采用恒 Q 调制包络的合成语音伪装检测方法

徐嘉[1]，简志华[1,2]，金宏辉[1]，吴超[1]
（1. 杭州电子科技大学通信工程学院，浙江 杭州 310018;
2. 浙江省数据存储传输及应用技术研究重点实验室，浙江 杭州 310018）

摘　要：针对传统的声学特征参数对合成语音伪装检测时存在的准确度低、未知类型合成语音检测效果较差、在噪声环境中表现欠佳的情况，提出了一种采用恒 Q 调制包络（constant Q modulation envelope，CQME）的合成伪装语音检测方法。该方法基于语音时域包络中包含的丰富信息，而合成语音与真实语音的包络在细节上存在较大差异，利用恒 Q 变换（constant Q transform，CQT）得到语音调制包络谱，并计算每个频率成分的均方根，获得 CQME 特征向量。再用该特征向量训练随机森林分类器，实现真伪语音的判别。实验结果表明，在 ASVspoof 2019 数据集上，CQME 特征训练的随机森林具有较高的检测性能，对未知类型的合成语音也具有较好的检测效果。并且在多种噪声条件下，该方法仍表现出较高的检测性能，具有很好的噪声鲁棒性。

关键词：合成语音；伪装语音检测；恒 Q 调制包络；随机森林
中图分类号：TP391.42
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2023187

A method of synthetic speech spoofing detection using constant Q modulation envelope

XU Jia[1], JIAN Zhihua[1,2], JIN Honghui[1], WU Chao[1]
1. School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
2. Key Laboratory of Data Storage and Transmission Technology of Zhejiang Province, Hangzhou 310018, China

Abstract: In response to the low accuracy of synthetic speech spoofing detection based on traditional acoustic feature parameters, poor detection performance for unknown types of synthetic speech, and performance degradation in noisy environments, a method for detecting spoofing synthetic speech was proposed using constant Q modulation envelope (CQME) . The motivation of the method was from the fact that the temporal envelope of speech contained abundant information and there was a big difference in detail between the envelope of synthetic speech and genuine speech. The modulation envelope spectrum of speech was obtained by employing constant Q transform (CQT), and the root mean square of each frequency component was calculated to derive the CQME feature vector. And then the CQME feature vector was used to train the random forest classifier for discriminating genuine speech from spoofing synthet-

收稿日期：2023−06−13；**修回日期：**2023−09−30
通信作者：简志华，jianzh@hdu.edu.cn
基金项目：国家自然科学基金资助项目（No.61201301，No.61772166，No.61901154）
Foundation Items: The National Natural Science Foundation of China (No.61201301, No.61772166, No.61901154)

ic speech. Experimental results demonstrate that the random forest trained with CQME features achieves high detection performance on the ASVspoof 2019 dataset and exhibites good detection efficacy for unknown types of synthetic speech. Furthermore, the proposed method shows high detection performance even under various noise conditions, having excellent noise robustness.

Key words: synthetic speech, spoofing speech detection, constant Q modulation envelope, random forest

0 引言

近年来，随着人工智能技术的发展，自动说话人验证（automatic speaker verification，ASV）系统得到了广泛应用，但该系统易受伪装语音的恶意攻击，安全性有待提高，因此伪装语音检测技术应运而生。伪装语音检测系统通过对输入语音进行特征提取和模型匹配，从而识别该语音是真实说话人的语音还是伪装语音。常见的伪装语音通过语音合成（speech synthesis，SS）、语音转换（voice conversion，VC）、重放攻击（replay attack）等技术实现[1]。其中，语音合成可以根据任意文本生成伪装语音，具有任意性的特点，且随着计算机硬件设备的发展，合成语音质量越来越高，对 ASV 系统构成了极大的威胁[2]。伪装语音检测可以识别恶意伪装攻击，提高系统的安全性能，因此针对合成语音的伪装检测方法具有广阔的应用前景。

