ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 41,No. 1,Jan. 2026,pp. 109－116
DOI:10. 16337/j. 1004-9037. 2026. 01. 007

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

面向Transformer语音识别模型的高迁移通用对抗样本生成方法

王　振，韩纪庆，何勇军，郑铁然，郑贵滨

（哈尔滨工业大学计算机科学与技术学院，哈尔滨 150001）

摘　要： Transformer模型的出现使得语音识别的正确率有了巨大提升。随着深度学习技术的发展，通过对抗样本来攻击语音识别系统，以了解该系统的脆弱性并进行完善，进而提高识别系统的鲁棒性。由于通用语音对抗样本对于任意语音都有效，更是受到了广泛关注，其关键问题是如何提高对抗样本的迁移性，进而实现高攻击成功率。本文利用Transformer类语音识别模型结构特征的相似性，通过使扰动后的语音与原始语音的中间层特征尽可能不同，以改变其中间层特征表示的规律，实现有效的通用对抗攻击。鉴于通用对抗样本需要利用与样本无关的底层声学信息，而与样本依赖的语义信息会抑制其性能，因而通过控制注意力梯度以减弱通用对抗样本对于语义上下文特征的学习，进而实现通用对抗样本的高迁移性。实验结果表明，本文所提出的通用对抗样本生成方法可以有效地提高对抗样本在Transformer类语音识别模型之间的迁移性。

关键词： 语音识别；对抗样本；黑盒攻击；注意力机制

中图分类号： TN912；TN18　　**文献标志码：** A

引用格式： 王振，韩纪庆，何勇军，等．面向Transformer语音识别模型的高迁移通用对抗样本生成方法[J]．数据采集与处理，2026，41(1)：109-116. WANG Zhen，HAN Jiqing，HE Yongjun，et al. Universal adversarial example generation method with high transferability for transformer-based speech recognition models[J]. Journal of Data Acquisition and Processing，2026，41(1)：109-116.

引　　言

语音识别技术能够将接受到的语音转录为文本，是语音人机交互系统的核心功能。近年来，Transformer模型的出现使得语音识别的正确率有了巨大提升。随着深度学习技术的发展，研究者开始研究通过对抗样本来攻击语音识别系统，以了解该系统的脆弱性并进行完善，进而提高识别系统的鲁棒性。这是近年来语音领域研究的一个热点问题[1]。对抗样本生成技术通过对预测样本故意添加细微的干扰，从而导致模型能以高置信度输出一个错误的目标类别结果[2]，进而达到攻击语音识别模型的目的。已有研究表明，通过修改音频信号能产生错误的类别结果，进一步证实了语音对抗样本的存在[3]。

语音对抗样本生成方法根据对语音识别模型信息掌握的程度不同，可分为白盒攻击和黑盒攻击。对前者，生成对抗样本时完全了解语音识别模型的结构和参数信息，能有目标的生成对抗样本[4-7]。对后者，由于无法掌握被攻击模型的信息，因此攻击难度更大。一些研究通过利用白盒模型作为代理模型，来生成具有迁移性的语音对抗样本[8-9]，即针对一种深度学习模型生成的对抗样本在迁移到另一个不同结构的模型后仍然具备一定的攻击能力，其攻击效果越好说明该对抗样本越具有高迁移性。也有

收稿日期：2024-03-30；**修订日期：**2024-10-12

一些工作基于查询黑盒模型的输出信息来指导生成语音对抗样本[10-11]。在现实场景中,攻击者通常难以获取识别模型的内部信息,因此黑盒攻击更具有实际的意义。然而,现有的黑盒攻击方法往往通过优化对抗样本生成的迭代过程[8-9],以减弱对抗样本在代理模型上的过拟合,从而提高迁移性。尽管这类方法适用于各种类型的语音识别模型,但对抗样本的迁移能力有限。

