第 49 卷第 5 期
2024 年 9 月

声 学 学 报
ACTA ACUSTICA

Vol. 49, No. 5
Sept., 2024

使用全局自注意 Teager 能量倒谱系数检测重放欺骗语音*

陈 铭 陈雪勤†

(苏州大学 电子信息学院 苏州 215006)

2023 年 6 月 25 日收到
2023 年 9 月 9 日定稿

摘要 提出了一种基于能量的前端特征提取方法, 旨在应对自动说话人验证系统中面临的重放攻击威胁。该方法实现了全频段上的可变分辨率, 以充分利用重放语音与真实语音在子带能量上的高鉴别非线性信息。首先, 通过采用 F-ratio 方法统计分析了多种录音和播放设备。接着, 根据统计结果在全频段上设计了一组滤波器, 旨在捕获高鉴别能量信息。最后, 利用 Teager 能量算子计算子带滤波信号的能量, 提出了全局自注意 Teager 能量倒谱系数 (GSTECC)。为了验证所提方法的有效性, 采用高斯混合模型作为分类器, 在 ASVspoof 2017 V2 和 ASVspoof 2021 PA 数据库上进行了一系列测试实验。实验结果表明, 相对于其他先进特征提取方法, 所提 GSTECC 特征在检测重放攻击方面表现出更优异的性能。

关键词 说话人验证, 重放攻击检测, 全局自注意特征, Teager 能量倒谱系数, 非线性滤波器组

PACS 数 43.72, 43.60 **DOI**：10.12395/0371-0025.2023106

Detection of replay spoof speech using global self-attentive Teager energy features

CHEN Ming CHEN Xueqin†

(*School of Electronic Information Engineering, Soochow University* Suzhou 215006)

Received Jun. 25, 2023
Revised Sept. 9, 2023

Abstract This paper proposes an energy-based front-end feature extraction method to address the threat of replay attacks in automatic speaker verification systems. This method achieves variable resolution over the entire frequency band to fully utilize the highly discriminative nonlinear information in sub-band energy between replayed speech and real speech. First, statistical analysis of various recording and playback devices is carried out by adopting the F-ratio method. Then, according to the statistical results, a set of filters on the whole frequency band is designed to capture high discriminative energy information. Finally, the Teager energy operator is used to calculate the energy of the sub-band filtered signal, and the global self-attentive Teager energy cepstral coefficients (GSTECC) is proposed. In order to verify the effectiveness of the proposed method, the Gaussian mixture model is used as the classifier, and a series of test experiments are conducted on the ASVspoof 2017 V2 and ASVspoof 2021 PA databases. Experimental results show that the proposed GSTECC feature performs better in detecting replay attacks compared to other advanced feature extraction methods.

Keywords Speaker verification, Replay attack detection, Global self-attention feature, Teager energy cepstral coefficients, Nonlinear filter bank

* 国家自然科学基金项目 (61340004) 资助
† 通讯作者: 陈雪勤, chenxueqin@suda.edu.cn

引言

自动说话人验证 (ASV) 系统旨在根据说话人的声音验证其身份, 是一项重要的生物特征识别技术。然而在实际应用中, ASV 系统面临着欺骗语音攻击的潜在威胁, 攻击大致可分为四种类型: 录音重放、人声模仿、语音转换和语音合成。其中, 录音重放是最常见、最易实施且威胁最大的一类攻击。该攻击方法仅需要简单的录音设备 (如手机、录音笔等) 记录原始说话者的声音, 然后通过播放设备进行声音重放, 无需任何专业技术知识即可实施[1-2]。随着高保真设备的便携化和普及化, 这种攻击方式严重威胁到 ASV 系统的安全性。因此, 开发能够检测重放欺骗语音的对策是至关重要的。近些年来, 许多前端特征已被提出用于检测重放语音, 并取得了一定效果。

捕捉重放语音的非线性特征是提高检测性能的一个研究方向。其中, Tapkir 等基于幂函数的非线性提出幂归一化倒谱系数 (PNCC) 和 Q-Log 归一化倒谱系数 (QLNCC)[3]。Kamble 等基于非线性的 Teager 能量算子提出了一系列特征, 包括 Teager 能量倒谱系数 (TECC) 和增强 Teager 能量倒谱系数 (ETECC) 等。这一类特征在捕捉混响和噪声抑制方面展现出不错的能力[4-7], 但该类特征并未重点研究相关频段上分辨率的重要性。

通过调节目标位置的分辨率来捕捉真实语音和重放语音之间差异性是特征提取的另一研究方向。其中, 常数 Q 倒谱系数 (CQCC) 作为 ASVspoof 2017 挑战赛的基线特征, 能更准确地捕捉真实语音和回放语音之间的差异信息[8]。该特征基于常数 Q 变换, 在低频处采用较高的频率分辨率, 高频处采用较高的时间分辨率。文献 [9-10] 进一步证实了在关键频段采用高分辨率对于重放检测有积极意义。文献 [11] 对梅尔频率倒谱系数 (MFCC)、线性频率倒谱系数 (LFCC)、逆梅尔频率倒谱系数 (IMFCC) 等特征做了重放检测的性能比较, 结果显示不同频率尺度下所设计的特征得到的检测效果不同且差异较大。

