ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 2,Mar. 2024,pp. 416－423
DOI:10. 16337/j. 1004-9037. 2024. 02. 014

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于自注意力机制的音频对抗样本生成方法

李珠海，郭　武

（中国科学技术大学语音及语言信息处理国家工程研究中心，合肥 230027）

摘　要：随着个人语音数据在网络上的传播以及自动说话人识别算法的发展，个人的声纹特征面对着泄露的风险。音频对抗样本可以在人耳主观听觉不变的前提下，使得自动说话人识别算法失效，从而保护个人的声纹特征。本文在典型的音频对抗样本生成算法FoolHD模型的基础上引入了自注意力机制来改进对抗样本生成，该方法称为FoolHD-MHSA。首先，使用卷积神经网络作为编码器来提取输入音频频谱的对抗扰动谱图；然后利用自注意力机制从全局角度提取扰动谱不同部分特征的关联特征，同时将网络聚焦到扰动谱中的关键信息、抑制无用信息；最后，使用解码器将处理后的扰动谱隐写到输入频谱中得到对抗样本频谱。实验结果表明，FoolHD-MHSA方法生成的对抗样本相比FoolHD方法有着更高的攻击成功率和平均客观语音质量评估（Perceptual evaluation of speech quality，PESQ）得分。

关键词：自注意力机制；对抗样本；说话人识别；深度神经网络

中图分类号：TN912.3　　**文献标志码：**A

Audio Adversarial Examples Generation Method Based on Self-attention Mechanism

LI Zhuhai，GUO Wu

（National Engineering Research Center of Speech and Language Information Processing，University of Science and Technology of China，Hefei 230027，China）

Abstract：With the widespread of personal speech and development of automatic speaker recognition algorithms，personal privacy protection is in a high-risk situation. Audio adversarial examples can protect personal voiceprint features through disabling automatic speaker recognition algorithms while the subjective hearing of the human ear remains unchanged. We improve the typical adversarial attacks algorithm FoolHD with multi-head self-attention mechanism，and we call it FoolHD-MHSA. First，convolutional neural networks are introduced as the encoder to extract adversarial perturbation spectrograms. Second，we use self-attention mechanism to extract correlation features of different parts of perturbation spectrogram from a global perspective ，focus the network on the important information and suppress the useless information. Finally，the processed perturbation spectrogram is steganographed into the input spectrogram with a decoder to get adversarial example spectrogram. Experimental results show that FoolHD-MHSA can generate adversarial examples with higher attack success rate and average PESQ score than FoolHD.

Key words: self-attention mechanism; adversarial examples; speaker recognition; deep neural network

收稿日期：2022-12-08；**修订日期：**2023-02-26

引　言

随着移动互联网的广泛应用,个人的数据在网络上广泛传播;另外,随着深度学习技术的快速发展,各种自动模式识别系统的性能得到了明显的提升。在说话人识别领域,以x-vector为代表的基于深度神经网络(Deep neural network,DNN)的说话人识别系统使得自动说话人识别进入了实用,从而很容易就可以自动获得个人的身份信息,导致个人隐私有可能泄露。在这种背景下说话人匿名研究开始得到关注,该研究旨在保护说话人身份信息的同时不破坏语音中的其他信息。例如,Meyer等[1]利用生成式对抗网络实现了保护说话人身份信息的同时不增加语音识别系统的词错误率。对抗样本也是一种说话人匿名的方法,采用合适的方法,在人耳主观听觉不变的情况下,对原始的语音进行少量的修改生成声纹对抗样本,使得自动说话人识别算法失效,既可以保证信息传播,又能保护个人隐私。

