计算机研究与发展
Journal of Computer Research and Development
DOI: 10.7544/issn1000-1239.202221034
61(3): 685−700, 2024

基于帧结构的语音对抗样本重点区域扰动分析

韩松莘[1,2] 郭松辉[1,2] 徐开勇[1,2] 杨 博[1] 于 淼[1]
[1](战略支援部队信息工程大学 郑州 450001)
[2](河南省信息安全重点实验室(战略支援部队信息工程大学) 郑州 450001)
(thordlos@foxmail.com)

Perturbation Analysis of the Vital Region in Speech Adversarial Example Based on Frame Structure

Han Songshen[1,2], Guo Songhui[1,2], Xu Kaiyong[1,2], Yang Bo[1], and Yu Miao[1]
[1](*Strategic Support Force Information Engineering University*, *Zhengzhou* 450001)
[2](*Henan Key Laboratory of Information Security* (*Strategic Support Force Information Engineering University*), *Zhengzhou* 450001)

Abstract At present, adversarial attacks on speech recognition models have typically involved adding noise to the entire speech signal, resulting in a wide perturbation range and introducing high-frequency noise. Existing research has attempted to reduce the perturbation range by designing optimization targets. However, controlling the transcription result requires adding perturbations to each frame, thus limiting further reduction in perturbation range. To address this issue, we propose a novel approach that examines the feature extraction process of speech recognition systems from a frame structure perspective. The study finds that framing and windowing determine the distribution of critical regions within the frame structure. Specifically, the weight of adding perturbation to each sampling point within the frame is influenced by its location. Based on the results of perturbation analysis on input features, we partition regions with shared attributes. Then we propose the adversarial example space measurement method and evaluation index to quantify the weight of sampling points for adversarial examples generation. We conduct cross-experiments by adding perturbations at different intervals within the frame, which enables us to identify key regions for perturbation addition. Our experiments on multiple models demonstrate that adding adversarial perturbation to vital regions can narrow the perturbation range, and provide a new perspective for generating high-quality audio adversarial examples.

Key words speech recognition; adversarial attack; input feature; perturbation analysis; adversarial example space metric

摘 要 目前针对语音识别模型的对抗攻击主要是在整条语音上添加噪声,扰动范围大且引入了高频噪声.现有研究在一定程度上缩小了扰动范围,但由于语音对抗攻击需要在每帧添加扰动实现对转录结果的控制,限制了扰动范围的进一步降低.针对此问题,从帧结构的角度研究了语音识别系统中的特征提取流程,发现分帧和加窗处理决定了帧结构中重点区域的分布,即帧内各采样点上添加扰动的重要性受采样点所处位置的影响.首先,根据对输入特征的扰动分析结果进行区域划分;然后,为了量化这些采样点对求解对抗样本的重要性,提出了对抗样本空间度量方法和相应的评价指标,并设计了在帧内不同区间上添加扰动的交叉实验,进而确定了扰动添加的重点区域;最后,在多个模型上进行了广泛的实验,表明

收稿日期: 2022-12-20; **修回日期**: 2023-05-04
基金项目: 国家自然科学基金项目(62176265)
This work was supported by the National Natural Science Foundation of China (62176265).
通信作者: 郭松辉(songhui.guo@outlook.com)

了在重点区域添加对抗扰动能够缩小扰动范围,为高质量语音对抗样本的生成提出新的角度.

关键词 语音识别;对抗攻击;输入特征;扰动分析;对抗样本空间度量

中图法分类号 TP309; TP391

基于深度学习的自动语音识别(automatic speech recognition, ASR)系统[1]能够将语音准确翻译为文本信息,深刻改变了人机交互方式[2].在智能家居或自动驾驶等交互场景中,ASR 系统接收语音并将其解释为相应的命令,为人们控制智能设备带来诸多便利.然而,目前主流的 ASR 系统已被发现存在潜在的安全隐患[3-4],即攻击者在语音中添加精心构造的扰动,将其以广播或播报形式播放[5],能够使目标设备执行恶意命令[6],严重威胁着受害者的隐私安全甚至人身安全.

