients (CQCC) [34, 35] used in [4, 36, 37], extended CQCC [38, 39], constant-Q magnitude-phase octave coefficients [40], and constant-Q statistics-plus-principal information coefficients [41].
- **Discrete Fourier transform (DFT) based features:** which include Mel frequency cepstral coefficients (MFCC) [4, 13, 36], mel filterbank slope [10], linear filterbak slope [10], and Q-log domain DFT-based mean normalized log spectral [42].
- **Variable length energy separation algorithm (VESA)-based features:** which include instantaneous frequency cosine coefficients based on VESA [6] and instantaneous amplitude cosine coefficients based on VESA [43].
- **Prediction cepstral coefficients-based features:** which include linear prediction cepstral coefficients residual part and linear prediction cepstral coefficients cepstrum [13, 19], frequency domain linear prediction [9].
- **Spectral centroid-based features:** which include subband spectral centroid frequency coefficients and subband spectral centroid magnitude coefficients [12] and spectral centroid deviation [16].
- **Phased-based features:** which include instantaneous frequency cosine coefficient [44, 45] and modified group delay cepstral coefficient [15].
- **Zero time windowing-based features:** zero time windowing cepstral coefficients [46, 47].
- **Single frequency filter-based features:** single frequency filter cepstral coefficients [3, 47].

In which, CQCC is the most widely used features in playback speech detection, in ASVspoof2017 and ASVspoof 2019 challenge, CQCC plus Gaussian mixture model (GMM) are used to form the baseline system by the organizers [2, 48]. The reason is that CQT is a long-term transform, and it can provide more frequency detail to capture playback information in playback speech detection compared with DFT.

Generally speaking, deep features perform better than hand-crafted design features in playback speech detection because more useful information for discriminating playback speech from genuine speech can be obtained by deep learning. However, deep features rely on training data heavily. That is to say, deep features only suit on the scope of training data. Further, if we want to study the property of playback speech, hand-crafted design features can be selected rather than deep features. The goal of the paper is to extract discriminative feature for playback speech detection, so hand-crafted design feature is studied. Therefore, our focus is how to extract hand-crafted discriminative features in this study.

Traditional hand-crafted features used in speech signal processing such as MFCC and CQCC are not designed for playback speech detection. In order to improve the performance of hand-crafted features to detect playback speech, we focus on designing discriminative features for

playback speech detection in this study. Considering three facts which are as following:

- CQT is a long-term transform and it can provide more frequency detail to capture the playback information compared with DFT.
- Many hand-crafted features such as CQCC and MFCC are extracted from log power spectrum that can be obtained from log magnitude spectrum (LMS).
- A feature can have more discriminative power to distinguish playback speech from genuine speech if the discriminative power between genuine speech and playback speech can be enhanced.

Therefore, in this study, LMS based on CQT is used as study object to investigating how to enhance the discriminative power between genuine speech and playback speech and then modified log magnitude spectrum can be obtained. Finally, by combining with octave segmentation and discrete cosine transform (DCT), hand-crafted features with more discriminative feature for playback speech detection can be obtained, we can call them as hand-crafted discriminative features.

1.2 Contributions of the work

The goal of the work is how to extract hand-crafted discriminative feature by enlarging the difference between genuine speech and playback speech for playback speech detection. There are mainly two contributions in this work.

We found that discriminative power between genuine speech and playback speech can be enhanced if LMS is added its variance or mean. Based on the findings, two methods are proposed to modify LMS and we refer them as variance-based modified log magnitude spectrum (VMLMS) and mean-based modified log magnitude spectrum (MMLMS). In which, LMS is obtained using CQT which is used to convert speech from the time domain into the frequency domain. It is the first contribution of the paper.

By combining VMLMS, octave segmentation and DCT, one new feature from VMLMS is obtained, namely, constant-Q variance-based octave coefficients (CVOC). In the same way, the other feature is obtained by combining MMLMS, octave segmentation, and DCT, which is named as constant-Q mean-based octave coefficients (CMOC). They are the second contribution of the paper.

The remainder of the paper is organized as follows. Section 2 introduces modified log magnitude spectrum. Section 3 introduces how to extract discriminative features. Sections 4 and 5 gives the experimental results and corresponding analysis on ASVspoof 2017 version 2.0 and ASVspoof 2019 physical access databases, respectively. Section 6 concludes the paper.

2 Proposed method I: modified log magnitude spectrum

In this section, in order to enhance the discriminative power between genuine speech and playback speech, two methods to modify LMS are proposed by analyzing discriminative power between genuine speech and playback speech, which are VMLMS and MMLMS. Here, Fisher's ratio [49] that is often used to measure discriminative power of two classes [50], is used to measure discriminative power between genuine speech and playback speech, its equation is as follows [11]:

$$F_{C_1C_2} = \frac{(\overline{C_1} - \overline{C_2})^2}{\sigma_{C_1}^2 + \sigma_{C_2}^2} \tag{1}$$

where C_1 and C_2 present two classes, $F_{C_1C_2}$ represents Fisher ratio between C_1 and C_2, $\overline{C_1}$ and $\overline{C_2}$ represent mean of C_1 and C_2, respectively, $\sigma_{C_1}^2$ and $\sigma_{C_2}^2$ represent variance of C_1 and C_2, respectively.

2.1 Variance-based modified log magnitude spectrum

We assume X_0 and Y_0 are a frame genuine speech magnitude spectrum and its corresponding playback speech magnitude spectrum, respectively, and K is frequency bin number, we can obtain

$$X_0 = \left\{x_1, x_2, ..., x_K\right\} \tag{2}$$

$$Y_0 = \left\{y_1, y_2, ..., y_K\right\} \tag{3}$$

In addition, we can obtain X_0 and Y_0 in log-scale, denoted as X and Y

$$X = \left\{\log(x_1), \log(x_2), ..., \log(x_K)\right\} \tag{4}$$

$$Y = \left\{\log(y_1), \log(y_2), ..., \log(y_K)\right\} \tag{5}$$

Supposing $\overline{X}$ and $\overline{Y}$ are means of X and Y, respectively, we can obtain

$$\overline{X} = \frac{\sum_{k=1}^{K} \log(x_k)}{K} \tag{6}$$

$$\overline{Y} = \frac{\sum_{k=1}^{K} \log(y_k)}{K} \tag{7}$$

Supposing σ_X^2 and σ_Y^2 are variance of X and Y, respectively, we can obtain

$$\sigma_X^2 = \frac{\sum_{k=1}^{K} (\log(x_k) - \overline{X})^2}{K} \tag{8}$$

$$\sigma_Y^2 = \frac{\sum_{k=1}^{K} (\log(y_k) - \overline{Y})^2}{K} \tag{9}$$

Supposing F_{XY} is Fisher's ratio between X and Y, according to Eq. (1), we can obtain

$$F_{XY} = \frac{(\overline{X} - \overline{Y})^2}{\sigma_X^2 + \sigma_Y^2} \tag{10}$$

Supposing X' and Y' satisfy:

$$X' = \left\{ \log(x_1) + \sigma_X^2, \log(x_2) + \sigma_X^2, ..., \log(x_K) + \sigma_X^2 \right\} \tag{11}$$

$$Y' = \left\{ \log(y_1) + \sigma_Y^2, \log(y_2) + \sigma_Y^2, ..., \log(y_K) + \sigma_Y^2 \right\} \tag{12}$$

The means of X' and Y', denoted as $\overline{X'}$ and $\overline{Y'}$, which are as follows:

$$\begin{aligned} \overline{X'} &= \frac{\sum_{k=1}^{K} (\log(x_k) + \sigma_X^2)}{K} \\ &= \overline{X} + \sigma_X^2 \end{aligned} \tag{13}$$

$$\begin{aligned} \overline{Y'} &= \frac{\sum_{k=1}^{K} (\log(y_k) + \sigma_Y^2)}{K} \\ &= \overline{Y} + \sigma_Y^2 \end{aligned} \tag{14}$$

