研究与开发

基于联合特征与随机森林的伪装语音检测

于佳祺[1]，简志华[1]，徐嘉[1]，游林[2]，汪云路[2]，吴超[1]
（1. 杭州电子科技大学通信工程学院，浙江 杭州 310018;
2. 杭州电子科技大学网络空间安全学院，浙江 杭州 310018）

摘　要：为了能较为全面地描述语音信号的特征信息，提高伪装检测率，提出了一种基于均匀局部二值模式纹理特征与常数 Q 倒谱系数声学特征相结合，并以随机森林为分类模型的伪装语音检测方法。利用均匀局部二值模式提取语音信号语谱图中的纹理特征矢量，并与常数 Q 倒谱系数构成联合特征，再用所获得的联合特征矢量训练随机森林分类器，从而实现了伪装语音检测。实验中，分别对其他特征参数以及支持向量机分类器模型所构建的几种伪装检测系统进行了性能对照，结果表明，所提联合特征与随机森林模型相结合的语音伪装检测系统具有最优的检测性能。

关键词：伪装语音检测；声学特征；纹理特征；均匀局部二值模式；随机森林
中图分类号：TP391.42
文献标志码：A
doi: 10.11959/j.issn.1000−0801.2022089

Spoofing speech detection algorithm based on joint feature and random forest

YU Jiaqi[1], JIAN Zhihua[1], XU Jia[1], YOU Lin[2], WANG Yunlu[2], WU Chao[1]
1. School of Communication Engineering, Hangzhou Dianzi University, Hangzhou 310018, China
2. School of Cyberspace Security, Hangzhou Dianzi University, Hangzhou 310018, China

Abstract: In order to describe the characteristic information of the speech signal more comprehensively and improve the detection rate of camouflage, a spoofing speech detection method based on the combination of uniform local binary pattern texture feature and constant Q cepstrum coefficient acoustic feature was proposed, which used random forest as the classifier model. The texture feature vector in the speech signal spectrogram was extracted by using the uniform local binary mode, and the joint feature was formed with the constant Q cepstrum coefficient. Then, the obtained joint feature vector was used to train the random forest classifier, so as to realize the camouflage speech detection. In the experiment, the performances of several spoofing detection systems constructed by other feature parameters and the support vector machine classifier model were compared, and the results show that the proposed speech spoofing detection system combined with the joint feature and the random forest model has the best performance.

Key words: spoofing speech detection, acoustic feature, texture feature, uniform local binary pattern, random forest

收稿日期：2022−01−02；**修回日期：**2022−05−15
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)

0 引言

自动说话人验证（automatic speaker verification，ASV）系统是通过对说话人语音信号进行分析并对说话人身份进行认证的技术。ASV 系统是一种无须直接接触便可完成识别的身份认证方式，检测设备成本低且便于操作[1-2]。虽然目前 ASV 系统的正确识别率高，但数据显示，以冒充目标说话人真实身份为目的的恶意欺骗攻击极大地降低了 ASV 系统的安全性。欺骗攻击的类型主要有语音合成、语音转换[3]、人为模仿与语音回放[4-5]。为了应对这些不同种类的欺骗攻击，需要提高说话人识别系统检测欺骗攻击的能力，使 ASV 系统具有反欺骗攻击的能力[6-7]。

伪装语音检测的研究重点是提取特征参数与建立欺骗检测模型，其中，特征提取主要是提取语音信号中的声学特征来描述目标语音特性[8]。目前的语音信号特征提取方法有很多，梅尔频率倒谱系数（Mel-frequency cepstral coefficient，MFCC）就是常用的声学特征之一，MFCC 是模仿人耳对不同频率的语音信号具有不同感知程度的听觉特性[9]。线性频率倒谱系数（linear frequency cepstral coefficient，LFCC）与 MFCC 的获取方法类似，但是滤波器组不是按照 Mel（梅尔）频率分布，而是使用线性频率。在 ASVspoof2019 挑战赛中，这两种特征参数都被 ASV 官方基线系统所选用。MFCC 与 LFCC 这两种特征在说话人验证中都有不错的表现，但是在欺骗检测中性能并不理想[10-12]。随着研究的深入，逐渐出现了其他针对欺骗语音检测的声学特征。Todisco 等[13]提出了基于常量 Q 变换（constant Q transform，CQT）的常量 Q 倒谱系数（constant Q cepstral coefficient，CQCC）。CQCC 能够提供可变的时间和频率分辨率，克服了其他声学特征时频分辨率均匀的缺点，且 CQT 能够更加有效地提取频谱的细节信息，这使得其在伪装语音检测中可以取得更好的效果。实验结果也表明，CQCC 在多数据集上有很好的泛化效果[14-15]。然而，这些特征参数都没有考虑频域特征与时域特征间的相关性。Massoud 等[16]借鉴图像领域的研究成果，使用卷积神经网络（convolutional neural network，CNN）直接对语音的梅尔频谱图进行识别分类，得到了很好的性能。也有学者在语谱图上提取特征并用于检测，实验结果表明都有更好的泛化性与鲁棒性[17]。欺骗检测模型有多种，深度神经网络（deep neural network，DNN）是常见的检测模型之一，它很适合做非线性映射的搜索，在伪装语音检测中有很好的表现，但需要较多的数据进行训练[18]。高斯混合模型（Gaussian mixture model，GMM）作为一种概率统计模型，也常用于语音分类与识别领域。支持向量机（support vector machine，SVM）可以通过解决二次优化问题实现二分类，有着强大的实用性与泛化能力。