上述特征提取方法基于非线性和关键频段提高分辨率的思路在重放语音检测中发挥了不错的效果。然而其中的多分辨能力大体上以高低频段来划分, 缺乏严谨的理论依据。本文运用 F-ratio 深入分析了若干种录放设备条件下的重放语音与真实语音的频谱关系, 在此基础上归纳提出全局自注意权重用来表示全频段中各个频率点的鉴别权重, 从而给出了基于非线性、多频带进行特征设计的理论依据。以全局自注意权重为依据设计了全局自注意滤波器组, 用来提取真实语音和重放语音之间的高鉴别频段信息, 同时利用 Teager 能量算子来捕捉重放语音的非线性失真。最终, 提出了一种有效检测重放欺骗语音的全局自注意 Teager 能量倒谱系数 (Global Self-attentive Teager Energy Cepstrum Coefficients, GSTECC) 特征。

1　全局自注意权重提取

1.1　重放语音信号的形成过程

重放攻击是指攻击者以特定方式将事先录制的语音样本再次播放给 ASV 系统, 以模仿合法用户的声音并试图获取系统的未经授权访问权限。因此, 为确保 ASV 系统的安全性, 必须确保系统具备检测并有效抵御重放语音攻击的能力。图 1 是录音重放语音检测模块, 其中重放语音攻击检测是二分类任务, 其目标是识别给定的语音信号是真实语音还是经录音设备录制后再经重放设备播放的重放语音。

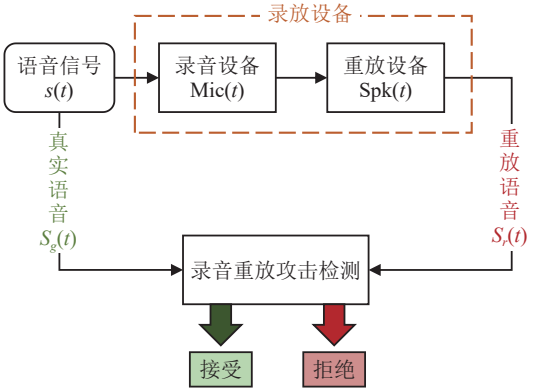

图 1　录音重放语音检测模块

根据图 1 可得真实语音 $S_g(t)$ 和重放语音 $S_r(t)$ 的关系为

$$S_r(t) = S_g(t) * h_{\mathrm{Mic}}(t) * h_{\mathrm{Spk}}(t), \tag{1}$$

其中, $h_{\mathrm{Mic}}(t)$, $h_{\mathrm{Spk}}(t)$ 分别为录音设备和重放设备的单位脉冲响应。由式 (1) 可知, 重放语音是由真实语音与录放设备的脉冲响应卷积而成, 将其转换至实倒谱域:

$$\boldsymbol{S}_r = \boldsymbol{S}_g + \boldsymbol{h}_{\mathrm{Mic}} + \boldsymbol{h}_{\mathrm{Spk}}, \tag{2}$$

其中, $\boldsymbol{S}_r$, $\boldsymbol{S}_g$, $\boldsymbol{h}_{\mathrm{Mic}}$, $\boldsymbol{h}_{\mathrm{Spk}}$ 分别表示重放语音、真实语音、录音设备和重放设备的脉冲响应的倒谱向量。由式 (2) 可知, 从重放语音中提取的特征会受到录放设备的影响。这意味着, 使用不同类型的录放设备

表 1 录音设备和重放设备详情

录音设备编号	录音设备详情	重放设备编号	重放设备详情
R01	Zoom H6 handy recorder	P01	All-in-one PC speakers
R02	BQ Aquaris M5 smartphone	P05	Beyerdynamic DT 770 PRO headphones
R03	Low-quality headset	P06	Dell laptop internal speakers
R04	Nokia Lumia 635 smartphone	P07	Dynaudio BM5A speaker
R05	Røde NT2 microphone	P08	HP Laptop internal speakers
R06	Røde smartLav + microphone	P09	VIFA M10MD-39-08 speaker
R07	Samsung Galaxy 7s smartphone	—	—

会对后续的分类检测产生不同程度的影响。

1.2 使用 F-ratio 分析不同录放设备的影响

研究在不同录放设备条件下 F-ratio 模式的变化, 辅助鉴别哪些频段具有更好的区分性能[12-15], 共选择了 10 种录放设备组合, 如表 1 所示。

F-ratio 的比值向量 $\boldsymbol{F}_{\text{ratio}}$ 定义为类间距离与类内方差, 其中类表示为真实语音和重放语音两类。对所采集的各帧语音信号进行快速傅里叶变换 (FFT), 并取模的平方作为样本向量:

$$\boldsymbol{X} = |\text{FFT}(\boldsymbol{x})|^2, \tag{3}$$

其中, $\boldsymbol{x} = \{x^1, x^2, \cdots, x^L\}$ 表示某一帧语音时域信号, L 为帧长; 样本向量 $\boldsymbol{X} = \{X^1, X^2, \cdots, X^M\}$ 表示对应帧语音信号在频域上各点的能量值, M 表示频域点数。

令 $\boldsymbol{X}_i$ 表示真实语音中的第 i 个样本向量, $\boldsymbol{X}_j$ 表示重放语音中的第 j 个样本向量, $\boldsymbol{\mu}_g$ 和 $\boldsymbol{\mu}_r$ 则分别表示真实语音和重放语音各自总样本向量的平均向量。N_g 和 N_r 分别代表真实语音和重放语音的样本总数, 则有

$$\boldsymbol{F}_{\text{ratio}} = \frac{(\boldsymbol{\mu}_g - \boldsymbol{\mu}_r)^2}{\dfrac{1}{N_g}\displaystyle\sum_{i=1}^{N_g}(\boldsymbol{X}_i - \boldsymbol{\mu}_g)^2 + \dfrac{1}{N_r}\sum_{j=1}^{N_r}(\boldsymbol{X}_j - \boldsymbol{\mu}_r)^2}, \tag{4}$$

