



研究与开发

采用局部相位量化的合成语音检测方法

徐嘉¹, 简志华^{1,2}, 金宏辉¹, 杨曼¹

(1. 杭州电子科技大学信工程学院, 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: 由于语音合成的便利性, 合成伪装语音对说话人认证系统的安全构成了很大的威胁。为了进一步提升说话人认证系统的伪装语音检测能力, 提出了一种利用语谱图频域信息的合成语音检测方法, 它通过局部相位量化算法对语谱图频域信息进行描述。首先, 将语谱图分为若干子块, 然后对每个子块进行局部相位量化, 经直方图统计分析后获得局部相位量化特征向量并将该特征向量作为随机森林分类器的输入特征, 实现合成语音检测。实验结果表明, 该方法进一步降低了合成语音检测系统的串联检测代价数值, 并且具有更强的泛化能力。

关键词: 说话人认证; 伪装攻击; 合成语音检测; 局部相位量化

中图分类号: TP391.42

文献标志码: A

doi: 10.11959/j.issn.1000-0801.2024024

A method for synthetic speech detection using local phase quantization

XU Jia¹, JIAN Zhihua^{1,2}, JIN Honghui¹, YANG Man¹

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: Due to the convenience of speech synthesis, synthesized disguised speech poses a great threat to the security of speaker verification systems. In order to further enhance the ability of detecting the camouflage to the speaker verification system, a method of synthetic speech detection was put forward using the information in spectral domain of the synthetic speech spectrogram. The method employed the local phase quantization (LPQ) algorithm to describe frequency domain information in the speech spectrogram. Firstly, the spectrogram was divided into several sub-blocks, and then the LPQ was performed on each sub-block. After the histogram statistical analysis, the LPQ feature vector was obtained and used as the input feature of the random forest classifier to realize the synthetic speech detection. The experimental results demonstrate that the proposed method further reduces tandem detection cost function (t-DCF) and has better generalization ability.

Key words: speaker verification, spoofing attack, synthetic speech detection, LPQ

收稿日期: 2023-11-27; 修回日期: 2024-01-07

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)

Foundation Items: The National Natural Science Foundation of China (No.61201301, No.61772166)



0 引言

近年来,随着人工智能的迅猛发展,自动说话人认证(automatic speaker verification, ASV)系统得到广泛的应用^[1],但该系统易受伪装语音的恶意攻击,安全性有待提高,伪装语音检测技术应运而生^[2]。伪装语音检测技术通过对输入语音进行特征提取和模型匹配,识别出该语音是真实说话人的语音还是伪装语音^[3]。常见的语音伪装方式有语音合成(speech synthesis, SS)、语音转换(voice conversion, VC)、重放攻击(replay attack)和话音模仿(voice mimicry)等技术。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,破坏性强,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^[4]。伪装语音检测可以识别恶意伪装攻击,提高 ASV 系统的安全性能,因此,研究合成语音伪装检测具有重要的应用价值。

伪装语音检测方法通过提取语音信号的特征参数信息实现伪装检测^[5]。传统的特征参数主要有两种,分别为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,将非线性问题转化为线性问题,更易处理,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^[6]。文献[7]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^[8]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。在文献[9]中,提出了 CQCC 特征,该特征在 ASVspoof 2015 数据集上 EER 为

0.255%。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[10]发现群时延特征保留了较多的共振峰结构,证明了其对于语音处理的鲁棒性。文献[11-12]表明,将语音信号的相位信息和幅频特性相结合,提取的特征参数可以很好地改善伪装语音检测的性能,并且可以更好地适应多种类型的伪装检测。由于声道模型不一致,不同说话人在语音的基频、共振峰、声道参数、相位等方面会有较大差异。另外,重放语音会引入设备噪声,转换语音对基频、相位等信息有不同程度的修改,因此在识别重放、转换或不同说话人的语音时,上述两类特征可以利用这些信息起到很好的区分作用。然而,随着合成语音算法和电子设备的进步发展,语音合成技术得到了长足的进步,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。利用这些特征进行合成语音伪装检测的准确率已经不能满足需求,而且对于未知类型的攻击,这种缺陷更加明显。

