文章编号：2095-2163(2021)11-0054-06 中图分类号：TN912.34 ；TP183 文献标志码：A

基于特征压缩和残差网络的语音重放检测

陈 露[1]，周 欣[1,2]，陈洪刚[1]，何小海[1]，王正勇[1]，卿粼波[1]
（1 四川大学 电子信息学院，成都 610065；2 中国信息安全测评中心，北京 100085）

摘 要：目前的语音重放攻击检测系统中，绝大部分性能良好的系统采用的特征和网络模型的数据量都很大，训练速度慢、对设备要求高。因此本文提出了一种基于 CQT（Constant Q Transform）变换的时间帧压缩方法，以减小特征尺寸和网络模型参数量，从而加快训练速度、降低设备要求。首先，将语音信号的 CQT 谱在时间帧维度上压缩，得到一维特征，成百倍地减少特征数据量；其次，对应设计一维小型残差网络模型，以辅助进一步减少数据量；最后，在 ASVspoof2019 挑战赛的 PA 数据集上训练并测试网络模型性能。实验结果表明，本文的特征提取算法和网络模型，相比挑战赛的基线系统以及其他特征-模型的性能有明显提升，t-DCF 为 0.105 1，EER 为 3.74%，并且训练速度快、设备要求低。

关键词：CQT 变换；语音重放攻击检测；特征提取；时间帧压缩；小型残差网络

Speech replay detection based on feature compression and residual network

CHEN Lu[1], ZHOU Xin[1,2], CHEN Honggang[1], HE Xiaohai[1], WANG Zhengyong[1], QING Linbo[1]
(1 College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China;
2 China Information Technology Security Evaluation Center, Beijing 100085, China)

【Abstract】 In the current speech replay attack detection systems, most of the systems with good performance adopt a large amount of features and network model data, so the training speed is slow and the equipment requirements are high. Therefore, a CQT (Constant Q Transform) transform time frame compression method is proposed in this paper to reduce feature size and network model parameters, speed up training, and reduce equipment requirements. Firstly, the CQT spectrum of speech signal is compressed in the dimension of time frame to obtain one-dimensional features, and the amount of feature data is reduced hundreds of times, which is the main method to reduce the amount of data in this paper. Then, a one-dimensional small residual network model is designed to further reduce the amount of data. Finally, the network model was trained and tested on the PA data set of ASVspoof2019 Challenge. The experimental results showed that the feature extraction algorithm and network model presented in this paper had significantly improved performance compared with the baseline system and other featured-model models in the challenge competition, with t-DCF of 0.1051 and EER of 3.74%, as well as fast training speed and low equipment requirements.

【Key words】 CQT transformation; speech replay attack detection; feature extraction; time frame compression; small residual network

0 引 言

语音重放攻击检测是一种判别真人发声和录音重放的生物识别技术。用于重放的语音样本是录制的真人发声，由于不需要专业的语音处理知识，获取容易且成本低，比如用手机录制、从音频中截取等。语音重放攻击给说话人识别与认证带来了严重威胁，因此研究语音重放攻击检测技术具有重要的现实意义。

说话人验证技术的广泛应用，引发人们对语音重放攻击检测的重视。重放攻击检测性能主要取决于特征提取和网络模型，国内外学者做了许多尝试和研究，如 Davis 提出的梅尔频率倒谱系数（Mel-scale Frequency Cepstrum Coefficients，MFCC）来提取按频率成不同分布的语音信号的能量，被广泛用于语音处理[1]；但是 MFCC 的傅里叶变换在低频段的低频率分辨率和高频段的低时间分辨率损失了一定信息，所以 Todisco 等人提出了常数 Q 倒谱系数（Constant Q Cepstral Coefficients，CQCC）[2]，被用作第三届自动说话人验证欺骗与对策挑战赛（Automatic Speaker Verification Spoofing and Countermeasures Challenge，ASVspoof2019 挑战赛）基线系统特征；然而这些特征都是语音信号的频谱幅度，所以清华-得意团队采用了以往被忽视的相位特征，提出了基于常数 Q 变换（Constant Q Transform，CQT）的群时延特征，取得了