根据攻击者是否知晓待攻击模型的参数可以将对抗攻击分为白盒攻击和黑盒攻击,此外根据对抗样本通过分类器的结果是否被预先设计还可以将对抗攻击分为目标攻击和非目标攻击[2],本文研究是针对白盒说话人识别模型的非目标对抗攻击。针对x-vector和i-vector两种主流的说话人识别系统,Kreuk等[3]和Li等[4]使用快速梯度符号方法(Fast gradient sign method,FGSM)[5]来进行攻击,均取得了成功。2017年,Carlini和Wagner[6]分别使用3种不同的距离度量L_0、L_2和L_∞实施了3种新的攻击,并在防御蒸馏网络上实现了100%的攻击成功率。2018年,Hayes等[7]设计了一种使用生成式网络生成图像通用对抗扰动的方法,实验结果表明,该方法要比基于梯度迭代方法更为有效。受到这一工作的启发,2020年Li等[8]利用生成式网络方法生成了针对说话人识别网络的通用对抗扰动,他们攻击的说话人识别网络是在TIMIT数据集上训练的SincNet网络。虽然该方法取得了95%以上的攻击成功率,但是生成的对抗样本只有3.0的平均客观语音质量评估得分[9]。综上所述,现有的大多数音频对抗样本生成方法虽然能实现较高的攻击成功率,但是生成的对抗扰动不具有很高的不可察觉性。

Shamsabadi等[10]注意到这个问题,提出了一种称为FoolHD的方法,该方法针对白盒说话人识别模型生成同时具有高攻击成功率和高不可察觉性的音频对抗样本。FoolHD方法使用编码器提取输入频谱的扰动谱信息,然后利用解码器将扰动谱信息隐写到输入频谱之中。该编码器和解码器均是基于卷积神经网络实现的,然而受到卷积核的限制,卷积神经网络只能提取局部特征,因此该方法没有利用到语音信号上下文之间具有的相关性信息,也没有对提取的扰动谱信息中的有用信息和无用信息进行区分。

为解决上述问题,本文在FoolHD方法的基础上引入注意力机制来对音频上下文之间的依赖性进行建模,同时将网络聚焦到扰动谱之中的关键部分以此来获得更有效的扰动谱表征,该方法被称为FoolHD-MHSA (FoolHD with multi-head self-attention)。首先,FoolHD-MHSA方法仍使用卷积神经网络作为编码器来提取输入音频频谱的对抗扰动谱图;然后利用自注意力机制即多头自注意力层结构[11]从全局角度提取扰动谱不同部分特征的关联特征,同时将网络聚焦到扰动谱中的关键信息、抑制无用信息;最后,使用解码器将处理后的扰动谱隐写到输入频谱中,再进行反变化得到对抗样本。实验针对在VoxCeleb数据集[12]上训练的x-vector说话人识别系统进行攻击,结果表明FoolHD-MHSA方法相比于FoolHD方法可获得更高的攻击成功率和平均客观语音质量评估(Perceptual evaluation of speech quality,PESQ)得分。

1　FoolHD方法

1.1　FoolHD方法流程

FoolHD方法的流程如图1所示。具体而言,对于单个样本首先将其变换到频域上,然后经过对抗

样本生成器得到对抗样本的频谱，之后通过反变化得到对抗样本。输入音频和对抗样本之间会计算一个由听觉损失和对抗损失构成的多目标损失函数[10]，听觉损失用于保证对抗样本与输入音频间的听觉相似度，对抗损失用于保证对抗样本会被说话人识别系统错误分类。