为了进一步提升系统对合成语音的检测性能,本文利用语谱图的时频信息,提出了一种采用局部相位量化(local phase quantization, LPQ)提取语音特征的合成语音检测(synthetic speech detection, SSD)方法。在语谱图的频率域内,低频反映了其基本轮廓,高频反映了其内部细节。考虑真实语音与合成语音的语谱图在轮廓和细节方面均存在一定差异,使用 LPQ 算法对语谱图进行纹理分析,利用语谱图的频域信息对合成语音进行检测,提高系统的检测性能。另外,通过麦克风采集语音信号时,可能存在电噪声、设备噪声等内部干扰信息,这些干扰信息的存在会造成语谱图图像模糊,从而影响对语谱图的纹理特征提取。而 LPQ 算法对图像模糊不敏感,提取的纹理特征性能更加稳定,对合成语音鉴别能力更强,非常适合合成语音检测。

1 LPQ 算法

图像频域处理是对图像蕴含的频域信息进行分析的过程,该过程一般借助傅里叶变换来进行。图像的频域信息揭示了图像灰度变化的程度,低频处灰度差异不明显,是构成图像大致结构的主要成分;高频处灰度差异较大,填充了图像的细节内容。LPQ 算法是一种在频域对图像进行纹理分析的算法,旨在通过局部邻域傅里叶变换的量化相位来描述图像。该算法最初用于模糊图像纹理描述,由 Ojansivu 等^[13]提出。

在图像处理时,模糊图像 $g(x)$ 可以看作原始图像 $f(x)$ 和点扩散函数 (point spread function, PSF) $h(x)$ 的卷积,表示为:

$$g(x) = f(x) * h(x) \quad (1)$$

其中, x 是坐标向量 $[x, y]^T$ 。若用 $G(u)$ 表示 $g(x)$ 的 DFT, $F(u)$ 表示 $f(x)$ 的 DFT, $H(u)$ 表示 $h(x)$ 的 DFT, u 是坐标向量 $[u, v]^T$, 则在频域内有如下关系式。

$$G(u) = F(u) \times H(u) \quad (2)$$

幅度和相位可分别表示为式 (3) 和式 (4)。

$$|G(u)| = |F(u)| \times |H(u)| \quad (3)$$

$$\angle G(u) = \angle F(u) + \angle H(u) \quad (4)$$

从式 (3) 和式 (4) 可以看出,图像模糊会对图像的幅度和相位产生影响。若 PSF 是中心对称的,则其相位 $\angle H(u)$ 只能为 0 或 π , 当 $H(u) \geq 0$ 时,相位 $\angle H(u)$ 为 0; 当 $H(u) < 0$ 时,

相位 $\angle H(u)$ 为 π 。则有:

$$\angle G(u) = \angle F(u), H(u) \geq 0 \quad (5)$$

因此在 $H(u) \geq 0$ 时,图像模糊不影响其原始相位。

基于上述模糊不变性,LPQ 算法将图像 $f(x)$ 中每个像素点 x 的 $M \times M$ 邻域 N_x 通过 STFT 变换至频域并提取其局部相位信息,该过程可表示为:

$$F(u, x) = \sum_{y \in N_x} f(x - y) e^{-j2\pi u^T y} \quad (6)$$

其中, $F(u, x)$ 表示像素点 x 在二维频率 u 处的 STFT 变换值。然后通过 4 个二维频率计算局部傅里叶系数: $u_1 = [a, 0]^T$, $u_2 = [0, a]^T$, $u_3 = [a, a]^T$ 和 $u_4 = [a, -a]^T$ 。其中, a 为一个极小的标量,通常取 $1/M$ 。计算后,对于像素点 x , 可得到向量:

$$F(x) = [F(u_1, x), F(u_2, x), F(u_3, x), F(u_4, x)] \quad (7)$$

$$G(x) = [\text{Re}\{F(x)\}, \text{Im}\{F(x)\}] \quad (8)$$

接着,根据 $G(x)$ 各分量的符号,使用简单的标量量化来统计相位信息:

$$q_j = \begin{cases} 1, & g_j \geq 0 \\ 0, & g_j < 0 \end{cases} \quad (9)$$

其中, g_j 为 $G(x)$ 的第 j 个分量。最终,根据式 (10) 计算得到各点的 LPQ 特征值。

$$\text{LPQ}(x) = \sum_{j=1}^8 q_j(x) 2^{j-1} \quad (10)$$

LPQ 算法处理流程如图 1 所示。

由于 LPQ 算法对每个像素点提取了 4 个局部

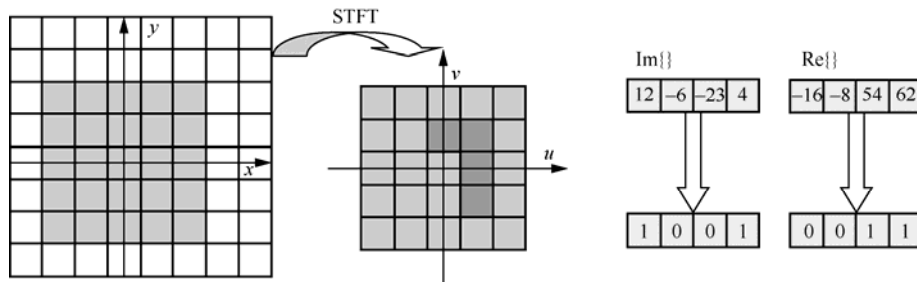


图 1 LPQ 算法处理流程



傅里叶系数,并分别量化其实部和虚部,因此,各点的 LPQ 特征由 8 位二进制数表示,LPQ 特征值共有 256 种不同的模式。在使用 LPQ 特征描述图像的纹理特性时,常将整幅图像的 LPQ 特征值映射至统计直方图。图像的 LPQ 特征图和直方图如图 2 所示。

近年来,LPQ 特征已被应用于磁共振成像疾病检测^[14]、指纹活体检测^[15]和非接触式人物识别^[16]等领域。考虑真实语音和合成语音在语谱图中具有轮廓和细节上的差异,且语谱图的频域信息与之对应,因此使用 LPQ 算法对语谱图进行处理,利用语谱图的频域信息对合成语音进行鉴别。

2 基于 LPQ 的合成语音检测方法

基于 LPQ 的 SSD 方法使用 LPQ 算法提取语谱图的频域信息作为纹理特征向量,从而检测合成语音。LPQ 提取流程如图 3 所示。

首先使用 STFT 获取语谱图。在真实语音与合成语音的语谱图中,轮廓与细节部分均存在差异,而语谱图的频域信息中蕴含着这些信息,因此将语音信号转换为语谱图进行特征提取。

获取语谱图后,考虑距离越远像素点之间的影响越小,将语谱图划分为无重叠的 4 个子块,以便充分使用语谱图的细节信息,从而更好地对其进行纹理分析。然后将三通道的 RGB 语谱图进行灰度处理,变为单通道的灰度图,降低了数据处理复杂度,同时保留了语谱图的纹理信息。灰度处理后,使用 LPQ 算法对每一子块进行频域纹理特征提取。在每个 3×3 邻域内,计算中心像素点的 LPQ 特征值,得到各子块的 LPQ 特征图。然后映射至统计直方图得到各子块的 LPQ 特征向量,长度为 256 维。再将各子块的 LPQ 特征向量串联即可得到语谱图的 LPQ 特征,长度为 $256 \times 4 = 1024$ 。各子块 LPQ 特征图和直方图如图 4 所示。

特征获取后,选用随机森林作为分类模型。随机森林是一种集成学习方法,通过组合多个决策树来构建分类器^[17],基于多个决策树的投票结果进行分类,对于含噪数据和异常值较为鲁棒。在构建决策树时,随机森林进行随机抽样和特征选择,增加决策树的多样性,可以有效减少模型过拟合的风险,提高整体模型的鲁棒性和泛化能力。并且由于构建决策树时进行特征选择,模型