目前大多语音对抗样本生成方法都是样本依赖的,即对抗样本生成方法对每个输入语音都单独产生对抗扰动。典型的工作有基于梯度符号[4,12]、目标优化[13]以及基于遗传算法[14-15]的生成方法。而在实际应用场景中,更希望生成的扰动能对任何样本都有效。为此,出现了通过条件变分自动编码器来合成不依赖于任何现有数据的对抗样本生成方法[16]。然而,该方法需要完全重新合成音频,并不适用于对已有语音进行扰动生成的情况。已有研究证实,存在适用于任意语音的通用对抗扰动(Universal adversarial perturbation,UAP)[17-18]。然而,由于语音长度不固定的特点会使通用对抗样本性能受到影响。为解决该问题,出现了使用定长通用噪声的重复回放来适应不同长度的输入语音,进而实现通用扰动的方法[19]。此外,有研究发现通过训练生成器模型也可得到通用对抗扰动[20]。与样本依赖的方法相比,通用对抗扰动则对大部分样本都有效,更易达到攻击的目的,同时也能大大提升扰动添加的速度。然而,由于语音具有时序依赖的特点,通用扰动对生成算法要求的更高且难以获得。同时,现阶段主流的语音识别模型大多基于Tranformer模块来实现,如Wav2vec2、WavLM和Data2vec等,其结构的相似性为研究高可迁移性的通用对抗样本提供了条件。鉴于模型特征中存在着相似性,通过攻击这种相似特征,可以实现通用对抗样本的高迁移性。

基于以上分析,本文使用白盒语音识别模型作为代理模型,通过利用模型特征的相似性来生成高度可迁移的通用语音对抗样本,以攻击不同的黑盒Transformer类语音识别模型。具体而言,鉴于Tranformer类语音识别模型中间层结构的相似性,因而对该结构特征的扰动更能适用于其他模型。因此,本文通过使扰动后的语音与原始语音的中间层特征尽可能不同,以改变其中间层特征表示的规律,进而实现有效的对抗攻击。此外,鉴于通用对抗样本需要利用非样本依赖的底层声学信息,而语音识别模型中注意力层包含的语义信息,会使对抗样本的迁移性受到影响。为此,通过控制注意力梯度以减弱通用对抗样本对于语义上下文特征的学习,进而实现通用对抗样本的高迁移性。实验验证了所提出方法的有效性。

1 基于高可迁移性的通用对抗样本生成方法

本文方法的整体框架如图1所示。其中上面部分为训练阶段,下面部分为应用推理阶段。在训练阶段,先生成随机的对抗样本并将其叠加到语音样本上,然后输入到语音识别模型中,通过扰动其中间层特征以及使其分类损失最大化,并在反向传播过程中控制注意力梯度,以优化更新通用对抗样本,进而实现其高迁移性。在推理阶段,输入语音直接加上通用对抗样本,输入到语音识别模型中,由于目标模型与代理模型中间层特征相似,最终识别结果出错率更大。

1.1 通用对抗样本生成方法

本文目的是找到一种通用的音频扰动,当将其添加到任何语音波形时,将以高概率导致语音识别模型在转录中出现错误,其优化目标可表示为

$$\begin{cases} \max_{\delta}\{P(\boldsymbol{D}(\boldsymbol{x}+\boldsymbol{\delta})\neq\boldsymbol{D}(\boldsymbol{x}))\}, \boldsymbol{x}\sim\boldsymbol{\mu} \\ \text{s.t. } \|\boldsymbol{\delta}\|_{\infty}\leqslant\varepsilon \end{cases} \tag{1}$$

式中:$\boldsymbol{\mu}$为输入样本分布,$\boldsymbol{x}$是从$\boldsymbol{\mu}$中随机采样的输入语音样本,$\boldsymbol{D}(\boldsymbol{x})$为代理白盒语音识别模型对语音输入$\boldsymbol{x}$的输出类别的概率值,$\boldsymbol{\delta}$为对抗性扰动,$\|\cdot\|_{\infty}$为无穷范数表示向量中绝对值最大的元素,约束条件保证扰动的无穷范数小于ε。