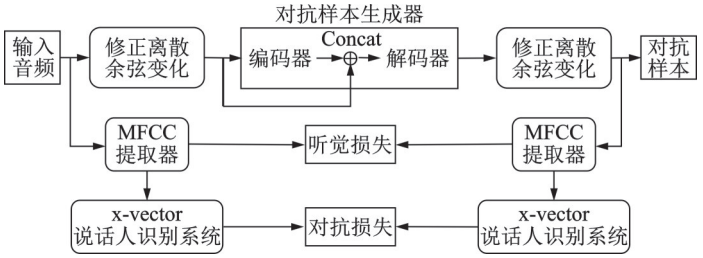

图1 FoolHD方法流程图

Fig.1 Flow chart of FoolHD

FoolHD方法考虑到人耳对声波不同的频率成分的敏感程度不同，因此对抗样本生成器并不是直接对输入音频在时域上进行处理，而是先使用修正的离散余弦变化(Modified discrete cosine transform，MDCT)[13]获得输入音频的频谱，再在频域上进行处理。由于短时傅里叶变化(Short time Fourier transform，STFT)的结果包含幅度和相位，在进行反变化时分别处理幅度和相位会导致重构误差[14]，因此采用变换结果是实数的MDCT来消除重构误差。FoolHD方法的对抗样本生成器由卷积神经网络构成，包含编码器和解码器两个部分，编码器由3层卷积层组成，解码器由4层卷积层组成，每个卷积层后还有批规整(Batch normalization，BN)层和随机失活层，所有卷积层的步长为1、填充为1。对抗样本生成器各层参数如表1所示，其中输入频谱维度为$T\times D\times 1$。对抗样本生成器根据输入音频经过MDCT后得到的频谱生成对抗样本的频谱，单通道的输入频谱经过编码器的处理变为64通道的谱，然后编码器输出的64通道谱会和输入频谱进行拼接得到通道数为65的谱，因此解码器层1输入通道数为65，拼接后的谱经过解码器最终又变为单通道谱，对该单通道谱进行修正离散余弦反变化就得到了对抗样本。

表1 对抗样本生成器各层参数

Table 1 Parameters for each layer of adversarial example generation network

对抗样本生成器层	(卷积核，输入通道数，输出通道数)	输出特征维度
编码器层1	(3×3,1,64)	$T\times D\times 64$
编码器层2	(3×3,64,64)	$T\times D\times 64$
编码器层3	(3×3,64,64)	$T\times D\times 64$
解码器层1	(3×3,65,64)	$T\times D\times 64$
解码器层2	(3×3,64,64)	$T\times D\times 64$
解码器层3	(3×3,64,64)	$T\times D\times 64$
解码器层4	(3×3,64,1)	$T\times D\times 1$

1.2 多目标损失函数

FoolHD使用多目标损失函数优化对抗样本生成器参数，目标函数包括听觉损失和对抗损失两个部分。

听觉损失使用到了梅尔频率倒谱系数(Mel frequency cepstrum coefficient，MFCC)间的余弦距离，输入音频$\boldsymbol{x}$的MFCC特征为$\boldsymbol{f}$，相应对抗样本$\hat{\boldsymbol{x}}$的MFCC特征为$\hat{\boldsymbol{f}}$，$\boldsymbol{f}_t$和$\hat{\boldsymbol{f}}_t$分别代表$\boldsymbol{f}$和$\hat{\boldsymbol{f}}$的第t帧特征($t=0,1,\cdots,T-1$)，听觉损失L_{P}定义为

$$L_{\mathrm{P}}(\boldsymbol{x},\hat{\boldsymbol{x}})=\sum_{t=0}^{T-1}\left(1-\frac{\boldsymbol{f}_t\hat{\boldsymbol{f}}_t}{\left\|\boldsymbol{f}_t\right\|\left\|\hat{\boldsymbol{f}}_t\right\|}\right) \tag{1}$$

对抗损失用于指导对抗样本生成器的参数向着使对抗样本被说话人识别网络错误分类的方向移动，对抗损失L_{A}定义为

$$L_{\mathrm{A}}(\boldsymbol{x},\hat{\boldsymbol{x}})=\hat{z}_y-\max_{\substack{i=1,2,\cdots,N\\ i\neq y}}\hat{z}_i \tag{2}$$

假设说话人识别系统$f(\cdot)$是一个N个说话人识别模型,输出层有N个节点,$z=f(x)$和$\hat{z}=f(\hat{x})$分别为输入音频x和对抗样本$\hat{x}$经过说话人识别系统输出的N维向量,每一维度的值可以代表分类为该节点对应说话人的概率,y为输入音频x通过说话人识别系统的分类结果,则有

$$y=\underset{i=1,2,\cdots,N}{\operatorname{argmax}}\ z_i \tag{3}$$

最终的多目标损失函数L可表示为

$$L(x,\hat{x})=L_{\mathrm{P}}\left(f,\hat{f}\right)+L_{\mathrm{A}}\left(f,\hat{f}\right) \tag{4}$$