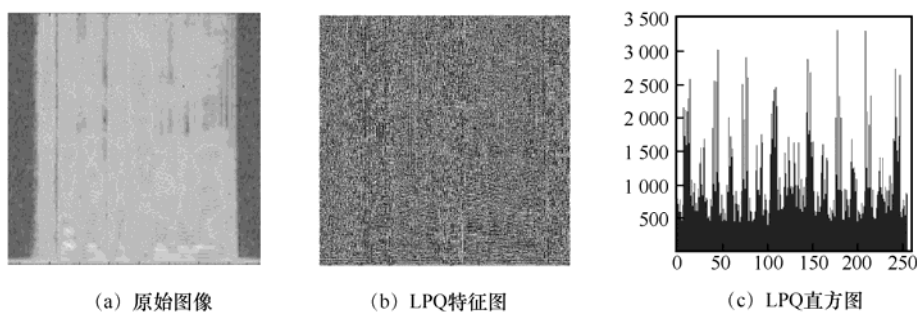


图2 图像的 LPQ 特征图和直方图

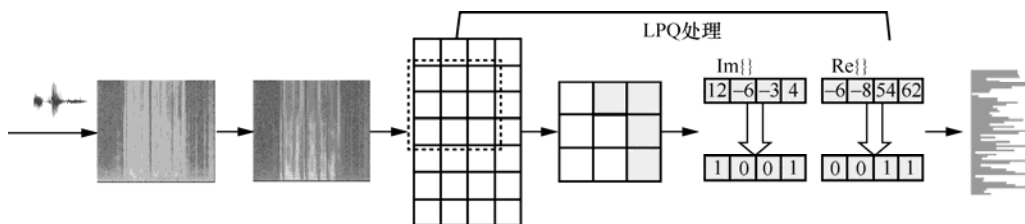


图3 LPQ 提取流程

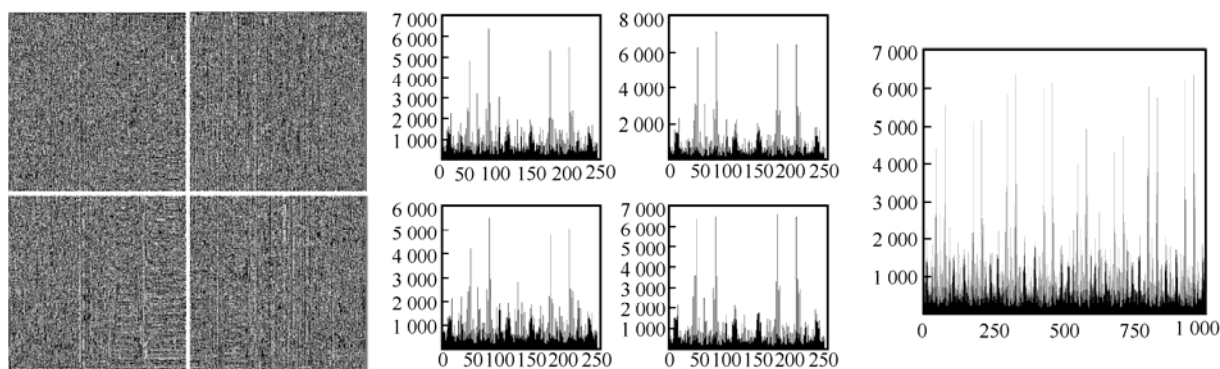


图4 各子块 LPQ 特征和直方图

计算的复杂性大大降低,多棵树并行计算也使模型计算效率大幅提升。

最终使用 LPQ 特征作为 SSD 系统的前端特征,随机森林模型作为后端分类器对合成语音进行鉴别,具体步骤如下。

步骤 1 选取训练集中真实语音和 A01~A04 类型的合成语音,构建训练样本库。

步骤 2 使用 STFT 获取各训练样本的语谱图。

步骤 3 将语谱图划分为 4 个子块,并将三通道的 RGB 语谱图转换为单通道的灰度图。

步骤 4 使用 LPQ 算法对各子块灰度图进行纹理分析,利用其频域信息得到各子块的 256 维 LPQ 特征向量。

步骤 5 将各子块的 256 维 LPQ 特征向量串联,得到各语音长度为 1 024 维的 LPQ 特征向量,

并将各样本特征向量集整理为训练特征矩阵。

步骤 6 为每个样本分配一个标签,真实语音标签为 1,合成语音标签为 0,得到各训练样本的标签序列。

步骤 7 利用训练特征矩阵和标签序列训练随机森林模型,得到 SSD 系统。

步骤 8 使用同样方法对待测语音提取 LPQ 特征向量,并输入 SSD 模型中进行鉴别。

步骤 9 若待测语音是真实语音,则允许接入系统,否则拒绝其接入。

基于 LPQ 的 SSD 整体流程如图 5 所示。

3 实验与结果

3.1 数据集

实验选用 ASVspoof 2019 数据库中的 LA 数

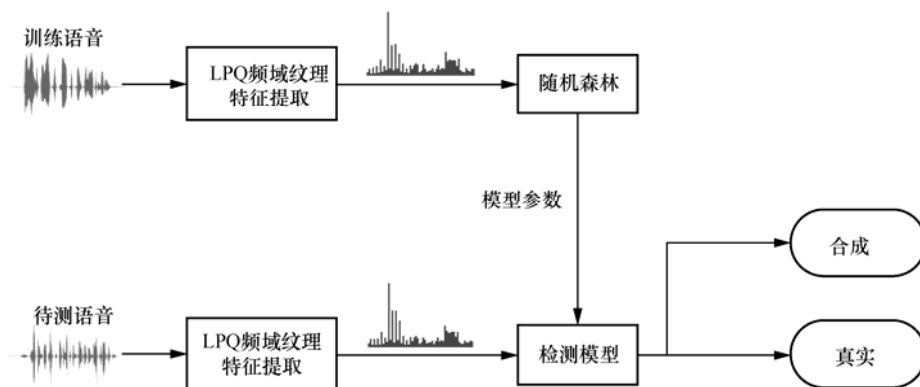


图5 基于 LPQ 的 SSD 整体流程



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,共包含了神经波形模型、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	转换	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 性能指标

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:

$$\text{t-DCF} = C_{\text{FRR}} \pi_{\text{tar}} P_{\text{FRR}}^{\text{CM}} + C_{\text{FAR}} (1 - \pi_{\text{tar}}) P_{\text{FAR}}^{\text{CM}} \quad (11)$$

$$P_{\text{FRR}}^{\text{CM}} = \frac{N_{\text{FR}}}{N_{\text{bonafide}}} \quad (12)$$

$$P_{\text{FAR}}^{\text{CM}} = \frac{N_{\text{FA}}}{N_{\text{spoof}}} \quad (13)$$

$$\pi_{\text{tar}} = \frac{N_{\text{bonafide}}}{N_{\text{bonafide}} + N_{\text{spoof}}} \quad (14)$$

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 $C_{\text{FRR}}=1$ 、 $C_{\text{FAR}}=10$ 、 $\pi_{\text{tar}}=0.9405$ 。

3.3 性能测试

LPQ 特征利用了语谱图的频域信息,以语谱图的傅里叶变换系数为处理对象,对各采样频点的相位进行标量量化,从而获得语谱图的 LPQ 特征。为了验证基于 LPQ 的 SSD 方法的有效性,分别提取语音信号的 MFCC 特征、LPCC 特征、LFCC 特征和 CQCC 特征,并提取反映语谱图纹理信息的 LBP (local binary pattern) 特征^[20]、CSLBP (center symmetric local binary pattern) 特征^[21]以及 LPQ 特征。将上述特征在 SVM 和 RF 模型做后端分类器时的检测情况进行比较,以 t-DCF 为评价标准。在实验中,用训练集中真实语音和 A01~A04 类型的合成语音作为训练样本,用开发集中真实语音和 A01~A04 类型的合成语音作为开发集样本,用评估集中真实语音和 A07~A16 类型的合成语音作为评估集样本。经实验,各系统的 t-DCF 对比结果见表 2。

表 2 各系统的 t-DCF 对比结果

特征	开发集		测试集	
	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
LBP	0.175 6	0.160 1	0.185 6	0.176 5
CSLBP	0.122 4	0.121 0	0.123 6	0.122 2
LPQ	0.094 1	0.082 5	0.094 5	0.082 7

