高技术通讯 2024 年 第 34 卷 第 10 期:1013-1023

doi:10.3772/j.issn.1002-0470.2024.10.001

基于倒谱特征数据增强的真实场景合成语音检测①

万 伊②* ** 李春国*** 杨飞然③* 杨 军* **

(* 中国科学院声学研究所 北京 100190)

(** 中国科学院大学 北京 100049)

(*** 东南大学信息科学与工程学院 南京 210096)

摘 要 现有合成语音检测系统在真实场景下性能损失严重。本文提出了一种基于频域掩蔽的倒谱特征数据增强方法。该方法对输入信号的线性滤波器组特征(LFBs)进行频域掩蔽,以引入符合真实场景的语音失真;计算掩蔽特征的线性频率倒谱系数(LFCC),以降低特征维度,提升检测性能。本文利用轻量级卷积神经网络(LCNN)、深度残差网络(ResNet)和一维卷积 Transformer 模型(OCT)建立了 3 种检测系统,用于验证所提方法的有效性。真实场景数据集上的实验结果表明,所提方法可使不同合成语音检测系统的等错误率(EER)相较无增强的基线降低 6.39% ~25.95%。将所提方法与基于音频编解码的增强技术相结合时,不同系统的 EER 比基线降低 31.71% ~42.47%,进一步提升了系统对真实场景的泛化能力,且性能优于现有数据增强方法。

关键词 合成语音检测;数据增强;真实场景;频域掩蔽;泛化能力

日趋成熟的语音合成(text-to-speech,TTS)与语音转换技术(voice conversion,VC)能够伪造出高质量的人类语音[1-3],或可统称为合成语音。这些合成语音的应用为日常生活提供了便利,但也存在被恶意使用的风险,例如利用合成语音攻击自动说话人认证(automatic speaker verification,ASV)系统导致用户财产损失和信息泄露,或用合成语音传播虚假信息导致他人名誉受损、引起社会恐慌[4]。开发独立的合成语音检测系统可以有效避免这些风险。

为推动合成语音检测技术的发展,多所高校与企业联合举办了 ASVspoof 挑战赛,其中的逻辑攻击(logical access,LA)任务为合成语音检测提供了高质量的纯净语音数据集[5-8]。当前合成语音检测的研究大多都基于 ASVspoof 挑战赛,在包含自然语音与合成语音的数据集上训练分类模型,并在包含未见数据的测试集上进行性能评估。目前主流的合成语音检测系统都是基于机器学习模型构建的,系统一般由前端特征提取与后端分类模型 2 个模块组成。特征提取模块通过语音信号处理或基于神经网络的方法对输入信号进行分析,旨在提取更具区分性的语音特征,以显示自然语音与合成语音的差异。后端分类模型则需要根据前端特征来判断输入的语音信号是自然语音还是合成语音。Sahidullah 等人[9]对多种语音信号特征进行了研究,发现相比于自然语音,合成语音通常缺少相位信息与高频细节信息,且无法很好地建模语音中的长时依赖关系。因此,语音信号的幅度谱特征、相位特征以及倒谱特征和其一阶、二阶动态差分系数都可作为检测合成语音的有效特征[10-16]。基于卷积运算的神经网络可处理语音信号的幅度谱等时频特征,卷积神经网

① 国家自然科学基金(62171438),北京市自然科学基金(4242013)和中国科学院声学研究所自主部署“前沿探索”类项目(QYTS202111)资助。

② 女,1995 年生,博士生;研究方向:音频信号处理,机器学习;E-mail: wanyi@mail.ioa.ac.cn。

③ 通信作者,E-mail: feiran@mail.ioa.ac.cn。

(收稿日期:2023-11-08)

络(convolutional neural network,CNN)、轻量级卷积神经网络(light convolutional neural network,LCNN)和深度残差网络(deep residual network,ResNet)等都在合成语音检测中被作为后端分类模型广泛使用[10-13, 15]。基于注意力机制的神经网络可以对语音信号中的长时依赖特征进行建模,因此也能够有效区分自然语音与合成语音[16-17]。

然而,近期研究显示,在从社交媒体等渠道收集到的真实场景数据集 In-the-Wild 上,基于纯净语音数据集开发的多类合成语音检测系统性能均大幅下降[18]。这表明真实环境更加复杂,而现有系统在真实场景下的泛化能力还亟待提升。数据增强是解决该问题的有效途径之一。数据增强方法通过对原始数据进行适当变换,以增加数据多样性、缓解过拟合、提升机器学习模型的泛化性能,在音频处理任务中得到了广泛使用[19-23]。在合成语音检测任务中,音频编解码及传输过程带来的音频失真是导致现有系统在真实场景中性能下降的一个重要原因[8, 24],因此,在真实场景下进行数据增强的主要目的是引入更符合实际的音频失真信息,同时促使模型学习到复杂环境下自然语音与合成语音的区分性特征。

频域掩蔽[20]是一种针对语音信号提出的数据增强方法。该方法计算简单,且无需任何额外数据,通过掩蔽语音信号经短时傅里叶变换(short-time Fourier transform,STFT)得到的功率谱或梅尔谱等特征中的部分频率信息即可实现数据增强。在基于纯净语音数据集的合成语音检测任务中,Chen 等人[25]对输入信号的线性滤波器组特征(linear filter banks,LFBs)进行了频域掩蔽以提升系统性能。对语音信号的 LFBs 或梅尔谱等基于滤波器组的特征进行离散余弦变换(discrete cosine transform,DCT),可得到相应的倒谱特征。相比语音信号的功率谱特征和基于滤波器组的特征,倒谱特征具有更低的维度,因此在合成语音检测任务中的应用更为广泛。然而,倒谱分析中的 DCT 操作会将语音信号滤波器组特征中代表语音频率信息的维度映射到另一特征空间,导致倒谱特征无法直接反映语音信号的频率信息。因此,目前频域掩蔽方法不能直接应用于倒谱特征的增强。