2 FoolHD-MHSA方法

2.1 FoolHD-MHSA方法流程

个人的声纹特征是通过整段语音表现出来的,FoolHD方法虽然通过叠加多层卷积来扩大感受野,但是对于整段语音的总体特性捕获仍有改进之处。与FoolHD方法不同,FoolHD-MHSA方法并非只使用卷积神经网路提取扰动谱信息,而是引入了自注意力机制对扰动谱进行进一步的处理以获得整段语音更为有效的扰动谱表征。FoolHD-MHSA集中在对抗样本生成器的改进上。改进后的对抗样本生成器结构如图2所示。自注意力机制源于自然语言处理领域的Transformer模型[11],通常用于提取长期依赖型特征,但是在捕捉局部低级特征上有所不足,这和卷积结构存在互补性。因此,本文设计的对抗样本生成器仍利用卷积层组成的编码器根据输入频谱提取帧级别扰动谱特征,利用卷积层的多个通道提取多层次的扰动谱。在此基础上希望能够利用多层次扰动谱之间和扰动谱不同帧特征之间的依赖性关系来对重要部分进行突出,对无用部分进行抑制,所以本文将编码器提取的扰动谱通过多头自注意力层处理,其核心部分是h个如图3所示的缩放点积注意力层[11]。在每个缩放点积注意力层中,维度为$C\times F\times D$的扰动谱经过3个不同的线性层得到3个不同的矩阵,分别记为$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$,矩阵$\boldsymbol{Q}$和$\boldsymbol{K}^{\mathrm{T}}$先进行矩阵乘法再乘以一个缩放系数,接着经过softmax函数进行归一化得到归一化系数矩阵,最后用归一化系数矩阵与$\boldsymbol{V}$进行矩阵乘法,上述过程可表示为

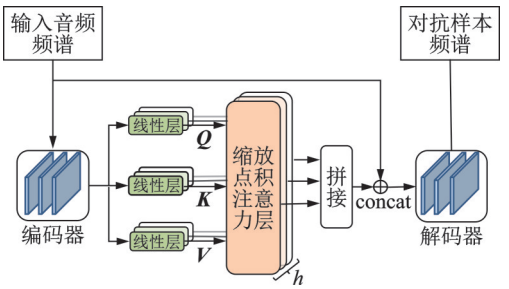

图2 对抗样本生成网络

Fig.2 Adversarial examples generation network

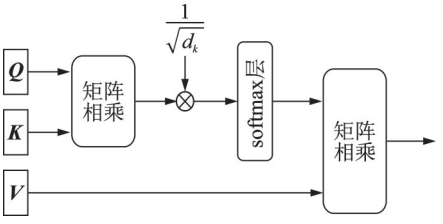

图3 缩放点积注意力层

Fig.3 Scaled dot-product attention layer

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

式中:参数d_k指$\boldsymbol{K}$中每帧特征的维度,本文选择$d_k=1\,024$的参数设置;$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$特征维度相同。

缩放点积注意力层计算的归一化系数矩阵用于对$\boldsymbol{V}$中每一帧的特征进行加权,网络通过学习权重系数自行为扰动谱中的重要部分赋予较高的权重、无用部分赋予较低的权重,从而达到突出有用信息并抑制无用信息的作用。h个缩放点积注意力层的输出会被拼接在一起,拼接层后跟一个线性层将h个扰动谱进行融合,采用多个缩放点积注意力层可以提取多个层次的注意力信息,进而获得更有效的扰动谱表征。经过多头自注意力层处理的扰动谱会和原始输入频谱进行拼接,这相当于扰动谱的通道数