The variances of X' and Y', denoted as $\sigma_{X'}^2$ and $\sigma_{Y'}^2$, which are as follows:

$$\begin{aligned} \sigma_{X'}^2 &= \frac{\sum_{k=1}^{K} (\log(x_k) + \sigma_X^2 - \overline{X'})^2}{K} \\ &= \frac{\sum_{k=1}^{K} (\log(x_k) + \sigma_X^2 - (\overline{X} + \sigma_X^2))^2}{K} \\ &= \frac{\sum_{k=1}^{K} (\log(x_k) - \overline{X})^2}{K} \\ &= \sigma_X^2 \end{aligned} \tag{15}$$

$$\begin{aligned} \sigma_{Y'}^2 &= \frac{\sum_{k=1}^{K} (\log(y_k) + \sigma_Y^2 - \overline{Y'})^2}{K} \\ &= \frac{\sum_{k=1}^{K} (\log(y_k) + \sigma_Y^2 - (\overline{Y} + \sigma_Y^2))^2}{K} \\ &= \frac{\sum_{k=1}^{K} (\log(y_k) - \overline{Y})^2}{K} \\ &= \sigma_Y^2 \end{aligned} \tag{16}$$

Supposing $F_{X'Y'}$ is Fisher's ratio between X' and Y', according to Eq. (1), we can obtain

$$\begin{aligned} F_{X'Y'} &= \frac{(\overline{X'} - \overline{Y'})^2}{\sigma_{X'}^2 + \sigma_{Y'}^2} \\ &= \frac{(\overline{X} + \sigma_X^2 - \overline{Y} - \sigma_Y^2)^2}{\sigma_X^2 + \sigma_Y^2} \\ &= \frac{(\overline{X} - \overline{Y} + \sigma_X^2 - \sigma_Y^2)^2}{\sigma_X^2 + \sigma_Y^2} \end{aligned} \tag{17}$$

Let

$$\begin{aligned} F_{mro} &= \frac{F_{X'Y'}}{F_{XY}} \\ &= \frac{\frac{(\overline{X} - \overline{Y} + \sigma_X^2 - \sigma_Y^2)^2}{\sigma_X^2 + \sigma_Y^2}}{\frac{(\overline{X} - \overline{Y})^2}{\sigma_X^2 + \sigma_Y^2}} \\ &= \frac{(\overline{X} - \overline{Y} + \sigma_X^2 - \sigma_Y^2)^2}{(\overline{X} - \overline{Y})^2} \\ &= (1 + \frac{\sigma_X^2 - \sigma_Y^2}{\overline{X} - \overline{Y}})^2 \end{aligned} \tag{18}$$

Let

$$F_{vrs} = \frac{\sigma_X^2 - \sigma_Y^2}{\overline{X} - \overline{Y}} \tag{19}$$

From Eqs. (18) and (19), we can see that F_{mro} is determined by F_{vrs} and then F_{vrs} is determined by σ_X^2, σ_Y^2, $\overline{X}$, and $\overline{Y}$. However, as these parameters are unknown, it is not possible to determine the value of F_1 and F_2 directly.

Therefore, statistical analysis methods can be used to obtained F_{vrs}. To this end, APSD [25] and AVspoof 2015 [26] are used here. There are two reasons behind selecting these two databases. One is that they are the two largest publicly available databases of genuine-playback speech utterances to date with 3600 and 5600 respectively. The other is that the former is designed for the purpose of replay speech detection and the latter for replaying spoofing detection and synthetic speech detection. The 3600 genuine-playback pairs utterances from APSD and 5600 genuine-playback pairs utterances from AVspoof 2015 can be used to obtain the statistics of F_{vrs} on different σ_X^2, σ_Y^2, $\overline{X}$, and $\overline{Y}$. The CQT is applied on utterances from the two databases to compute F_{vrs} on σ_X^2, σ_Y^2, $\overline{X}$ and $\overline{Y}$ frame by frame. Finally, average F_{vrs} can be obtained, denoted as $\overline{F_{vrs}}$. In the same way, average values of σ_X^2, σ_Y^2, $\overline{X}$, and $\overline{Y}$ can be denoted as $\overline{\sigma_X^2}$, $\overline{\sigma_Y^2}$, $\overline{\overline{X}}$, and $\overline{\overline{Y}}$.

Table 1 shows the statistics value of $\overline{F_{vrs}}$ on APSD and AVspoof 2015. From Table 1, it can be observed that $\overline{F_{vrs}}$ is above 0 not only for APSD but also for AVspoof 2015. According to the relationship between F_{mro} and F_{vrs} in Eqs. (18) and (19), we can know that the statistics value of $\overline{F_{mro}}$ is above 1 on the two databases. Further, we can know that $F_{X'Y'}$ is larger than F_{XY}.

Table 1 Statistics value of $\overline{F_{vrs}}$ on APSD and AVspoof 2015

Database	$\overline{\sigma_X^2}$	$\overline{\sigma_Y^2}$	$\overline{\overline{X}}$	$\overline{\overline{Y}}$	$\overline{F_{vrs}}$
APSD	1.07	0.83	−3.33	−3.80	0.51
AVspoof 2015	6.22	5.41	−19.27	−21.42	0.38

The above discussion leads to the findings that discriminative power between X' and Y' is greater than discriminative power between X and Y. In addition, from the comparison between Eqs. (4) and (11), Eqs. (5) and (12), in order to enhance the discriminative power between genuine speech and playback speech, a method to modify LMS is proposed. The VMLMS can be obtained by adding LMS and its variance. Figure 1(a) shows the framework how to obtain VMLMS on the basis of LMS.

2.2 Mean-based modified log magnitude spectrum

Supposing X'' and Y'' satisfy

$$X'' = \left\{ \log(x_1) + \overline{X}, \log(x_2) + \overline{X}, ..., \log(x_K) + \overline{X} \right\} \quad (20)$$

$$Y'' = \left\{ \log(y_1) + \overline{Y}, \log(y_K) + \overline{Y}, ..., \log(y_K) + \overline{Y} \right\} \quad (21)$$

The means of X'' and Y'', denoted as $\overline{X''}$ and $\overline{Y''}$, which are as follows:

$$\begin{aligned} \overline{X''} &= \frac{\sum_{k=1}^{K}(\log(x_k) + \overline{X})}{K} \\ &= \frac{\sum_{k=1}^{K}\log(x_k)}{K} + \overline{X} \\ &= 2\overline{X} \end{aligned} \quad (22)$$

$$\begin{aligned} \overline{Y''} &= \frac{\sum_{k=1}^{K}(\log(y_k) + \overline{Y})}{K} \\ &= \frac{\sum_{k=1}^{K}\log(y_k)}{K} + \overline{Y} \\ &= 2\overline{Y} \end{aligned} \quad (23)$$

The variances of X'' and Y'', denoted as $\sigma_{X''}^2$ and $\sigma_{Y''}^2$, which are as follows:

$$\begin{aligned} \sigma_{X''}^2 &= \frac{\sum_{k=1}^{K}(\log(x_k) + \overline{X} - \overline{X''})^2}{K} \\ &= \frac{\sum_{k=1}^{K}(\log(x_k) + \overline{X} - 2\overline{X})^2}{K} \\ &= \frac{\sum_{k=1}^{K}(\log(x_k) - \overline{X})^2}{K} \\ &= \sigma_X^2 \end{aligned} \quad (24)$$

$$\begin{aligned} \sigma_{Y''}^2 &= \frac{\sum_{k=1}^{K}(\log(y_k) + \overline{Y} - \overline{Y''})^2}{K} \\ &= \frac{\sum_{k=1}^{K}(\log(y_k) + \overline{Y} - 2\overline{Y})^2}{K} \\ &= \frac{\sum_{k=1}^{K}(\log(y_k) - \overline{Y})^2}{K} \\ &= \sigma_Y^2 \end{aligned} \quad (25)$$