可以看出,在所有对比特征中,使用 LPQ 特征的 SSD 系统 t-DCF 最低,这说明 LPQ 特征对合成语音的表征能力最强。主要原因是,传统声学特征中往往只包含语音信号的时域或频域某一方面的信息,其蕴含的声学特性虽然具有一定的辨别能力,但整体来讲对语音信号的描述较为片

面。而 LPQ 特征是在语谱图的基础上进行分析提取的,本身就蕴含着语音信号的时域、频域和能量信息,再对其进行频域特征提取,通过频域相位精准地描述语谱图的轮廓与细节差异,可以更加深入地描述语音特性,因此 t-DCF 远低于传统声学特征。

而在各纹理特征中,LPQ 也表现了其显著优势。与 CSLBP-RF 系统相比,LPQ-RF 系统的 t-DCF 降低了 32.32%。其原因是,CSLBP 特征是对语谱图的空域信息进行处理,根本上是对语谱图中各像素点的灰度进行分析,反映了各像素点的灰度数值大小和位置关系。在 CSLBP 特征中,真实语音和合成语音语谱图的差异仅表现在灰度值的大小和位置变化。而 LPQ 特征是对语谱图的频域信息进行处理,根本上是对语谱图中各像素点间灰度变化的程度进行分析,反映了图像的梯度信息^[22]。并且,通过傅里叶变换,可以将语谱图分解为低频分量和高频分量,更加准确地反映真实语音与合成语音语谱图的区别所在,因此检测性能更优。

另外,系统训练时无法将所有类型的合成语音考虑在内,若系统在面对已知类型合成攻击时表现较好,却在面对未知类型的合成攻击时表现不佳,则会对系统稳定性造成影响,从而影响 ASV 系统的安全性。因此,对各系统在开发集和评估集中的 t-DCF 差值进行比较。差值越小,系统泛化能力越强,各系统评估集与开发集 t-DCF 之差如图 6 所示。在使用传统声学特征的 SSD 系统中,评估集与开发集 t-DCF 之差最小为 0.013 9。LBP-RF 系统评估集与开发集 t-DCF 之差为 0.016 4, CSLBP-RF 系统评估集与开发集 t-DCF 之差为 0.001 2,而 LPQ-RF 系统评估集与开发集 t-DCF 之差仅为 0.000 2。这说明,当系统并未学习过待测语音的同种内在特性时,LPQ-RF 系统的检测更具有稳定性,不易受语音合成技术不同的影响,具有更强的泛化能力。

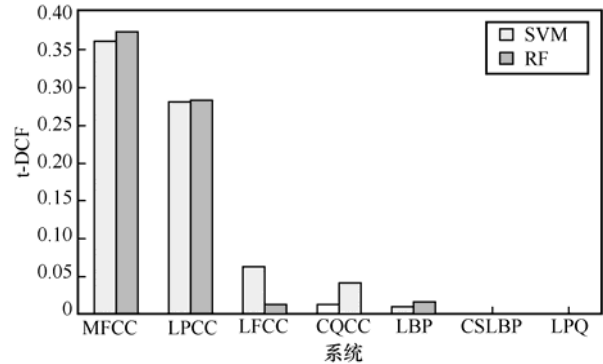


图 6 各系统评估集与开发集 t-DCF 之差

考虑本文主要利用语谱图的纹理特征检测合成语音,为了分析系统对各类型合成语音的检测能力,各系统使用 LBP、CSLBP 和 LPQ 等纹理特征对开发集和评估集中不同类型合成语音分别进行检测,其 t-DCF 比较结果见表 3 和表 4。

表 3 各系统对开发集中不同类型合成语音进行检测的 t-DCF 比较结果

类型	SVM			RF		
	LBP	CSLBP	LPQ	LBP	CSLBP	LPQ
A01	0.080 7	0.100 4	0.033 2	0.073 2	0.105 3	0.032 5
A02	0.375 3	0.204 6	0.248 9	0.269 0	0.178 6	0.206 4
A03	0.088 4	0.082 1	0.012 4	0.177 0	0.105 2	0.041 2
A04	0.143 1	0.102 7	0.081 6	0.120 4	0.093 0	0.050 2