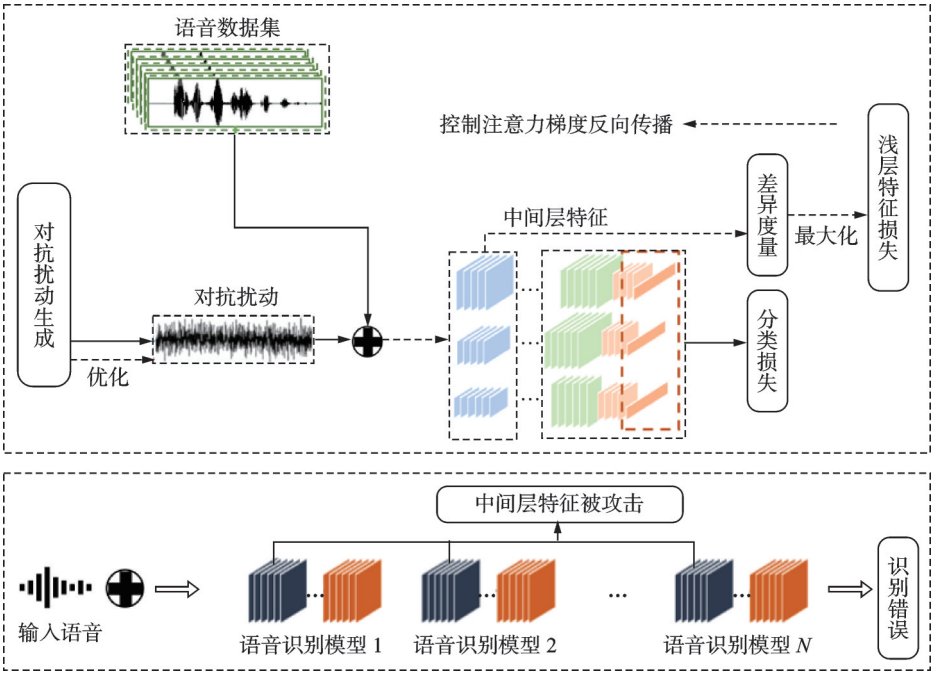

图1　基于高可迁移性的通用对抗样本生成方法框架

Fig.1　Framework of a general adversarial example generation method with high transferability

为了解决上述问题,采用文献[16]提出的通用对抗性扰动算法来寻找通用对抗性扰动,在每次迭代中,需要解决以下优化问题

$$\begin{cases}\Delta\boldsymbol{\delta}_i \leftarrow \arg\min_{\boldsymbol{r}}\|\boldsymbol{r}\|_\infty \\ \text{s.t. } \mathrm{CER}(C(\boldsymbol{x}_i+\boldsymbol{\delta}+\boldsymbol{r}), C(\boldsymbol{x}_i)) > t\end{cases} \tag{2}$$

式中:$\boldsymbol{r}$为每一批样本迭代中产生的扰动,将所有样本产生的扰动叠加起来即为通用对抗扰动,CER(·)表示字符错误率用于衡量两个文本序列之间的差异程度,$C(\cdot)$表示识别模型输出的字符,约束条件表示每批样本生成的对抗扰动$\boldsymbol{r}$的字符错率需要大于t,即攻击成功率。

为提升通用对抗样本的可迁移性,使其针对Transformer类的语音识别模型有更好的性能,拟通过利用模型特征的相似性来生成高度可迁移的通用语音对抗样本。

1.2　中间层模块特征攻击方法

鉴于对Transformer类的语音识别模型,其中间层特征一般描述语音信号底层的局部声学信息,同一语音的这些特征在不同结构的深度语音模型中具有相似性,同时对中间层特征的扰动也更能适用于其他模型。因此,本文通过使扰动后的语音与原始语音在中间层网络的特征尽可能不同,以改变其特征表示的规律性或分布,进而使得识别结果产生错误。

考虑到对同一深度神经网络,不同中间层描述的局部声学信息也不同,它们对扰动效果的影响也不同。本文拟通过加权方式来引入多层特征,以改变对扰动效果有影响的浅层特征表示的规律性,进一步提高代理模型生成的通用扰动迁移到黑盒模型上的攻击成功率。