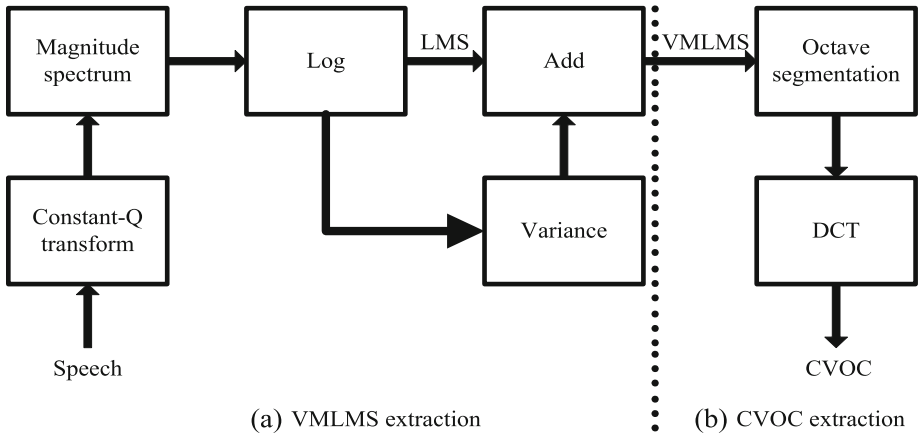

Fig. 1 Schematic diagram of constant-Q variance-based octave coefficients extraction, including **a** VMLMS extraction and **b** CVOC extraction

Supposing $F_{X''Y''}$ is the Fisher's ratio between X'' and Y'', according to Eq. (1), we can obtain

$$\begin{aligned} F_{X''Y''} &= \frac{(\overline{X''} - \overline{Y''})^2}{\sigma_{X''}^2 + \sigma_{Y''}^2} \\ &= \frac{(2\overline{X} - 2\overline{Y})^2}{\sigma_X^2 + \sigma_Y^2} \\ &= \frac{4(\overline{X} - \overline{Y})^2}{\sigma_X^2 + \sigma_Y^2} \\ &= 4F_{XY} \end{aligned} \tag{26}$$

From Eq. (26), we can see that $F_{X''Y''}$ is four times of F_{XY}. In other words, discriminative power between X'' and Y'' is greater than discriminative power between X and Y. Hence, the other method to modify LMS is proposed. MMLMS can be obtained by adding LMS and its mean. Figure (2)(a) shows the framework how to obtain MMLMS on the basis of LMS.

3 Proposed method II: hand-crafted discriminative features extraction

In this section, CVOC and CMOC extraction is introduced. Figures 1 and 2 show the block diagram of CVOC and CMOC extraction, respectively.

From Fig. 1, it can be seen that it consists of two parts: (a) VMLMS extraction and (b) CVOC extraction, in which CVOC is obtained on the basis of VMLMS. Further, there are five modules in VMLMS extraction, which are CQT, magnitude spectrum, log, variance, and add. There are two modules in CVOC extraction on the basis of VMLMS, which are octave segmentation and DCT.

From Fig. 2, it can be observed that it consists of two parts: (a) MMLMS extraction and (b) CMOC extraction, in which CMOC is obtained on the basis of MMLMS. Further, there are five modules in MMLMS extraction, which are CQT, magnitude spectrum, log, mean, and add. There are two modules in CMOC extraction on the basis of MMLMS, which are octave segmentation and DCT.

The module of CQT is used to convert speech from the time domain into the frequency domain. Magnitude spectrum is used to obtain magnitude spectrum on the basis of CQT. Log is used to obtained LMS. The modules of variance (mean) and add are used to obtained VMLMS (MMLMS) on the basis of LMS. Octave segmentation is used to segment MLMS frequency bins into blocks according to octave. The DCT is used to extract principal information of every block. Next, CQT, octave segmentation, and DCT will be introduced in detail.

3.1 Constant-Q transform

The CQT was proposed [51, 52]. Here, Q is defined as the ratio of center frequency to bandwidth, which is as Eq. (27), in which, f_m is center frequency and δ_f is the bandwidth.

$$Q = \frac{f_m}{\delta_f} \tag{27}$$

where f_m represents mth frequency bin and it obeys

$$f_m = f_1 2^{\frac{m-1}{B}} \tag{28}$$

where f_1 is the center frequency of the lowest-frequency bin, B is the number of bins in every octave.

From Eq. (28), we can see that every frequency bin has different frequency bandwidth, the more k, the more bandwidth. This is different from the frequency region in DFT in which every frequency bin has the equal frequency bandwidth.

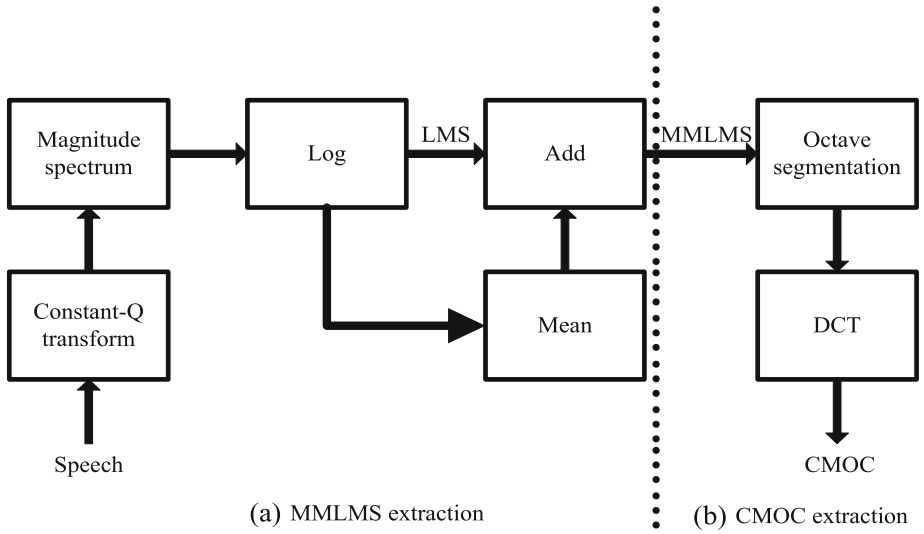

Fig. 2 Schematic diagram of constant-Q mean-based octave coefficients extraction, including **a** MMLMS extraction and **b** CMOC extraction

For a discrete time domain signal $x(n)$, supposing $Y(m,n)$ is its CQT, which is defined as

$$Y(m,n) = \sum_{j=n-\lfloor \frac{N_m}{2} \rfloor}^{n+\lfloor \frac{N_m}{2} \rfloor} x(j) a_m^*(j-n-\frac{N_m}{2}) \tag{29}$$

where $m = 1, 2, ...K$ is the frequency bin index, N_m are the variable window lengths, $a_m^*(n_1)$ denotes the complex conjugate of $a_m(n)$, and $\lfloor \bullet \rfloor$ denotes rounding toward negative infinity. The basic functions $a_m(n)$ are complex-valued time-frequency atoms and are defined by

$$a_m(n) = \frac{1}{C} v(\frac{n}{N_m}) \exp[\, i(2\pi n \frac{f_m}{f_s} + \phi_m)] \tag{30}$$

where f_m is the center frequency of m-th bin, f_s is the sampling rate, and $v(t)$ is a window function (e.g., Hanning window). ϕ_m is a phase offset. C is a scaling factor and

$$C = \sum_{n'=-\lfloor \frac{N_m}{2} \rfloor}^{\lfloor \frac{N_m}{2} \rfloor} v(\frac{n' + \frac{N_m}{2}}{N_m}) \tag{31}$$

3.2 Octave segmentation and discrete cosine transform

In our previous work, octave segmentation was proposed to segment magnitude-phase spectrum [40] and octave power spectrum [53]. In this work, octave segmentation is used here to segment VMLMS (MMLMS) into un-overlapped blocks according to octave. After octave segmentation, every block has B frequency bins. And then DCT is used to extract principal information of every block. Next, we will take VMLMS as an example to show how to calculate the final coefficients.