由C变为$C+1$,接着使用卷积层构成的解码器将多层次的扰动谱融合到输入频谱中,得到的谱图会被作为对抗样本的频谱。本文提出的FoolHD-MHSA方法的编码器参数配置与FoolHD方法保持一致,同时将FoolHD解码器中的解码器层3输出通道数改为1并去除解码器层4后作FoolHD-MHSA的解码器。

2.2 说话人识别模型及对抗样本实验设置

攻击的说话人识别模型是基于时延神经网络(Time delay neural network, TDNN)的x-vector系统,网络主要结构包括5个基于TDNN的帧处理层、1个统计池化层、基于线性层的2个段处理层和输出层[15]。网络输入采用MFCC声学特征,提取29维的MFCC声学特征作为输入,帧长为25 ms、帧移为10 ms,使用语音活动检测剔除非语音帧。本文在Pytorch[16]上实现了上述说话人识别模型,并使用VoxCeleb数据集进行训练,该数据集包含有多个不同种族、口音、职业和年龄的7 363个说话人[12],训练好的模型将会作为一个预训练模型。

在后续对抗样本生成实验中,从VoxCeleb数据集中随机选择100个说话人并为每个说话人随机选择10段话,使用算法生成的对抗样本来实现对x-vector系统的攻击。由于使用太少的数据训练神经网络容易造成过拟合现象,因此本文从VoxCeleb数据集中额外选择了150个说话人,这150个说话人将会和前面选择的用于测试的100个说话人作为训练x-vector系统的说话人集合,构建基于这250人之上的闭集说话人识别,直接采用神经网络输出来进行目标说话人判决,如式(3)所示。在训练过程中,使用这250个说话人的所有数据对前述的预训练模型进行微调(Fine-tune),所有训练数据均被降采样到8 kHz的采样率同时被截短至4 s。说话人识别网络按照1e−5的学习率训练了100轮,使用交叉熵作为目标函数,使用Adam优化器更新网络参数,最终在训练集上达到了98.5%的识别准确率。

FoolHD和FoolHD-MHSA都是迭代式的对抗样本生成方法而不是训练-测试的框架,每段原始音频会经过反复迭代来最小化目标函数以获得相应的对抗样本,因此评估算法性能时只需要测试集。

3 实验设置与结果分析

3.1 实验数据与评价指标

如前所述,为了测试不同对抗样本生成算法对x-vector系统的攻击能力,本文从250个说话人中随机选择了100个说话人并为每个说话人随机选择10段音频,分别使用FoolHD方法和FoolHD-MHSA方法生成测试集中这1 000段音频的对抗样本。

本文的评价指标包括攻击成功率和平均PESQ得分。为了评估攻击方法的有效性,使用攻击成功率这一评价指标,输入音频$\boldsymbol{x}$通过说话人识别模型为$f(\cdot)$的分类结果为t,相应的对抗样本$\hat{\boldsymbol{x}}$的分类结果为$\hat{t}$,由于执行的是非目标对抗攻击,所以当$t\neq\hat{t}$时即可认为攻击成功,统计生成的1 000个对抗样本攻击成功的比例即可计算攻击成功率。为了评估所添加的对抗扰动的不可察觉性,使用平均PESQ得分这一指标,这是一种客观平均意见得分(Mean opinion score, MOS)值的评价方法。先将原始信号与失真信号经过电平调整再通过滤波器滤波,将两个信号在时间上对齐后通过感知模型计算平均扰动值D_{ind}和平均非对称扰动值A_{ind},然后按照式(6)计算PESQ得分[9],其中$\alpha_0=4.5$、$\alpha_1=-0.1$、$\alpha_2=-0.0309$。得到的结果一般在1~4.5之间,得分越大表示失真越小,最后统计1 000对输入音频和对抗样本之间的PESQ得分即可计算出平均PESQ得分。