利用音频编解码(coder-decoder,Codec)算法进行数据增强的方法被广泛用于真实场景下的语音识别、说话人识别等任务中。Vu 等人[21]根据输入语音信号的时频谱失真程度及其音频平均意见得分(mean opinion score,MOS)将 Codec 算法分为 3 类,并对原始输入数据进行增强,用以训练电话语音识别系统。Ferras 等人[22]提出了一个包含多种 Codec 算法及噪声的大型开源模拟器,用于训练语音退化条件下的说话人识别系统。Hailu 等人[23]通过改变不同 Codec 算法的采样率、比特率以及采样位深对 4 种语言的小型数据集进行增强,提升了端到端自动语音识别系统的性能。在基于 ASVspoof 挑战赛的合成语音检测任务中,使用 Codec 算法进行数据增强的方法也有所应用[24,26-27],证实该方法可以在算法模拟的语音失真场景下提升合成语音检测系统的性能,但其在真实场景中的有效性还有待探究。此外,Tomilov 等人[13]还通过设计有限长脉冲响应(finite impulse response,FIR)滤波器对合成语音检测系统的输入信号进行滤波,来模拟不同 Codec 算法的幅频响应以实现数据增强。

针对现有合成语音检测系统在真实场景下性能恶化的问题,本文提出了一种基于频域掩蔽的倒谱特征数据增强方法。该方法将频域掩蔽用于具有更低维度和更优检测性能的倒谱特征数据增强中,以提升合成语音检测系统在真实场景下的泛化能力。所提方法先掩蔽系统输入信号 LFBs 的一段连续频段信息,再提取其倒谱特征及动态差分系数,以利用频域掩蔽为输入特征引入更符合真实场景的语音失真信息,实现倒谱特征的在线数据增强。所提方法可以使不同架构的合成语音检测系统在真实场景 In-the-Wild 数据集上的等错误率(equal error rate,EER)相较未进行数据增强时下降 6.39% ~25.95%,显著提升了系统的泛化能力。此外,本文还将所提方法与基于 Codec 算法的数据增强方法相结合,用于真实场景的合成语音检测任务中,使不同架构系统在 In-the-Wild 数据集上的 EER 比未使用数据增强时降低了 31.71% ~42.47%,进一步提升了现有合成语音检测系统在真实场景下的检测性能。本文将所提方法与现有数据增强方法在真实场景数据集

上进行了性能对比,实验结果也显示出所提方法的优势。

1　基于数据增强方法的合成语音检测系统构建

图 1 给出了基于数据增强方法的合成语音检测系统框架。本文中的合成语音检测系统包括数据增强、前端特征提取与后端分类模型 3 个模块,分为训练与测试 2 个阶段。数据增强模块包含 2 种方法,第 1 种是本文所提的基于频域掩蔽的倒谱特征在线增强方法,第 2 种是基于 Codec 算法的离线增强方法。本节首先介绍所提方法,然后给出了基于 Codec 算法的数据增强方法及其实现步骤。在合成语音检测系统的训练阶段,首先选取每条输入数据对应的任一 Codec 增强数据,提取其倒谱特征,并用所提方法进行数据增强,最后将增强后的倒谱特征输入后端模型并计算相应的语音得分。在系统测试阶段,其输入为测试数据,无须考虑数据增强,直接提取其倒谱特征并输入到训练好的后端模型中,再计算测试分数以判定测试数据的类别。

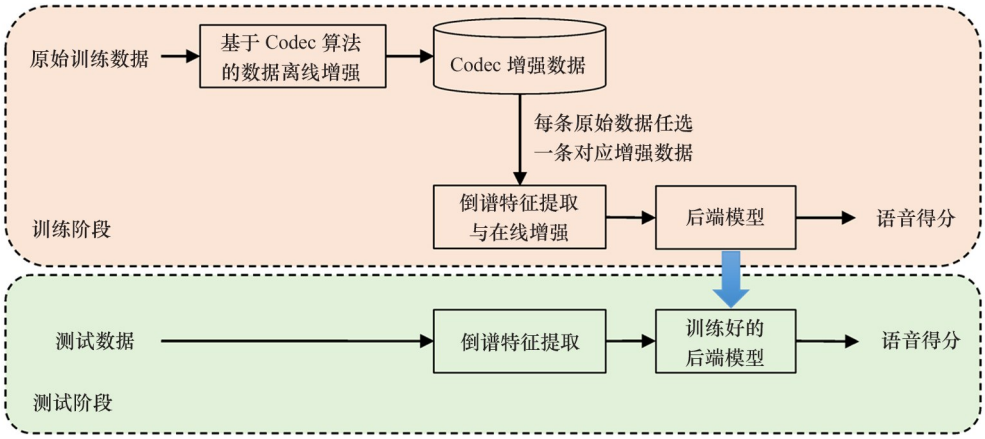

图 1　基于数据增强方法的合成语音检测系统框架

1.1　基于频域掩蔽的倒谱特征数据增强方法

线性频率倒谱系数(linear frequency cepstral coefficients,LFCC)是合成语音检测中使用最为广泛的倒谱特征,采用等间隔的线性频率三角滤波器组对语音信号的对数功率谱特征进行滤波,对全频带信息同等关注,更有利于合成语音的检测。LFCC 在基于高斯混合模型或神经网络的合成语音检测系统上均获得了良好的检测性能[8,12-13,16,28]。合成语音检测中所用的 LFCC 包含静态特征与其一阶、二阶动态差分系数。本节首先介绍频域掩蔽方法的流程,然后对本文所提基于频域掩蔽的倒谱特征增强方法及其在 LFCC 上的实现进行详细的描述。

频域掩蔽方法[20]需作用于语音信号的对数功率谱或基于滤波器组的特征上,这些特征均能直接反映语音信号的频率信息。设特征频点数为 Q,掩蔽参数为 q,则频域掩蔽方法的掩蔽操作流程如下。