本文在语谱图的基础上，通过均匀局部二值模式（uniform local binary pattern，ULBP）分析并提取其纹理特征，然后与 CQCC 声学特征进行联合，提出了一种联合特征进行欺骗检测的方法。纹理特征作为描述语音信号的一种重要特征参数，可以反映出语音信号语谱图中的排列规则与重复出现的局部模式，可以描述语谱图的表面特性，并且具有良好的抗噪声性能[19]。考虑到联合特征的维数过高问题，引入主成分分析（principal component analysis，PCA）算法对特征矢量进行降维处理，很好地解决了联合特征维数过大的问题。同时考虑到联合特征与分类器的匹配问题，选取随机森林（random forest，RF）模型用于伪装语音与真实语音的分类。RF 能够根据各个特征矢量的重要性程度进行评估，更能应对特征数值差异大的联合特征矢量，在处理联合特征时有更高的匹配度，能得到更好的分类效果[20]。

1　联合特征提取

1.1　ULBP 算法

局部二值模式（local binary pattern，LBP）是通过比较目标图像相邻像素之间灰度值大小来描述目标的纹理特征[21-22]，首先选取一个灰度值作为比较标准，通常选取图像的中心点像素的灰度值 g_c，此时形成了一个半径为 R 的圆形范围，散落在这个圆上的各个像素点的灰度值大小表示为 $g_i(i=1,\cdots,N)$，圆心处所代表的像素点的 LBP 值表示为 $\mathrm{LBP}_{N,R}$，再将所有相邻像素点的灰度值 g_i 与 g_c 进行比较。如果满足 $g_i-g_c>0$，则将该点编码为 1；如果满足 $g_i-g_c<0$，则将该点编码为 0。规定左上角作为第一个数字并顺时针依次记录，于是可以得到一组由 0 和 1 组成的序列，再将这组二进制数转化为十进制，这样就完成了一个像素点的 LBP 值求解[23]，LBP 求解过程示例如图 1 所示。

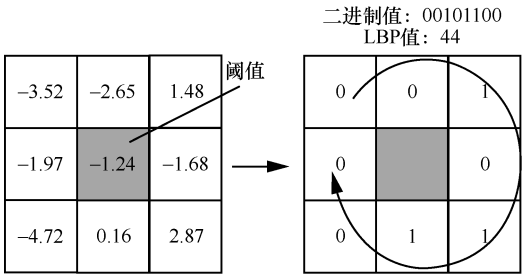


图 1　LBP 求解过程示例

用 LBP 算法对待测的每段语音信号的语谱图进行纹理分析，为提升分类效果，先对语谱图进行分块处理。采用 4×4 的格式将整个语谱图均分为 16 块，每块分别进行 LBP 特征提取。提取 LBP 时选取采样半径 $R=1$、采样点数 $N=8$。$\mathrm{LBP}_{8,1}$ 模式中，通过图 1 中 3×3 的 LBP 处理语谱图之后，每个像素点的值都位于 0～255，利用统计直方图的方法对每个像素点的值进行统计，每块语谱图就得到一个 1×256 维的特征矢量。

对于 3×3 的 LBP 来说，1×256 维的特征矢量里有很多维是空的，这样就增加了许多无用的信息和计算量。因此，采用 ULBP 进行压缩处理，大幅度减少了 LBP 纹理特征矢量维数。对 3×3 LBP 的每个 8 位二进制数，ULBP 将 0 与 1 之间跳变超过 2 次就归为同一类，这样处理后，LBP 直方图维度大幅度减少。3×3 的 LBP 原本 256 种情况就变成了 59 种情况，然后通过统计直方图统计这 59 种可能的数值，便可得到 1×59 维的特征矢量。对整个语谱图的 16 块区域进行同样的操作，就得到一个 16×59 维的特征矩阵，ULBP 纹理特征矢量提取过程如图 2 所示。