基金项目：四川省科技计划项目（2018HH0143）；四川省教育厅项目（18ZB0355）。

作者简介：陈 露（1997-），女，硕士研究生，主要研究方向：语音识别、深度学习；周 欣（1985-），男，博士研究生，主要研究方向：数据挖掘、自然语言处理；陈洪刚（1991-），男，博士，副研究员、硕士生导师，主要研究方向：图像/视频处理；何小海（1964-），男，博士，教授，主要研究方向：图像处理、网络通信等；王正勇（1969-），女，博士，副教授，主要研究方向：图像处理、模式识别等；卿粼波（1982-），男，博士，副教授，主要研究方向：多媒体通信与信息系统、嵌入式系统等。

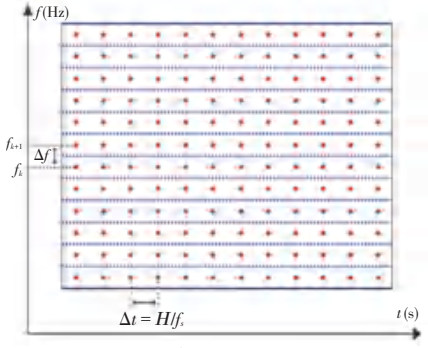
收稿日期：2021-07-30

ASVspoof2019 挑战赛第一[3]；在网络模型方面，Lai 等人引入注意力机制，取得了不错的效果[4]。但是大多数性能良好的系统的实现，对实验设备提出了极高的要求，模型训练的速度很慢。对此，本文尝试以减小特征为主要手段减小数据量、减小网络模型为辅助措施，在保证重放攻击检测高性能的同时，加快训练速度、降低实验设备要求。

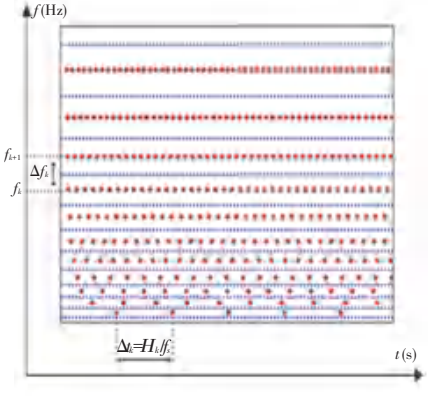
ASVspoof2019 挑战赛是由多个世界领先的研究机构组织发起的，是目前针对虚假语音鉴别规模最大且最全面的挑战赛[5]，其中的 PA（Physical access）数据集专门针对重放攻击检测。因此本文采用 ASVspoof2019 PA 数据集进行实验。

1　特征提取及压缩

人耳对声音的频率是指数敏感的，短时傅里叶变换（Short-Time Fourier Transform，STFT）的线性频谱不能较好地拟合音频，CQT 变换的时频特性同样是呈指数分布的，可以一一对应声音的频率点；另一方面，STFT 的时间分辨率 Δt 和频率分辨率 Δf 是固定的，而 CQT 时频分辨率可变，在整个频段内能同时满足高时间分辨率 Δt_k 和高频率分辨率 Δf_k，即低频段的频率分辨率更高、高频段的时间分辨率更高，如图 1 所示。



（a）STFT



（b）CQT

图 1　STFT 和 CQT 时频分辨率对比图

Fig. 1　The comparison of STFT and CQT time - frequency resolution

其中 H、H_k 表示短时窗口长度；f_k 表示滤波器中心频率；对应图 1 中的红点，k 是频段序号；f_s 代表采样频率。因此，本文提取的特征是基于 CQT 变换的。

离散时间信号 $x(n)$ 的 CQT 变换如式（1）：

$$X^{CQ}(k,n)=\sum_{j=n-\lfloor N_k/2\rfloor}^{n+\lfloor N_k/2\rfloor}x_j a_k^*(j-n+N_k/2) \tag{1}$$

其中，$k=1,2,\cdots,K$ 是频带号；$\lfloor * \rfloor$ 表示向下取整运算；$*$ 代表任意表达式。$a_k(n)$ 定义为式（2）：

$$a_k(n)=\frac{1}{C}\left(\frac{n}{N_k}\right)\exp\left(i\left(2\pi n\frac{f_k}{f_s}+\Phi_k\right)\right) \tag{2}$$