$$\text{PESQ}=\alpha_0+\alpha_1 D_{\text{ind}}+\alpha_2 A_{\text{ind}} \tag{6}$$

3.2 实验结果及分析

分别使用FoolHD方法和FoolHD-MHSA方法生成测试集中1 000段音频的对抗样本,生成对抗样

本时进行100次迭代,FoolHD-MHSA方法中多头自注意力层的头数设置h=8,结果如表2所示。对比表2中的实验结果,在相同的迭代次数下,本文提出的FoolHD-MHSA方法相比于FoolHD方法在攻击成功率上由96.0%提升到99.3%,平均PESQ得分由4.19提升到4.34,这说明相比于FoolHD方法,基于自注意力机制的FoolHD-MHSA方法能够生成具有更高攻击性和不可察觉性的对抗扰动,同时也证明了本文引入的注意力模块的有效性。

为了探究FoolHD-MHSA方法的多头自注意力层中缩放点积注意力层数目h对生成的对抗样本的影响,分别选择h=2、4、8、12的不同设置生成对抗样本,实验结果如表3所示。对比表3中的实验结果,缩放点积注意力层数目h的增加虽然对于生成对抗样本的平均PESQ得分没有明显的影响,但是会使攻击成功率逐渐上升,这说明多个缩放点积注意力层可以提取多个层次的注意力信息,获得更有效的扰动谱表征,进而提高攻击成功率。

此外,本文进一步增大测试集的规模并将FoolHD和FoolHD-MHSA算法在新的测试集上进行对比。新的测试集额外选择100个说话人与已有的100个说话人组成测试集说话人集合,仍然为每个说话人随机选择10段音频用于测试,因此总共包含了200个说话人的2 000段音频。为了进一步说明本文所提算法的有效性,选择了目前在各个领域都得到广泛应用的FGSM算法和其迭代版本的PGD算法[5]作为对比。实验结果如表4所示。对比表4中的实验结果,在更大规模的测试集上,相比于FoolHD,FoolHD-MHSA仍能生成具有更高的攻击成功率和更好的音质水平的对抗样本,而且FoolHD-MHSA在攻击成功率和PESQ得分两个指标上均明显优于FGSM。虽然PGD在攻击成功率上有着一定优势,但是在PESQ得分上与FoolHD-MHSA存在着27.9%的差距。以上的实验结果进一步说明了本文所提FoolHD-MHSA算法的有效性。

表2 FoolHD和FoolHD-MHSA方法实验结果

Table 2 Experiment results of FoolHD and FoolHD-MHSA

对抗攻击方法	迭代次数 M	攻击成功率/%	平均PESQ得分
FoolHD	100	96.0	4.19
FoolHD-MHSA	100	99.3	4.34

表3 不同缩放点积注意力层数下的实验结果

Table 3 Experiment results under different number of scaled dot-product attention layer

缩放点积注意力层数h	迭代次数 M	攻击成功率/%	平均PESQ得分
2	100	96.6	4.32
4	100	97.5	4.33
8	100	99.3	4.34
12	100	99.5	4.34

表4 更大规模测试集上FGSM、PGD、FoolHD和FoolHD-MHSA算法的实验结果

Table 4 Experiment results of FGSM, PGD, FoolHD and FoolHD-MHSA on larger scale test set

对抗攻击方法	迭代次数 M	攻击成功率/%	平均PESQ得分
FGSM	—	62.9	3.24
PGD	10	100	3.36
FoolHD	100	95.7	4.17
FoolHD-MHSA	100	98.6	4.30

为进一步比较FoolHD和FoolHD-MHSA方法的差异,绘制了原始音频以及两种方法生成的对抗样本、对抗扰动的短时傅里叶变换谱,结果如图4所示。对抗扰动按照式(7)计算。在绘制对抗扰动语谱图时,先按照式(7)在时域上将对抗样本和原始音频相减得到对抗扰动,然后对时域上的对抗扰动进行短时傅里叶变换绘制其语谱图。