1.2　联合特征

考虑到声学特征与纹理特征在欺骗检测中各有优势，使用 CQCC 声学特征与 ULBP 纹理特征联合的方式用于欺骗检测。在欺骗攻击场景中，

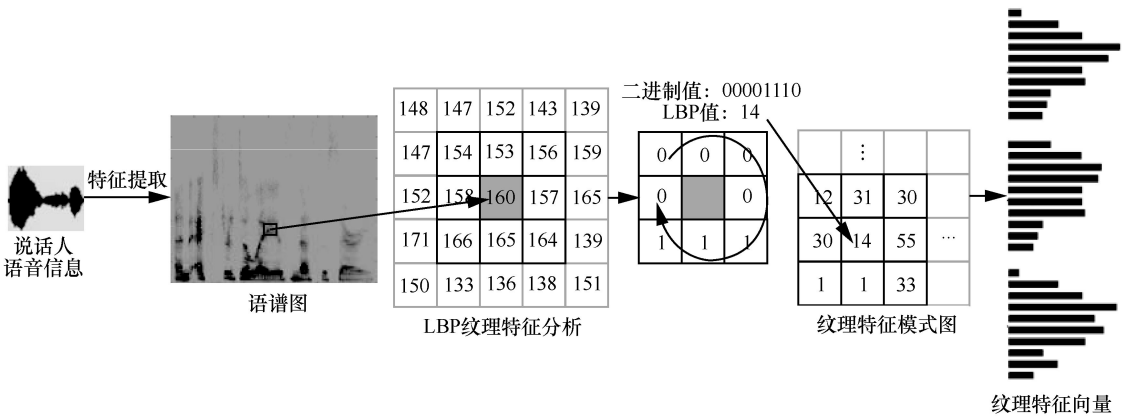


图 2　ULBP 纹理特征矢量提取过程

联合特征带有更多的语音信息，有更好的表现。考虑到特征参数维度过大，导致欺骗检测系统计算量大而影响系统的实时性，同时声学特征矢量与纹理特征矢量中存在信息冗余。因此，采用主成分分析算法分别对CQCC与ULBP特征进行处理[24]，达到降维的效果，然后再将降维后的特征进行拼接，从而生成联合特征，降维的具体流程如下。

步骤 1 输入由 M 个 D 维矢量构成的数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_i,\cdots,\boldsymbol{x}_M\}$，将 $\boldsymbol{X}$ 中每一个矢量 $\boldsymbol{x}_i$ 减去均值矢量，即 $\tilde{\boldsymbol{x}}_i=\boldsymbol{x}_i-\frac{1}{M}\sum_{j=1}^{M}\boldsymbol{x}_j$，得到去中心化的数据集 $\tilde{\boldsymbol{X}}=\{\tilde{\boldsymbol{x}}_1,\tilde{\boldsymbol{x}}_2,\cdots,\tilde{\boldsymbol{x}}_i,\cdots,\tilde{\boldsymbol{x}}_M\}$。

步骤 2 通过计算得到 $\tilde{\boldsymbol{X}}$ 的协方差矩阵 $\frac{1}{M}\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}}^{\mathrm{T}}$（T表示矩阵转置），为了进一步得到这个矩阵特征，再对这个协方差矩阵进行操作，对该矩阵按照特征值进行分解。分解后可以得到特征值，所得最大的 D' 个特征值会对应 D' 个特征矢量，将它们用 $\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_{D'}$ 表示，并将它们按顺序排列，如此，就出现了一个特征矢量矩阵 $\boldsymbol{W}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_{D'}\}$。

步骤 3 降维处理数据集 $\tilde{\boldsymbol{X}}$ 中的每个样本矢量 $\tilde{\boldsymbol{x}}_i$，即 $\boldsymbol{z}_i=\boldsymbol{W}^{\mathrm{T}}\tilde{\boldsymbol{x}}_i,\ i=1,2,\cdots,M$，可获得降维后的 D' 维数据集 $\boldsymbol{Z}=\{\boldsymbol{z}_1,\boldsymbol{z}_2,\cdots,\boldsymbol{z}_M\}$。

在实验中，首先计算一段 L 帧的语音信号 $x(n)$ 的语谱图，使用前述的ULBP算法对语谱图进行分析并得到 16×59 维的纹理特征矩阵 $\boldsymbol{ULBP}$。同时将每帧语音信号的 $\boldsymbol{CQCC}$ 特征矢量提取出来，该矢量的维数取60，这样对于一段 L 帧的语音信号来说，经过处理后得到的 $\boldsymbol{CQCC}$ 特征矢量的维数为 $60\times L$ 维。再使用PCA分别对矩阵 $\boldsymbol{ULBP}$ 和 $\boldsymbol{CQCC}$ 进行降维处理，并取 $N'=1$，这样就分别得到 16×1 维的 $\boldsymbol{ULBP}'$ 和 60×1 维的 $\boldsymbol{CQCC}'$。最后将 60×1 维的 $\boldsymbol{CQCC}'$ 和 16×1 维的 $\boldsymbol{ULBP}'$ 首尾拼接，这样可获得 76×1 维的联合特征矢量。联合特征提取流程如图3所示。