其中, $\boldsymbol{F}_{\text{ratio}}$ 的维度与样本向量 $\boldsymbol{X}$ 的维度一致, 均为频域的点数。将频域各点转换为对应的频率值, 得到的模式曲线见图 2。

如图 2 所示, 在不同频段上使用不同录放设备的 F-ratio 模式曲线呈现出差异, 这是由于录放设备的不同所导致, 而这种变化是导致模型泛化性能不佳的原因之一。然而, 依然可以观察到一些子频段的模式曲线之间表现出较高的一致性, 这表明这些子频段的鉴别能力不容易受到录放设备变化的影响。因此, 在特征设计阶段, 若能加强对稳定性与鉴别能力相对较高的子频段的关注, 减少对其他容易受录放设备影响的子频段的关注度, 有助于降低特征受录放设备变化影响的程度, 提高系统的稳定性。下文将依据 F-ratio 曲线模式的一致性和幅值的大小设计一个权重参数用以表达各频段上需要赋予的注意力程度。

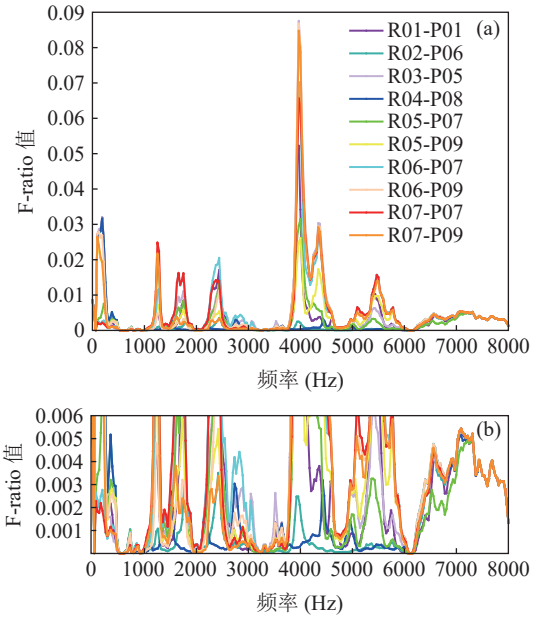

图 2 10 种录放设备下的 F-ratio 模式曲线 (a) F-ratio 曲线; (b) 0~0.006 幅度区间的放大曲线

1.3 全局自注意权重

式 (4) 定义了 F-ratio 模式曲线 (图 2)。根据不同模式曲线的特性, 本节定义自注意权重对不同频率点的重要性进行建模:

$$W_i = \frac{\log_{10}\left(k \cdot \dfrac{m_i}{s_i}\right)}{\log_{10}\left(\max\left(k \cdot \dfrac{m_i}{s_i}\right)\right)}, \quad i \in \{1, 2, 3, \cdots, M\}, \tag{5}$$

其中, W_i 代表第 i 个频率点上的自注意权重; m_i 是图 2 中第 i 个频率点上所有 F-ratio 值的均值; s_i 是图 2 中第 i 个频率点上所有 F-ratio 值的标准差; 其中 k 为调节因子, 为确保权重值在本研究中均为非负数, k 取值 1.0022。总数为 M 个频率点的自注意权重组成的 M 维列向量 $\boldsymbol{W}$, 称作全局自注意权重矢量:

$$\boldsymbol{W}=[W_1,W_2,W_3,\cdots,W_M]^{\mathrm{T}}. \tag{6}$$

具体而言, 自注意权重 $\boldsymbol{W}$ 被定义为与各频率点上各条 F-ratio 值的均值成正比, 与标准差成反比。这样的定义旨在表示当图 2 中的各条模式曲线越接近且 F-ratio 值都尽可能高时, 权重也会相应增加。当模式曲线越接近时, 意味着不同录放设备之间的 F-ratio 值的标准差越小, 而当 F-ratio 值都尽可能高时, 则表示在不同录放设备的 F-ratio 值的平均值越大。通过这种方式, 便于将特征提取重点放在模式曲线一致性较强且鉴别能力较高的频率上。

全频段全局自注意权重如图 3 所示, 这些权重受到了 F-ratio 模式曲线的一致性以及 F-ratio 平均值的共同影响。由实验所得 F-ratio 值均处于 10 的负幂次量级 (10^{-2}~10^{-10}), 因此分母的标准差对权重的影响更为显著, 而均值的大小对权重的贡献相对较小。由图 2(a) 和图 3 可知, F-ratio 模式曲线的一致性越好 (标准差越小), 对应的权重越大。在此基础上, 更大的均值也对应了更高的权重。图 2(b) 将 0~0.006 幅度区间的 F-ratio 曲线进行了放大展示, 可见 F-ratio 模式曲线一致性较好的几个频段对应了相对高的权重值, 而具有高一致性和高均值的频段如 6180~8000 Hz 频段显示出更高的权重。综上所述, 全局自注意权重图有效地反映了全频段上鉴别信息的分布情况, 各频率点的鉴别能力可以通过自注意权重来描述, 为本文的特征设计奠定了基础。