For $Y(m,n)$, after log operation, we can obtain $\log(|Y(m,n)|)$. Considering $\log(|Y_{VMLMS}(m,n)|)$ is the modified $\log(|Y(m,n)|)$. Further, after octave segmentation is supplied on $\log(|Y_{VMLMS}(m,n)|)$, it can be written as

$$\log(|Y_{VMLMS}(m,n)|) = \left\{ block_1, block_2, ..., block_R \right\} \tag{32}$$

where R represents total octave number, and it satisfies

$$R = \frac{K}{B} \tag{33}$$

After DCT is employed on every block. For every block DCT result, the former Z dimensions as selected as feature (Z is a positive integer), we can obtain CVOC of $x(n)$, denoted as $CVOC_{x(n)}$.

$$CVOC_{x(n)} = \left\{ C_1(0), C_1(z), ..., C_R(0), C_R(z) \right\} \tag{34}$$

where z is from 1 to Z-1 and

$$C_1(0) = \sqrt{\frac{1}{B}} \sum_{b=1}^{B} block_1 \tag{35}$$

$$C_1(z) = \sqrt{\frac{2}{B}} \sum_{b=1}^{B} \text{block}_1 \cos\left\{ \frac{(2b-1)z\pi}{2B} \right\} \tag{36}$$

$$C_R(0) = \sqrt{\frac{1}{B}} \sum_{b=(R-1)\times B+1}^{R\times B} \text{block}_R \tag{37}$$

$$C_R(z) = \sqrt{\frac{2}{B}} \sum_{b=(R-1)\times B+1}^{R\times B} \text{block}_R \cos\left\{ \frac{(2b-1)z\pi}{2B} \right\} \tag{38}$$

4 Studies on ASVspoof 2017

In this section, CVOC and CMOC are evaluated on ASVspoof 2017 corpus version 2.0 (ASVspoof 2017 V2) for playback speech detection.

4.1 Database introduction

ASVspoof 2017 corpus was released after ASVspoof 2017 challenge [1, 30] . However, the organizers found some zero-value samples and silence in ASVspoof 2017 will affect the result of playback speech detection. In 2018, the organizers updated ASVspoof 2017 by removing those zero-value samples and silence, and named the correct version as ASVspoof 2017 V2 [30]. It is constituted by three subsets: training data, development, and evaluation data, Table 2 gives some details of ASVspoof 2017 V2.

4.2 Evaluation rule and experimental setup

In ASVspoof 2017 challenge, participants are allowed to pool training data with development data together to train a final model. Equal error rate (EER) is used as evaluation metric. According to ASVspoof 2017 challenge rule, two types models are trained, one is used to evaluate the performance of the proposed features on evaluation set, wherein 4724 utterances from training and development

Table 2 Number of speakers and the corresponding number of genuine and playback utterances in the training, development, and evaluation sets in ASVspoof 2017 V2

Subset	Number			
	Speakers	Utterances	Genuine	Spoofed
Training	10	3014	1507	1507
Development	8	1710	760	950
Evaluation	24	13,306	1298	12,008

set are used; the other is used to evaluate the performance of the proposed features on development set, wherein 3014 utterances from training set are used.

In CQT, there are several important parameters, which will affect the final performance. They are the number of bins in a octave (B), octave number (R), sampling period that is used for re-sampling to transform octave power spectrum into linear power spectrum [34], gamma, respectively. In the process of CMOC and CVOC, all the parameters in CQT are set according to [34], in which, B is set as 96, sampling period is set as 16, gamma is set as 3.3026, and R is set as 9, which means that there are 9 octaves in CQT. In addition, in CMOC extraction, Z is set as 12, which means there are 12 coefficients obtained from every block after DCT is applied on it. Therefore, the static dimension of CMOC is set as 108. Since our previous works [38, 40] have shown static features will degrade the performance in playback speech detection, only dynamic features are used in this study. For different feature combinations of CMOC and CVOC dynamic features, D and A represent delta and acceleration, respectively.

In this study, similar to our previous playback speech detection studies [38, 40, 41], deep neural network (DNN) is selected as a suitable classifier because we found that DNN based systems can give better performance. The reason may be that DNN has both a classifier function and feature-learning ability [54]. Computational network Toolkit [55] is used to train DNN, which is used as classifier in our experiments. In addition, in the DNN training process, stochastic gradient descent is used. In our experiments, a series of four-layer DNN classifiers are trained for different feature combinations of CMOC and CVOC, which have two hidden layers with 512 nodes at every layer and output layer with 2 nodes and the input nodes number constituted by 11-frame context window of the input feature vector. In other words, for different feature combining of CVOC and CMOC dynamic features, the input nodes are different, for example, for CVOC-A, the input node number is 108 $\times$11 (including left five frames and right five frames), while for CMOC-DA , the input node number is 216 $\times$11.

All the DNNs trained in our experiments follow the same method, which consists of the following: (1) The training criterion is cross-entropy with softmax. (2) sigmoid network is used for the hidden layers training. (3) Mean and variance normalization is supplied on the input data. (4) In DNN training, stochastic gradient descent is used. (5) The learning rate is set as 0.8 for the first epoch, 3.2 for from second to fifteenth epochs, and 0.08 for sixteenth to twenty-fifth, in DNN training, there are totally 25 epochs. (6) The minibatch size is set as 256 for the first eopchs and 1024 for the rest epochs. (7) 0.9 is set for the momentum.

4.3 Experiment results and analysis

Table 3 gives the experimental results on ASVspoof 2017 V2 development set using dynamic features of CMOC and CVOC. From Table 3, two conclusions can be obtained: (1) For CMOC, CMOC-A can give the best performance on ASVspoof 2017 V2 development set, then followed by CMOC-DA and CMOC-D. (2) For CVOC, CVOC-DA performs better than CVOC-A, and then CVOC-A performs better than CVOC-D on ASVspoof 2017 V2 development set.

Table 4 gives the experimental results on ASVspoof 2017 V2 evaluation set using different dynamic features of CMOC and CVOC. From Table 4, several conclusions can be drawn: (1) For CMOC, CMOC-DA gives the best performance on ASVspoof 2017 V2 evaluation set, then followed by CMOC-D and CMOC-A. (2) For CVOC, CVOC-DA performs better than CVOC-D and then CVOC-D performs better than CVOC-A on ASVspoof 2017 V2 evaluation set. (3) Comparing Table 3 with Table 4, it can be seen that CMOC-A performs the best on development set while CMOC-DA on evaluation set, also it can be observed that CVOC-DA gives the best performance on development and evaluation set. (4) CVOC-DA performs better than CMOC-DA on ASVspoof 2017 V2 evaluation set. As mentioned above, CVOC and CMOC are obtained by applying octave segmentation plus DCT on VMLMS and MMLMS, respectively. Further, VMLMS is obtained by statistical analysis method while MMLMS is obtained by maths formula. Though we cannot compare their discriminative power using Fisher's ratio directly, we can say that CVOC has more discriminative power than CMOC on ASVspoof 2017 V2 evaluation set from the experimental result.

4.4 Comparison with modified log magnitude spectrum

In this subsection, modified log magnitude spectrum, namely, MMLMS and VMLMS, their performance is compared with corresponding CMOC and CVOC on ASVspoof 2017 V2 evaluation set. Table 5 gives the comparison with modified log magnitude spectrum on ASVspoof 2017 V2 evaluation set in terms of EER. In which, DNN is also used to model MMLMS and VMLMS, respectively. From Table 5, it can be seen that CMOC performs better than MMLMS and then CVOC performs better than VMLMS on ASVspoof 2017 V2 evaluation set,

Table 3 Experimental results (EER(%)) on ASVspoof 2017 V2 development set using dynamic features of CMOC and CVOC

Feature combinations	CMOC	CVOC
D	23.16	19.49
A	16.24	17.75
DA	22.70	17.41

Table 4 Experimental results (EER(%)) on ASVspoof 2017 V2 evaluation set using dynamic features of CMOC and CVOC

Feature combinations	CMOC	CVOC
D	15.37	12.81
A	18.28	16.03
DA	14.16	11.46

respectively. The reason is that more discriminative information can be obtained by applying octave segmentation plus DCT on the modified spectrums, which can make EER reduce 16.46% and 12.25%, respectively.