伪装语音检测方法通过提取语音信号的特征参数实现伪装检测。传统的特征参数主要有两种：基于声道谱的倒谱特征和相位特征。倒谱特征利用同态信号处理的思想，将非线性问题转化线性问题，更易处理，比如梅尔频率倒谱系数（Mel frequency cepstrum coefficient，MFCC）、线性预测倒谱系数（linearity predicts cepstrum coefficients，LPCC）、恒 Q 倒谱系数（constant Q cepstral coefficients，CQCC）等[3]。文献[4]比较了 MFCC 等特征的检测性能，在 ASVspoof 2019 数据库的 LA 数据集上，利用 MFCC 进行伪装语音检测的等错误率（equal error rate，EER）为 9.33%。Nagakrishnan 等[5]比较了 MFCC、LPCC 等特征在伪装语音检测中的有效性，并在此基础上提出了有效矩形带宽与 LPC 结合的（equivalent rectangular bandwidth，ERB-PLPC）特征。文献[6]提出了 CQCC 特征，该特征在 ASVspoof 2015 数据集上 EER 为 0.255%。这类特征虽然在伪装语音检测方向上取得了一定的进步，但由于该类特征多为针对某种特定算法合成的语音而提取，在面对使用不同算法的合成语音时，准确性有待提高，泛化能力不佳。另一类是相位特征，如相对相移（relative phase shift，RPS）和修正群时延（modified group delay，MGD）等。文献[7]发现群时延特征保留了较多的共振峰结构，证明了其对语音处理的鲁棒性。文献[8]比较了 MGD 和 RPS 两种特征的检测性能，在使用 MGD 进行伪装语音检测时，EER 为 4.918 7%；在利用 RPS 做特征时，EER 为 4.473 0%。由于人耳对语音信号的相位感知不敏感，各类算法在生成语音时，往往忽视对相位信息的重建。模仿语音由于声道模型不一致，在语音的基频、共振峰、声道参数、相位等方面会有较大差异；重放语音会引入设备噪声，造成相位变化；转换语音对基频、相位等信息均有不同程度的修改。相位特征利用了真伪语音在相位上的差异信息，可以很好地分辨真伪语音，在一段时间内取得了较好的检测效果。但是随着合成语音算法和电子设备的发展，语音合成技术取得了长足的进步，利用语音基频、相位等参数作为系统输入生成的合成语音在语音参数方面与真实语音相差无几。因此，利用相位特征进行合成语音伪装检测的准确率已经不能满足需求，对于未知类型的攻击，这种缺陷更加明显。近年来，调制包络在语音处理领域得到广泛应用。Drullman 等[9]

将调制包络应用在语音处理领域，发现它对语音理解的性能很好。Lu 等[10]利用调制包络和精细结构实现了自动语音识别。Ding 等[11]通过比较语音和音乐信号中的调制频谱（modulation spectrum，MS）特征，即时间调制包络傅里叶变换的均方根，发现可以将其看作语音信号的内在特征，且该特征会因为语音产生条件的不同而发生变化，但该特征在合成语音伪装检测中暂无应用。与真实语音相比，合成语音只模仿到语音时域包络的大致轮廓，细节部分被忽略，由于任何波形都可被看作不同频率正弦波的叠加，将语音信号的时域包络变换到频域，即利用语音信号的调制包络，可以更加直观地看出真伪语音时域包络的差异。

本文提出了一种采用恒 Q 调制包络（constant Q modulation envelope，CQME）作为特征向量的合成语音伪装检测方法，CQME 在传统调制频谱的基础上改进而来，首先将语音信号划分为不同频段，并提取各频段的时域包络，然后用恒定 Q 变换（constant Q transform，CQT）取代传统的傅里叶变换来提取语音信号的调制包络谱，计算每个频率成分的均方根后，得到用于训练随机森林（random forest，RF）的特征向量，最后用训练好的随机森林模型实现合成语音的伪装检测。该方法利用真伪语音包络谱的差异性，并利用时频分辨率可变的CQT取代傅里叶变换来获取恒Q调制包络谱特征，更符合语音信号的特性，包含的信息更加全面，对未知类型的合成语音检测具有很好的检测性能。

1　恒 Q 调制包络

1.1　恒 Q 调制包络提取过程

在合成语音中，语音信号时域包络的大致轮廓被重建，但细节信息被忽略[12]，因此与真实语音相比，合成语音的时域包络存在差异。

传统调制频谱通过傅里叶变换将语音信号的时域包络变换到频域[11]，并计算其均方根。本文提出的 CQME 在传统调制频谱的基础上改进，不再使用傅里叶变换，而是通过 CQT 将语音信号时域包络变换到频域，傅里叶变换中频率分辨率保持不变，得到的频谱在频率上是线性分布的，然而语音信号的频率并非线性分布，与傅里叶变换频谱上的频点无法一一对应，CQT 具有可变的时频分辨率，更适合语音信号的处理，可以更好地将时域包络映射至频域。

CQME 提取流程如图 1 所示，语音信号首先通过 Mel（梅尔）滤波器组分频。人耳对声音的感知是非线性的，Mel 滤波器组模拟人耳的这一特性，对语音信号分频[13]，得到时域上的各路窄带信号 $\boldsymbol{x}_k(n), k=1,2,\cdots,K$，本文中 K 取 32。然后对每一路窄带信号求 Hilbert（希尔伯特）变换从而得到其幅度包络，表示为：

$$\boldsymbol{s}_k(n)=\boldsymbol{x}_k(n)+\mathrm{j}\tilde{\boldsymbol{x}}_k(n),\ k=1,2,\cdots,K \tag{1}$$

$$\boldsymbol{m}_k(n)=\left|\boldsymbol{s}_k(n)\right|=\sqrt{\boldsymbol{s}_k^2(n)},\ k=1,2,\cdots,K \tag{2}$$