其中，f_s 是采样频率；Φ_k 是频带 k 的相位偏移；$a_k^*(n)$ 是其共轭复数。尺度因子 C 为式（3）：

$$C=\sum_{l=\lfloor N_k/2\rfloor}^{\lfloor N_k/2\rfloor}w\left(\frac{l+N_k/2}{N_k}\right) \tag{3}$$

频带 k 的中心频率 f_k 直接对应音符，可以根据十二平均律，由公式（4）计算：

$$f_k=f_1 2^{\frac{k-1}{B}} \tag{4}$$

其中，f_1 是最低频带的中心频率。待分析频段 $f_{\min}\sim f_{\max}$ 分割为包含 $\lfloor \log_2(f_{\max}/f_{\min})\rfloor$ 个指数分布的八音度，每个八音度再分割为 B 个频带，B 决定了时频分辨率的权衡。

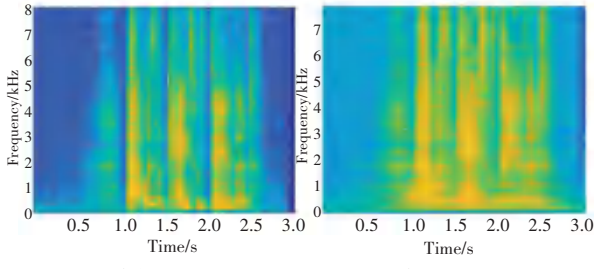
Q 因子定义如式（5）：

$$Q=\frac{f_k}{\delta_k}=\frac{f_k}{f_{k+1}-f_k}=(2^{1/B}-1)^{-1} \tag{5}$$

可见，对于频带 k 的所有频率，Q 均为常数，并且公式（1）和公式（2）中的动态窗口长度 N_k 可以通过公式（6）计算：

$$N_k=\frac{f_s}{f_k}Q \tag{6}$$

STFT 和 CQT 频谱图如图 2 所示。在 CQT 谱的低频段的分辨率也很高，可以清楚地看到不同基频的变化，而 STFT 谱的低频区分性则较差。



（a）STFT　　（b）CQT

图 2　STFT 和 CQT 频谱图

Fig. 2　The spectrum of STFT and CQT

根据十二平均律,一般只用 27.5~4 185.6 Hz 范围内的音[6],而一些研究学者证明高频段(4~8 kHz)含有检测录音攻击的有效信息[7-8]。再结合 CQT 变换的计算,本文选取的频段(15.625~8 000 Hz)用于特征提取是合理的,选取的待分析频段包含 9 个八音度。

对一段语音信号进行 CQT 变换得到一个二维 (M,N) 频谱。其中,M 与最低频率 $f_{\min}$、最高频率 $f_{\max}$ 及每个八度音的频带个数 B 有关,式(7):

$$M = B\lfloor \log_2(f_{\max}/f_{\min}) \rfloor \tag{7}$$

N 则由原始音频长度 H 和帧移 hop_length 决定,式(8):

$$N = \lceil H/hop\_length \rceil \tag{8}$$

其中,$\lceil * \rceil$ 表示向上取整,$*$ 代表任意表达式。

本文选取的频带个数 B 为 96,因此将在 864(96×9=864)个频率上对语音信号进行 CQT 分解,每条语音得到一个 864×N(在 ASVsppof2019 PA 数据集中 N 的平均值约为 300)二维 CQT 谱,这个数据量是很大的。由此带来了对录音攻击检测网络训练设备的极高要求,而且输入特征数据量过大,网络训练速度很慢。

由公式(7)可知,M 由频段范围和时频分辨率决定,在该维度上的数据减小是以丢失音频信息为代价的,因此,本文提出了 CQT 谱的时间帧压缩(以下均简写为 CQT_Z),即在上述 N 所代表的维度上减少数据量。