2 伪装语音检测

2.1 随机森林分类算法

随机森林采用集成学习的思想，将多个弱学习器组成一个强学习器。随机森林通过随机选取数据样本来形成多个决策树从而形成森林结构，每一棵树都会得出一个分类结果。原则上，随机森林算法在进行分类时，使用票数占少的需要遵从票数占多的规则进行投票分配，整个森林系统的分类结果应以票数最高的分类结果为准。RF的训练流程如下。

步骤 1 从语音库中随机选取真、伪语音，假设共有 S 个语音段，从每段语音提取一个76维的联合特征矢量，则共有 S 个矢量样本，数据集由这些样本组成。然后从数据集中进行有放回地随机抽取，得到用于训练决策树的训练集样本，训练集样本为 $S'(S'\leqslant S)$ 个矢量样本。在该过程中可能存在被重复提取与没有被提取的样本。

步骤 2 每个样本包含 T 个属性，即联合特征矢量的维数。当决策树开始分裂时，从 T 个属性中随机选择 T' 个属性，其中，T' 的数量应远小于 T。使用基尼指数作为分裂策略选取这 T' 个属性的分裂属性，即分裂规则就是在待选属性中找出信息增益高于平均值的属性，然后找出信息增

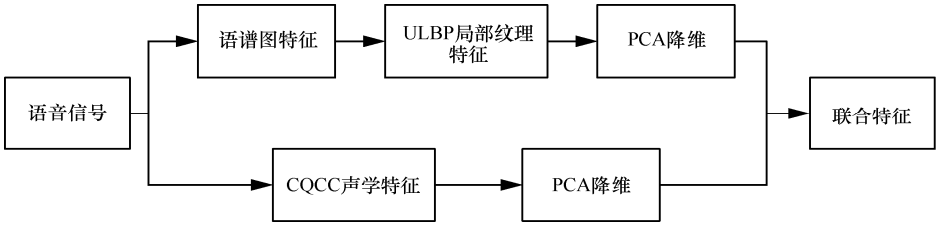


图3 联合特征提取流程

益率最高的属性。

步骤 3　对决策树的节点进行分裂，直到所有可能的值都已被使用，停止生长，从而能最大限度生长增加决策树，避免剪枝。这样便可得到一棵决策树，重复 F 次这个过程，生长增加决策树的数量，形成随机森林。

2.2　伪装语音检测方法

首先，提取出语音信号的语谱图，并确保语谱图纹理清晰，将语谱图转换成灰度图，通过统计直方图得到 ULBP 纹理特征。同时根据特征联合的方式，将 ULBP 纹理特征与 CQCC 声学特征进行联合，即从两个方面分析语音信号。将一段任何时长的语音信号经过整个联合特征提取流程后，转换成一个 CQCC-ULBP 联合特征矢量，并用于训练随机森林分类模型。在对随机森林分类模型完成训练后，得到对应的最佳决策树参数，再对待检测的语音进行测试，然后根据每棵树所给出的投票情况给出判决结果。使用随机森林用于分类时，每棵树的权重相同且互不相关，依据投票的情况给出最后结论。选取随机森林分类算法来训练联合特征实现语音信号的特征分类时，使用随机森林对提取的真实语音与欺骗语音数据集所得到的联合特征向量进行训练，再对待认证语音集进行测试。因此，便可以得到一个基于联合特征与随机森林的伪装语音检测系统，基于联合特征与随机森林的伪装语音检测系统流程如图 4 所示。

3　实验与结果

3.1　数据集

实验使用的语音库是 Interspeech 在 2019 年举办的 ASVspoof 挑战赛中所使用的逻辑访问（logical access，LA）场景数据集。ASVspoof2019LA 数据库基于语音克隆工具包（voice cloning tool kit，VCTK）语料库提取，是一个在消声暗室中以 16 kHz 的采样率录制的多人英语语音数据库。ASVspoof2019LA 语音库中的伪装语音由语音转换和语音合成两种伪装方式生成，伪装方式 A01-A19 的具体信息详见文献[25]。同时选取 ASVspoof2015 语音库进一步对实验结果进行验证。ASVspoof2015 语音库中的欺骗攻击语音由语音转换和语音合成两种伪装方式生成，伪装语音 S1-S10 的生成信息详见文献[26]。

3.2　性能评价方法

欺骗检测系统的性能评价方法通常有两种，一种是等错误率（equal error rate，EER），另一种是串联检测代价函数。EER 是权衡错误拒绝率（false rejection rate，FRR）与错误接受率（false acceptance rate，FAR）的指标，EER、FRR 与 FAR 的数值关系表示为 $P_{\mathrm{EER}} = P_{\mathrm{FRR}} = P_{\mathrm{FAR}}$。而 EER 忽略了对 P_{FAR} 或者 P_{FRR} 的需求不同时的场景，由于欺骗检测系统对 P_{FAR} 的容忍度要小于 P_{FRR}，在对安全性要求高的场景中，P_{FAR} 的数值应尽可能降