$$\text{对抗扰动} = \text{对抗样本} - \text{原始音频} \tag{7}$$

对比图4中的语谱图,FoolHD方法(图4(b))、FoolHD-MHSA方法(图4(c))生成的对抗样本语谱

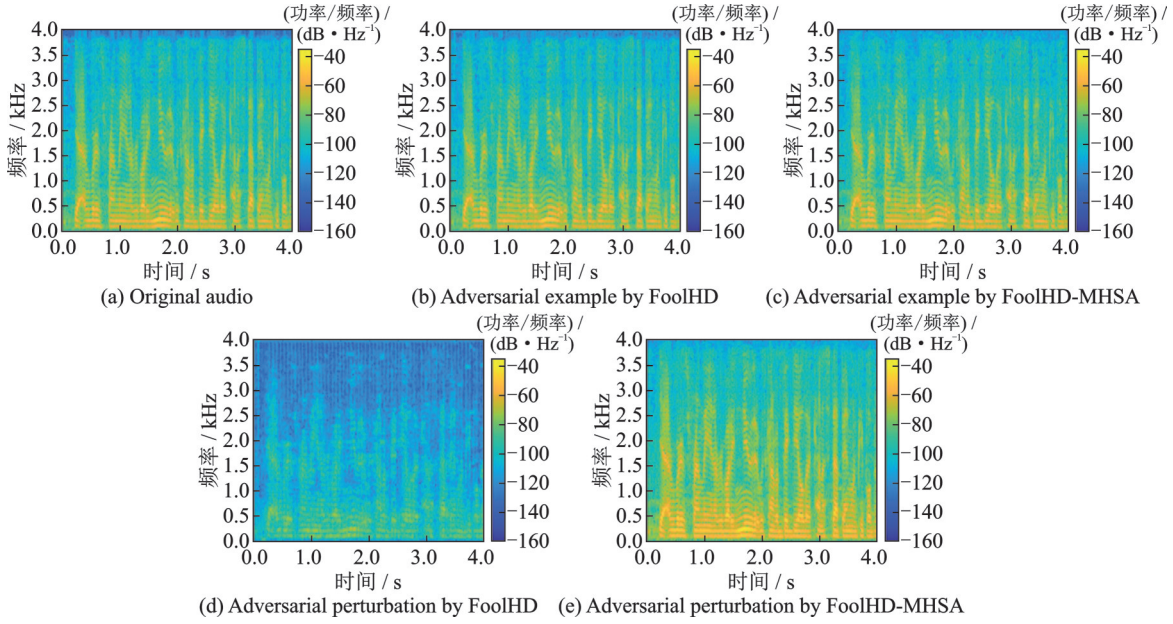

图4 FoolHD和FoolHD-MHSA方法的原始音频、对抗样本和对抗扰动的短时傅里叶变换谱

Fig.4 STFT spectrograms of original audio, adversarial examples and adversarial perturbations for FoolHD and FoolHD-MHSA

图和原始音频的语谱(图4(a))具有高度的相似性,语谱图结构的相似性保证了原始音频和对抗样本之间的听觉相似性。然而,对比两种方法生成的对抗扰动的语谱图,可以发现这两种方法保证对抗样本和原始音频之间语谱图结构相似的原理不同:FoolHD方法生成的对抗扰动(图4(d))语谱图相比于FoolHD-MHSA方法生成的对抗扰动(图4(e))语谱图所具有能量更低,因此对原始音频谱结构改变不大,而FoolHD-MHSA生成的对抗扰动和原始音频谱结构仍具有很高的相似性,所以在将对抗扰动添加到原始音频之后仍能保证谱结构的相似性,这说明了FoolHD-MHSA方法能够利用全局信息生成扰动谱来保持添加对抗扰动前后谱结构的相似度,进而保证听觉上的相似性。

4 结束语

本文提出了一种基于自注意力机制的音频对抗样本生成方法,引入注意力机制来对音频上下文之间的依赖性进行建模,同时将网络聚焦到扰动谱之中的关键部分以此来获得更有效的扰动谱表征。在VoxCeleb数据集中1 000段音频上的实验结果表明,基于自注意力机制的FoolHD-MHSA方法相比于FoolHD方法可取得更高的攻击成功率和平均PESQ得分,从而证明了本文引入的自注意力模块的有效性。语谱图的分析表明,FoolHD-MHSA方法生成的对抗样本和对抗扰动都能和原始音频在谱结构上保持良好的相似性。

参考文献:

[1] MEYER S, TILLI P, DENISOV P, et al. Anonymizing speech with generative adversarial networks to preserve speaker privacy[C]//Proceedings of 2022 IEEE Spoken Language Technology Workshop (SLT). Doha, Qatar: IEEE, 2023: 912-919.