基于代理模型中间层特征的通用对抗语音样本优化目标可表示为

$$\begin{cases}\max_{\boldsymbol{\delta}} \mathbb{E}_{x\sim\mu}\sum_{l=1}^{k}\lambda_l\dfrac{\|\boldsymbol{D}_l(\boldsymbol{x}+\boldsymbol{\delta})-\boldsymbol{D}_l(\boldsymbol{x})\|_2}{\|\boldsymbol{D}_l(\boldsymbol{x})\|_2}+\mathrm{CELoss}(\boldsymbol{D}_{x+\Delta}(\boldsymbol{x}+\boldsymbol{\delta}), \boldsymbol{1}_y) \\ \text{s.t. } \|\boldsymbol{\delta}\|_\infty \leqslant \varepsilon\end{cases} \tag{3}$$

式中：$\boldsymbol{D}_l(\cdot)$表示代理白盒语音模型中间层l的输出；λ_l为权重系数；第1项利用Transformer类模型中间层特征具有相似性，使得扰动后的特征尽可能与原始语音不同；第2项CELoss(·)为深度语音模型的交叉熵损失函数，计算对抗语音样本与全1类别向量$\boldsymbol{1}_y$的交叉熵，使模型输出结果尽可能出错。同时，约束条件保证扰动尽可能小。

此外，由于某些层的特征可能比其他层的特征更鲁棒，难以对其进行扰动，因此对每层特征施加的扰动权重应该不同。由于模型特征的鲁棒性体现在模型的输入受到扰动后输出的变化程度，输出变化越小说明模型的鲁棒性越好，而输出的变化程度可以体现在加入扰动后反向传播变化的大小。为此，对鲁棒性差的中间层设计更大的权重，使对抗扰动能针对鲁棒性差的中间层进行攻击，相应的自适应权重系数表示为

$$\lambda_l = \exp\left(\|\nabla \boldsymbol{D}_l(\boldsymbol{x}+\boldsymbol{\delta})\| / \left\|\sum\nolimits_l \nabla \boldsymbol{D}_l(\boldsymbol{x}+\boldsymbol{\delta})\right\|\right) \tag{4}$$

式中$\|\nabla \boldsymbol{D}_l(\boldsymbol{x}+\boldsymbol{\delta})\|$为中间层$l$在加入对抗扰动后的反向传播梯度的模。该值越大说明该层特征越容易被扰动，即越不鲁棒，因此需要使用较大的权重进行训练，以提升对抗样本的攻击成功率。在实验中发现，该权重需要较大才能对中间层进行足够的扰动，为此设计时使用了指数型函数，适当放大该权重系数。

1.3 基于非语义信息的梯度反向传播控制方法

在Transformer类的语音识别模型中，单个多头注意力层使用了多个自注意机制，其中每个自注意学习其上下文特征。通过投射来自多个头部的连接输出，多头注意力层能组合来自不同表示子空间的信息，其表示语音样本的语义信息。鉴于通用对抗样本生成更需要利用非样本依赖的非语义信息，因此，注意力层中的语义信息会使通用对抗样本的迁移性受到影响。本文通过控制注意力梯度以减弱通用对抗样本对于语义上下文特征的学习，进而实现通用对抗样本的高迁移性。

给定一个输入嵌入$\boldsymbol{Z}\in\mathbf{R}^{N\times K}$，$N$表示输入序列的长度，$K$表示每个输入元素的嵌入维度，查询、键和值权重$\boldsymbol{W}^{\mathrm{q}}$、$\boldsymbol{W}^{\mathrm{k}}$、$\boldsymbol{W}^{\mathrm{v}}\in\mathbf{R}^{K\times P}$，$P$表示注意力计算特征维度，则自注意力机制为

$$\boldsymbol{A} = \mathrm{softmax}\left(\boldsymbol{Z}\boldsymbol{W}^{\mathrm{q}}(\boldsymbol{Z}\boldsymbol{W}^{\mathrm{k}})^{\mathrm{T}}/\sqrt{P}\right) \tag{5}$$