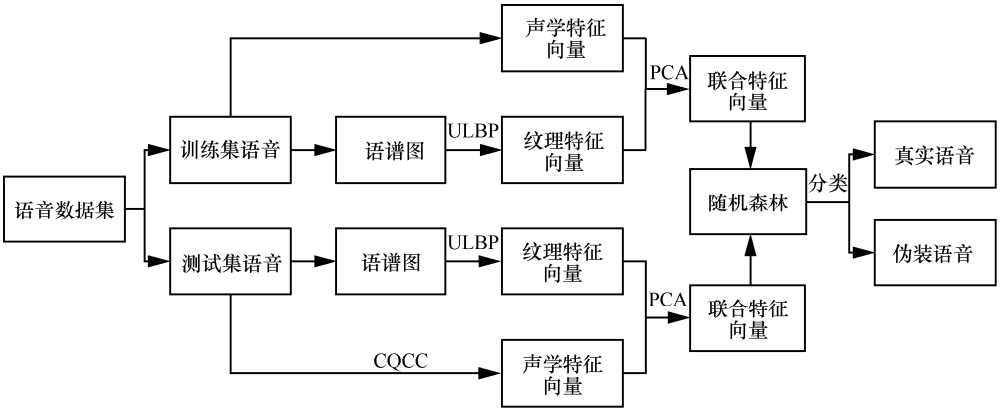


图 4　基于联合特征与随机森林的伪装语音检测系统流程

低。同时在反欺骗攻击的任务中，当反欺骗攻击系统与 ASV 系统级联时，EER 往往不能作为可靠的性能评价标准，若单纯考虑二分类错误的 EER 是不足以衡量整体反欺骗攻击系统与 ASV 系统的，因为真实语音被提前判错，则 ASV 系统将可能失去一段真实目标语音；同理，若欺骗语音被反欺骗攻击系统判为真实，则 ASV 系统将面临欺骗语音的攻击，同样会导致 ASV 系统 EER 下降。因此，Cheng 等[27]提出串联检测代价函数（tandem detection cost function，t-DCF），其综合考虑二分类决策和说话人识别，同时考虑反欺骗攻击系统与 ASV 串联顺序或两者并联的关系。反欺骗攻击的任务本质是为说话人识别服务的，假如欺骗语音成功欺骗了反欺骗攻击系统，就加重了 ASV 系统的负担；假如说话人识别判断了一句欺骗语音是目标的话，那就要依赖反欺骗系统能成功分类出来。t-DCF 也被 ASVspoof20l9 挑战赛选为主要的评价标准，表示为：

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

其中，先验概率 π_{tar}=0.940 5，ASV 代价值 C_{FRR}=1、C_{FAR}=10。

3.3 伪装语音检测系统性能测试

选取 ASVspoof2019LA 语音库中的语音样本用于实验，随机选取了 5 850 条语音用于系统性能测试，其中有 5 000 条语音作为训练集，850 条语音作为测试集。

提取 60 维的 CQCC，最大频率为 8 000 Hz，采样频率 f_s=16 kHz，每个八度音阶包含的频带数为 96，重抽样采样阶段的采样间隔为 16，CQCC 特征参数包括 30 维 CQCC 系数、30 维一阶差分系数。同时提取训练集中每一个语音信号的 36 位 MFCC 特征作为对照，其中包括 12 维 MFCC 系数、12 维一阶差分系数、2 维二阶差分系数。使用 PCA 降维方法处理，任何一段语音就可以表示为 36×1 维的 MFCC 特征矢量，同样也可以表示为 60×1 维的 CQCC 特征矢量，再将这两种特征矢量集分别用于训练 SVM 系统与 RF 系统。用待测语音对训练完成的分类模型进行测试，并对实验结果进行比较分析，性能评价标准使用 t-DCF 参数，应对不同欺骗攻击时 MFCC 与 CQCC 特征在 SVM 与 RF 系统中的 t-DCF 值见表 1。

表 1 应对不同欺骗攻击时 MFCC 与 CQCC 特征在 SVM 与 RF 系统中的 t-DCF 值

伪装类型	SVM		RF	
	MFCC	CQCC	MFCC	CQCC
A01	0.599	0.017	0.438	0.011
A02	0.325	0.014	0.22	0.004
A03	0.415	0.248	0.22	0.112
A04	0.414	0.153	0.326	0.018
A05	0.272	0.097	0.242	0.094
A06	0.234	0.126	0.171	0.112
A07	0.503	0.007	0.343	0.007
A08	0.359	0.024	0.253	0.033
A09	0.457	0.063	0.364	0.028
A10	0.368	0.003	0.208	0.022
A11	0.466	0.024	0.244	0.015
A12	0.428	0.093	0.375	0.015
A13	0.282	0.007	0.255	0.003
A14	0.382	0.014	0.339	0.052
A15	0.298	0.07	0.199	0.06
A16	0.417	0.037	0.244	0.018
A17	0.332	0.09	0.212	0.123
A18	0.28	0.248	0.202	0.211
A19	0.265	0.107	0.179	0.092
平均值	0.373	0.076	0.265	0.054