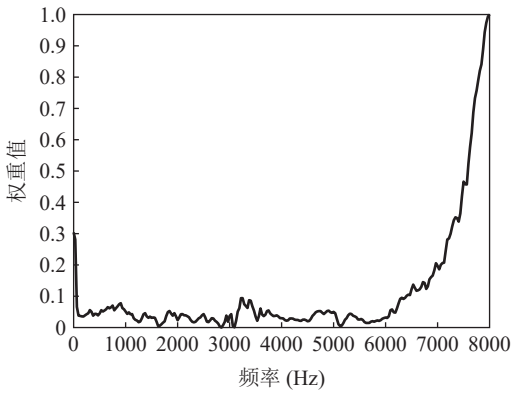

图 3　全局自注意权重

2　全局自注意 Teager 能量倒谱系数

2.1　全局自注意滤波器组

在重放语音检测任务中, 为了更精确地提取到真实语音和重放语音之间的区别性特征, 设计一个有效的滤波器组的主要原则是增强具有强鉴别能力的信息, 并抑制具有较弱鉴别能力的信息。由图 3 可知, 鉴别信息在不同频带上表现出不均匀性, 其鉴别频率与一般频率之间并不呈线性关系。因此, 本文提出了一种非线性全局频率尺度变换, 通过采用频率尺度规整方法[15-17] 来实现这一目标。与传统的梅尔频率尺度或逆梅尔尺度不同, 该方法是围绕全局自注意权重向量进行设计, 其频率分辨率的高低与自注意权重值的大小相对应, 这意味着将更多的注意力集中在高鉴别频率区域上。

2.1.1　非线性全局频率尺度变换

在权重向量中, 各频率点的权重大小反映了该点应该受到的关注程度。为了便于后续设计滤波器组的各个中心频率与带宽, 本小节将线性频率变换到新的频率尺度上。

首先, 通过计算各点的自注意权重与总自注意权重的比值, 将全局自注意权重向量进行尺度变换:

$$R_j=\begin{cases}0, & j=1,\\ \dfrac{W_j}{\sum \boldsymbol{W}-W_1}\cdot\dfrac{f_s}{2}, & j\in\{2,3,\cdots,M\},\end{cases} \tag{7}$$

其中, W_j 为第 j 个自注意权重, $\boldsymbol{W}$ 是全局自注意权重向量, R_j 表示变换后第 j 个点的值, 其大小与对应点的权重成正比, 为了使频率尺度变换前后的频率范围保持一致, 将变换公式中的系数设计为 $f_s/2$, f_s 是语音信号的采样频率。

进一步, 通过对 R_j 累加求和计算出新的频率尺度上各点的刻度:

$$F_j=\sum_{i=1}^{j}R_j,\quad j\in\{1,2,3,\cdots,M\}, \tag{8}$$

$$\boldsymbol{F}=[F_1,F_2,F_3,\cdots,F_M]^{\mathrm{T}}, \tag{9}$$

其中, F_j 表示新的频率尺度上的第 j 个刻度值, 矢量 $\boldsymbol{F}$ 代表新的频率尺度 (下文简称 F 尺度)。如图 4(a) 所示, F 尺度下各段曲线的斜率大小与权重成正比关系。虚线展示着 F 尺度下的均匀划分, 而在线性频率域下对应着不同的疏密程度, 这意味着高权重即高斜率处得到的关注度高。因此高鉴别频段具有了更高的频率分辨能力, 将分配更多的滤波器组, 在这些关键频率段上可以更准确地捕捉细微的频率变化, 从而提高了频域分析的性能。

2.1.2　全局自注意滤波器组设计

传统的特征参数提取方法通常使用三角滤波器, 然而三角滤波器的下降趋势较为陡峭, 缺乏平滑性, 可能会对相邻子带之间的联系造成一定的影响。相比之下, 高斯滤波器具有较高的平滑性, 在时频联合分辨率方面具有最佳性能, 既能有效地进行频率选