其中，$\boldsymbol{s}_k(n)$ 为 $\boldsymbol{x}_k(n)$ 的解析信号，$\tilde{\boldsymbol{x}}_k(n)$ 为 $\boldsymbol{x}_k(n)$ 的 Hilbert 变换，$\boldsymbol{m}_k(n)$ 为第 k 个窄带信号的幅度包络，对 $\boldsymbol{m}_k(n)$ 进行对数非线性处理后，通过 CQT[14] 将其变换到频域：

$$\boldsymbol{M}_k^{\mathrm{CQT}}(l)=\frac{1}{N_l}\sum_{n=0}^{N_l-1}\mathrm{lb}(\boldsymbol{m}_k(n))\boldsymbol{w}_{N_l}(n)\mathrm{e}^{-\mathrm{j}\frac{2\pi Q}{N_l}n},\quad l=1,2,\cdots,L \tag{3}$$

其中，$\boldsymbol{M}_k^{\mathrm{CQT}}(l)$ 为第 k 个分频信号的第 l 个频率分量的 CQT 系数，$\boldsymbol{w}_{N_l}(n)$ 是长度为 N_l 的汉明窗，且窗长 N_l 随频率而变化，窗长 N_l 的计算式如下：

$$N_l=\left\lceil Q\times\frac{f_\mathrm{s}}{f_l}\right\rceil,\ l=1,2,\cdots,L \tag{4}$$

其中，f_s 为采样频率，本文中取值为 8 kHz。Q 为一个恒定的常数，等于中心频率与滤波器带宽的比值，Q 值计算如下：

$$Q=\frac{f_l}{\triangle f_l}=\frac{f_l}{f_{l+1}-f_l} \tag{5}$$

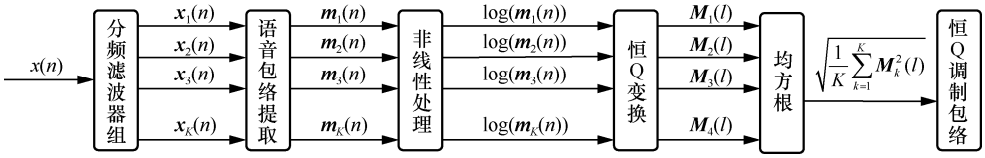

图 1　CQME 提取流程

其中，$\triangle f_l$ 表示第 l 个频率分量的滤波器带宽，f_l 表示第 l 个频率分量，可通过式（7）得到：

$$f_l = f_{\min} \times 2^{\frac{l}{B}},\ l = 1,2,\cdots,L \tag{6}$$

$$L = \left\lceil B \times \mathrm{lb}\left(\frac{f_{\max}}{f_{\min}}\right)\right\rceil \tag{7}$$

其中，L 表示 CQT 变换谱中的频率分量数，B 表示每倍频程中的 bins 数目，本文中 B 取 12，$f_{\max}$ 表示处理信号的最高频率，$f_{\min}$ 表示处理信号的最低频率。考虑说话人语音信号的频率范围一般为 60～3 400 Hz，将最高频率设置为 3 400 Hz，最低频率设置为 60 Hz，计算可得频率分量数 L 为 70。

计算得到 CQT 系数后，求得每一个频率成分的均方根，即可得到每个语音的 CQME 特征向量，大小为 70×1：

$$\mathrm{CQME}(l) = \sqrt{\frac{1}{K}\sum_{k=1}^{K} \boldsymbol{M}_k^{2\mathrm{CQT}}(l)},\ l = 1,2,\cdots,L \tag{8}$$

1.2　恒 Q 调制包络性能分析

为了分析真伪语音恒 Q 调制包络的差异性，在 ASVspoof 2019 数据库的 LA 数据集中，对训练集和开发集中的真实语音以及 A01、A02、A03、A04 这 4 种不同类型的合成伪装语音进行 CQME 特征提取，不同类型语音的 CQME 轮廓如图 2 所示。由图 2 可以看出，真实语音的 CQME 特征轮廓具有一致性，但与合成语音的 CQME 特征轮廓存在差异。在真实语音的特征轮廓曲线中，幅度一直处于下降状态，但在 1 600～2 200 Hz 范围内，幅度变化明显减缓。A02 类型合成语音的特征轮廓与真实语音最为相似，但该类型语音的轮廓曲