由表 1 中的 t-DCF 值可以看出，在伪装语音检测中，MFCC 的检测结果较差。MFCC 虽然能很好地反映人耳的听觉机理，在说话人验证系统中可以取得较好的性能，然而在伪装语音检测时并不能很好地辨别出真实语音与欺骗语音的区别，由于欺骗语音与真实语音的语音内容十分相似，难以区分，欺骗检测性能较差。相比而言，CQCC 是针对伪装语音检测所使用的声学特征，避免了时频分辨率均匀的缺点，更能在伪装语音检测中代表语音特征，相比 MFCC 有更好的检测效果。同时，在对语音 MFCC 与 CQCC 两种特征进行分类时，SVM 与 RF 的性能表现差异不大，

t-DCF 值相差比较相近，RF 略微要好一些。

实验提取语谱图纹理特征，使用 ULBP 算法提取训练集中语音信号的 ULBP 特征矢量，使用 PCA 对 ULBP 特征、CQCC 特征和 MFCC 特征进行降维处理并得到联合特征，再将联合特征矢量分别用于训练 SVM 与 RF 系统，将所有训练的 SVM 系统与 RF 系统在测试集中进行测试，在应对不同欺骗攻击时两种联合特征在 SVM 与 RF 系统中的 t-DCF 值如图 5 所示。

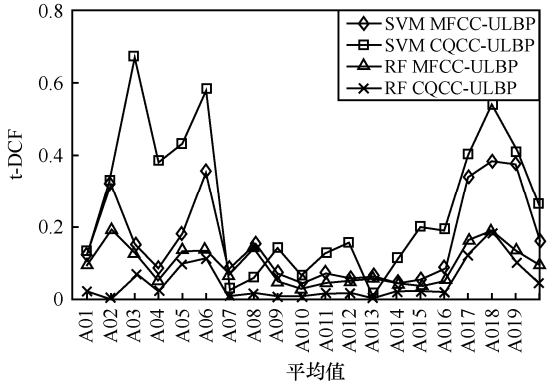


图 5　在应对不同欺骗攻击时两种联合特征在 SVM 与 RF 系统中的 t-DCF 值

通过对比图 5 与表 1 中的实验数据发现，基于 MFCC-ULBP 特征矢量的检测系统明显优于基于 MFCC 特征矢量的检测系统。同样地，基于 CQCC-ULBP 特征矢量的检测系统明显优于基于 CQCC 特征矢量的检测系统。因为联合特征中包含语音信号中所携带的能量与纹理特征，比传统声学特征更具有代表性。同时也发现，采用 CQCC-ULBP 联合特征的伪装语音检测方法具有最佳的检测效果。在分类器方面，使用 SVM 与 RF 模型分别对 MFCC-ULBP 与 CQCC-ULBP 两种联合特征训练时，通过 RF 模型训练特征的检测效果明显优于 SVM。使用 RF 模型进行伪装语音检测时，采用的联合特征用于伪装语音检测的系统性能整体上都提高了检测效果。但在使用 SVM 对联合特征进行伪装语音检测时，系统检测性能在部分伪装种类中会有一定程度的下降。在处理普通的二分类问题时，SVM 具有优秀的性能与泛化能力。但在伪装语音检测实验场景中，真实语音样本数量应普遍少于欺骗语音样本数量，并且由于真实语音与欺骗语音样本同等重要，故不宜在实验前对数据进行预处理，而数据预处理可有效地提升 SVM 在二分类数据上的泛化能力。但 RF 在进行训练和分类时都不需要进行数据预处理。

同时，实验也选取 ASVspoof2015 语音库中的语音样本用于实验来进一步验证实验中的结论，仍然随机选取 5 850 条语音用于系统性能测试，其中有 5 000 条语音用于训练作为训练集，850 条语音用于测试作为测试集。将该数据集中语音样本在本文所提出的伪装语音检测方法进行验证，使用联合特征的提取方式提取该语音数据集中语音的特征参数，将得到的真伪语音特征参数在 RF 与 SVM 中进行训练，所有训练的 SVM 系统与 RF 系统在测试集中进行测试，将各类特征矢量在各个伪装语音检测系统上进行测试，应对不同欺骗攻击时各类特征在 SVM 与 RF 系统中的 t-DCF 值如图 6 所示。