4.5 Comparison with Gaussian mixture model

In this subsection, the performance of CMOC-DA and CVOC-DA using the DNN will be compared with the corresponding performance using GMM as the model of CMOC-DA and CVOC-DA on ASVspoof 2017 V2. Table 6 shows that the comparison with GMM on ASVspoof 2017 V2 evaluation set in terms of EER, in which, the mixture of the GMM is 512. From Table 6, several conclusions can be obtained: (1) For CMOC-DA, the EER can increase from 14.16% to 31.33%, which increases by 121.26%. (2) For CVOC-DA, the EER can increase from 11.46% to 30.56%, which increases by 166.67%. (3) From the performance comparison, we can say that DNN can perform better than GMM on ASVspoof 2017 V2 evaluation set for CMOC-DA and CVOC-DA, the reason is that DNN has feature learning ability as well as classification, it also confirms that consideration of DNN for our studies is useful.

4.6 Comparison with some commonly used features

In this section, some commonly used features, for example, MFCC and CQCC are compared and with CVOC and CMOC on ASVspoof 2017 V2 evaluation set. In addition, considering the modules of variance or add are removed from Fig. 1 or the modules of mean and add are removed from Fig. 2, the obtained feature can be named as constant-Q octave coefficients (COC). It can be used to compare the performance with CVOC and CMOC to show the role of VMLMS and MMLMS in CVOC and CMOC.

Table 7 gives the performance comparison among MFCC-DA, CQCC-DA, COC-DA, CMOC-DA, and CVOC-DA on ASVspoof 2017 V2 evaluation set in terms of EER. In which, MFCC-DA, CQCC-DA, and COC-DA have their respective DNN classifiers. From Table 7, it can be seen that (1) the performance of CQCC-DA, COC-DA, CMOC-DA, and CVOC-DA is better than MFCC-DA. The reason is that MFCC-DA is based on DFT which is a short-term transform while the other four features are based on CQT which is a long-term transform. CQT can provide more frequency details. (2) Both CVOC-DA and CMOC-DA perform better than COC-DA, which means that our proposed VMLMS and MMLMS have more discriminative power toward playback speech. In addition, it also confirms that our idea is correct and effective. (3) The performance of CVOC-DA and CMOC-DA is better than CQCC-DA and COC-DA, the reason is that modified log magnitude spectrum is used the two feature extraction.

Table 5 Comparison with modified log magnitude spectrum on ASVspoof 2017 V2 evaluation set in terms of EER (%)

Feature	EER	Features	EER
MMLMS-DA	16.95	CMOC-DA	14.16
VMLMS-DA	13.06	CVOC-DA	11.46

Table 6 Comparison with GMM on ASVspoof 2017 V2 evaluation set in terms of EER (%)

Feature	Model	EER
CMOC-DA	DNN	14.16
	GMM	31.33
CVOC-DA	DNN	11.46
	GMM	30.56

4.7 Comparison with some other known systems

Table 8 gives the comparison with some known systems based on hand-crafted features on ASVspoof 2017 V2 evaluation set. In which, logE represents logarithm energy, qDFTspe represents Q-log domain DFT-based mean normalized log spectral [42], eCQCC represents extended CQCC [38], CMPOC represents constant-Q magnitude-phase octave coefficients [40] and CQSPIC represents constant-Q statistics-plus-principal information coefficients [41].

From Table 8, it can be seen that the performance of our systems are better than some other known systems. The reason may be that discriminative features are used our systems. However, our systems are a little worse than the system based on qDFTspe [42]. In addition, feature combination SDA perform the best in [30] while feature combination DA performs the best in our system.

Table 7 Comparison with some commonly used features on ASVspoof 2017 V2 evaluation set in terms of EER (%)

Feature	EER
MFCC-DA	23.79
CQCC-DA	15.18
COC-DA	14.26
CMOC-DA	14.16
CVOC-DA	11.46

Table 8 Comparison with some known systems based on hand-crafted features on the ASVspoof 2017 V2 evaluation set in terms of EER(%)

Feature	Classifier	EER
CQCC-SDA [30]	GMM	15.33
(CQCC-logE)-SDA [30]	GMM	12.24
CQCC-SDA [30]	I-vector	15.63
(CQCC-logE)-SDA [30]	I-vector	12.93
qDFTspe [42]	GMM	11.19
CQCC-SDA	DNN	32.46
CQCC-DA	DNN	15.18
CMPOC-DA [40]	DNN	14.99
CMPOC-D [40]	DNN	14.93
eCQCC-DA [38]	DNN	13.38
CQSPIC-DA [41]	DNN	11.09
CMOC-DA	DNN	14.16
CVOC-DA	DNN	11.46

The reason is that cepstral mean and variance normalization (CMVN) is applied on feature in [30] and the feature distribution has been changed while CMVN is not applied on our feature. We also found that CQSPIC performs better than CVOC and CMOC, the reason is that CQSPIC is a combined feature, it has spectral principal information, subband information, and short-term spectral statistical information while our CVOC and CMOC only has spectral principal information.

5 Studies on ASVspoof 2019 physical access

5.1 Database introduction and evaluation metric

In this section, CMOC and CVOC are evaluated on ASVspoof 2019 physical access [48], which was released in 2019 for ASVspoof 2019 challenge, some details are given in Table 9. In which, the corpus has three subset, train, development, and evaluation set. According to ASVspoof 2019 challenge rule, tandem detection cost function (t-DCF) [56] and EER are used as the primary and secondary metric, respectively, which is the same as the previous works [57–64].

Table 9 Details of ASVspoof 2019 physical access corpus

Subset	# Speakers		# Utterances	
	Male	Female	Bonafide	Spoof
Training	8	12	5400	48,600
Development	8	12	5400	24,300
Evaluation	30	37	18,089	13,4630

Table 10 Experimental results (t-DCF and EER(%)) on ASVspoof 2019 physical access development set using dynamic features of CMOC and CVOC

Feature	Feature combinations	t-DCF	EER
CMOC	D	0.235	13.686
	A	0.198	10.906
	DA	0.220	12.705
CVOC	D	0.234	12.737
	A	0.165	8.430
	DA	0.232	12.665

5.2 Experimental results and analysis

Table 10 gives the experimental results on ASVspoof 2019 physical access development set using dynamic features of CMOC and CVOC. From Table 10, according to t-DCF or EER, two conclusions can be obtained: (1) For CMOC, CMOC-A can give the best performance on ASVspoof 2019 physical access development set, then followed by CMOC-DA and CMOC-D. (2) For CVOC, CVOC-A performs better than CVOC-DA, and then CVOC-DA performs better than CVOC-D on ASVspoof 2019 physical access development set.

Table 11 gives the experimental results on ASVspoof 2019 physical access evaluation set using different dynamic features of CMOC and CVOC. From Table 11, several conclusions can be drawn: (1) For CMOC, CMOC-A gives the best performance on ASVspoof 2019 physical access evaluation set, then followed by CMOC-DA and CMOC-D. (2) For CVOC, CVOC-A performs better than CVOC-DA and then CVOC-DA performs better than CVOC-D on ASVspoof 2019 physical access evaluation set. (3) Comparing Table 10 with Table 11, it can be seen that CMOC-A and CVOC-A perform the best on ASVspoof 2019 physical access development and evaluation set. (4) CVOC-A performs better than CMOC-A on ASVspoof 2019 physical access development and evaluation set. Which also confirms that CVOC-A has more discriminative ability than CMOC-A, the same as on ASVspoof 2017 evaluation set.

Table 11 Experimental results (t-DCF and EER(%)) on ASVspoof 2019 physical access evaluation set using dynamic features of CMOC and CVOC

Feature	Feature combinations	t-DCF	EER
CMOC	D	0.225	13.614
	A	0.208	11.447
	DA	0.212	12.582
CVOC	D	0.221	12.610
	A	0.178	9.269
	DA	0.213	12.029

Table 12 Comparison with modified log magnitude spectrum on ASVspoof 2019 physical access evaluation set in terms of t-DCF and EER (%)

Feature	t-DCF	EER	Feature	t-DCF	EER
MMLMS-A	0.357	16.330	CMOC-A	0.208	11.447
VMLMS-A	0.306	14.703	CVOC-A	0.178	9.269

5.3 Comparison with modified log magnitude spectrum

In this subsection, modified log magnitude spectrum, the performance of MMLMS and VMLMS is compared with their corresponding CMOC and CVOC on ASVspoof 2019 physical access evaluation set. Table 12 gives the comparison with modified log magnitude spectrum on ASVspoof 2019 physical access evaluation set in terms of EER. In which, DNN is also used to model MMLMS-A and VMLMS-A, respectively. From Table 12, it can be seen that CMOC-A, CVOC-A perform much better than corresponding MMLMS-A and VMLMS-A on ASVspoof 2019 physical access evaluation set in terms of t-DCF or EER, respectively. The reason is that more discriminative information can be obtained by applying octave segmentation plus DCT on the modified spectra.