前人的工作提出了一系列针对深度神经网络的攻击方案.这些攻击以损失[7](*loss*)或适应度(*fitness*)函数值[8]为目标进行迭代优化,将生成的特殊扰动添加到原始语音上,改变语音识别模型对语音的转录结果,实现对 ASR 系统的攻击.已有研究证明,如果攻击者能够完全获取目标模型的网络参数(白盒攻击),则能以接近 100% 的攻击成功率[9]使目标 ASR 将一段语音转录为攻击者设置的任意文本.当前,在语音识别领域,对于对抗攻击的研究主要分布在降低扰动感知度[10-12]、实时扰动[13]、通用扰动[14]等方向.现有的语音对抗攻击通常在整条语音上添加扰动,引入了高频噪声,易被人耳察觉.而降低扰动感知度研究的普遍做法是设计优化目标,将对抗扰动集中到人类不易注意到的音频区域[11, 15-16](听觉掩蔽区域).但是,这样做一方面会降低对抗样本的鲁棒性,另一方面计算过程依赖输入语音,不能适用于生成通用对抗扰动.为了解决该问题,Liu 等人[9]将扰动点的数量因素引入到音频对抗样本的生成中,限制在部分采样点上添加扰动,将扰动比例降低至 75%,但受限于 ASR 模型转录的上下文特征,能够降低的采样点比例有限,且没有给出采样点选取策略.

针对扰动范围难以进一步降低的问题,本文分析了 ASR 网络模型的特征提取过程[17-18],发现每帧语音中,在不同位置上添加扰动能够对特征造成不同程度的影响,这些影响能够通过正向传播改变神经网络的决策.通过分析特征提取中对生成对抗样本的影响因素,可以筛选出对于生成对抗样本更重要的采样点[19],从而进一步降低扰动范围.基于深度学习的语音识别系统通用框架如图 1 所示.

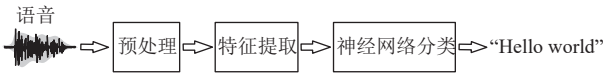

Fig. 1 General framework for ASR systems

图 1 ASR 系统的通用框架

预处理模块对原始语音进行剪切、滤波操作,以消除语音信号的静默和突兀噪声部分;特征提取模块中 ASR 系统将语音信号分帧,并以帧为单位提取信号特征,特征类型包含梅尔频率倒谱系数[20](mel-scale frequency cepstral coefficients, MFCC)、FBank[21](filter bank)特征和自动提取的高维特征等;神经网络对这些特征进行分类后,ASR 系统将每帧的分类结果组合解码,最后输出语音信号对应的转录文本.经典 ASR 系统如 DeepSpeech[22], Kaldi[23]等提取原始语音的 MFCC 特征, Lingvo[24]提取 FBank 特征,洪青阳等人[25]总结了上述特征的计算关系.如图 2 所示,本文依据 ASR 系统中的计算对图进行了简化修改.

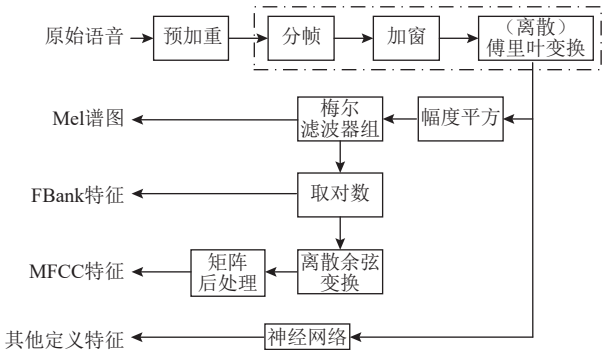

Fig. 2 General feature extraction process in ASR system

图 2 ASR 系统中一般特征提取流程

上述定义特征的共同点在于对原始语音进行分帧、加窗处理,然后以帧为单位进行离散傅里叶变换(discrete Fourier transform, DFT),以准确提取频域信息.如图 3 所示,分帧通常采用交叠分段方法,保证相邻 2 帧间相互重叠一部分,使得帧与帧之间能平滑过渡.加窗即每帧乘以一个窗口函数,增加每帧头尾端的连续性,减少频谱泄漏.在主流的 ASR 系统实现中,多采用汉宁窗.