对于时间平稳信号的处理,可以直接对整个信号进行频谱变换(此时 $N=1$),而不用担心变换时信号的频率轮廓会随时间的推移而丢失,但实际的音频信号是非平稳的,所以在 CQT 变换中对音频信号分帧是必要的,也因此带来了时间帧数 N 的增加,且 N 与帧移成反比。既然 N 只是在时间上的分片处理,那么对每个音频的 CQT 谱在时间帧的维度上求和压缩,并不会影响整个音频的频率成分及含量,因此这种压缩处理是合理的。通常为了便于对各条数据进行批量处理,会在对原始语音求频谱等变换之前,将每条语音进行填充或截断成相同的长度,使得最终每条语音得到的时间帧数 N 一致,至少要保证每个 $batch$ 中的数据尺寸一致。但是由此也会带来问题,若将所有语音填充到最长语音的长度,越短的语音加入越多或重复或空白无用的数据;若把每条语音填充、截断成适当长度,截断会损失一些语音信息。而本文的算法则不需要将语音处理成相同长度,可以避免填充的重复或空白数据带来的无用数据量的增加,以及截断带来的语音信息的损失。综上可预测压缩后的 CQT 谱对本文的录音攻击检测是有效的。

CQT 压缩处理如式(9):

$$CQT_Z = \left(\sum_{n=1}^{N} \log_2(|X^{CQ}(f_k,n)| + 10^{-20}) - mean \right) / SD \tag{9}$$

其中,$mean$、SD 分别代表 $\sum_{n=1}^{N} \log_2(|X^{CQ}(f_k,n)| + 10^{-20})$ 的均值和标准差,$n = 1,2,\cdots,N$ 代表时间帧。

CQT 谱的数据量经压缩后大大减少,在此随机抽取 PA 数据集里的音频,以本文实验设置得到的特征压缩前后的数据量大小,见表 1。

表 1 音频特征压缩前后数据量对比

Tab. 1 The comparison of data volume before and after audio feature compression

音频特征文件名	压缩前 CQT 特征文件大小	压缩后 CQT_Z 特征文件大小
PA_D_0000001.npy	1 007 KB	5 KB
PA_D_0019859.npy	852 KB	5 KB
PA_T_0039651.npy	578 KB	5 KB
PA_E_0137329.npy	615 KB	5 KB

未压缩的 CQT 特征文件平均约为 600 KB,是本文压缩特征 CQT_Z 的百倍。在本文的实验设置下,压缩特征 CQT_Z 的 $batchsize$ 为 32,未压缩特征只能设为个位数,而减小的 $batchsize$ 会影响模型训练效率,除非使用更大的显存和内存设备。

2 面向重放攻击检测的残差网络模型

Hornik 在 1989 年就证明了只要一个包含足够多神经元的隐层,多层前馈神经网络就能以任意精度逼近任意复杂的连续函数。然而,实践证明对于更复杂、更不易辨别的训练对象和训练目标来说,浅层的神经网络学习到的低层次简单特征远远不够,可以采用深度神经网络,以期学习到更复杂的特征。但是盲目地加深网络,网络饱和之后会发生损失不降反升的退化现象,此时,残差网络应运而生。

近年来,愈加高质量的录音设备给重放攻击检测带来了更大的挑战,因此需要更深层次的神经网络学习更有效的复杂特征。对于本文的目的在于保证高水平的重放攻击检测效果的同时,大幅提高网络模型训练速度、降低实验设备需求,因此选择 50 层的残差网络作为基础网络训练模型。

对应本文提出的一维特征 CQT_Z,参照 resnet50

网络模型，设计了一维处理模块的 resnet50_1D，具体网络结构及参数设计见表 2。

表 2　resnet50 和 resnet50_1D 网络结构对比

Tab. 2　The comparison of network structure between resnet50 and resnet50_1D

layer_name	resnet50	resnet50_1D
conv1	7×7,64,stride 2	15,64,stride 2
	3×3maxpool,stride 2	3 maxpool_1D,stride 2
conv2	$\begin{bmatrix}1\times1,64\\3\times3,64\\1\times1,256\end{bmatrix}\times3$	$\begin{bmatrix}7,16\\11,16\\7,64\end{bmatrix}\times3$
conv3	$\begin{bmatrix}1\times1,128\\3\times3,128\\1\times1,512\end{bmatrix}\times4$	$\begin{bmatrix}7,32\\11,32\\7,128\end{bmatrix}\times4$
conv4	$\begin{bmatrix}1\times1,256\\3\times3,256\\1\times1,1024\end{bmatrix}\times6$	$\begin{bmatrix}7,64\\11,64\\7,256\end{bmatrix}\times6$
conv5	$\begin{bmatrix}1\times1,512\\3\times3,512\\1\times1,2048\end{bmatrix}\times3$	$\begin{bmatrix}7,128\\11,128\\7,512\end{bmatrix}\times3$
–	average pool, fc softmax	average pool, fc, logsoftmax
Params_size	89.66 MB	19.65 MB