(1)从 $[0,Q-q]$ 中随机选择一点 q_0 作为掩蔽区间的起点。

(2)从 $[0,q)$ 中随机选择一个值 τ 作为掩蔽区间的长度。

(3)使用恒定值对待掩蔽特征频率通道上的 $[q_0,q_0+\tau]$ 这一连续频段进行掩蔽。

(4)输出掩蔽后的特征。

其中掩蔽参数 q 实际代表了在线数据增强过程中掩蔽区间长度的上限,而随机选取的方式也能够有效缓解后端分类模型对训练数据的过拟合。

图 2 给出了本文所提基于频域掩蔽的倒谱特征数据增强方法流程,所提取的倒谱特征为 LFCC。倒谱分析中的 DCT 操作导致频域掩蔽方法不能直接作用于 LFCC,因此所提方法通过线性频率三角滤波器组对语音信号的功率谱进行滤波,先计算得到信号的 LFBs。LFBs 能够直接反映语音信号的频率信息,因此可用频域掩蔽方法实现数据增强。利用对数运算与DCT对掩蔽后LFBs进行倒谱分析,

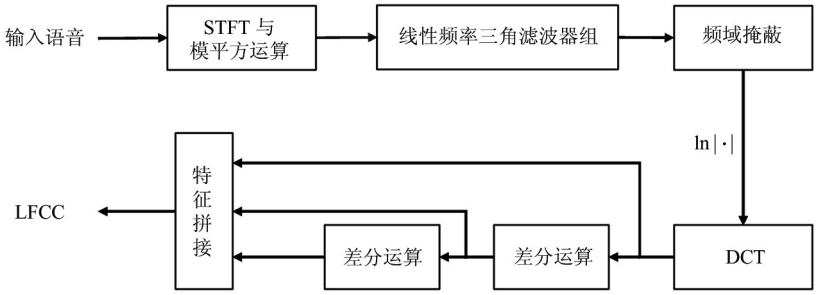

图 2　基于频域掩蔽的 LFCC 特征提取与数据增强方法流程图

即可得到增强后的语音信号 LFCC 特征。本文所提方法通过在输入信号特征提取模块的滤波与 DCT 操作之间加入频域掩蔽,为信号的 LFCC 引入了更符合实际场景的频域失真信息,以提升场景失配时模型的泛化能力。

对于系统的输入语音信号 $\boldsymbol{x}$,首先利用窗函数 $\boldsymbol{w}$ 对其进行短时傅里叶变换,则其短时谱特征第 n 帧第 k 个频点的特征值可表示为

$$\boldsymbol{X}_n(k) = \sum_{i=0}^{K-1} \boldsymbol{x}(i+nF)\boldsymbol{w}(i)\,\mathrm{e}^{-j\frac{2\pi}{K}iK}, \quad 0 \leqslant k \leqslant K-1 \tag{1}$$

其中, K 为傅里叶变换的点数, F 为帧移。设线性频率三角滤波器组中滤波器的个数为 M, 则对于第 n 帧信号的幅度谱 $|\boldsymbol{X}_n(k)|$, 第 m 个滤波器输出的语音信号 LFBs 的特征值为

$$\boldsymbol{S}_n(m) = \sum_{k=0}^{N/2} |\boldsymbol{X}_n(k)|^2 \boldsymbol{H}_m(k), \quad 0 \leqslant m \leqslant M-1 \tag{2}$$

其中 $\boldsymbol{H}_m(k)$ 为第 m 个滤波器的第 k 阶系数。利用零值对 LFBs 进行频域掩蔽,则掩蔽后的特征值为

$$\begin{cases} \hat{\boldsymbol{S}}_n(m) = 0 & q_0 \leqslant m < q_0 + \tau \\ \hat{\boldsymbol{S}}_n(m) = \boldsymbol{S}_n(m) & \text{其他} \end{cases} \tag{3}$$

其中 q_0 和 τ 分别为掩蔽区间起点与区间长度。通过对数运算与 DCT 计算掩蔽后 LFBs 所对应的 LFCC 静态特征,其特征值为

$$\boldsymbol{c}_n(d) = \sum_{m=0}^{M-1} \ln(\hat{\boldsymbol{S}}_n(m))\cos\left(\frac{\pi d(m+0.5)}{M}\right), \quad 0 \leqslant d \leqslant D-1 \tag{4}$$

其中 D 为 DCT 后保留的倒谱系数的阶数,且 $D \leqslant M$。LFCC 静态特征值对应的一阶与二阶动态差分系数分别为

$$\boldsymbol{\delta}_n(d) = \frac{\sum_{l=1}^{(L-1)/2} i(\boldsymbol{c}_{n+l}(d) - \boldsymbol{c}_{(n-l)}(d))}{2\sum_{l=1}^{(L-1)/2} l^2}, \quad 0 \leqslant d \leqslant D-1 \tag{5}$$

$$\boldsymbol{\delta}'_n(d) = \frac{\sum_{l=1}^{(L-1)/2} i(\boldsymbol{\delta}_{n+l}(d) - \boldsymbol{\delta}_{(n-l)}(d))}{2\sum_{l=1}^{(L-1)/2} l^2}, \quad 0 \leqslant d \leqslant D-1 \tag{6}$$

其中, L 为计算差分系数时选择的连续帧的数目, l 为帧索引。

1.2　基于 Codec 算法的数据增强方法

在实际应用中,Codec 算法导致的语音质量损失会使得自然语音与合成语音更加难以区分[8, 24]。因此,本文将基于 Codec 算法的数据增强方法用于真实场景合成语音检测任务中,并将所提基于频域掩蔽的倒谱特征数据增强方法与其相结合来进一步提升系统的跨场景泛化能力。