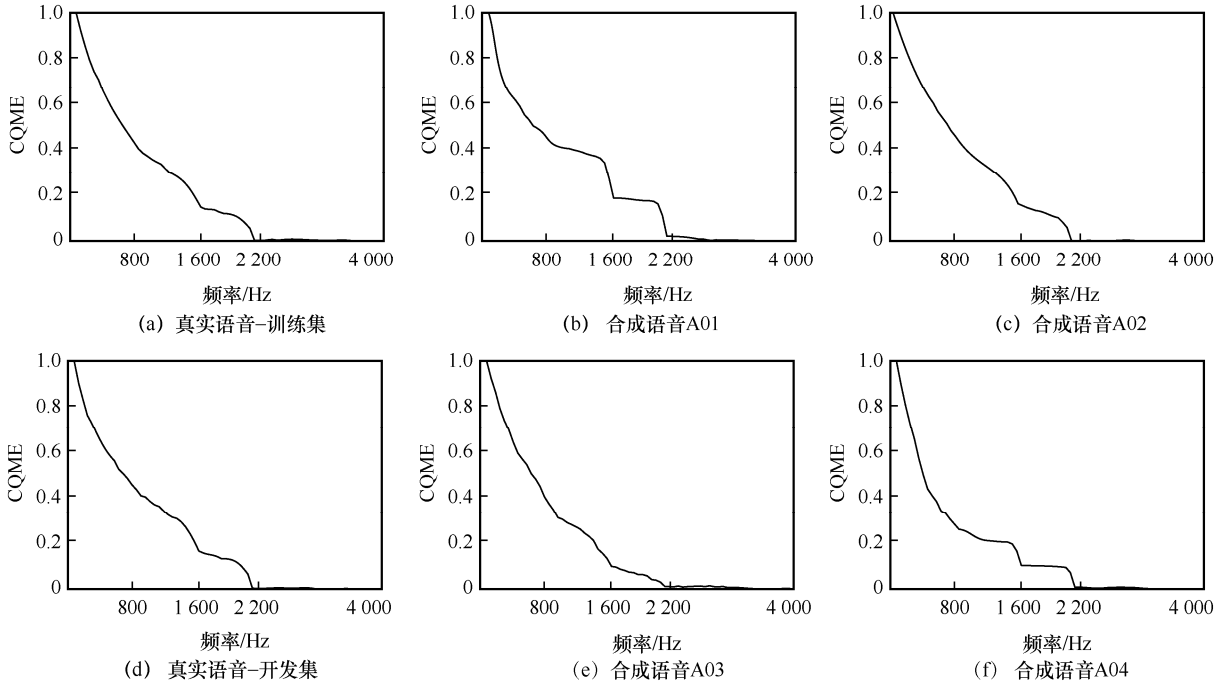

图 2　不同类型语音的 CQME 轮廓

线更加平滑，轮廓图中没有明显的抖动。在 A03 类型合成语音的特征轮廓曲线中，幅度则一直处于下降状态，且下降趋势没有发生明显的变化。而对于 A01 和 A04 这两种类型的合成伪装语音来说，特征轮廓与真实语音差异较大，在 800～1 400 Hz 和 1 600～2 200 Hz 频率范围内，这两种类型的合成语音均产生了非常缓慢的幅度变化，而真实语音仅在 1 600～2 200 Hz 幅度变化有所减缓。若将特征图中幅度变化趋势发生明显改变的点看作特征轮廓的转折点，可以发现在真实语音和 A02 类型合成语音中只有 2 个转折点，A03 类型无转折点，而在 A01 和 A04 两种类型的合成语音中，存在 4 个转折点。因此，CQME 特征对合成伪装语音具有较好的分辨能力。

2　基于恒 Q 调制包络的伪装语音检测

本文提出的基于恒 Q 调制包络的合成伪装语音检测系统包括特征提取和模型匹配两部分，基于 CQME 的伪装语音检测系统整体框图如图 3 所示。

首先提取训练语音的 70 维 CQME 特征向量，作为用于随机森林训练的特征向量，并给每个训练语音指定一个标签 $y_i \in \{0,1\}, i=1,2,\cdots,N$，其中，0 代表伪装语音，1 代表真实语音，$N$ 为语音的总数量。

然后利用这 N 个样本的特征向量来构建随机森林分类模型。随机森林利用了集成学习的原理，由若干个不同的决策树组成，这些决策树之间互不影响，当待分类样本输入时，森林中的每棵决策树分别对其进行分类，在所有决策树的分类结果中最多的一类就是随机森林最终的分类结果。