从表 2 可以看出，尽管一维模型卷积核的参数比二维模型多，但是 resnet50_1D 在每个卷积、归一化中采用的输出通道数仅有 resnet50 的 1/4，综合计算 resnet50_1D 的参数量只有 19.65 MB，远小于 resnet50 的 89.66 MB。因此在训练过程中，需要计算、保存并更新的参数量大大减少，有利于加快训练速度，在内存、显存等条件有限的情况下也能满足实验需求。

3　实验设置与结果

3.1　实验设置

本次实验在 Windows10 x64 操作系统下完成，具体的开发和测试环境见表 3。

表 3　实验开发和测试环境

Tab. 3　The development and test environment of experiment

环境	型号/版本
CPU	Intel i5-4590 3.3GHz 4 核
内存	16G
GPU	NVIDIA GTX960　4G
CUDA	10.1.120
Pytorch	1.7.0+cu101
Python	3.7.9
IDE	PyCharm2020.1.2 x64 专业版

ASVspoof2019 的 PA 数据集分为 3 个子集：训练集54 000条语音、验证集 29 700 条语音、测试集134 730条语音，每个子集都包含真实语音和重放语音，其中重放语音共包含 9 种攻击类型[9]。

不同于以往的评估标准，为了与自动说话人验证（Automatic Speaker Verification，ASV）结合更紧密，ASVspoof2019 首次采用新的以 ASV 为中心的度量标准，即串联决策成本函数（Tandem Detection Cost Function，t-DCF）[10] 为主要评价指标，等错误率（Equal Error Rate，ERR）为次要指标。

3.2　实验结果及分析

按照上述的实验相关设置，用压缩特征 CQT_Z 在 resnet50_1D 网络上训练了重放攻击检测模型。将在验证集和测试集上来展示重放攻击检测的训练结果，验证集和测试集在训练好的模型上得到的真实语音/重放语音分数分布直方图如图 3 和图 4 所示，评价指标 t-DCF 如图 5 所示。

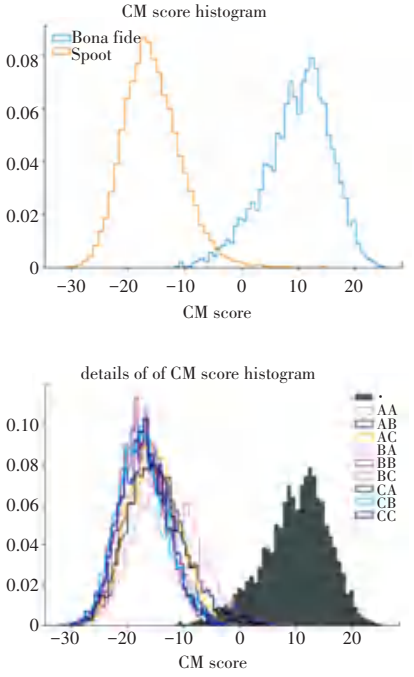


图 3　验证集分数统计直方图

Fig. 3　The score statistical histogram of verification set

横、纵坐标分别代表分数和分数密度，蓝色、黄色分别代表真实语音和重放语音检测分数的分布，蓝色和黄色曲线交集越小，不同颜色的曲线越窄、越集中，检测效果越好，越容易划分阈值以区分真实语音和重放语音。图 3 和图 4 中靠后侧图是详细的、不同攻击类型语音的分数统计直方图，-表示真人发声的语音，两位字母组合，如 AA，表示不同攻击类型的重放语音，共 9 种。