分帧和加窗操作将导致帧内不同区域采样点对计算离散傅里叶变换的贡献是不等价的,因此在各点上添加扰动对频域信息的影响也不均衡.主要体现在:1)在分帧结构的非重叠区间上添加扰动只会

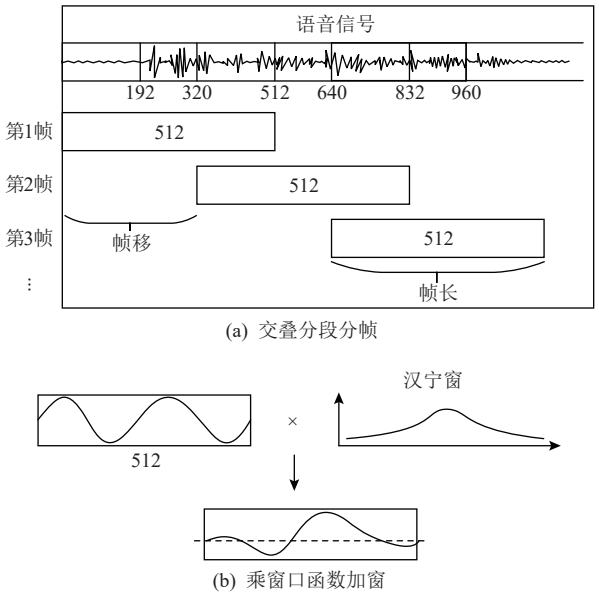

Fig. 3 Diagram of framing and windowing

图 3 分帧与加窗处理示意图

直接影响单帧频域信息，而在重叠区间添加扰动会同时改变包含该重叠区间的相邻 2 帧的频域信息；2）帧片段和窗口函数相乘将导致同等扰动水平下，在帧内中间区域添加扰动比在头尾两端添加扰动对 DFT 的影响更大.

当前语音对抗样本研究中，在反向传播更新对抗性扰动阶段，均在整条语音范围内添加扰动[4]，而没有考虑到上述特性. 为了进一步降低扰动范围，本文研究的主要问题包含：1）帧内不均衡结构存在于 DFT 的计算过程中，直接影响的是神经网络的输入，是否能通过神经网络影响语音识别的转录结果；2）要将扰动范围限制在重点区域的采样点上，需要分析上述单类影响因素叠加时对语音识别结果的综合影响，并给出其分布规律.

本文分别对分帧、加窗进行理论分析，提出单因素影响下添加扰动的位置与求解对抗样本之间的规律. 在此基础上，根据各影响因素在帧内的分布设计了交叉实验，并提出度量方法和相应的评价指标：将潜在可求解的对抗样本规模定义为对抗样本空间，并以白盒攻击方式攻击目标模型，基于语音对抗攻击扰动幅值和求解难度反相关的特性，以条件衰减的方式对对抗性扰动进行迭代和优化，通过统计不同幅度水平下成功攻击的次数，作为对对抗样本空间的近似度量. 我们在 LibriSpeech 数据集[26] 上对交叉试验组进行测试，实验结果证明了对抗样本空间随耦合作用、位置权重、区间长度变化的一般规律，并提出了最重要的扰动范围分布，约占总采样点的 40%. 另外，我们在讨论中证明了在不受耦合作用影响时，对抗样本空间和位置权重正相关. 最后，讨论了本文提出的方法用于度量对抗样本空间时的客观性.

本文的主要贡献包括 3 个方面：

1）完成了分帧、加窗结构中单个影响因素扰动 DFT 特征的理论分析，提出了分帧结构下耦合作用导致对抗样本空间缩减，加窗结构下对抗样本空间和位置权重正相关的分布规律.