根据其应用场景,Codec 算法可分为 2 类,一类用于实时语音通话,另一类用于音视频文件存储。多媒体领域的标准化组织为不同场景制定了多种音频编码与压缩标准。例如国际电信联盟为固定电话、IP 电话等实时通话场景制定了 G.711、G.726、G.722、G.723.1 等标准,其中 G.711 有 a-law 和 μ-law 这 2 种形式;欧洲电信标准组织为数字移动通信制定了全球移动通信系统(global system for mobile communications,GSM)标准;运动图像专家组为音视频存储制定了 MPEG-1 Layer3、MPEG-AAC 等标准。由互联网工程任务组进行标准化的 Opus,能处理比特率从 6 kbps 至 510 kbps 的音频,可以用于实时通话场景和音频文件的压缩存储。表 1 给出了本文在进行离线数据增强时所使用的 10 种 Codec

算法及其对应的中间文件格式。本文所用 Codec 算法均借助 FFmpeg 工具[29]实现,算法比特率最低为 6.3 kbps,最高为 128.0 kbps,涵盖了不同的语音编码质量。为了便于训练时的统一处理,最终得到的增强语音均转换为 FLAC(free lossless audio codec)格式,即一种无损压缩的音频编码格式。

表 1　本文所用 Codec 算法列表

Codec	采样率/kHz	比特率/kbps	中间文件格式
GSM	8	13.0	GSM
G.726	8	32.0	WAV
G.723.1	8	6.3	WAV
G.722	16	64.0	WAV
a-law	16	128.0	WAV
μ-law	16	128.0	WAV
Opus	16	16.0	OGG
Opus	16	64.0	OGG
AAC	16	64.0	AAC
MP3	16	128.0	MP3

基于 Codec 算法的离线数据增强方法具体步骤如下。

(1) 从原始数据集中读取一条语音。

(2) 遍历表 1 中的 Codec 算法,并根据算法对输入语音采样率的需求对语音进行重采样。

(3) 使用 Codec 算法处理语音,生成对应的 10 个中间文件。

(4) 将得到的中间文件统一重采样至 16 kHz,并转换为 FLAC 格式。

(5) 存储离线语音数据。

基于 Codec 算法的增强数据使用方式已在图 1 中给出。在训练阶段,首先从表 1 中随机选择一种 Codec 算法;然后从离线增强数据集中读取输入数据对应的增强数据,提取其 LFCC 并利用所提方法进行在线增强;最后将增强后的倒谱特征输入后端模型计算语音得分。

1.3　合成语音检测系统后端模型

本文在 3 种不同架构的合成语音检测后端模型上验证所提数据增强方法的性能,包括 LCNN[12]、ResNet[30]以及一维卷积 Transformer 模型(one-dimensional convolutional Transformer,OCT)[16]。

1.3.1　LCNN

LCNN 使用最大特征图激活(max-feature-map,MFM)操作来对重要特征进行选择与学习[31-32]。结合语音信号 LFCC 与 LCNN 模型的合成语音检测系统在 ASVspoof 2019 挑战赛的合成语音检测任务中取得了最优性能[12],在后续的 ASVspoof 挑战赛中也作为基线模型之一[8]。本文遵循基线模型的结构,这样能够更好地评估数据增强方法的性能。

1.3.2　ResNet

ResNet 使用短路连接的残差模块,很大程度上缓解了深层网络性能退化的问题。在合成语音检测任务中,残差网络及其变体也得到了广泛的应用[15,24,33-34]。本文选择 18 层 ResNet 模型作为验证数据增强方法性能的后端模型之一。

1.3.3　OCT

Transformer[35]是一种完全基于自注意力机制的自然语言处理模型,对语音信号中的长期依赖特征具有优秀的建模特性。OCT[16]是近期在 ASVspoof 挑战赛的合成语音检测任务上获得先进性能的模型之一。它集成了一维卷积模块与 Transformer 模型,利用 Transformer 编码器对语音的长时关联特征进行提取,并利用卷积模块提升模型的泛化能力。OCT 模型具有更少的参数量和更优的检测性能,但它在真实场景下仍会出现严重的性能退化现象。因此,本文将 OCT 作为后端模型之一,以探究所提数据增强方法在真实场景下对现有合成语音检测系统性能的影响。

2　实验设置

2.1　数据集

为验证本文算法的有效性,实验采用 ASVspoof 2019-LA 所提供的训练集与验证集对模型进行训练与超参数调节,并在真实场景 In-the-Wild 数据集上对系统进行性能评估。ASVspoof 2019-LA 测试集上的测试结果则用于对比。表 2 给出了本文所用数据集的划分与组成方式。

ASVspoof 2019-LA 数据集[7]是近年来合成语音

检测中最为常用的基准数据集。数据集中仅包含纯净语音数据,且数据均录制于半消声室环境中,无明显噪声,也未经过任何 Codec 算法处理。其中合成语音数据由自然语音通过表 2 中相应编号的 TTS 或 VC 算法处理得到。该数据集的训练集与验证集包含了挑战赛中编号为 A01 ~ A06 的 TTS 或 VC 算法,测试集中包含了编号为 A07 ~ A19 的 TTS 或 VC 算法。A07 ~ A19 算法中有 2 种为训练集中出现过的已知算法,其余均为未知算法。用于所提方法性能评估的数据来自近期发布的 In-the-Wild 数据集[18]。该数据集中的语音数据完全来自于真实场景,在发布与传播的过程中就已经由不同的 Codec 算法进行了处理,但其 Codec 算法以及合成语音的处理算法均为未知。In-the-Wild 数据集中的语音数据具有更多样的声学特征,更具挑战性,在该数据集上的测试结果能够更真实地反映合成语音检测系统在实际应用中的性能。本文所用数据集的语音数据采样率均为 16 kHz。

表 2　本文所用数据集列表

数据集名称	ASVspoof 2019-LA			In-the-Wild
	训练集	验证集	测试集	
数据集类型	训练集	验证集	纯净语音测试集	真实场景测试集
自然语音数目	2 580	2 548	7 355	19 963
合成语音数目	22 800	22 296	63 882	11 816
说话人数目	20	20	68	54
TTS/VC 算法编号	A01 ~ A06	A01 ~ A06	A07 ~ A19	未知
Codec 编号	—	—	—	未知
语音数据格式	FLAC			WAV