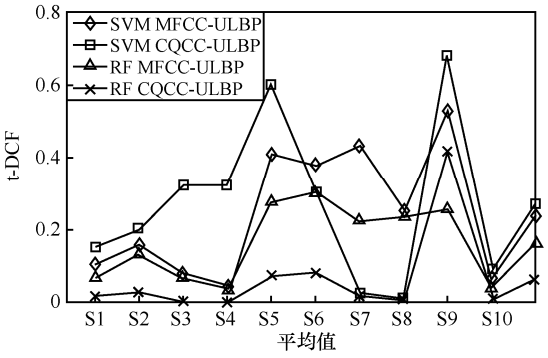


图 6　应对不同欺骗攻击时各类特征在 SVM 与 RF 系统中的 t-DCF 值

从图 6 中的实验结果可以看出，在 ASVspoof 2015 数据集中，基于 CQCC-ULBP 的联合特征与随机森林的伪装语音检测模型在整体上实现了最佳的分类性能。在使用声学特征对 S2 类型欺骗攻击进行分类时，t-DCF 参数的值普遍很大，因为 S2 类型是改变声学特征的生成的伪装语音，更容

易破坏使用声学特征识别的系统，而联合特征弥补了这一点，在应对 S2 类型欺骗攻击时检测效果较好。在应对 S3、S4 类型语音合成欺骗攻击时，各系统都有不错的表现，并且联合特征得到了最佳的效果。但在应对 S9 类型欺骗攻击时，对联合特征的检测性能造成了一定影响，t-DCF 参数的值明显增加。这是由于 S9 类型的语音转换攻击，几乎不改变语谱图的声纹特征，导致纹理特征识别效果不好。纹理特征表现不佳，影响了联合特征的整体性能。同时从图 5 可以看出，相同条件下采用联合特征与 RF 模型进行伪装语音检测时的性能要优于采用联合特征与 SVM 模型进行检测的效果。

同时，实验还对随机森林与支持向量机的分类效率进行了测试，特征参数采用的是 CQCC-ULBP 联合特征参数，SVM 与 RF 平均执行时间见表 2，RF 相较于 SVM 平均效率有很大的提升。在处理大规模的训练样本数据时，SVM 算法的处理效率不如随机森林算法，主要是由于 SVM 在确定求解支持向量时使用的方法是二次规划方法，这种求解方式在求解过程中会涉及 m 阶矩阵的运算（m 表示样本个数），假如 m 的值过大，会导致矩阵的存储过程与计算处理过程占用大量的机器内存，从而使运算的持续时间过长，致使 SVM 检测系统的效率降低。

表 2　SVM 与 RF 平均执行时间

检测系统	平均用时/s
SVM	1 007.142
RF	0.106 7

4　结束语

为了改善基于传统声学特征参数的伪装语音检测系统的性能，提出了一种利用 ULBP 算法在语谱图中提取纹理特征并与 CQCC 声学特征进行联合的伪装语音检测方法。在该方法中，分别使用 PCA 将一段语音的 ULBP 特征参数矩阵和 CQCC 特征矢量序列进行压缩，然后进行联合，成为一个矢量。同时，将该联合矢量所构成的语音特征参数集训练 RF 分类器，就可以得到伪装语音检测系统。实验结果表明，联合特征可以更加全面地描述语音信号的特征信息，便于分类检测，本文采用随机森林作为分类器与 ULBP-CQCC 联合特征参数进行匹配具有最优的检测性能。

参考文献：

[1] GOMEZ-ALANIS A, GONZALEZ-LOPEZ J A, PEINADO A M. A kernel density estimation based loss function and its application to ASV-spoofing detection[J]. IEEE Access, 2020, 8: 108530-108543.

[2] 彤娅峰, 陈晨, 陈德运, 等. 基于贝叶斯主成分分析的 i-vector 说话人确认方法[J]. 电子学报, 2021, 49(11): 2186-2194.
RONG Y F, CHEN C, CHEN D Y, et al. Bayesian principal component analysis for I-vector speaker verification[J]. Acta Electronica Sinica, 2021, 49(11): 2186-2194.

[3] LI N, MAK M W, CHIEN J T. Deep neural network driven mixture of PLDA for robust i-vector speaker verification[C]//Proceedings of 2016 IEEE Spoken Language Technology Workshop. Piscataway: IEEE Press, 2016: 186-191.

[4] ALEGRE F, JANICKI A, EVANS N. re-assessing the threat of replay spoofing attacks against automatic speaker verification[C]//Proceedings of 2014 International Conference of the Biometrics Special Interest Group (BIOSIG). Piscataway: IEEE Press, 2014: 1-6.

[5] 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.

[6] NAUTSCH A, WANG X, EVANS N, et al. ASVspoof 2019: spoofing countermeasures for the detection of synthesized, converted and replayed speech[J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2021, 3(2): 252-265.

[7] 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.
REN Y Z, LIU C Y, LIU W Y, et al. A survey on speech forgery and detection[J]. Journal of Signal Processing, 2021, 37(12): 2412-2439.

[8] YU H, TAN Z H, MA Z Y, 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.