2）研究了对序列模型的对抗样本空间度量，提出了基于扰动水平迭代衰减的对抗样本空间度量方法和相应的评价指标，以探索复合因素作用下对抗样本空间随扰动位置的分布规律.

3）根据 ASR 中的分帧类型，设计了限制扰动范围的交叉实验. 以降低对抗样本上的整体噪声能量为目标，提出了基于帧结构的重点区域扰动范围. 通过在多个模型上进行实验验证，证明了帧同步结构的模型中对抗样本空间主要受耦合作用影响而缩减，为高质量语音对抗样本的生成提出新的角度.

1 相关工作

根据扰动作用阶段，将相关工作分为对定义特征的扰动分析，以及添加扰动对神经网络的影响分析，并介绍了它们的相关应用.

1.1 定义特征的扰动分析研究

早期，在神经网络研究和算力水平发展薄弱的阶段，主流的语音识别工具采用特征提取和模式识别方法将语音转录为文本. 该阶段对特征扰动分析的研究目的集中在提升识别的准确率和对噪声的鲁棒性. Breithaupt 等人[27] 通过对 DFT 特征进行扰动分析，发现基于模式识别的 ASR 对方差较大的噪声更敏感，并应用特征平滑来降低该误差. Ravindran 等人[28] 在特征提取前对信号进行低通滤波，减小重叠影响，并在信道中产生更平滑的包络信号，从而提升了 MFCC 的噪声鲁棒性. 针对分帧、加窗结构，文献 [29] 通过将窗口函数导致的加权最优解和迭代求解到的帧内局部最优解相结合，解决助听器噪声消除最优解问题. 该方案证明了加权重叠相加（weighted overlap-add, WOLA）结构本身不会对语音去噪能力造成严重限制，通过在 FFT（fast Fourier transform）域应用迭代方案，可以计算出 WOLA 权重并据此产生在声学上无法与干净语音区分的滤波噪声. 从而，本文将对抗性扰动

噪声集中至权重较高的重点区域也具备一定的可行性, 但神经网络的重点区域权重计算方法与模式识别不同, 仍需探讨神经网络模型中输入重点区域的分布规律. 文献 [30] 研究了每个 MFCC 特征提取步骤的影响, 分析出 MFCC 特征向量的输入信噪比(SNR)与输出扰动界之间的关系, 通过实验验证: 即使在输入信号中添加信噪比约等于 0 的扰动, 其频谱覆盖率也能达到 98% 以上.

1.2 降低对抗样本感知度研究

随着对深度学习的进一步研究, 基于深度神经网络的 ASR 大幅度提升了识别准确率, 但同时引入了新的安全风险. 文献 [31] 指出深度神经网络易被添加在原始数据中的微小扰动影响而做出错误分类, 这种错误被攻击者利用后能够执行带有恶意意图的目标攻击. 在语音识别中, 由于对抗攻击对扰动后每帧的转录结果都有一定要求, 目前所有针对语音识别的目标攻击主要基于迭代优化方式进行求解[4, 32]. 而降低扰动感知度的研究主要通过设计优化目标, 将对抗扰动集中至人耳不易感知到的频域内. 文献 [11, 16] 提出一种优化目标, 利用心理声学掩蔽和频率掩蔽现象, 将对抗性扰动集中至人耳不易注意到的区域, 从而降低扰动感知度, 但仍在语音数据的全局范围内添加了噪声, 且增加了迭代所需时间. Eisenhofer 等人[15] 反向利用了掩蔽原理, 使求解到的对抗样本极易被人耳感知, 从而提升模型对对抗攻击的鲁棒性.