2.2　评价指标

EER 是合成语音检测任务中的通用评价指标[7]。更低的 EER 代表了系统具有更好的检测性能。在合成语音检测任务中,系统的输出是语音分别属于自然语音与合成语音的概率 P_{bona} 和 P_{spoof},输入语音的得分由这 2 种概率计算得到,可表示为

$$s_{\text{score}} = \ln P_{\text{bona}} - \ln P_{\text{spoof}} \tag{7}$$

根据式(7)可以看出,合成语音检测系统对于自然语音会给出较高得分,而对于合成语音会给出较低得分。

从合成语音检测系统的错误来源分析,其错误率主要有 2 种。将合成语音误判为自然语音的错误率称为错误接受率(false accept rate,FAR),可表示为 $P_{\text{FAR}}(\theta)$,将自然语音误判为合成语音的错误率称为错误拒绝率(false reject rate,FRR),可表示为 $P_{\text{FRR}}(\theta)$。其中,θ 代表合成语音检测系统进行分类判决时的分数阈值。合成语音检测系统的 2 种错误率可分别表示为

$$P_{\text{FAR}}(\theta) = \frac{\text{得分大于等于}\,\theta\,\text{的合成语音数目}}{\text{测试集中合成语音总数目}} \tag{8}$$

$$P_{\text{FRR}}(\theta) = \frac{\text{得分小于}\,\theta\,\text{的自然语音数目}}{\text{测试集中自然语音总数目}} \tag{9}$$

最终,合成语音检测系统将对比输入语音得分 s_{score} 与判决阈值 θ 的大小,并据此对输入语音进行分类。具体而言,系统将得分大于等于 θ 的输入语音判别为自然语音,反之则判别为合成语音。在该判决规则中,θ 的选取是影响系统性能的关键因素之一。在式(8)中,测试集合成语音总数目为固定值,θ 取值越大,得分大于等于 θ 的合成语音数目就越少,因此 $P_{\text{FAR}}(\theta)$ 为 θ 的单调递减函数。同理,式(9)中测试集自然语音总数目固定,θ 取值越大,得分小于 θ 的自然语音数目就越多,可知 $P_{\text{FRR}}(\theta)$ 为 θ 的单调递增函数。根据上述函数的单调性质可知,$P_{\text{FAR}}(\theta) = P_{\text{FRR}}(\theta)$ 时合成语音检测系统的 2 种错误率可同时达到最低,此时的系统错误率即定义为 EER,可表示为 P_{EER},对应的系统判决阈值为 θ_{EER}。P_{EER} 的定义式如下:

$$P_{\text{EER}} = P_{\text{FAR}}(\theta) = P_{\text{FRR}}(\theta), \theta = \theta_{\text{EER}} \tag{10}$$

2.3　实验环境及参数设置

本文中的合成语音检测系统基于 Python 3.9 实

现,使用 PyTorch 1.13 作为深度学习框架。训练在 Nvidia RTX 3090 GPU 上进行。对于 3 种后端模型本文均采用 Focal Loss[36](FL)作为损失函数进行训练,它能够缓解不同类别数据量不平衡带来的影响,其表达式为

$$L_{\mathrm{FL}} = -\alpha(1 - P(y \mid x))^{\gamma}\ln(P(y \mid x)) \quad (11)$$

其中 $P(y \mid x)$ 为输入的语音信号 x 属于类别 y 的概率。损失函数的参数设置为 $\alpha = 0.75$, $\gamma = 2.0$。优化器选择 AdamW[37],优化器参数 $\beta_1 = 0.900$, $\beta_2 = 0.999$, 权重衰减值设为 0.01。训练过程的批处理大小设为 64。对于 LCNN 和 ResNet 后端,训练的初始学习率为 0.001,训练周期为 100,学习率预热周期为 10。分别在每 10 个和每 20 个周期后,将学习率衰减为之前的 0.5 倍。OCT 模型遵循原论文的训练策略[16],设置初始学习率为 0.000 8,训练周期为 300,学习率预热周期为 30,之后采用余弦函数进行学习率衰减。对于所提基于频域掩蔽的数据增强方法,设置其掩蔽参数 $q_f = 20$, 掩蔽恒定值为零值。本文中 LFCC 特征提取的相关参数参考 ASVspoof 挑战赛的基线[8],设 $M = 70$, $D = 20$, $L = 5$, STFT 中对语音信号进行处理的窗函数选择汉宁窗,窗长为 20 ms,帧移为 10 ms。

2.4　对比方法实验设置

为探究掩蔽与 DCT 操作在特征提取与在线数据增强模块中的顺序对系统性能的影响,本文为图 2中所提的数据增强方法设置了对比方法。对比方法的流程如图 3 所示。相比所提方法,对比方法先对滤波模块输出的语音信号的 LFBs 进行对数运算与离散余弦变换,得到相应的 LFCC 静态特征,然后对其特征通道进行掩蔽来实现数据增强。但 LFCC 静态特征无法直接反映语音信号的频率信息,因此对其特征通道采取掩蔽操作的方法并没有明确的物理意义,也无法反映实际中的语音失真。对比方法的掩蔽操作流程与掩蔽参数设置与所提方法保持一致,不同的是该方法的掩蔽操作是在特征通道上进行的。

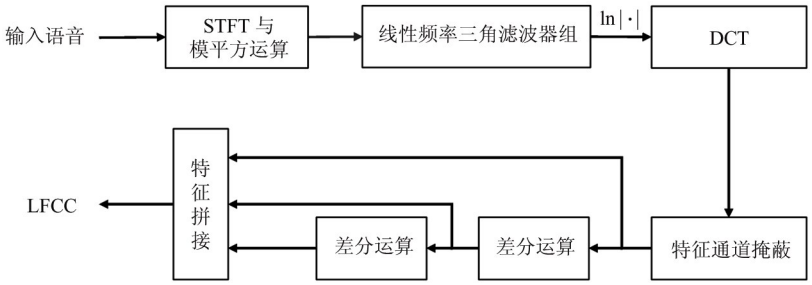

图 3　基于特征通道掩蔽的 LFCC 特征提取与数据增强方法流程图

3　实验结果与分析

本节首先对基于 3 种不同架构后端的合成语音检测系统在仅有纯净语音的场景和真实场景下分别进行了性能评估,以便分析现有系统的跨场景泛化能力,然后在真实场景数据集上对本文所提数据增强方法的有效性进行了实验验证与分析,最后还探究了训练数据集使用方式对系统在真实场景下检测性能的影响。