5.4 Comparison with Gaussian mixture model

In this subsection, the performance of CMOC-A and CVOC-A using the DNN will be compared with the corresponding performance using GMM as the model of CMOC-A and CVOC-A on ASVspoof 2019 physical access. Table 13 shows that the comparison with GMM on ASVspoof 2019 physical access evaluation set in terms of EER, in which, the mixture of the GMM is 512. From Table 13, several conclusions can be obtained: (1) For CMOC-A, the t-DCF increases from 0.208 to 0.411, which increases by 97.60%. In addition, the EER can increase from 11.447% to 21.128%, which increases by 84.57%. (2) For CVOC-A, the t-DCF increases from 0.178 to 0.379, which increases by 107.30%. In addition, the EER can increase from 9.269% to 18.850%, which increases by 103.37%. (3) From the performance comparison, we can say that DNN can perform better than GMM on ASVspoof 2019 physical access evaluation set for CMOC-A and CVOC-A.

Table 13 Comparison with GMM on ASVspoof 2019 physical access evaluation set in terms of t-DCF and EER (%)

Feature	Model	t-DCF	EER
CMOC-A	DNN	0.208	11.447
	GMM	0.411	21.128
CVOC-A	DNN	0.178	9.269
	GMM	0.369	18.850

Table 14 Comparison with some commonly used features on ASVspoof 2019 physical access evaluation set in terms of t-DCF and EER (%)

Feature	t-DCF	EER
MFCC-A	0.427	21.227
CQCC-A	0.389	21.193
eCQCC-A [38]	0.331	14.118
CQSPIC-A [41]	0.204	10.690
COC-A	0.209	11.608
CMOC-A	0.208	11.447
CVOC-A	0.178	9.269

5.5 Comparison with some commonly used features

Table 14 gives the performance comparison among MFCC-A, CQCC-A, COC-A, CMOC-A, eCQCC-A, CQSPIC-A, and CVOC-A on ASVspoof 2019 physical access evaluation set in terms of t-DCF and EER. In which, eCQCC represents extended CQCC (eCQCC) [38], CMPOC represents constant-Q magnitude-phase octave coefficients [40] and CQSPIC represents constant-Q statistics-plus-principal information coefficients [41]. In addition, MFCC-A, CQCC-A, eCQCC-A, CQSPIC-A, and COC-A have their respective DNN classifiers. From Table 14, according to t-DCF or EER, it can be seen that (1) The performance of CQCC-A, COC-A, eCQCC-A, CQSPIC-A, CVOC-A, and CMOC-A is better than MFCC-DA. The reason is that MFCC-DA is based on DFT which is a short-term transform while the other four features are based on CQT which is a long-term transform. CQT can provide more frequency details. (2) Both CMOC-A and CVOC-A perform better than COC-A, which also confirms that our proposed VMLMS and MMLMS have more discriminative power than LMS toward playback speech. (3) Similar to the performance between CVOC and eCQCC on ASVspoof 2017 V2, CVOC also give better performance than eCQCC on ASVspoof 2019 physical access evaluation set. It means that CVOV has more discriminative ability than eCQCC on the two databases. (4) It is surprising to found that CVOC-A performs better than CQSPIC-A on ASVspoof 2019 physical access evaluation set unlike the comparison between them on ASVspoof 2017 V2 evaluation set. The reason may be that CVOC can extract more discriminative information than CQSPIC on ASVspoof 2019 physical access evaluation set. (5) The performance of CMOC-A and CVOC-A is better than CQCC-A and COC-A, the reason is that modified log magnitude spectrum is used the two feature extraction.

5.6 Comparison with some other known systems

Table 15 gives the comparison with some known systems based on hand-crafted features on ASVspoof 2019 physical access evaluation set. In which, LFCC represents linear frequency cepstral coefficients. From Table 15, it can be

Table 15 Comparison with some known systems on ASVspoof 2019 physical access evaluation set in terms of t-DCF and EER (%)

Feature	Model	t-DCF	EER
CQCC [48]	GMM	0.245	11.04
LFCC [48]	GMM	0.302	13.54
CMOC-A	DNN	0.208	11.447
CVOC-A	DNN	0.178	9.269

seen that the performance of our systems are better than the two known systems. The reason is that discriminative features are used our systems.

6 Conclusion

This paper addresses the problem how to extract handcrafted discriminative features for playback speech detection. Two methods to obtain modified log magnitude spectrum are proposed by analyzing the discriminative power between genuine speech and playback speech using Fisher's ratio. Then, CVOC and CMOC are extracted by using octave segmentation and DCT on the basis of VMLMS and MMLMS, respectively. The experimental results on ASVspoof 2017 V2 and ASVspoof 2019 physical access databases show that both CVOC and CMOC perform better than some commonly used features because VMLMS and MMLMS can enhance the discriminative power between genuine speech and playback speech. In addition, CVOC can perform better than CMOC on the two databases, which means that CVOC has more discriminative power than CMOC. The EER of CVOC on ASVspoof 2017 V2 evaluation set can reach 11.46%, and the t-DCF on ASVspoof 2019 physical access evaluation set can achieve 0.165. It is somewhat surprising to find that the proposed method can work so well. Future work can explore how far this idea can be extended.

Abbreviations

ASV: Automatic speaker verification; APSD: Authentic and playback speech database; CQT: Constant-Q transform; ResNet: Residual network; CQCC: Constant-Q cepstral coefficients; DFT: Discrete Fourier transform; MFCC: Mel frequency cepstral coefficients; VESA: Variable length energy separation algorithm; GMM: Gaussian mixture model; LMS: Log magnitude spectrum; DCT: Discrete cosine transform; VMLMS: Variance-based modified log magnitude spectrum; MMLMS: Mean-based modified log magnitude spectrum; CVOC: Constant-Q variance-based octave coefficients; CMOC: Constant-Q mean-based octave coefficients; EER: Equal error rate; DNN: Deep neural network; COC: Constant-Q octave coefficients; loge: Logarithm energy; qDFTspe: Q-log domain DFT-based mean normalized log spectra; eCQCC: Extended CQCC; CMPOC: Constant-Q magnitude-phase octave coefficients; CQSPIC: Constant-Q statistics-plus-principal information coefficients; CMVN: Cepstral mean and variance normalization; t-DCF: Tandem detection cost function

Authors' contributions

JY and LX designed the idea. BR and YJ did the experiments. All the authors contributed to the writing of this work. In addition, both JY and LX have equal contributions for the work and are equally first authors. All author(s) read and approved the final manuscript.

Funding

This work was supported by the Programmatic Grant A1687b0033 through the Singapore Government's Research, Innovation and Enterprise 2020 Plan (Advanced Manufacturing and Engineering domain), Shanghai Sailing Program (no. 19YF1402000), the Fundamental Research Funds for the Central Universities (no. 2232019D3-52), the National Natural Science Foundation of China (no. 61601248), and the Initial Research Funds for Young Teachers of Donghua University.

Availability of data and materials

The datasets used and analyzed during the current study are available online. ASVspoof 2017 V2 dataset is available from (https://datashare.is.ed.ac.uk/handle/10283/3055). ASVspoof 2019 physical access dataset is available from (https://datashare.is.ed.ac.uk/handle/10283/3336).

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Electrical and Computer Engineering, National University of Singapore, Singapore, Singapore. [2]College of Information Science and Technology, Donghua University, Shanghai China. [3]Microsoft Search Technology Center Asia, Suzhou, China. [4]Electronics and Information School, Nantong University, Nantong, China.