文献 [33] 分析了语音特征提取流程中的 MFCC 计算过程, 在 MFCC 特征向量中生成对抗性噪声, 并将其注入到语音数据中, 具有扰动感知小且生成速度快的优点. Abdullah 等人[34] 分析了语音特征提取流程中的 DFT 计算过程, 通过删除其中强度低于设定阈值的分量, 并利用反变换从剩余的分量中构建一个新的语音, 以较小的扰动实现对转录结果的修改. 文献 [33−34] 的共同点在于扰动感知小且生成速度快, 然而, 这些方法只能用于无目标攻击. Liu 等人[9] 认为, 将扰动集中至某一频率或某一时间段内会破坏对抗样本的鲁棒性, 因此提出了采样点攻击, 限制只在部分语音采样点上添加扰动, 最高将扰动范围降低至 75%. 本文结合对 DFT 特征的扰动分析, 探索添加对抗扰动的重点区域分布规律, 在重点区域上添加扰动, 以进一步降低该扰动比例.

2 攻击模型

本文攻击方法的目标是探索不同扰动范围下对抗样本的潜在求解规模, 即对抗样本空间. 最终求解的对抗性扰动幅度越小, 对抗样本的质量越高. 但由于优化算法的效率限制, 难以求解全局最优解, 本文以有限次数迭代中的最优结果进行对抗样本空间的相对比较. 为了减小误差, 我们采用白盒攻击的方式, 在完全访问目标模型网络参数的条件下进行对抗样本生成, 更新扰动的示意图如图 4 所示.

针对原始语音“Set alarm”, 要添加对抗性扰动使目标模型将其识别为“Open the door”. 在正向传播阶段, 攻击者首先向目标模型查询原始语音到目标转录的梯度[35], 这个过程需要访问模型的网络参数和 *loss* 值[36] 来计算 *loss* 减小的梯度信息. 在反向传播阶段, 攻击者利用梯度信息更新对抗性扰动, 并将部分对抗性扰动添加到原始语音上, 添加扰动的范围选择见实验设计部分. 该方法从梯度信息到对抗性扰动的计算规则较简单, 且不考虑到 LSTM 网络中每帧添加扰动后模型决策结果 ***logits*** 的改变对后续帧

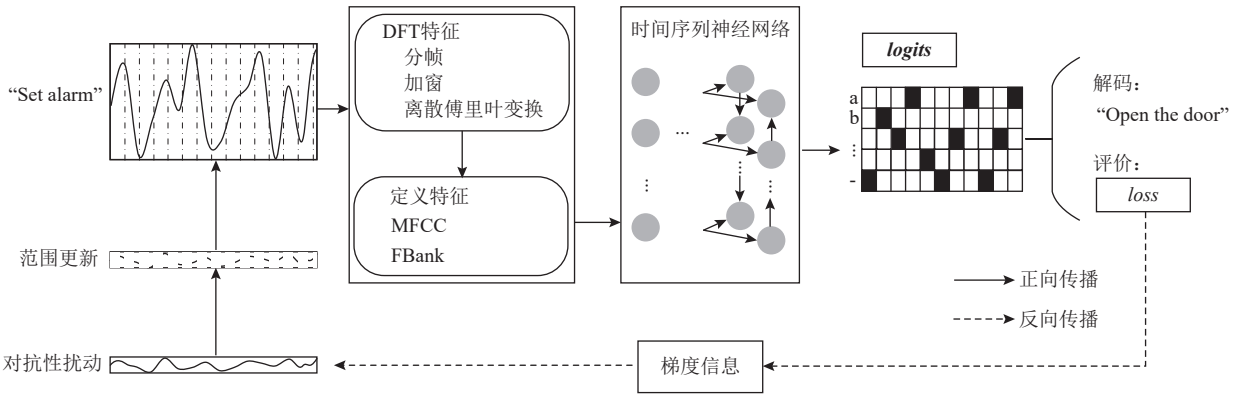

Fig. 4 Flow diagram of the attack model

图 4 攻击模型流程图

的影响，难以经过单次迭代实现目标攻击，所以我们设置短步、多次迭代的策略进行对抗样本求解.

3 单影响因素扰动分析

为了充分利用语音的短时平稳特性，语音信号的特征提取方法中普遍包含由分帧、加窗和离散傅里叶变换组成的短时分析技术. 分帧结构中存在的帧间层叠，加窗结构中所乘窗口函数的曲线随位置变化，使得同一段扰动添加在帧内不同位置时，能在不同程度上影响该帧的短时分析结果，从而导致神经网络对该帧及相邻帧的识别结果发生变化. 为了界定出对于求解对抗样本最重要的扰动区域，本文首先对特征提取流程进行扰动分析.