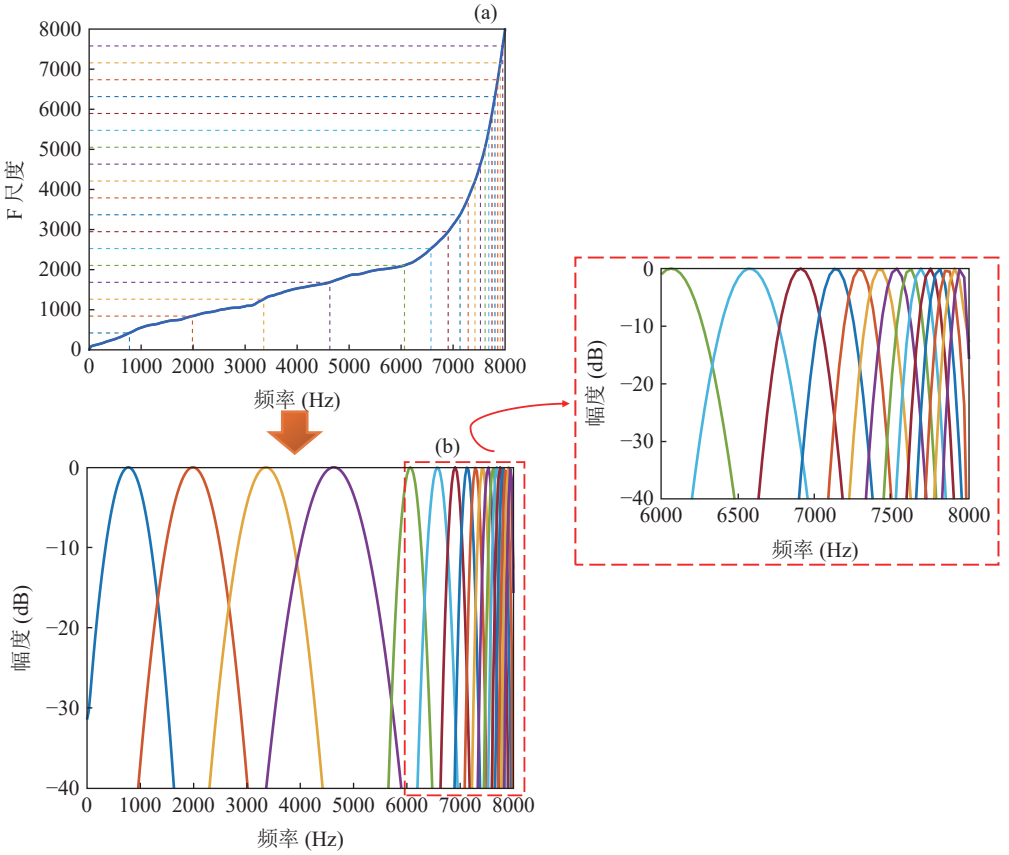

图 4 基于非线性全局频率尺度的全局自注意滤波器组 (a) 非线性全局频率尺度变换; (b) 全局自注意滤波器组

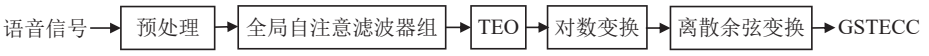

图 5 全局自注意 Teager 能量倒谱系数特征提取流程

择, 又能减小信号失真的影响。因此, 本文采用余弦调制的高斯函数来建模滤波器组的形状, 其脉冲响应 $h_n(t)$ 和频率响应 $H_n(f)$ 分别为

$$h_n(t)=\cos 2\pi\mu_n t\cdot\exp\left(-\frac{t^2}{2\delta_n^2}\right), \tag{10}$$

$$H_n(f)=\frac{\sqrt{2\pi}\delta_n}{2}\left[\exp\left(-2\pi^2\delta_n^2(f_n-\mu_n)^2\right)+\exp\left(-2\pi^2\delta_n^2(f_n+\mu_n)^2\right)\right]. \tag{11}$$

$h_n(t)$ 表示第 n 个滤波器在时间 t 时的脉冲响应, 其中 μ_n 是根据非线性全局频率尺度选择滤波器的中心频率, 参数 δ_n 控制子带滤波器的带宽。

综上可得到基于非线性全局频率尺度映射的全局自注意滤波器组, 如图 4 所示。在全频段上滤波器组的中心频率分布是通过非线性全局频率尺度转换曲线进行划分的, 用以提取出原始语音和重放语音之间的高度可辨别区域。

2.2 基于 Teager 能量算子的特征提取

Teager 能量算子 (TEO) 是一种非线性微分算子, 反映信号的瞬时变化, 可以用于捕捉重放语音因录放设备而导致的非线性失真。对于单分量离散时间信号 $x(n)$, TEO 定义为

$$E_n=\psi_d\{x(n)\}=x^2(n)-x(n-1)x(n+1). \tag{12}$$

TEO 针对单分量信号, 因此适用于窄带信号。如使用 2.1.2 节中设计的全局自注意滤波器组处理语音信号, 以获取高鉴别窄带信号, 之后便可对这些窄带信号进行 TEO 能量估计。

在上述基础上, 本文提出了一种基于全局自注意机制的 Teager 能量倒谱系数特征提取方法, 其流程如图 5 所示。首先, 对输入语音信号进行预加重处理, 其本质是通过一个高通滤波器用来增强高频成分, 其传递函数表示为 $H(z)=1-\alpha z^{-1}$, 其中 α 为预加重系数, 一般 $0.9<\alpha<1$; 随后经过预加重处理的语音信号被输入到全局自注意滤波器组, 以获得高鉴别窄带滤波信号组; 接着对每个高鉴别窄带信号计算 Teager 能量, 以捕捉重放语音中的非线性失真; 最后经过对数变换和离散余弦变换得到 GSTECC

特征。

重放语音过程会导致语音信号能量的失真, 通过全局自注意滤波器组对高鉴别区域的加强关注进一步增强真实语音和重放语音之间的区别, 同时使用 Teager 能量算子捕捉信号能量的非线性失真, 从而可更精确地提取真实语音与重放语音之间的差异特征。

3　实验结果与分析

3.1　数据库

ASVspoof 2017 V2、ASVspoof 2019 PA 和 ASVspoof 2021 PA 是三个广泛用于重放语音检测的数据库。其中 ASVspoof 2017 V2 (训练集、验证集和测试集) 和 ASVspoof 2021 PA (测试集) 是真实的重放语音数据库, 而 ASVspoof 2019 PA 是模拟重放语音数据库[18]。因此, 本实验的主要数据集基于 ASVspoof 2017 V2, 同时使用 ASVspoof 2021 PA 的测试集进行跨数据库的特征算法评估。 ASVspoof 2017 V2 数据库和 ASVspoof 2021 PA 测试集的详细信息见表 2。

表 2　ASVspoof 2017 V2 数据库和 ASVspoof 2021 PA 数据库

数据库	数据集	真实语音	重放语音
ASVspoof 2017 V2	训练集	1507	1507
	验证集	760	950
	测试集	1298	12008
ASVspoof 2021 PA	测试集	94068	627264