Received: 29 May 2019 Accepted: 10 March 2020
Published online: 07 April 2020

References

1. T. Kinnunen, M. Sahidullah, M. Falcone, L. Costantini, R. G. Hautamaki, D. Thomsen, S. Achintya, Z.-H. Tan, H. Delgado, M. Todisco, N. Evans, V. Hautamaki, K. A. Lee, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. RedDots replayed: a new replay spoofing attack corpus for text-dependent speaker verification research, (2017), pp. 5395–5399. https://doi.org/10.1109/icassp.2017.7953187
2. T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, K. A. Lee, in *Annual Conference of the International Speech Communication Association*. The ASVspoof 2017 challenge: assessing the limits of replay spoofing attack detection, (2017), pp. 2–6. https://doi.org/10.21437/interspeech.2017-1111
3. K. N. R. K. R. Alluri, S. Achanta, S. R. Kadiri, S. V. Gangasheetty, A. K. Vuppala, in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Sff anti-spoofer: IIIT-H submission for automatic speaker verificatio spoofing and countermeasures channlenge 2017, (2017), pp. 107–111. https://doi.org/10.21437/interspeech.2017-676
4. Z. Chen, Z. Xie, W. Zhang, X. Xu, in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ResNet and model fusion for automatic spoofing detection, (2017), pp. 102–106. https://doi.org/10.21437/interspeech.2017-1085
5. S. Jelil, R. K. Das, S. R. M. Prasanna, R. Sinha, in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Spoof detection using source, instantaneous frequecny and cepstral features, (2017), pp. 22–26. https://doi.org/10.21437/interspeech.2017-930
6. H. A. Patil, M. R. Kamble, T. B. Patel, M. Soni, in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Novel variable length teager engery separation based on instantaneous frequency features for replay detection, (2017), pp. 12–16. https://doi.org/10.21437/interspeech.2017-1362
7. A. G. AlanÃs, A. M. Peinado, J. A. Gonzalez, A. Gomez, in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. A deep identity representation for noise robust spoofing detection, (2018), pp. 676–680. https://doi.org/10.21437/interspeech.2018-1909
8. G. Suthokumar, V. Sethu, C. Wijenayake, E. Ambikairajah, in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Modulation dynamic features for the detection of replay attacks, (2018), pp. 691–695. https://doi.org/10.21437/interspeech.2018-1846
9. B. Wickramasinghe, S. Irtza, E. Ambikairajah, J. Epps, in *19th Annual Conference of the International Speech Communication Association*

(INTERSPEECH). Frequency domain linear prediction features for replay spoofing attack detection, (2018), pp. 661–665. https://doi.org/10.21437/interspeech.2018-1574
10. M. S. Saranya, H. Murthy, in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Decision-level feature switching as a paradigm for replay attack detection, (2018), pp. 686–690. https://doi.org/10.21437/interspeech.2018-1494
11. L. Li, Y. Chen, D. Wang, T. F. Zheng, in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. A study on replay attack and anti-spoofing for automatic speaker verification, (2017), pp. 92–96. https://doi.org/10.21437/interspeech.2017-456
12. R. Font, J. M. EspÃn, M. J. Cano, in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Experimental analysis of features for replay attack detection-results on the asvspoof 2017 challenge, (2017), pp. 7–11. https://doi.org/10.21437/interspeech.2017-450
13. M. Withowski, S. Kacprasko, P. Zelasko, K. Kowlczyk, J. Galka, in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Audio replay attack detection using high-frequency features, (2017), pp. 27–31. https://doi.org/10.21437/interspeech.2017-776
14. G. Lavrentyeva, S. Novoselov, E. Malykh, A. Kozlov, O. Kudasher, V. Shchemelinin, in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Audio replay attack detection with deep learning framework, (2017), pp. 82–86. https://doi.org/10.21437/interspeech.2017-360
15. D. Li, L. Wang, J. Dang, M. Liu, Z. Oo, S. Nakagawa, H. Guan, X. Li, in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Multiple phase information combination for replay attacks detection, (2018), pp. 656–660. https://doi.org/10.21437/interspeech.2018-2001
16. T. Gunendradasan, B. Wickramasinghe, N. P. Le, E. Ambikairajah, J. Epps, in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Detection of replay-spoofing attacks using frequency modulation features, (2018), pp. 636–640. https://doi.org/10.21437/interspeech.2018-1473
17. H. B. Sailor, M. R. Kamble, H. A. Patil, in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Auditory filterbank learning for temporal modulation features in replay spoof speech detection, (2018), pp. 666–670. https://doi.org/10.21437/interspeech.2018-1651
18. M. Kamble, H. Tak, H. A. Patil, in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Effectiveness of speech demodulation-based features for replay detection, (2018), pp. 641–645. https://doi.org/10.21437/interspeech.2018-1675
19. W. Cai, H. Wu, D. Cai, M. Li, in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. The DKU replay detection system for the ASVspoof 2019 challenge on data augmentation, feature representation, clssifier and fusion, (Graz, Austria, 2019), pp. 1023–1027. https://doi.org/10.21437/interspeech.2019-1230
20. R. Bialobrzeski, M. Kosmider, M. Matuszewski, M. Plata, A. Rakowski, in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Robust Bayesian and light neural networks for voice spoofing detection, (Graz, Austria, 2019), pp. 1028–1032. https://doi.org/10.21437/interspeech.2019-2676
21. G. Lavrentyva, S. Novoselov, A. Tseren, M. Volkova, A. Gorlanov, A. Kozlos, in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. STC antispoofing systems for the ASVspoof2019 challenge, (Graz, Austria, 2019), pp. 1033–1037. https://doi.org/10.21437/interspeech.2019-1768
22. Y. Yang, H. Wang, H. Dinkel, Z. Chen, S. Wang, Y. Qian, K. Yu, in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. The SJTU robust anti-spoofing system for the ASVspoof 2019 challenge, (Graz, Austria, 2019), pp. 1038–1042. https://doi.org/10.21437/interspeech.2019-2170
23. J.-w. Jung, H.-j. Shim, H.-S. Heo, H.-J. Yu, in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Replay attack detection with complementary high-resolution information using end-to-end for the ASVspoof 2019 challenge, (Graz, Austria, 2019), pp. 1083–1087. https://doi.org/10.21437/interspeech.2019-1991
24. W. Shang, M. Stevenson, in *Proceddings of Canadian Conference onElectrical and Computer Engineering*. A preliminary study of factors affecting the performance of a playback attack detector, (2008), pp. 459–464. https://doi.org/10.1109/ccece.2008.4564576
25. Z. Wang, Q. He, X. Zhang, H. Luo, Z. Su, in *Journal of South China University of Technology (Natural Science Edition)*. Playback attack detection based on channel pattern noise, (2011), pp. 1708–1713
26. S. K. Ergunay, E. Khoury, A. Lazaridis, S. Marcel, in *Proceedings of the Seventh IEEE International Conference on Biometric Theory, Application Ans Systems*. On the vulnerability of speaker verification to realistic voice spoofing, (2015), pp. 1–6. https://doi.org/10.1109/btas.2015.7358783
27. W. Shang, M. Stevenson, in *IEEE International Conference on Acoustic, Speech and Signal Processing*. Score normalization in playback attack detection, (2010), pp. 1678–1681. https://doi.org/10.1109/icassp.2010.5495503
28. Z. Wang, G. Wei, Q. He, in *Proceedings of the 2011 International Conference on Machine Learning and Cybernetics, vol. 39*. Channel pattern noise based on playback attack detection algorithm for speaker recognition, (2011), pp. 5–12. https://doi.org/10.1109/icmlc.2011.6016982
29. C. Wang, Y. Zou, S. Liu, W. Zheng, in *IEEE Second International Conference on Multimedia Big Data*. An efficient learning based smartphone playback attack detection using GMM supervector, (2016), pp. 385–389. https://doi.org/10.1109/bigmm.2016.14
30. H. Delgado, M. Todisco, M. Sahidullah, N. Evans, T. Kinnunen, K. Lee, J. Yamagishi, in *Speaker and Language Recognition Workshop*. ASVspoof 2017 version 2.0: meta-data analysis and baseline enhancements, (2018), pp. 296–303. https://doi.org/10.21437/odyssey.2018-42
31. K. Sriskandaraja, V. Sethu, E. Ambikairajah, in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Deep siamese architecture based replay detection for secure voice biometric, (2018), pp. 671–675. https://doi.org/10.21437/interspeech.2018-1819
32. F. Tom, M. Jain, P. Dey, in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. End-to-end audio replay attack detection using deep convolutional networks with attention, (2018), pp. 681–685. https://doi.org/10.21437/interspeech.2018-2279
33. X. Cheng, M. Xu, T. F. Zheng, in *Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Replay detection using cqt-based modified group, (Lanzhou, China, 2019), pp. 540–545. https://doi.org/10.1109/apsipaasc47483.2019.9023158
34. M. Todisco, H. Delgado, N. Evans, in *Speaker and Language Recognition Workshop*. A new feature for automatic speaker verification anti-spoofing: constant Q cepstral coefficients, (2016). https://doi.org/10.21437/odyssey.2016-41
35. M. Todisco, H. Delgado, N. Evans, Constant Q cepstral coefficients: a spoofing countermeasure for automatic Speaker verification. Comput. Speech. Lang. **45**, 516–535 (2017)
36. Z. Ji, Z.-Y. Li, P. Li, M. An, S. Gao, D. Wu, F. Zhao, in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Ensemble learning for countermeasure of audio repaly spoofing attack in asvspoof2017, (2017), pp. 87–91. https://doi.org/10.21437/interspeech.2017-1246
37. X. Wang, Y. Xiao, X. Zhu, in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Feature selection based on CQCCs for automatic speaker verification spoofing, (2017), pp. 32–36. https://doi.org/10.21437/interspeech.2017-304
38. J. Yang, R. K. Das, H. Li, in *Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Extended constant-Q cepstral coefficients for detection of spoofing attacks, (2018), pp. 1024–1029. https://doi.org/10.23919/apsipa.2018.8659537
39. R. K. Das, J. Yang, H. Li, in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Long range acoustic features for spoofed speech detection, (Graz, Austria, 2019), pp. 1058–1062. https://doi.org/10.21437/interspeech.2019-1887
40. J. Yang, L. Liu, Playback speech detection based on magnitude-phase spectrum. Electron. Lett. **54**(14), 901–903 (2018)
41. J. Yang, C. You, Q. He, in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Feature with complementarity of statistics and principal information for spoofing detection, (2018), pp. 651–655. https://doi.org/10.21437/interspeech.2018-1693
42. J. Alam, G. Bhattacharya, P. Kenny, in *In Speaker and Language Recognition Workshop (ODYESSEY)*. Boosting the performance of spoofing detection