式中：$\boldsymbol{A}\in\mathbf{R}^{N\times N}$表示注意力权重，那么这个自注意力层的输出为

$$\boldsymbol{Z}_{\mathrm{out}} = \boldsymbol{A}(\boldsymbol{Z}\boldsymbol{W}^{\mathrm{v}}) \tag{6}$$

因此输出$\boldsymbol{Z}_{\mathrm{out}}$相对于输入$\boldsymbol{Z}$的梯度可以展开为

$$\frac{\partial \boldsymbol{Z}_{\mathrm{out}}}{\partial \boldsymbol{Z}} = (\boldsymbol{E}\otimes\boldsymbol{A})\frac{\partial(\boldsymbol{Z}\boldsymbol{W}^{\mathrm{v}})}{\partial \boldsymbol{Z}} + ((\boldsymbol{Z}\boldsymbol{W}^{\mathrm{v}})^{\mathrm{T}}\otimes\boldsymbol{E})\frac{\partial \boldsymbol{A}}{\partial \boldsymbol{Z}} \tag{7}$$

式中：$\boldsymbol{E}$为单位矩阵，“$\otimes$”表示克罗内克内积。通过控制注意力梯度反向传播来减弱对于样本语义的学习，进而提高通用对抗样本的可迁移性，即

$$\frac{\partial \boldsymbol{Z}_{\mathrm{out}}}{\partial \boldsymbol{Z}} = (\boldsymbol{E}\otimes\boldsymbol{A})\frac{\partial(\boldsymbol{Z}\boldsymbol{W}^{\mathrm{v}})}{\partial \boldsymbol{Z}} + \alpha((\boldsymbol{Z}\boldsymbol{W}^{\mathrm{v}})^{\mathrm{T}}\otimes\boldsymbol{E}\partial)\frac{\partial \boldsymbol{A}}{\partial \boldsymbol{Z}} \tag{8}$$

式中$\alpha<1$为控制注意力梯度传播的系数，其迫使扰动仅通过使用特征表示来利用网络，同时保证网络中反向传播连贯，而不是利用针对样本的注意力特征。

2 实验设置与结果

2.1 实验设置

本文在wav2vec2和WavLM两个语音识别模型上进行实验。wav2vec2是一种自监督语音识别模型，通过与噪声混合并预测掩盖的部分来进行训练。它采用编码器-解码器架构和Transformer解码器，

能够从未标记数据中自动学习音频表示,并在基准测试中展现出优异性能。WavLM是一种基于自监督学习的语音到文本的转换模型,由Facebook AI Research提出。它的目标是将音频波形转化为对应的文本序列,实现语音转写的任务。WavLM的训练过程是通过利用大量未标记的音频数据进行自动学习。它使用自监督学习的方法,首先将原始音频波形通过编码器模型转化为特征表示,然后利用Transformer解码器模型生成对应的文本序列。为了提高模型效果,WavLM采用了对比目标函数,即预测音频特征的相邻上下文,以帮助模型学习音频和文本之间的对应关系。此外,使用LibriSpeech作为实验数据,它是从LibriVox项目的有声书籍中提取的英语口语语料库,包含超过1 000 h的语音录音,约有28 000个独立的说话者。本文从LibriSpeech选取10 h的音频进行训练和测试。本文的通用对抗性扰动的长度固定为48 000个采样点,相当于16 kHz下3 s的音频,扰动大小信噪比SNR=30 dB。通用对抗性扰动使用前文介绍的方法进行训练,依次对模型的不同中间层特征进行测试。

本文采用文献[16]的UAP作为基线系统,该方法框架与2.1节中一致,其优化目标为

$$\begin{cases}\max_{\delta}\mathrm{CELoss}(D(\boldsymbol{x}+\boldsymbol{\delta}),\boldsymbol{1}_y)\\ \text{s.t. }\|\boldsymbol{\delta}\|_{\infty}\leqslant\varepsilon\end{cases}\tag{9}$$