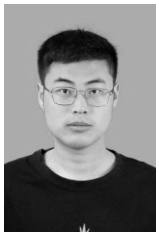
[9] PAUL D, PAL M, SAHA G. Novel speech features for improved detection of spoofing attacks[C]//Proceedings of 2015 Annual IEEE India Conference. Piscataway: IEEE Press, 2015: 1-6.

[10] HIDAYAT R, BEJO A, SUMARYONO S, et al. Denoising speech for MFCC feature extraction using wavelet transformation in speech recognition system[C]//Proceedings of 2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE). Piscataway: IEEE Press, 2018: 280-284.

[11] ÖZSÖNMEZ D B, ACARMAN T, PARLAK İ B. Optimal

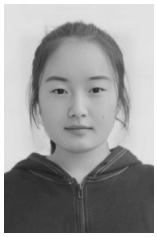
classifier selection in Turkish speech emotion detection[C]// Proceedings of 2021 29th Signal Processing and Communications Applications Conference (SIU). Piscataway: IEEE Press, 2021: 1-4.
[12] PENG X, LU C Y, YI Z, et al. Connections between nuclear-norm and frobenius-norm-based representations[J]. IEEE Transactions on Neural Networks and Learning Systems, 2018, 29(1): 218-224.
[13] TODISCO M, DELGADO H, EVANS N. Constant Q cepstral coefficients: a spoofing countermeasure for automatic speaker verification[J]. Computer Speech & Language, 2017 (45): 516-535.
[14] SARANYA S, BHARATHI B, KAVITHA S. An approach to detect replay attack in automatic speaker verification system[C]//Proceedings of 2018 International Conference on Computer, Communication, and Signal Processing (ICCCSP). Piscataway: IEEE Press, 2018: 1-5.
[15] YE Y C, LAO L J, YAN D Q, et al. Detection of replay attack based on normalized constant Q cepstral feature[C]//Proceedings of 2019 IEEE 4th International Conference on Cloud Computing and Big Data Analysis. Piscataway: IEEE Press, 2019: 407-411.
[16] MASSOUDI M, VERMA S, JAIN R. Urban sound classification using CNN[C]//Proceedings of 2021 6th International Conference on Inventive Computation Technologies (ICICT). Piscataway: IEEE Press, 2021: 583-589.
[17] LI P H, LI Y Y, LUO D C, et al. Speaker identification using FrFT-based spectrogram and RBF neural network[C]//Proceedings of 2015 34th Chinese Control Conference (CCC). Piscataway: IEEE Press, 2015: 3674-3679.
[18] WANG J, HAN Z Y. Research on speech emotion recognition technology based on deep and shallow neural network[C]//Proceedings of 2019 Chinese Control Conference (CCC). Piscataway: IEEE Press, 2019: 3555-3558.
[19] 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.
[20] K L, DABHADE S B, RODE Y S, et al. Identification of breast cancer from thermal imaging using SVM and random forest method[C]//Proceedings of 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI). Piscataway: IEEE Press, 2021: 1346-1349.
[21] TAO Y, HE Y Z. Face recognition based on LBP algorithm[C]//Proceedings of 2020 International Conference on Computer Network, Electronic and Automation (ICCNEA). Piscataway: IEEE Press, 2020: 21-25.
[22] OJALA T, PIETIKAINEN M, MAENPAA T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24(7): 971-987.
[23] FAUDZI S A A M, YAHYA N. Evaluation of LBP-based face recognition techniques[C]//Proceedings of 2014 5th International Conference on Intelligent and Advanced Systems (ICIAS). Piscataway: IEEE Press, 2014: 1-6.
[24] WANG L L. Research on distributed parallel dimensionality reduction algorithm based on PCA algorithm[C]//Proceedings of 2019 IEEE 3rd Information Technology, Networking, Electronic and Automation Control Conference. Piscataway: IEEE Press, 2019: 1363-1367.
[25] 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.
[26] WU Z Z, KINNUNEN T, EVANS N, et al. ASVspoof 2015: the first automatic speaker verification spoofing and countermeasures challenge[C]//Proceedings of Interspeech 2015. ISCA: ISCA, 2015.
[27] CHENG X L, XU M X, ZHENG T F. Replay detection using CQT-based modified group delay feature and ResNeWt network in ASVspoof 2019[C]//Proceedings of 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). Piscataway: IEEE Press, 2019: 540-545.

[作者简介]



于佳祺（1997− ），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伪装检测、特征提取与分析。

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，主要研究方向为语音转换、伪装语音检测、声纹识别等。



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及检测。

游林（1966− ），男，博士，杭州电子科技大学网络空间安全学院教授、硕士生导师，主要研究方向为生物信息处理、信息安全、密码学等。

汪云路（1980− ），女，博士，杭州电子科技大学网络空间安全学院讲师，主要研究方向为音频信息处理、信息隐藏。

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