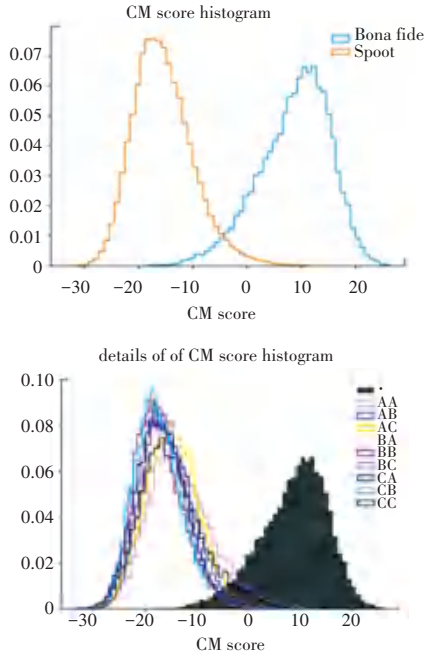
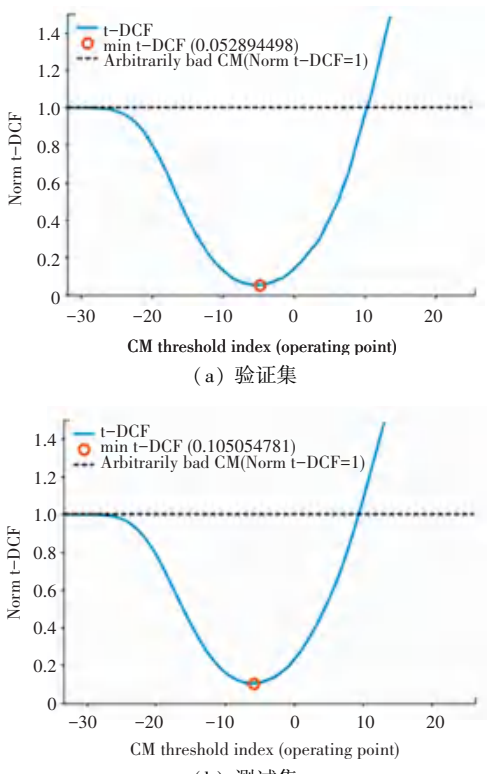


图 4 测试集分数统计直方图

Fig. 4 The score statistical histogram of test set

从图 3 和图 4 的分数直方图、图 5 的 t-DCF 曲线可以看出,本次实验训练的网络模型能使蓝色曲线和黄色曲线重叠区域很小,且每种颜色的形状较窄、集中,t-DCF 达到了较好的指标。同时,通过观察靠后的详细分数统计直方图,发现 xA 类型(用高质量设备录制)的录音识别率较低,这给人们今后的进步提供了目标。



(a) 验证集

(b) 测试集

图 5 t-DCF 曲线

Fig. 5 The t-DCF curves

有了实验结果的上述定性分析,还需定量表征。表 4 记录了其他特征-网络和本文特征-网络在 ASVspoof2019 PA 的验证集和测试集上的结果。

表 4 其他特征-网络和本文特征-网络在 ASVspoof2019 PA 的测试结果

Tab. 4 The test results of other feature-networks and this feature-networks on ASVspoof2019 PA

特征-模型	验证集		测试集	
	t-DCF ↓	EER(%) ↓	t-DCF ↓	EER(%) ↓
LFCC-GMM 基线系统[5]	0.255 4	11.96	0.301 7	13.54
CQCC-GMM 基线系统[5]	0.195 3	9.87	0.245 4	11.04
CQCC-ResNet[11]	-	-	0.107	4.43
(CQEPIC-SD)-GMM[12]	-	-	0.137	6.97
SI-IMFCC[13]	0.115 6	4.23	0.172 4	6.72
本文 CQT_z-resnet50_1D	0.052 9	1.89	0.105 1	3.74

相比 ASVspoof2019 的两个基线系统,在测试集中,t-DCF 效能分别改善了 65.16%、57.17%,EER 分别改善了 72.38%、66.12%,也明显优于其他特征-网络系统。

本文算法训练用时小于 3 h,网络模型较小,在保证高效检测性能0.105 1的 t-DCF 和 3.74%的 EER 的同时,有较快的网络训练和测试速度。可见,语音 CQT 频谱的时间帧压缩方法在重放攻击检测中是高效的。

4 结束语

本文在语音重放攻击检测中对指数敏感的语音 CQT 谱进行时间帧压缩,得到特征丰富但数据量极小的一维特征,同时对应设计一维处理模块的残差网络模型,在 ASVspoof2019 PA 公开数据集上取得了很好的效果,模型训练和测试速度很快、实验设备要求较低,证明语音 CQT 频谱的时间帧压缩方法在

(下转第 63 页)