systems of replay attacks using Q-logarithm domain feature normalization, (2018), pp. 393–398. https://doi.org/10.21437/odyssey.2018-55
43. M. Kamble, H. A. Patil, in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Novel variable length energy separation algorithm using instantaneous amplitude features for replay detection, (2018), pp. 646–650. https://doi.org/10.21437/interspeech.2018-1687
44. S. Jelil, R. K. Das, S. R. M. Prasanna, R. Sinha, in *18th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Spoof detection using source, instantaneous frequency and cepstral features, (2017), pp. 22–26. https://doi.org/10.21437/interspeech.2017-930
45. R. K. Das, H. Li, in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. Instantaneous phase and excitation source features for detection of replay attacks, (Honolulu, Hawaii, 2018), pp. 1030–1037. https://doi.org/10.23919/apsipa.2018.8659789
46. K. N. R. K. R. Alluri, A. K. Vupala, Replay spoofing countermeasures using high spectro-temporal resolutional features. Int. J. Speech Technol., 1–11 (2019). https://doi.org/10.1007/s10772-019-09602-z
47. K. N. R. K. R. Alluri, A. K. Vupala, in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. IIIT-H spoofing countermeasures for automatic speaker verification spoofing and countermesures challenge 2019, (Graz, Austria, 2019), pp. 1043–1047. https://doi.org/10.21437/interspeech.2019-1623
48. M. Todisco, X. Wang, V. Vestman, M. Sahidullah, H. Delgado, A. Nautsch, J. Yamagishi, N. Evas, T. Kinnunen, K. A. Lee, in *20th Annual Conference of the International Speech Communication Association*. ASVspoof 2019: Future horizons in spoofed and fake audio detection, (2019), pp. 1008–1012
49. J. J. Wolf, Efficient acoustic parameter for speaker recognition. J. Acoust. Soc. Am. **51**(6B), 2044–2056 (1972)
50. D. Paul, M. Pal, G. Saha, Spectral features for synthetic speech detection. IEEE J. Sel. Top. Signal Process. **11**, 605–617 (2017)
51. J. Youngberg, S. Boll, in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Constant-q signal analysis and synthesis, (1978), pp. 375–378. https://doi.org/10.1109/icassp.1978.1170547
52. J. C. Brown, Calculation of a constant Q spectral transform. J. Acoust. Soc. Am. **89**(1), 425–434 (1991)
53. J. Yang, C. You, Q. He, in *19th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Feature with complementarity of statistics and principal information for spoofing detection, (2018), pp. 651–655. https://doi.org/10.21437/interspeech.2018-1693
54. F. Seide, G. Li, X. Chen, D. Yu, in *IEEE Workshop on Automatic Speech Recognition and Understanding*. Feature engineering in context-dependent deep neural networks for conversational speech transcription, (2011), pp. 24–29. https://doi.org/10.1109/asru.2011.6163899
55. F. Seide, A. Agarwal, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. CNTK:Microsoft's open-source deep learning toolkit, (2016), pp. 2135–2135. https://doi.org/10.1145/2939672.2945397
56. , in *The Speaker and Language Recognition Workshop*. t-DCF: a detection cost function for tandem assessment of spoofing countermeasures and automatic speaker verification, (2018), pp. 312–319. https://doi.org/10.21437/odyssey.2018-44
57. C.-I. Lai, N. Chen, J. Villaba, N. Dehak, in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. ASSERT:Anti-spoofing with squeeze-excitation and residual networks, (Graz, Austria, 2019), pp. 1013–1017. https://doi.org/10.21437/interspeech.2019-1794
58. B. Chettri, D. Stoller, V. Morfi, M. A. M. Ramirez, E. Benetos, B. L. Sturm, in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Ensemble models for spoofing detection in autoamtic speaker verification, (Graz, Austria, 2019), pp. 1018–1022. https://doi.org/10.21437/interspeech.2019-2505
59. R. Li, M. Zhao, Z. Li, L. Li, Q. Hong, in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Anti-spoofing speaker verification system with multi-feature integration and multi-task learning, (Graz, Austria, 2019), pp. 1048–1052. https://doi.org/10.21437/interspeech.2019-1698
60. J. Willoams, J. Rownika, in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Speech replay detection with x-vector attack emneddings and spectral features, (Graz, Austria, 2019), pp. 1053–1057. https://doi.org/10.21437/interspeech.2019-1760
61. S.-Y. Chang, K. C. Wu, C.-P. Chen, in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Transfer-representation learning for detecting spoofing attacks with converted and synthesized speech in automatic speaker verification system, (Graz, Austria, 2019), pp. 1063–1067. https://doi.org/10.21437/interspeech.2019-2014
62. A. Gomez-Alanis, A. M. Peinado, J. A. Gonzalez, A. M. Gomez, in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. A light convolutional GRU-RNN deep feature extractor for ASV spoofing detection, (Graz, Austria, 2019), pp. 1068–1072. https://doi.org/10.21437/interspeech.2019-2212
63. M. Alzanto, Z. Wang, M. B. Srivastava, in *20th Annual Conference of the International Speech Communication Association (INTERSPEECH)*. Deep residual neural networks for audio spoofing detection, (Graz, Austria, 2019), pp. 1078–1082. https://doi.org/10.21437/interspeech.2019-3174
64. R. K. Das, J. Yang, H. Li, in *Automatic Speech Recognition and Understanding Workshop (ASRU)*. Long range acoustic and deep features perspective on ASVspoof 2019, (Sentosa Island, Singapore, 2019), pp. 1018–1025. https://doi.org/10.1109/asru46091.2019.9003845

Publisher's Note