构建随机森林模型时，核心思想是自举汇聚（bootstrap aggregating，Bagging），首先森林中的每棵决策树都会在大小为 N 的训练集中进行 N 次随机且有放回的抽样，用得到的 N 个训练样本作为这棵树的训练集，这样既保证每棵树的训练集是不同的，从而可以训练出不同的决策树；也保证每棵树的训练样本是有交集的，避免决策树的分类结果过于片面[15]。然后在每个节点处，根据基尼系数选取适当的特征对训练样本进行划分，从而使决策树不断向下分裂，每个节点分裂时会在 70 维特征中随机的选取 m 维特征[16]，在本实验中，m 取值为 8。m 值的选取取决于随机森林的袋外错误率（out-of-bag error，OOB），在构建决

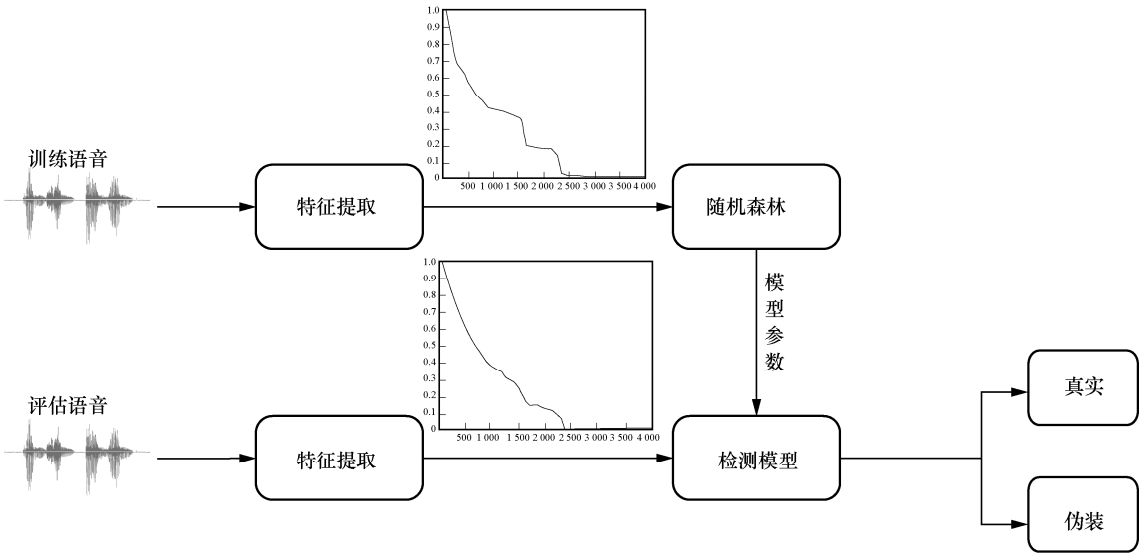

图 3　基于 CQME 的伪装语音检测系统整体框图

策树时，采用了随机且有放回抽取训练样本的方法，这对于每棵树来说，大约有 36.8%的样本没有参与这棵决策树的构建过程，这些样本就是这棵树的OOB样本[17]，计算每个样本在被作为OOB样本的树中的分类情况，然后投票决定该样本的分类结果，最后分类错误的个数与样本总数的比值即袋外错误率，最小袋外错误率对应的 m 值即最优的 m 值。

在上述构建过程中，每个节点不断分裂，直到所有可能的值都已经被使用，不再继续分裂，就得到一棵决策树。重复若干次即可得到一个训练好的随机森林模型，然后提取待测语音的CQME 特征向量，将该特征输入训练好的随机森林模型中，多棵决策树分别对其进行判断，并最终投票决定该待测语音的分类结果，从而分辨真伪语音。

基于恒 Q 调制包络的伪装语音检测方法整体流程如下。

步骤 1 利用分频滤波器组处理语音信号，使其分解为多路窄带信号，以模仿人耳滤波器对声音信号的处理过程。

步骤 2 通过 Hilbert 变换计算各窄带信号的解析信号。

步骤 3 取解析信号的模，即所求的时间包络。

步骤 4 将时间包络通过 CQT 变换至频域，以分析其随时间波动的内在特性。

步骤 5 变换得到 CQT 系数后，计算各频率分量的均方根，即可得到 CQME 特征向量。

步骤 6 根据语音真伪给定标签 1 或−1。

步骤 7 利用训练语音的特征向量和数据标签来训练随机森林。

步骤 8 待测语音输入后，按上述方法提取 CQME 特征向量，并输入随机森林中进行判别。

步骤 9 若分类器鉴别为真，则允许访问；否则拒绝访问。

3 实验与结果

3.1 数据集

实验选用 ASVspoof 2019 数据库中的 LA 数据集进行评估，其中包含真实语音、合成语音和转换语音。实验仅采用数据集中的真实语音和合成语音。数据集分为训练集、开发集和评估集，3 个子集之间无交叉，ASVspoof 2019 数据库中的 LA 数据集见表 1。该数据集在 VCTK 语料库的基础上生成，共包含神经波形模型、声码器等 17 种语音合成算法[18]。