本文方法是“即插即用”的,可以和其他提高迁移性的方法[20-22]叠加使用,进一步提升通用对抗扰动的迁移性。本文采用多种迭代更新方法进行实验测试,以验证方法的有效性。快速梯度符号法(Fast gradient sign method, FGSM)通过最大化损失函数一步生成对抗样本。迭代快速梯度符号法(Iterative FGSM, IFGSM)在FGSM方法的基础上实现可变的迭代方法。投影梯度下降法(Projected gradient descent, PGD)通过随机生成以及多次迭代的方法实现对抗攻击。在本实验中主要针对Transformer不同模块的鲁棒性进行测试。

使用词错率(Word error rate,WER)进行评价对抗样本性能,它是一种用于评估自动语音识别系统性能的度量指标。WER是通过比较识别结果和参考文本之间的差异来计算,衡量的是识别结果中与参考文本不匹配的词的数量。对抗攻击后WER越大说明该对抗样本性能越好。

2.2　性能对比

使用上述实验设置,将wav2vec2作为代理模型,WavLM作为目标模型进行测试,通过对模型中不同中间层进行攻击,从而生成通用语音对抗样本的迁移性能,结果如表1所示。从表1可以看出,本文提出的方法与FGSM迭代方法结合对比UAP方法词错率有较大提升,说明本文方法生成的通用对抗样本在迁移到另一个Transformer语音识别模型上,仍有较大的对抗攻击能力,即迁移性强。此外,通过对模型中不同中间层进行攻击,从而生成通用语音对抗样本的迁移性能,结果如表2所示。其中,Transformer_1_output、Transformer_5_output和Transformer_12_output分别表示使用中间层攻击方法攻击模型中Transformer模块中第1层、第5层和第12层的结果。可以看出,不同层的鲁棒性不同,其扰动后结果也不同。Transformer_1_ attention、Transformer_5_attention和Transformer_12_attention分别表示使用中间层攻击方法攻击模型中Transformer模块自注意力机制第1层、第5层和第12层的结果。相比直接攻击模块输出,攻击自注意力层效果更好。Weight_all_attention表示引入本文提出的基于扰动梯度的权重对每层注意力层进行加权扰动的方法,可以看出其迁移后攻击成功率有较大提升。最后,$\alpha=0$和$\alpha=0.1$分别表示在反向传播中控制

表1　不同攻击方式的WER对比

Table 1　WER comparison of different attack methods　%

方法	WER
Without attack	5.00
UAP_PGD	53.01
UAP_I-FGSM	55.41
UAP_FGSM	59.00
Proposed_FGSM	**80.61**

注意力梯度的权重分别设置为0和0.1的情况。可以看出,控制注意力梯度可以有效提升通用对抗样本对于语音识别模型的迁移性和攻击成功率。并且,当$\alpha=0$时,即完全舍弃注意力层的梯度,其性能并不是最佳的,这可能是由于反向传播梯度不连贯造成的影响。而当$\alpha=0.1$时,既能减弱对于样本语义的学习,又能保证反向传播的连贯性,其性能最佳。

表2 不同参数攻击的WER对比

Table 2 WER comparison of attacks with different parameters %

方法	WER
Transformer_1_output	60.01
Transformer_5_output	58.11
Transformer_12_output	63.32
Transformer_1_attention	63.04
Transformer_5_attention	59.25
Transformer_12_attention	66.08
Weight_all_attention	71.03
Weight_all_attention $\alpha=0$	76.27
Weight_all_attention $\alpha=0.1$	**80.61**

3 结束语

本文针对通用语音对抗样本在语音识别模型中的迁移性问题,提出了面向Transformer类语音识别模型的特殊攻击方法。通过利用Transformer模型结构的相似性对模型中间层特征进行攻击,提高了通用语音对抗样本的可迁移性。同时,在反向传播过程中控制注意力梯度大小,能够引导模型关注特定的特征,从而生成更具迁移性的对抗样本。实验结果表明,本文所提出的方法显著提高了模型之间的迁移性,使对抗样本在不同模型上都能实现高攻击成功率。此外,本文所提出的方法可以与现有的高迁移方法相结合,以进一步提升通用语音对抗样本的迁移性和攻击成功率。

参考文献:

[1] 张雄伟,张星昱,孙蒙, 等. 说话人验证系统攻击方法的研究现状及展望[J]. 数据采集与处理, 2021, 36(5): 831-849.
ZHANG Xiongwei, ZHANG Xingyu, SUN Meng, et al. Attack methods in speaker verification system: The state of the art and prospects[J]. Journal of Data Acquisition and Processing, 2021, 36(5): 831-849.