3.1 帧重叠分析

事实上，ASR 中的分帧方式可以被分为 2 类，我们分别称为 Ⅰ 类分帧和 Ⅱ 类分帧，它们的主要区别在于重叠区间的分布不同. 其中，Ⅰ 类分帧方式存在非重叠区间，相邻帧间的相关性较小，减少了后续特征提取与神经网络分类的计算量. 为了描述方便，以帧移为单位，根据重叠程度的不同，本文将整条语音分为甲、乙 2 类扰动区间. 如图 5 所示，Ⅰ 类分帧方式中帧重叠比例$\mu<0.5$，帧移较长，存在部分区间乙，其中的采样点只被用来计算单帧的 DFT 特征. 在原始语音上添加扰动时，如果扰动范围属于区间乙，则该扰动直接影响区间所属帧的 DFT 特征；而扰动范围属于重叠区间甲时，会同时影响相邻 2 帧的 DFT 特征.

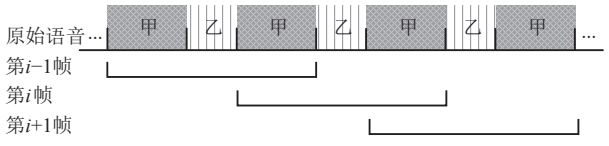

Fig. 5 Interval categories of class-Ⅰ framing

图 5 Ⅰ类分帧的区间种类

这种相邻帧之间共用部分数据的情形属于外部耦合，本文将在重叠范围上添加扰动同时影响多帧 DFT 特征的现象称作耦合作用. 并有理由相信，发生在输入空间的影响能够通过神经网络，对求解对抗样本造成影响. 分析如下：生成对抗样本即求解让神经网络做出目标误分类的理想最小扰动，我们以$C(\cdot)$表示神经网络分类器，σ表示激活函数，$\boldsymbol{w}$，$\boldsymbol{b}$分别表示神经网络的权重和偏置，在分析过程中忽略序列模型神经网络 ***logits*** 中前一帧决策结果对后一帧的影响. 以相邻 2 帧上对抗样本的求解为例，$\boldsymbol{s}_i$代表第i帧的原始语音信号，$\boldsymbol{\delta}_i$为添加在第i帧的局部扰动，t_i表示神经网络对第i帧识别结果的目标分类. 在乙区间上添加扰动，即求解

$$\begin{cases} C(\sigma(\boldsymbol{w}(\boldsymbol{s}_{i+1}+\boldsymbol{\delta}_{i+1})+\boldsymbol{b}))=t_{i+1}, \\ C(\sigma(\boldsymbol{w}(\boldsymbol{s}_i+\boldsymbol{\delta}_i)+\boldsymbol{b}))=t_i. \end{cases} \tag{1}$$

不考虑序列模型的帧间影响，式 (1) 可理解为分别求解 2 个分类任务中的对抗样本，其解空间互不影响. 而在相邻 2 帧的重叠区间甲上添加扰动，即求解式 (2)：

$$\begin{cases} C(\sigma(\boldsymbol{w}(\boldsymbol{s}_{i+1}+\boldsymbol{\delta}_i)+\boldsymbol{b}))=t_{i+1}, \\ C(\sigma(\boldsymbol{w}(\boldsymbol{s}_i+\boldsymbol{\delta}_i)+\boldsymbol{b}))=t_i. \end{cases} \tag{2}$$

扰动$\boldsymbol{\delta}_i$需满足使相邻 2 帧同时实现目标攻击，即求解当前区间上使各自帧实现目标攻击对抗性扰动的交集，从而导致解空间的缩减.