表 1 ASVspoof 2019 数据库中的 LA 数据集

子集	说话人数目		语音数目		
	男性	女性	真实	伪装	合成
训练集	8	12	2 580	22 800	15 200
开发集	4	6	2 548	22 296	14 864
评估集	21	27	7 355	63 882	49 140

3.2 评估方法

实验选用串联检测代价函数（tandem detection cost function，t-DCF）作为评估标准[19]，t-DCF 计算式为：

$$\text{t-DCF} = C_{\text{FRR}}\pi_{\text{tar}}P_{\text{FRR}}^{\text{CM}} + C_{\text{FAR}}(1-\pi_{\text{tar}})P_{\text{FAR}}^{\text{CM}} \tag{9}$$

$$P_{\text{FRR}}^{\text{CM}} = \frac{N_{\text{FR}}}{N_{\text{bonafide}}} \tag{10}$$

$$P_{\text{FAR}}^{\text{CM}} = \frac{N_{\text{FA}}}{N_{\text{spoof}}} \tag{11}$$

$$\pi_{\text{tar}} = \frac{N_{\text{bonafide}}}{N_{\text{bonafide}} + N_{\text{spoof}}} \tag{12}$$

其中，C_{FRR} 表示错误拒绝一个真实语音的代价，C_{FAR} 表示错误接受一个伪装语音攻击的代价，$P_{\text{FRR}}^{\text{CM}}$ 和 $P_{\text{FAR}}^{\text{CM}}$ 分别表示错误拒绝率和错误接受率，N_{FR} 表示被错误拒绝的真实语音的个数，N_{bonafide} 表示真实语音的总数量，N_{FA} 表示被错误接受的伪装语音个数，N_{spoof} 表示伪装语音的总数量，π_{tar} 表示目标说话人出现的先验概率，在本实验中即真实语音随机出现的先验概率。t-DCF 将两种错误发生的代价以

及真伪语音的先验概率考虑进去，用来评估伪装语音检测系统比等错误率（equal error rate，EER）更合理，在 ASVspoof 2019 挑战赛中已给出各参数，其中 $C_{\mathrm{FRR}}=1$、$C_{\mathrm{FAR}}=10$、$\pi_{\mathrm{tar}}=0.9405$。

3.3　性能测试

为了比较不同 K 值即不同分频段数对 CQME 特征的影响，对不同分频段数 CQME 特征的检测性能进行比较，用不同 K 值的 CQME 特征进行检测的 t-DCF 见表 2，可以看出在将语音分为 32 个频段时，CQME 特征在 SVM 和随机森林上都取得了最优的检测效果，后续实验中分频数 K 均取 32。

表 2　用不同 K 值的 CQME 特征进行检测的 t-DCF

K	SVM	随机森林
8	0.703 1	0.362 6
16	0.202 6	0.136 6
32	0.113 3	0.062 4
64	0.122 6	0.078 0
128	0.123 2	0.081 7
256	0.107 9	0.096 2

为证实 CQME 的有效性，利用已有的伪装语音检测系统与其进行比较。在本实验中分别提取语音信号的 36 维 MFCC 特征（12MFCC、12△MFCC、12△△MFCC）、12 维 LPCC 特征、LFCC 特征和 CQCC 特征，比较这几个特征以及传统 MS 特征和 CQME 在 SVM 和随机森林做分类器时的检测代价，各系统的 t-DCF 比较见表 3。

表 3　各系统的 t-DCF 比较

特征	开发集		评估集	
	SVM	RF	SVM	RF
MFCC	0.380 0	0.357 1	0.743 8	0.731 3
LPCC	0.349 4	0.211 6	0.631 1	0.495 4
LFCC	0.124 6	0.122 8	0.187 3	0.136 7
CQCC	0.159 0	0.130 0	0.172 9	0.172 0
MS	0.164 8	0.084 5	0.189 3	0.098 5
CQME	0.098 2	0.060 1	0.113 3	0.062 4