[2] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[C]//Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015). San Diego, CA, USA: OpenReview.net, 2015: 1-11.

[3] VAIDYA T, ZHANG Y, SHERR M, et al. Cocaine noodles: Exploiting the gap between human and machine speech recognition[C]//Proceedings of the 9th USENIX Workshop on Offensive Technologies (WOOT). Washington, DC, USA: USENIX Association, 2015: 10-11.

[4] YUAN X, HE P, ZHU Q, et al. Adversarial examples: Attacks and defenses for deep learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824.

[5] YUAN X, CHEN Y, ZHAO Y, et al. CommanderSong: A systematic approach for practical adversarial voice recognition [C]// Proceedings of the 27th USENIX Security Symposium (USENIX Security 18). Baltimore, MD, USA: USENIX Association, 2018: 49-64.

[6] ZHANG X, ZANG X, ZOU X, et al. Towards generating adversarial examples on combined systems of automatic speaker verification and spoofing countermeasure[J]. Security and Communication Networks, 2022(1): 1-12.

[7] CISSE M, ADI Y, NEVEROVA N, et al. Houdini: Fooling deep structured visual and speech recognition models with adversarial examples[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS 2017). Long Beach, CA, USA: Curran Associates, 2017: 6980-6990.

[8] ZHANG Y K, JIANG Z Y, VILLALBA J, et al. Black-box attacks on spoofing countermeasures using transferability of adversarial examples[C]//Proceedings of the 21st International Conference of the International Speech Communication Association (Interspeech 2020). Shanghai, China: ISCA, 2020: 4238-4242.

[9] CHEN Y, YUAN X, ZHANG J, et al. Devil's whisper: A general approach for physical adversarial attacks against commercial black-box speech recognition devices[C]//Proceedings of the 29th USENIX Security Symposium (USENIX Security 20). Boston, MA, USA: USENIX Association, 2020: 2667-2684.

[10] MA C, CHEN L, YONG J H. Simulating unknown target models for query-efficient black-box attacks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2021: 11835-11844.

[11] KREUK F, ADI Y, CISSE M, et al. Fooling end-to-end speaker verifi- cation with adversarial examples[C]//Proceedings of 2018 IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2018: 1962-1966.

[12] CARLINI N, WAGNER D, COMMUNICATION N A B T, et al. Audio adversarial examples: Targeted attacks on speech-to-text[C]//Proceedings of 2018 IEEE Security and Privacy Workshops. [S.l.]: IEEE, 2018: 1-7.

[13] HARIK G R, LOBO F G, GOLDBERG D E. The compact genetic algorithm[J]. IEEE Transactions on Evolutionary Computation, 1999, 3(4): 287-297.

[14] TAORI R, KAMSETTY A, CHU B, et al. Targeted adversarial examples for black box audio systems[C]//Proceedings of 2019 IEEE Security and Privacy Workshops. [S.l.]: IEEE, 2019: 15-20.

[15] QU X, WEI P, GAO M, et al. Synthesising audio adversarial examples for automatic speech recognition[C]//Proceedings of the 8th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD 2022). Washington, DC, USA: ACM, 2022: 1430-1440.

[16] NEEKHARA P, HUSSAIN S, PANDEY P, et al. Universal adversarial perturbations for speech recognition systems[C]//Proceedings of the 20th Annual Conference of the International Speech Communication Association (Interspeech 2019). Graz, Austria: ISCA, 2019: 481-485.

[17] ABDOLI S, HAFEMANN L G, RONY J, et al. Universal adversarial audio perturbations[EB/OL]. (2019-08-08). https://arxiv.org/abs/1908.03173.