表 4 各系统对评估集中不同类型合成语音进行检测的 t-DCF 比较结果

类型	SVM			RF		
	LBP	CSLBP	LPQ	LBP	CSLBP	LPQ
A07	0.101 7	0.123 6	0.065 1	0.107 4	0.114 1	0.062 3
A08	0.092 3	0.109 0	0.067 7	0.102 4	0.110 5	0.065 4
A09	0.115 4	0.111 1	0.064 6	0.138 7	0.101 7	0.060 1
A10	0.206 9	0.104 1	0.118 2	0.189 8	0.111 0	0.093 6
A11	0.186 9	0.090 9	0.064 9	0.257 7	0.100 6	0.093 5
A12	0.181 4	0.124 4	0.076 5	0.158 8	0.114 2	0.085 0
A13	0.191 7	0.122 3	0.107 0	0.128 0	0.111 4	0.081 9
A14	0.179 3	0.138 4	0.077 4	0.167 0	0.144 0	0.071 2
A15	0.409 5	0.164 6	0.197 2	0.352 5	0.161 9	0.134 6
A16	0.172 6	0.147 0	0.101 1	0.154 7	0.144 3	0.080 2



整体来看,使用 LPQ 特征和 RF 分类器的方法在 SSD 任务上有更好的表现。但从表 3 中可以发现,在检测 A02 类型的合成语音时,LPQ-RF 系统的 t-DCF 略高于 CSLBP-RF 系统,这说明语谱图的空间信息对该类型的合成语音表征能力略强。另外,各个系统对 A02 类型合成语音进行检测的 t-DCF 数值要显著高于其他类型,这是由于 A02 类型的合成语音使用了基于神经网络的语音合成算法,并将 MFCC、基频和非周期性系数作为 WORLD 声码器的输入,语音合成时保留了较多的纹理信息,因此利用纹理特征检测该类型合成语音的效果不佳。

从表 4 中还可以看出,不论哪种类型的合成语音,LPQ-RF 系统均取得了最低的 t-DCF,但对不同类型语音进行纹理特征检测的性能改善程度不同。从表 4 中可以看出,对于 A11 和 A15 类型的合成语音来说,使用 CSLBP 特征或 LPQ 特征可以有效地降低 SSD 系统的 t-DCF,且 LPQ 降低幅度更大。A11 类型的合成语音使用 Griffin-Lim 算法来生成波形^[23],仅通过幅度谱来对相位谱进行估计,语谱图的像素位置与频域相位信息与真实语音差异较大,因此 LPQ 特征可以有效地改善对该类型语音的检测性能。而 A15 类型的语音整体算法与 A14 类型相同,但 A15 类型语音使用了 WaveNet 声码器生成波形,A14 类型合成语音使用 STRAIGHT 声码器。A15 的性能提升显著高于 A14,这说明使用 WaveNet 声码器生成的语谱图中,缺少位置与相位的相关信息,因此使用 LPQ 特征可以降低其 t-DCF。然而在语音合成领域,WaveNet 声码器在许多语音合成技术中得到了应用,因此使用 LPQ 特征可以有效地将合成语音检出。整体来看,LPQ-RF 系统对各类型合成语音具有较好的辨别能力,有效地降低了合成语音的 t-DCF。

4 结束语

为了进一步提升 SSD 系统的检测性能,本文

提出了一种基于 LPQ 的 SSD 方法。该方法利用了语谱图的频域信息,使用 LPQ 算法对语谱图进行纹理分析,获取的纹理特征可以更好地表征真实语音与合成语音的差异,进一步降低了系统的 t-DCF 值,提升了系统的检测性能。实验结果表明,基于 LPQ 的 SSD 方法的 t-DCF 比传统的声学特征以及反映语谱图纹理信息的 LBP 特征和 CSLBP 特征都要低,而且在几种纹理特征中,LPQ 的泛化能力也最强。

参考文献:

- [1] REN Y Q, PENG H P, LI L X, et al. Generalized voice spoofing detection via integral knowledge amalgamation[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023(31): 2461-2475.
- [2] CHENG P, ROEDIG U. Personal voice assistant security and privacy—a survey[J]. Proceedings of the IEEE, 2022, 110(4): 476-507.
- [3] 徐剑, 简志华, 于佳祺, 等. 采用完整局部二进制模式的伪装语音检测[J]. 电信科学, 2021, 37(5): 91-99.
XU J, JIAN Z H, YU J Q, et al. Completed local binary pattern based speech anti-spoofing[J]. Telecommunications Science, 2021, 37(5): 91-99.
- [4] 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. Synthetic spoofing speech detection method based on center-symmetric local binary pattern[J]. Telecommunications Science, 2023, 39(1): 72-78.
- [5] 陈佳, 章坚武, 张浙亮. 基于上下文信息与注意力特征的欺骗语音检测[J]. 电信科学, 2023, 39(2): 92-102.
CHEN J, ZHANG J W, ZHANG Z L. Spoof speech detection based on context information and attention feature[J]. Telecommunications Science, 2023, 39(2): 92-102.
- [6] 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.
- [7] 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.
- [8] 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.
- [9] TODISCO M, HÉCTOR DELGADO, EVANS N. Constant Q cepstral coefficients: a spoofing countermeasure for automatic speaker verification[J]. Computer Speech & Language, 2017(45): 516-535.
- [10] LOWEIMI E, BARKER J, HAIN T. Statistical normalisation of phase-based feature representation for robust speech recognition[C]//Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2017: 5310-5314.
- [11] YANG J C, WANG H J, DAS R K, et al. Modified magnitude-phase spectrum information for spoofing detection[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021(29): 1065-1078.
- [12] KIM J, BAN S M. Phase-aware spoof speech detection based on Res2net with phase network[C]//Proceedings of the ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2023: 1-5.
- [13] OJANSIVU V, HEIKKILA J. Blur insensitive texture classification using local phase quantization[J]. Lecture Notes in Computer Science, 2008(5099): 236-243.
- [14] MACIN G, TASCI B, TASCI I, et al. An accurate multiple sclerosis detection model based on exemplar multiple parameters local phase quantization: ExMPLPQ[J]. Applied Sciences, 2022, 12(10): 4920-4929.
- [15] RASWA F H, KINART I Y, PULUNGAN R, et al. Fingerprint liveness detection using denoised-bayes shrink wavelet and aggregated local spatial and frequency features[C]//Proceedings of the 2022 International Conference on Machine Learning and Cybernetics (ICMLC). Piscataway: IEEE Press, 2022: 103-108.
- [16] CHAA M, AKHTAR Z, LATI A. Contactless person recognition using 2D and 3D finger knuckle patterns[J]. Multimedia Tools and Applications, 2022, 81(6): 8671-8689.
- [17] 刘琳岚, 高声荣, 舒坚. 基于随机森林的链路质量预测[J]. 通信学报, 2019, 40(4): 202-211.
- LIU L L, GAO S R, SHU J. Link quality prediction based on random forest[J]. Journal on Communications, 2019, 40(4): 202-211.
- [18] WANG X, YAMAGISHI J, TODISCO M, et al. ASVspoof 2019: a large-scale public database of synthesized, converted and re-played 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] LU F X, HUANG J. An improved local binary pattern operator for texture classification[C]//Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE Press, 2016: 1308-1311.
- [21] XIONG Z X, LIU M L, GUO Q. Finger vein recognition method based on center-symmetric local binary pattern[C]//Proceedings of the 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA). Piscataway: IEEE Press, 2019: 262-266.
- [22] 朱长水, 丁勇, 袁宝华, 等. 融合 LBP 和 LPQ 的人脸识别[J]. 南京师大学报(自然科学版), 2015, 38(1): 104-107, 112.
- ZHU C S, DING Y, YUAN B H, et al. Face recognition based on local binary pattern and local phase quantization[J]. Journal of Nanjing Normal University (Natural Science Edition), 2015, 38(1): 104-107, 112.
- [23] GRIFFIN D, LIM J. Signal estimation from modified short-time Fourier transform[J]. IEEE Transactions on Acoustics, Speech, and Signal Processing, 1984, 32(2): 236-243.

[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副教授, 主要研究方向为语音转换、伪装语音检测、声纹识别以及语音隐私保护等。

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。

杨曼 (2000-), 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。