3.2　分类模型

考虑到高斯混合模型 (GMM) 在说话人识别系统中的优秀表现, 且其具有快速收敛和易于训练的优点[19], 本文选择将其作为后端分类器。GMM 阶数设置为 512 阶, 测试时根据下式计算每个测试语音的对数似然比 $\Lambda(X)$:

$$\Lambda(X) = \log P(X|\theta_g) - \log P(X|\theta_s), \tag{13}$$

其中, $\log P(X|\theta_g)$ 和 $\log P(X|\theta_s)$ 分别表示测试语音对应真实语音 GMM 模型和重放语音 GMM 模型的平均对数似然度。重放攻击检测系统使用 $\Lambda(X)$ 来判断测试语音是否为真实语音。如果 $\Lambda(X)$ 值大于设定阈值 θ, 则判决测试语音为真实语音; 否则为重放语音。当阈值 θ 越大时, 误判为真实语音的可能性, 即错误接受率 (FAR) 会降低, 但是误判为重放语音的可能性, 即错误拒绝率 (FRR) 会增加。相反地, 当阈值 θ 越小时, FRR 会降低, 但是 FAR 会增加。系统的性能评价指标为等错误率 (EER), 即当 FAR 和 FRR 相等时, 对应的 $\Lambda(X)$ 值即为 EER。重放攻击检测系统的性能越好, EER 值越小, 也就是重放攻击检测的能力越强。

3.3　ASVspoof 2017 V2 数据库的实验结果

3.3.1　滤波器数量的选择

通过在验证集上测试不同数量的全局自注意滤波器组, 主要目标是确定适当的滤波器组数量, 以优化模型在测试集上的性能, 并确保避免在测试集上出现过拟合情况, 同时提高模型的泛化能力。实验结果如图 6 所示, 在验证集上当滤波器组中有 18 个子带滤波器时, 等错误率达到最低点。随着子带滤波器数量的增加, 等错误率也随之增加。因此, 可以推断随着滤波器数量增加, 滤波器之间的重叠会导致鉴别信息的丢失, 从而降低检测性能。故最终选择 18 个滤波器, 并将其应用于测试集以获得最佳性能结果。

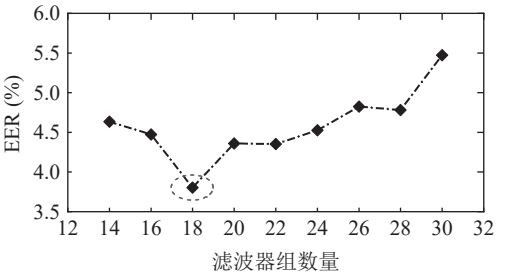

图 6　全局自注意滤波器组数量变化在验证集上的等错误率

3.3.2　特征归一化前后性能比较

对 GSTECC 特征与现有特征 CQCC、MFCC、IMFCC、LFCC、AFCC 和 TECC, 在相同的实验条件下进行了比较。CQCC(90 维)、MFCC(39 维)、IMFCC(39 维)、LFCC(39 维)、AFCC(39 维)、TECC(15 维) 和 GSTECC(15 维) 的等错误率如表 3 所示, 各特征参数在重放攻击检测中的检测误差权衡 (DET) 曲线如图 7 所示。由表 3 可知, 相对于其他特征, GSTECC 特征在测试集上获得了较低的等错误率, 仅为 20.20%。

此外, 在特征提取之后采用倒谱均值和方差归一化 (CMVN) 来提高性能和鲁棒性。在相同实验条件下, 评估了 CQCC + CMVN(90 维)、MFCC + CMVN(54 维)、IMFCC + CMVN(54 维)、LFCC + CMVN(54 维)、AFCC + CMVN(54 维)、TECC + CMVN(54 维) 和 GSTECC + CMVN(54 维), 其等错误率如表 4 所示, 各特征参数经过 CMVN 后在重放攻击检测中的 DET 曲线如图 8 所示。由表 4 可见, 使用 CMVN 后

表 3 不同特征的重放攻击检测等错误率 (%)

特征	测试集
CQCC	29.35
MFCC	35.75
IMFCC	31.68
LFCC	33.43
AFCC	24.22
TECC	25.26
GSTECC	**20.20**

表 4 不同特征使用 CMVN 后的重放攻击检测等错误率 (%)

特征	测试集
CQCC + CMVN	19.28
MFCC + CMVN	29.47
IMFCC + CMVN	18.54
LFCC + CMVN	18.14
AFCC + CMVN	17.67
TECC + CMVN	13.27
GSTECC + CMVN	**11.14**

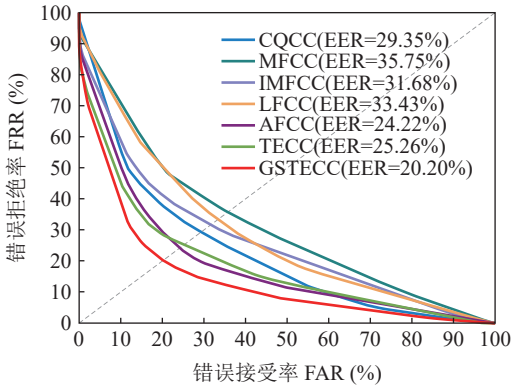

图 7 各特征参数的 DET 曲线

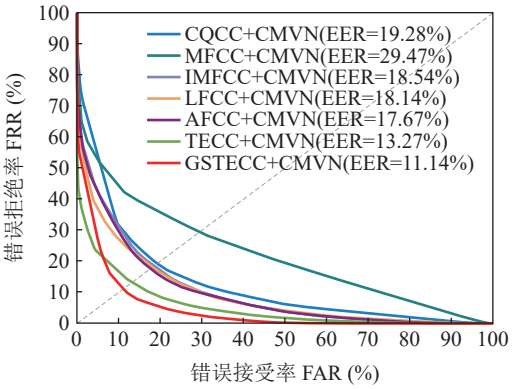

图 8 各特征参数经 CMVN 后的 DET 曲线

的 GSTECC 特征相较于其他特征, 在测试集上同样获得了相对较低的 EER 值, 仅为 11.14%。

选择 MFCC、IMFCC、LFCC 和 AFCC 作为对比特征是因为它们在特征提取过程中使用了不同的频率尺度来提取能量谱信息, 而 TECC 是在全频段上线性提取子带滤波信号并计算 Teager 能量。相比之下, 本文提出的 GSTECC 特征使用非线性全局频率尺度来重点提取高鉴别能量信息, 并使用 Teager 算子来捕捉非线性能量信号, 而不是对全频段赋予相同的权重来提取语音信息。