[2] 陈佳豪, 白炳松, 王冬华, 等. 面向语音识别系统的对抗样本攻击及防御综述[J]. 小型微型计算机系统, 2022, 43(3): 466-474.

CHEN Jiahao, BAI Bingsong, WANG Donghua, et al. Adversarial attacks and countermeasures for speech recognition system

[J]. Journal of Chinese Computer Systems, 2022, 43(3): 466-474.

[3] KREUK F, ADI Y, CISSE M, et al. Fooling end-to-end speaker verification with adversarial examples[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE, 2018: 1962-1966.

[4] LI X, ZHONG J, WU X, et al. Adversarial attacks on GMM i-vector based speaker verification systems[C]//Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain: IEEE, 2020: 6579-6583.

[5] JATI A, HSU C C, PAL M, et al. Adversarial attack and defense strategies for deep speaker recognition systems[J]. Computer Speech & Language, 2021, 68: 101199.

[6] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy. San Jose, USA: IEEE, 2017: 39-57.

[7] HAYES J, DANEZIS G. Learning universal adversarial perturbations with generative models[C]//Proceedings of 2018 IEEE Security and Privacy Workshops (SPW). San Francisco, USA: IEEE, 2018: 43-49.

[8] LI J, ZHANG X, JIA C, et al. Universal adversarial perturbations generative network for speaker recognition[C]//Proceedings of 2020 IEEE International Conference on Multimedia and Expo (ICME). London, United Kingdom: IEEE, 2020: 1-6.

[9] HU Y, LOIZOU P C. Evaluation of objective quality measures for speech enhancement[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2007, 16(1): 229-238.

[10] SHAMSABADI A S, TEIXEIRA F S, ABAD A, et al. FoolHD: Fooling speaker identification by highly imperceptible adversarial disturbances[C]//Proceedings of 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Toronto, Canada: IEEE, 2021: 6159-6163.

[11] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA: Neural Information Processing Systems(NIPS), 2017: 6000-6010.

[12] NAGRANI A, CHUNG J S, ZISSERMAN A. Voxceleb: A large-scale speaker identification dataset[C]//Proceedings of Conference of the International Speech Communication Association. Stockholm, Sweden: ISCA, 2017: 2616-2620.

[13] ZHANG S, DOU W, YANG H. MDCT sinusoidal analysis for audio signals analysis and processing[J]. IEEE Transactions on Audio, Speech, and Language Processing, 2013, 21(7): 1403-1414.

[14] KREUK F, ADI Y, RAJ B, et al. Hide and speak: Towards deep neural networks for speech steganography[C]//Proceedings of Conference of the International Speech Communication Association(INTERSPEECH). Shanghai, China: ISCA, 2020: 4656-4660.

[15] SNYDER D, GARCIA-ROMERO D, SELL G, et al. X-vectors: Robust DNN embeddings for speaker recognition[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Calgary, Canada: IEEE, 2018: 5329-5333.

[16] PASZKE A, GROSS S, MASSA F, et al. Pytorch: An imperative style, high-performance deep learning library[C]//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: NIPS, 2019: 8026-8037.

作者简介:

李珠海(2001-),男,硕士研究生,研究方向:声纹识别、声纹对抗,E-mail:snowsea@mail. ustc. edu. cn。

郭武(1973-),**通信作者**,男,博士,副教授,研究方向:语音信号处理,E-mail:guowu@ustc.edu.cn。

(编辑:王静)