3.1　合成语音检测系统跨场景泛化性能分析

表 3 给出了在不同场景测试集上基于语音信号 LFCC 和不同架构后端的 3 种合成语音检测系统的性能。该实验仅使用 ASVspoof 2019-LA 的原始训练集进行模型训练。相比于仅包含纯净语音数据的 ASVspoof 2019-LA 测试集,不同系统在真实场景 In-the-Wild 数据集上的性能均严重退化。这表明,现有的合成语音检测系统在不同场景下存在着巨大的性能差异,不同系统在真实场景下的表现甚至不如

表 3　3 种合成语音检测系统在不同场景测试集上的性能

系统名称	EER/%	
	ASVspoof 2019-LA 测试集	In-the-Wild
LFCC + LCNN	6.55	67.47
LFCC + ResNet	4.92	69.21
LFCC + OCT	1.06	54.02

随机预测的结果。同时,实验结果也表明基于纯净语音数据训练的模型尚无法直接应用于真实环境中。

3.2　所提方法在真实场景合成语音检测任务中的性能分析与对比

表 4 给出了使用不同数据增强方法训练得到的系统在 In-the-Wild 数据集上的测试结果,加粗的数据分别表示不同系统所获得的最佳性能。使用无掩蔽方法与原始训练数据的合成语音检测系统,其 EER 即为表 3 中 In-the-Wild 数据集上的结果,这里将其列于表 4 中以便参考。

首先对本文所提基于频域掩蔽的倒谱特征数据增强方法进行性能分析。从表 4 可以看出,用仅包含纯净语音的原始数据集进行训练的 LCNN、ResNet 和 OCT 模型,在使用本文所提倒谱特征频域掩蔽方法时,相应系统的 EER 值相较于无掩蔽分别降低了 25.95%、16.02% 和 6.39%。这表明所提方法能够显著提升不同合成语音检测系统在真实场景下的泛化能力。表 4 的结果还表明,将所提基于频域掩蔽的倒谱特征增强方法与基于 Codec 的数据增强方法相结合,可以使 LCNN、ResNet 和 OCT 模型对应系统的最低 EER 值相较不使用任何数据增强时分别降低 42.42%、42.47% 和 31.71%,进一步提升了系统在真实场景下的泛化能力与检测性能。此外,表 4的结果也显示出基于 Codec 算法的数据增强方法在真实场景合成语音检测任务中的有效性。

表 4　使用不同数据增强方法时不同合成语音检测系统在 In-the-Wild 数据集上的性能

系统名称	训练数据集	EER/%		
		无掩蔽	所提倒谱频域掩蔽方法	对比掩蔽方法
LFCC + LCNN	原始数据集	67.47	41.52	66.57
	Codec 增强数据集	31.73	25.05	32.37
LFCC + ResNet	原始数据集	69.21	53.19	72.75
	Codec 增强数据集	30.96	26.74	30.12
LFCC + OCT	原始数据集	54.02	47.63	49.91
	Codec 增强数据集	43.28	22.31	33.41

其次,本文对所提方法和 2.4 节中介绍的对比方法进行了比较与分析,以探讨掩蔽操作与 DCT 操作在合成语音检测系统特征提取模块中的顺序对于系统检测性能的影响,其实验结果也列于表 4 中。本文对比了不同系统与训练数据设置下所提方法与对比方法的检测性能。从表 4 可以看出,在真实场景下,对比掩蔽方法对合成语音检测系统的性能提升较小,甚至还可能会使系统的 EER 增大。而所提方法考虑到了 DCT 导致的倒谱特征无法直接反映信号频率信息的问题,因此能够比对比方法更好地模拟信号可能存在的频率失真。表 4 中的实验结果显示,相较于对比方法,所提方法均获得了更低的 EER。这表明在利用掩蔽操作对倒谱特征进行数据增强时,需要考虑 DCT 对语音特征频率通道进行变换所带来的影响,合理选择掩蔽操作在特征提取模块中的位置,才能引入更符合实际的语音失真信息。同时对比实验的结果也证实了所提方法在真实场景合成语音检测任务中的有效性与可靠性。根据上述分析,本文认为在进行数据增强时,需要结合实际的测试场景,而不是盲目利用增强方法对原始数据进行变换,否则可能会干扰系统判决,对系统的检测性能产生负面影响。

本文还比较了所提方法与现有基于语音特征的数据增强方法在 In-the-Wild 数据集上的性能。为保证实验的公平性,现有方法分别在本文所用的 3 种不同后端模型上进行了实现,实验结果在表 5 中给出。基于 FIR 滤波器的数据增强方法是对原始信号进行滤波,来模拟不同 Codec 算法所带来的语音失真[13]。在训练阶段,利用随机选择的低通或高通 FIR 滤波器对合成语音检测系统的输入数据进行增强,并提取其 LFCC 作为后端模型的输入。FIR 滤波器的设计方法遵循文献[13],两类滤波器被选择的概率均为 50%。由于 FIR 滤波器方法与基于 Codec 的方法原理类似,都是将 Codec 算法的失真信

息引入训练数据,因此在不同后端模型上实现该方法时本文所用的训练数据为纯净语音数据。文献[25]直接将频域掩蔽后的语音信号的 LFBs 作为后端模型的输入,并未进一步提取倒谱特征,也没有加入动态系数。为了保证对比的公平性,在不同后端模型上实现该方法时,本文均用 Codec 增强数据作为训练数据。