[18] XIE Y, SHI C, LI Z H, et al. Real-time, universal, and robust adversarial attacks against speaker recognition systems[C]//Proceedings of ICASSP 2020—2020 IEEE International Conference on Acoustics, Speech and Signal Processing. [S.l.]: IEEE, 2020: 1738-1742.

[19] XIE Y, LI Z, SHI C, et al. Enabling fast and universal audio adversarial attack using generative model[J].Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(16): 14129-14137.

[20] HE B, JIA X, LIANG S, et al. SA-attack: Improving adversarial transferability of vision-language pre-training models via selfaugmentation[EB/OL]. (2023-12-08). https://arxiv.org/abs/2312.04913.

[21] LU D, WANG Z, WANG T, et al. Set-level guidance attack: Boosting adversarial transferability of vision-language pre-training models[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. [S.l.]: IEEE, 2023: 102-111.

[22] LIN J, SONG C, HE K, et al. Nesterov accelerated gradient and scale invariance for adversarial attacks[C]//Proceedings of the 8th International Conference on Learning Representations (ICLR 2020). Addis Ababa, Ethiopia: OpenReview.net, 2020: 1-14.

作者简介:

王振(1999-),男,博士研究生,研究方向:语音对抗样本, E-mail:wangzhenhit-wh@163.com。

韩纪庆(1964-),**通信作者**,男,教授,研究方向:语音信号处理、音频信息分析,E-mail: jqhan@hit.edu.cn。

何勇军(1980-),男,教授,研究方向:语音信号处理、图像处理,E-mail: holy-wit@163.com。

郑铁然(1972-),男,教授,研究方向:语音信号处理、音频信息处理,E-mail: zhengtieran@hit.edu.cn。

郑贵滨(1973-),男,副教授,研究方向:音频信号处理、音频检索,E-mail: zhengguibin@ hit.edu.cn。

(编辑:刘彦东)

Universal Adversarial Example Generation Method with High Transferability for Transformer-Based Speech Recognition Models

WANG Zhen, HAN Jiqing*, HE Yongjun, ZHENG Tieran, ZHENG Guibin

(College of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China)

Abstract: In recent years, the emergence of the Transformer model has significantly enhanced the accuracy of automatic speech recognition technology. This research aims to address the critical security vulnerabilities in Transformer-based automatic speech recognition systems by enhancing the transferability of universal speech adversarial examples. While Transformer models have significantly advanced speech processing, their susceptibility to universal adversarial perturbations remains a major concern. To exploit these weaknesses effectively, we propose a novel attack framework that leverages the structural commonalities of Transformer architectures. First, we implement a feature-level disruption strategy that maximizes the dissimilarity between perturbed and original speech within the middle-layer representations. By altering these latent representation patterns, the attack successfully shifts the internal decision boundaries of models. Second, given that sample-dependent semantic information often inhibits the generalization of universal noise, we introduce an attention gradient control mechanism. This mechanism strategically weakens the gradients associated with semantic context features, forcing the perturbation to capture underlying, sample-independent acoustic vulnerabilities instead. Finally, experimental evaluations conducted on LibriSpeech demonstrate the superior performance of the proposed method. The results indicate that our approach achieves an average word error rate of 80.6% across multiple target models, representing a 36.6% improvement in transferability compared to existing baseline universal attacks. These findings conclude that the targeted manipulation of middle-layer features combined with the suppression of semantic dependencies is a highly effective strategy for cross-model adversarial threats.

Highlights:

1. Propose a novel framework of universal speech adversarial attacks that maximizes middle-layer feature dissimilarity to exploit the structural similarities inherent in Transformer-based speech recognition models.
2. Introduce a targeted attention gradient control mechanism to decouple sample-independent acoustic features from sample-dependent semantic context, significantly boosting attack transferability.
3. Achieve a substantial increase in universal attack success rates across diverse Transformer architectures, outperforming traditional universal perturbation methods.

Key words: speech recognition; adversarial examples; black-box attack; attention mechanism

Received: 2024-03-30; **Revised:** 2024-10-12

***Corresponding author**, **E-mail:** jqhan@hit.edu.cn.