从表 3 的实验结果来看，本文提出的用 CQME 特征向量训练随机森林分类器的伪装检测方法取得了最好的检测性能。在评估集中，使用 CQME 特征进行伪装检测的检测代价比 MFCC 降低了 91.47%，比 LPCC 降低了 87.40%，比 LFCC 降低了 54.35%，比 CQCC 降低了 63.72%，与传统的 MS 特征相比，也有 36.65%的下降。这是因为合成语音在进行包络重构时，包络的细节部分被忽略，利用真伪语音包络细节的差异提取语音信号的调制包络谱可以实现真伪语音的识别。并且与传统 MS 特征相比，CQME 利用 CQT 将时域包络变换到频域，变换时频率是以 2 为底的对数，更加符合语音信号的特点，中心频率与滤波器带宽比值保持不变，在低频时具有较高的频率分辨率，在高频具有较高的时间分辨率，更适合语音信号的特性。另外，对语音进行 CQME 特征提取后，数据中存在边界值，这些边界值会对决策边界的选取产生影响，而随机森林算法在计算过程中随机选择样本和特征，由多棵决策树投票决定分类结果，避免对训练数据过拟合，因此取得了更好的结果。

6 种特征的 t-DCF 比较如图 4 所示，比较了使用本文所提方法检测未知的合成语音（未参与训练）攻击时，几种特征的检测效果。从图 4 可知，在检测未知攻击 A08 时的 t-DCF 明显比检测其他几种类型语音的 t-DCF 略高，这是因为语音信号是由不同振幅、频率的纯音叠加而成的复合音，且基频反映了复合音中的重复频率。而 A08 类型的伪装语音利用语音信号的基频与相位信息作为输入，采用神经元滤波波形模型[20]生成语音波形，较多地保留了原始波形的包络信息，因此与其他几种类型的合成语音相比，检测效果稍差。在检测其他类型的未知攻击时，CQME 特征均较其他特征表现出良好的性能。整体上看，对语音信号提取 CQME 特征相比传统的 MFCC 特征、LPCC 特征、LFCC 特征、CQCC 特征以及传统 MS 特征在检测效果上均有较大的提升。

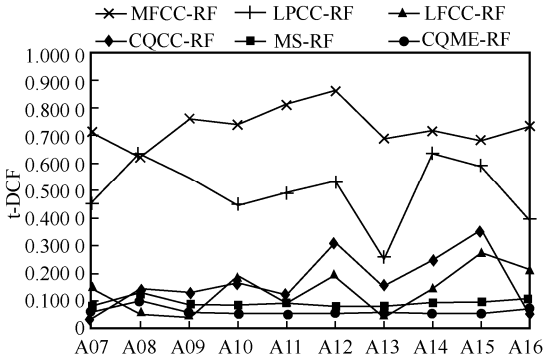

图 4　6 种特征的 t-DCF 比较

在实际应用场景中，设备收集的语音并非纯净的，而是由语音和噪声叠加而成的，若伪装语音检测系统没有良好的抗噪性能，检测准确性将受到极大的挑战。为了验证所提方法的抗噪性能，分别给语音加入 babble 噪声、volvo 噪声和 pink 噪声，并设定信噪比为 0，此时信号的平均功率与噪声的平均功率相同，干扰较强。各噪声条件下系统的 t-DCF 比较见表 4，经验证，在此强干扰条件下，用 CQME 特征进行合成伪装语音检测，babble 噪声下的 t-DCF 为 0.095 0，volvo 噪声下的 t-DCF 为 0.072 5，pink 噪声下的 t-DCF 为 0.089 5。虽然与纯净语音相比，加噪情况下的检测性能有所下降，但是所提方法仍具有较好的伪装语音检测性能。

表 4　各噪声条件下系统的 t-DCF 比较

分类器	CQME 噪声鲁棒性			
	无噪	babble 噪声	volvo 噪声	pink 噪声
SVM	0.113 3	0.157 4	0.130 8	0.153 2
RF	0.062 4	0.095 0	0.072 5	0.089 5

4　结束语

本文提出了一种基于恒 Q 调制包络特征的伪装语音检测方法，该方法利用真实语音与合成语音包络细节的差异性，提取语音信号的包络并通过恒定 Q 变换映射至频域得到语音的恒 Q 调制包络谱，计算均方根后得到 CQME 特征向量，后端利用随机森林对特征向量训练和预测，实现了真伪语音的判别。该方法有效地改善了系统检测性能，降低了合成语音的检测代价，对未知类型合成语音检测具有明显优越性。实验表明，在 ASVspoof 2019 数据集上，用 CQME 做特征参数，随机森林做分类器构建的合成伪装语音检测系统具有非常好的检测性能，且具有较好的噪声鲁棒性。

参考文献：

[1] TAN C B, HIJAZI M H A, KHAMIS N, et al. A survey on presentation attack detection for automatic speaker verification systems: state-of-the-art, taxonomy, issues and future direction[J]. Multimedia Tools and Applications, 2021, 80(21-23): 32725-32762.