根据表 3 和表 4 的结果, 无论是在 CMVN 归一化前还是归一化后, 本文提出的 GSTECC 特征相对于其他特征表现出较低的等错误率, 其归一化后特征参数的等错误率低于未经归一化的情况, 但相应地特征维度也在增加。

3.3.3 多种录放设备下的性能比较

ASVspoof 2017 V2 数据库提供了录放配置的详细描述, 在收集重放语音的过程中, 使用了 25 种不同的录音设备 (标记为 R01-R25) 和 26 种不同的播放设备 (标记为 P01-P26)。根据文献 [20] 中的设备质量分类标准, 测试集中的重放语音设备质量可以分为高、中、低三个等级。由表 3 和表 4 可知, 与其他特征相比, GSTECC 在不同录放设备配置下的总体性能表现出较低的等错误率。因此, GSTECC 特征能够更好地检测不同的录放配置。

分析和讨论多个特征在各个录放配置下的个体等错误率。测试集中不同特征在不同录音设备下的等错误率如图 9 所示。与其他特征相比, 在大多数录音设备下, GSTECC 特征表现出较低的等错误率。这表明 GSTECC 在不同录音设备上具有更好的性能, 可更有效地检测重放语音。

不同特征在各个播放设备下的等错误率如图 10 所示。与录音设备情况类似, 所提出的 GSTECC 特征在大部分播放设备下相比其他特征具有较低的等错误率。这意味着 GSTECC 特征在不同播放设备配置下的鲁棒性更好, 可以更准确地辨别重放语音。

使用所有不同重放配置的多个特征在不同威胁级别下的等错误率如图 11 所示。高级威胁很难检测到, 这是因为攻击者可能使用专业的音频设备来制造更逼真的重放语音。随着威胁程度的增加, 即攻击的复杂程度和伪装技巧的提高, 检测重放语音的难度也相应增加, 这会导致等错误率随着威胁程度的增加而增加。然而, 与其他特征相比, GSTECC 在所有威胁级别下都展现出较低的等错误率。这表明 GSTECC 特征对于各种威胁情况下的重放语音具有更强的辨别能力。

综上所述, GSTECC 特征在不同录音设备、播放设备和威胁级别下表现出更优异的性能, 具有更好的鲁棒性, 可用于有效地检测重放语音。

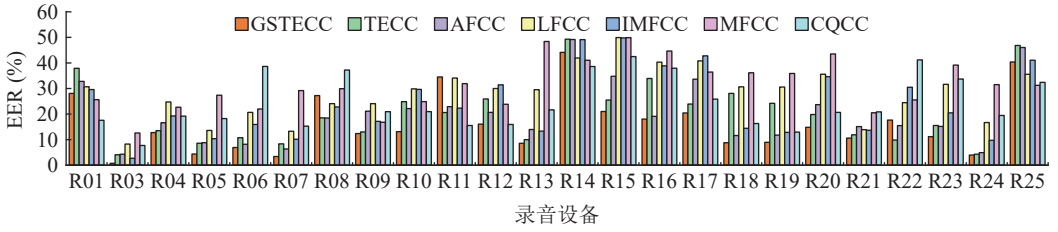

图 9　不同特征在不同录音设备下的个体等错误率

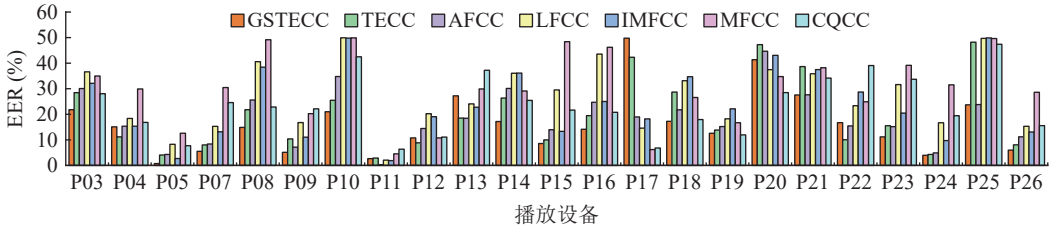

图 10　不同特征在不同播放设备下的个体等错误率

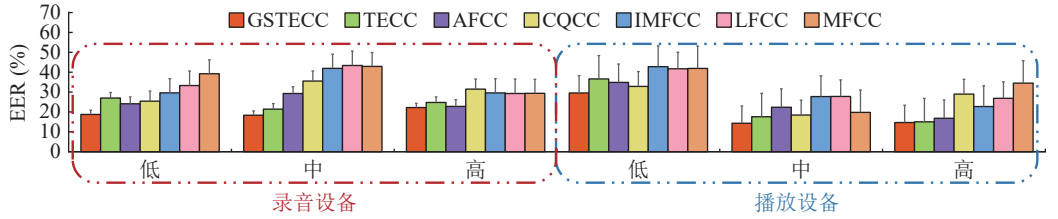

图 11　不同特征在不同威胁级别的录放设备下的等错误率

3.4　ASVspoof 2021 PA 数据库的实验结果

为了验证 GSTECC 特征在不同数据库上的检测效果, 本文在 ASVspoof 2021 PA 数据库上进行了实验, 用于评估其在不同数据库上的检测能力, 并将其性能与其他常用于重放语音检测任务的传统特征(如 CQCC、MFCC、IMFCC 和 LFCC) 进行了比较。各个特征参数在重放攻击检测方面的性能表现如表 5 所示, GSTECC 特征显示出相对较低的等错误率, 进一步验证了其在不同数据库上的有效性。