从表 5 看出,在真实场景数据集上,所提方法在 3 种不同模型上的性能均优于现有方法。所提方法与文献[25]方法的性能对比可以证实,利用 DCT 提取输入信号 LFBs 对应的 LFCC 并加入动态差分系数,可以提升合成语音检测系统在真实场景下的检测性能。

表 5　使用所提方法与现有数据增强方法时不同合成语音检测系统在 In-the-Wild 数据集上的性能

数据增强方法	EER/%		
	LCNN	ResNet	OCT
LFCC + FIR 滤波[13]	34.18	37.85	30.49
LFB + 频域掩蔽 + Codec[25]	37.41	31.29	28.50
LFCC + 倒谱频域掩蔽 + Codec(本文方法)	**25.05**	**26.74**	**22.31**

3.3　真实场景下训练数据集使用方式对系统性能的影响

在机器学习任务中,增强数据通常会与原始数据一起用于模型的训练。但由于 ASVspoof 所提供的原始训练数据均为纯净语音,与真实场景的测试数据并不匹配,因此在训练中使用原始数据所带来的失配问题,可能会对系统在真实场景下的性能产生影响。

为探究上述问题,本文在模型训练阶段加入了纯净语音数据,并在基于不同后端与不同掩蔽方法的合成语音检测系统上进行了性能测试。表 6 给出了纯净数据与 Codec 增强数据同时使用时不同系统获得的检测性能。对比表 6 与表 4 的实验结果,相比仅使用原始数据训练的系统,数据量扩充后不同系统的性能均得到了提升,但相比仅使用 Codec 增强数据训练的系统,其性能反而有所下降。导致这种现象的可能原因是 In-the-Wild 数据集中的数据完全来自真实场景,并不包含在半消声室中录制的纯净语音,所以原始训练数据的使用反而会影响系统性能。因此,本文建议在真实场景任务中使用基于 Codec 方法得到的离线增强数据对模型进行训练,既能够提升合成语音检测系统在真实场景下的泛化能力,也能节约训练成本。此外,结合表 4 与表 6 的实验结果可以看出,相比无掩蔽和对比掩蔽方法,所提基于频域掩蔽的倒谱特征数据增强方法在不同的训练集设置下均取得了更低的 EER,显示出该方法的性能优势。

表 6　同时使用原始数据和 Codec 增强数据训练时不同合成语音检测系统在 In-the-Wild 数据集上的性能

系统名称	EER/%		
	无掩蔽	所提倒谱频域掩蔽方法	对比掩蔽方法
LFCC + LCNN	45.58	26.72	40.81
LFCC + ResNet	36.00	29.55	31.95
LFCC + OCT	50.62	24.13	37.16

4　结 论

现有的合成语音检测系统在真实场景数据集上性能恶化,表明当前研究距离实际应用还有较大距离。针对上述问题,本文提出了一种基于频域掩蔽的倒谱特征数据增强方法,用于提升合成语音检测系统在真实场景下的泛化能力。所提方法先对语音信号 LFBs 的随机连续频段进行掩蔽,为合成语音检测系统的前端特征引入了更符合实际场景的频域失真信息,然后再对其进行倒谱分析与动态差分系

数提取,有效降低了前端特征的维度,同时更有利于提升系统对合成语音的检测性能。本文利用 LCNN、ResNet 以及基于注意力机制的 OCT 模型构建了 3 种不同架构的合成语音检测系统,用于对所提基于频域掩蔽的倒谱特征数据增强方法进行性能测试。实验结果表明,所提方法能够显著提升合成语音检测系统在真实场景数据集上的泛化能力,且在不同后端模型与不同训练集设置下均优于对比方法。将本文所提方法与基于 Codec 算法的数据增强方法相结合,可使合成语音检测系统在真实场景数据集上的检测性能得到进一步提升。与其他已有方法进行比较,本文方法得到的实验结果也具有性能优势与竞争力。此外,实验结果证实,在训练阶段使用原始的纯净语音会影响合成语音检测系统在真实场景下的检测性能,因此本文建议利用基于 Codec 算法得到的增强数据作为训练数据来进行系统的构建。

参考文献

[1] OORD A, DIELEMAN S, ZEN H, et al. Wavenet: a generative model for raw audio[EB/OL]. (2016-09-19)[2023-07-24]. http://arxiv.org/pdf/1609.03499.

[2] WANG Y, SKERRY-RYAN RJ, STANTON D, et al. Tacotron: towards end-to-end speech synthesis [EB/OL]. (2017-04-06)[2023-07-24]. https://arxiv.org/pdf/1703.10135.

[3] ARIK S Ö, CHRZANOWSKI M, COATES A, et al. Deep voice: real-time neural text-to-speech[C]//Proceedings of the 34th International Conference on Machine Learning. Sydney, Australia: ACM, 2017:195-204.

[4] 任延珍, 刘晨雨, 刘武洋, 等. 语音伪造及检测技术研究综述[J]. 信号处理, 2021,37(12):2412-2439.

[5] WU Z, KINNUNEN T, EVANS N, et al. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge[C]//The 16th Annual Conference of the International Speech Communication Association. Dresden, Germany: ISCA, 2015:2037-2041.

[6] TODISCO M, WANG X, VESTMAN V, et al. ASVspoof 2019: future horizons in spoofed and fake audio detection[C]//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019:1008-1012.

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

[8] YAMAGISHI J, WANG X, TODISCO M, et al. ASVspoof 2021: accelerating progress in spoofed and deepfake speech detection[C]//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. Online: ISCA, 2021:47-54.

[9] SAHIDULLAH M, KINNUNEN T, HANILCI C. A comparison of features for synthetic speech detection[C]//Proceedings of the 16th Annual Conference of the International Speech Communication Association. Dresden, Germany: ISCA, 2015:2087-2091.

[10] ZHANG C, YU C, HANSEN J. An investigation of deep-learning frameworks for speaker verification antispoofing[J]. IEEE Journal of Selected Topics in Signal Processing, 2017,11(4):684-694.