[2] 徐嘉, 简志华, 金宏辉, 等. 基于中心对称局部二值模式的合成伪装语音检测方法[J]. 电信科学, 2023, 39(1): 72-78.
XU J, JIAN Z H, JIN H H, et al. A method for synthetic spoofing speech detection based on center-symmetric local binary pattern [J]. Telecommunications Science, 2023, 39(1): 72-78.

[3] MITTAL A, DUA M. Automatic speaker verification systems and spoof detection techniques: review and analysis[J]. International Journal of Speech Technology, 2021, 25(1): 105-134.

[4] ALZANTOT M, WANG Z, SRIVASTAVA M B. Deep residual neural networks for audio spoofing detection[C]//Proceedings of 20th Annual Conference of the International Speech Communication Association 2019 (INTERSPEECH 2019). Graz, Austria: ISCA, 2019: 1078-1082.

[5] NAGAKRISHNAN R, REVATHI A. Generic speech based person authentication system with genuine and spoofed utterances: different feature sets and models[J]. Multimedia Tools and Applications, 2021, 81(1):1179-1208.

[6] TODISCO M, HÉCTOR D, EVANS N. Constant Q cepstral coefficients: a spoofing countermeasure for automatic speaker verification[J]. Computer Speech & Language, 2017(45):516-535.

[7] RAJAN P, PARTHASARATHI S, MURTHY H A. Robustness of phase based features for speaker recognition[C]//Proceedings of 10th Annual Conference of the International Speech Communication Association 2009 (INTERSPEECH 2009). Brighton: ISCA, 2009: 2299-2302.

[8] SARATXAGA I, SANCHEZ J, WU Z, et al. Synthetic speech detection using phase information[J]. Speech Communication, 2016(81): 30-41.

[9] DRULLMAN R, FESTEN J M, PLOMP R. Effect of temporal envelope smearing on speech reception[J]. The Journal of the Acoustical Society of America, 1994, 95(2): 1053-1064.

[10] LU X, UNOKI M, NAKAMURA S. Sub-band temporal modulation envelopes and their normalization for automatic speech recognition in reverberant environments[J]. Computer Speech & Language, 2011, 25(3):571-584.

[11] DING N, PATEL A D, CHEN L, et al. Temporal modulations in speech and music[J]. Neuroscience & Biobehavioral Reviews, 2017(81): 181-187.

[12] NING Y, HE S, WU Z, et al. A review of deep learning based speech synthesis[J]. Applied Sciences, 2019, 9(19):4050.

[13] 林朗, 王让定, 严迪群, 等. 基于逆梅尔对数频谱系数的回放语音检测算法[J]. 电信科学, 2018, 34(5): 90-98.
LIN L, WANG R D, YAN D Q, et al. A playback speech detection algorithm based on log inverse Mel-frequency spectral coefficient[J]. Telecommunications Science, 2018, 34(5): 90-98.

[14] BROWN J C. Calculation of a constant Q spectral transform[J]. Journal of the Acoustical Society of America, 1998, 89(1): 425-434.

[15] HAMSA S, SHAHIN I, IRAQI Y, et al. Emotion recognition from speech using wavelet packet transform cochlear filter bank and random forest classifier[J]. IEEE Access, 2020(8): 96994-97006.

[16] CHEN L, SU W, FENG Y, et al. Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction[J]. Information Sciences, 2020(509): 150-163.

[17] RAMOSAJ B, PAULY M. Consistent estimation of residual variance with random forest out-of-bag errors[J]. Statistics & Probability Letters, 2019(151):49-57.

[18] WANG X, YAMAGISHI J, TODISCO M, et al. ASVspoof 2019: a large-scale public database of synthesized, converted and replayed speech[J]. Computer Speech & Language, 2020(64): 101114.

[19] KINNUNEN T, DELGADO H, EVANS N, et al. Tandem assessment of spoofing countermeasures and automatic speaker verification: fundamentals[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020(28): 2195-2210.

[20] WANG X, TAKAKI S, YAMAGISHI J. Neural source-filter-based waveform model for statistical parametric speech synthesis[C]//2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2019: 5916-5920.

[作者简介]

徐嘉（1998− ），女，杭州电子科技大学通信工程学院硕士生，主要研究方向为语音伪装检测。

简志华（1978− ），男，博士，杭州电子科技大学通信工程学院副教授、硕士生导师，浙江省数据存储传输及应用技术研究重点实验室教师，主要研究方向为语音转换、伪装语音检测、声纹识别等。

金宏辉（1999− ），男，杭州电子科技大学通信工程学院硕士生，主要研究方向为语音转换和伪装检测。

吴超（1988− ），男，博士，杭州电子科技大学通信工程学院讲师、硕士生导师，主要研究方向为导航信号处理及欺骗干扰检测。