表 5　不同特征的重放攻击检测 EER(%)

特征	测试集
CQCC	38.07
MFCC	46.07
IMFCC	42.54
LFCC	46.97
GSTECC	**36.24**

4　结论

本文设计了一种全局自注意滤波器组, 用于甄选出具有较强鉴别能力的子带信号, 并运用 Teager 能量算子捕捉这些子带信号的非线性能量, 进而提出了 GSTECC 特征。基于 ASVspoof 2017 V2 数据库的实验结果显示, GSTECC 特征相对于其他常见特征具有更出色的性能。此外, 在测试集中的不同录放配置下进行了不同特征的检测效果评估, 与其他特征相比, GSTECC 特征在不同威胁条件下的 EER 也相对较低。另外, 本文还在 ASVspoof 2021 PA 数据库上进行了泛化性测试实验, 结果表明, GSTECC 特征同样表现出更好的检测性能。

参　考　文　献

1　Evans N W, Kinnunen T, Yamagishi J. Spoofing and countermeasures for automatic speaker verification. Interspeech, Lyon, France, 2013: 925−929

2　Alegre F, Janicki A, Evans N. Re-assessing the threat of replay spoofing attacks against automatic speaker verification. International Conference of the Biometrics Special Interest Group, IEEE, Darmstadt, Germany, 2014: 1−6

3　Tapkir P A, Kamble M R, Patil H A, *et al*. Replay spoof detection using power function based features. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, IEEE, Honolulu, HI, USA, 2018: 1019−1029

4　Kamble M R, Tak H, Patil H A. Amplitude and frequency modulation-based features for detection of replay spoof speech. *Speech Commun.*, 2020; **125**: 114−127

5　Kamble M R, Patil H A. Detection of replay spoof speech using teager energy feature cues. *Comput. Speech Lang.*, 2021; **65**: 101140

6 Therattil A, Gupta P, Chodingala P K, *et al*. Teager energy based-detection of one-point and two-point replay attacks: Towards cross-database generalization. The Speaker and Language Recognition Workshop (Odyssey 2022), Beijing, China, 2022: 47−54

7 Patil A T, Acharya R, Patil H A, *et al*. Improving the potential of enhanced Teager energy cepstral coefficients (ETECC) for replay attack detection. *Comput. Speech Lang.*, 2022; **72**: 101281

8 Todisco M, Delgado H, Evans N. Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Comput. Speech Lang.*, 2017; **45**: 516−535

9 Alluri K R, Achanta S, Kadiri S R, *et al*. SFF anti-spoofer: IIIT-H submission for automatic speaker verification spoofing and countermeasures challenge 2017. Interspeech, Stockholm, Sweden, 2017: 107−111

10 汤爽, 张二华, 唐振民. 基于小波包的回放语音检测算法. 计算机与数字工程, 2022; **50**(2): 238−242

11 Font R, Espín J M, Cano M J. Experimental analysis of features for replay attack detection-results on the ASVspoof 2017 Challenge. Interspeech, Stockholm, Sweden, 2017: 7−11

12 Li L, Chen Y, Wang D, *et al*. A study on replay attack and anti-spoofing for automatic speaker verification. Interspeech, Stockholm, Sweden, 2017: 92−96

13 Liu M, Wang L, Dang J, *et al*. Replay attack detection using magnitude and phase information with attention-based adaptive filters. IEEE International Conference on Acoustics, Speech and Signal Processing, Brighton, UK, 2019: 6201−6205

14 陈树丽, 张学帅, 张鹏远, 等. 静音掩蔽和频域分段的音频指纹检索算法. 声学学报, 2022; **47**(4): 531−540

15 Liu M, Wang L, Dang J, *et al*. Replay attack detection using variable-frequency resolution phase and magnitude features. *Comput. Speech Lang.*, 2021; **66**: 101161

16 郭星辰, 俞一彪. 具有仿冒攻击检测的鲁棒性说话人识别. 计算机科学, 2022; **49**(S1): 531−536

17 俞一彪, 袁冬梅, 薛峰. 一种适于说话人识别的非线性频率尺度变换. 声学学报, 2008; **33**(5): 450−455

18 Xu L, Yang J, You C H, *et al*. Device features based on linear transformation with parallel training data for replay speech detection. *IEEE/ACM Trans. Audio Speech Lang. Process.*, 2023; **31**: 1574−1586

19 姜涛, 韩纪庆, 郑铁然. 基于高斯混合模型移动因子补偿的说话人识别方法. 声学学报, 2011; **36**(6): 658−664

20 Delgado H, Todisco M, Sahidullah M, *et al*. ASVspoof 2017 Version 2.0: Meta-data analysis and baseline enhancements. The Speaker and Language Recognition Workshop (Odyssey 2018), Les Sables d'Olonne, France, 2018: 296−303