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

[12] LAVRENTYEVA G, NOVOSELOV S, TSEREN A, et al. STC antispoofing systems for the ASVspoof2019 challenge[C]//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019:1033-1037.

[13] TOMILOV A, SVISHCHEV A, VOLKOVA M, et al. STC antispoofing systems for the ASVspoof2021 challenge[C]//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. Online: ISCA, 2021:61-67.

[14] YU H, TAN Z, MA Z, et al. Spoofing detection in automatic speaker verification systems using DNN classifiers and dynamic acoustic features[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018,29(10):4633-4644.

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

[16] LI C, YANG F, YANG J. The role of long-term dependency in synthetic speech detection[J]. IEEE Signal Processing Letters, 2022,29:1142-1146.

[17] TAK H, JUNG J, PATINO J, et al. End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection[C]//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. Online: ISCA, 2021:1-8.

[18] MüLLER N M, CZEMPIN P, DIECKMANN F, et al. Does audio deepfake detection generalize? [C]//Proceedings of the 23rd Annual Conference of the International Speech Communication Association. Incheon, Korea: ISCA, 2022:2783-2787.

[19] 袁文浩, 娄迎曦, 梁春燕, 等. 利用生成噪声提高语音增强方法的泛化能力[J]. 电子学报, 2019,47(4):791-797.

[20] PARK D S, CHAN W, ZHANG Y, et al. SpecAugment: a simple data augmentation method for automatic speech recognition[C]//Proceedings of the 20th Annual Conference of the International Speech Communication Association. Graz, Austria: ISCA, 2019:2613-2617.

[21] VU T L, ZENG Z, XU H, et al. Audio codec simulation based data augmentation for telephony speech recognition[C]//Proceedings of 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference. Lanzhou, China: IEEE, 2019:198-203.

[22] FERRAS M, MADIKERI S, MOTLICEK P, et al. A large-scale open-source acoustic simulator for speaker

recognition[J]. IEEE Signal Processing Letters, 2016, 23(4):527-531.
[23] HAILU N, SIEGERT I, NüRNBERGER A. Improving automatic speech recognition utilizing audio-Codecs for data augmentation[C]//Proceedings of IEEE 22nd International Workshop on Multimedia Signal Processing. Tampere, Finland: IEEE, 2020:1-5.
[24] COHEN A, RIMON I, AFLALO E, et al. A study on data augmentation in voice anti-spoofing[J]. Speech Communication, 2022,141:56-67.
[25] CHEN T, KUMAR A, NAGARSHETH P, et al. Generalization of audio deepfake detection[C]//Proceedings of the Speaker and Language Recognition Workshop (Odyssey 2020). Tokyo, Japan: ISCA, 2020:132-137.
[26] CHEN X, ZHANG Y, ZHU G, et al. UR channel-robust synthetic speech detection system for ASVspoof 2021[C]//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. Online: ISCA, 2021:75-82.
[27] WANG X, QIN X, ZHU T, et al. The DKU-CMRI system for the ASVspoof 2021 challenge: vocoder based replay channel response estimation[C]//Proceedings of 2021 Edition of the Automatic Speaker Verification and Spoofing Countermeasures Challenge. Online: ISCA, 2021:16-21.
[28] LI X, LI N, WENG C, et al. Replay and synthetic speech detection with res2net architecture[C]//Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. Toronto, Canada: IEEE, 2021:6354-6358.
[29] TOMAR S. Converting video formats with FFmpeg[J]. Linux Journal, 2006,146:10.
[30] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the 29th IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE, 2016:770-778.
[31] WU X, HE R, and SUN Z. A lightened CNN for deep face representation[EB/OL]. (2015-11-09)[2023-07-24]. https://arxiv.org/pdf/1511.02683.
[32] GOODFELLOW I, WARDE-FARLEY D, MIRZA M, et al. Maxout networks[C]//Proceedings of the 30th International Conference on Machine Learning. Atlanta, USA: PMLR, 2013:1319-1327.
[33] YANG J, WANG H, DAS R, et al. Modified magnitude-phase spectrum information for spoofing detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021,29:1065-1078.
[34] ZHANG Y, JIANG F, DUAN Z. One-class learning towards synthetic voice spoofing detection[J]. IEEE Signal Processing Letters, 2021,28:937-941.
[35] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st Annual Conference on Neural Information Processing Systems. Long Beach, USA: Curran Associates Inc., 2017:6000-6010.
[36] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE International Conference on Computer Vision. Venice, Italy: IEEE, 2017:2999-3007.
[37] LOSHCHILOV I, HUTTER F. Decoupled weight decay regularization[EB/OL]. (2017-11-14)[2023-07-24]. https://arxiv.org/pdf/1711.05101.

Real scene synthetic speech detection based on cepstral feature data augmentation

WAN Yi[*,**], LI Chunguo[***], YANG Feiran[*], YANG Jun[*,**]
([*]Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190)
([**]University of Chinese Academy of Sciences, Beijing 100049)
([***]School of Information Science and Engineering, Southeast University, Nanjing 210096)

Abstract

The performance of existing synthetic speech detection systems is significantly degraded in real scenarios. This paper proposes a data augmentation method for cepstral features via frequency masking. First, linear filter banks (LFBs) of the input signal are masked on frequency channels for realistic speech distortion. Then, the linear frequency cepstral coefficients (LFCC) of the masked features are calculated to reduce the feature dimensionality and improve the detection performance. Using light convolutional neural network (LCNN), deep residual network (ResNet) and one-dimensional convolutional Transformer (OCT), three detection systems are established to verify the effectiveness of the proposed method. Experiments on the real scene datasets show that the proposed method can reduce the equal error rate (EER) of different synthetic speech detection systems by 6.39% -25.95% compared with the baseline without augmentation. The proposed method with the codec-based augmentation can reduce the EER of different systems by 31.71% -42.47% compared with the baseline, which further improves the generalization ability of the systems in real scenarios, and outperforms the existing data augmentation methods.

Key words: synthetic speech detection, data augmentation, real scenes, frequency masking, generalization ability