Ⅱ 类分帧方式中的所有区间都是重叠区间，但重叠程度有所差异. 我们同样以帧移为单位，按重叠程度将其分为甲、乙区间. 如图 6 所示，Ⅱ 类分帧中重叠比例$\mu>0.5$且帧移较小，可以跟踪语音信号的连续性，并且不会遗漏帧边缘处的突然变化.

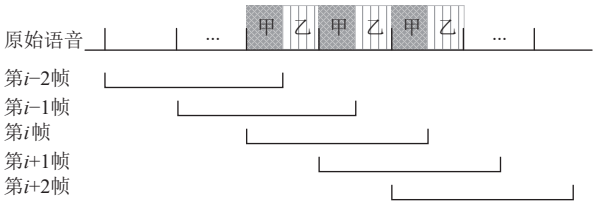

Fig. 6 Interval categories of class-Ⅱ framing

图 6 Ⅱ类分帧的区间种类

由于所有的重叠区间都是帧移的一部分，以帧移为单位划分扰动区间可以不考虑重叠关系，并扩展至整个语音序列. 在以上 2 类分帧方式中，以帧移为单位的区间划分总结如表 1 所示.

Table 1 Practical Interval Categories and Characteristics Summary

表 1 实际区间类型及特点总结

区间类型	Ⅰ 类分帧	Ⅱ 类分帧
甲	重叠区间 受耦合作用影响	非重叠区间 不受耦合作用影响
乙	重叠区间 受耦合作用影响最大	重叠区间 受耦合作用影响

在重叠区间上扰动，首先对相邻帧的 DFT 特征产生直接影响，然后经神经网络的前向传播改变模型决策. 后续实验将会证明，特征提取结构对神经网络的输入产生的影响能够作用于对抗样本空间.

3.2 位置权重分析

在本节分析中，我们定义符号上标表示特征类型，下标表示区间范围. 语音信号被读取到数字空间后以离散数值形式存储，定义原始语音信号为$\boldsymbol{s}(n)$，扰动信号为$\boldsymbol{\delta}(n)$，n为采样点序号，则对抗样本

$$\boldsymbol{x}(n)=\boldsymbol{s}(n)+\boldsymbol{\delta}(n). \tag{3}$$

在特征提取过程中，预加重操作能够在一定程度上弥补高频部分的损耗，提升模型识别准确率，因而其在音频特征提取中被广泛应用，在时域上对抗样本的预加重为

$$\boldsymbol{x}^{\mathrm{p}}(n)=\boldsymbol{x}(n)-\alpha\boldsymbol{x}(n-1), \tag{4}$$

其中滤波器系数α是一个常数，且$0.9<\alpha<1$. 随后，对抗样本被分帧、加窗. 为了区分重叠部分和非重叠部分，本文定义符号为：帧移N；重叠比例μ；单帧长度$N+\mu N$；第i帧信号$\{\boldsymbol{x}^{\mathrm{p}}_{i,\,[0,N]}(n),\ \boldsymbol{x}^{\mathrm{p}}_{i+1,\,[0,\mu N]}(n)\}$. 第$i$帧信号由帧移$\boldsymbol{x}^{\mathrm{p}}_{i,\,[0,N]}(n)$和重叠(overlap)$\boldsymbol{x}^{\mathrm{p}}_{i+1,\,[0,\mu N]}(n)$两部分拼接而成. 加窗即每帧信号乘以窗口函数，第$i$帧的汉宁窗特征为

$$\boldsymbol{x}^{\boldsymbol{w}}_{i,\,[0,N+\mu N]}(n)=\begin{cases}\boldsymbol{x}^{\mathrm{p}}_{i}(n)\boldsymbol{w}_{[0,N+\mu N]}(n),\ n\in[0,N],\\ \boldsymbol{x}^{\mathrm{p}}_{i+1}(n)\boldsymbol{w}_{[0,N+\mu N]}(n),\ n\in[N,N+\mu N].\end{cases} \tag{5}$$

这里的窗口函数

$$\boldsymbol{w}_{[0,N+\mu N]}(n)=(1-a)-a\cos\left(\frac{2\pi n}{N+\mu N-1}\right),$$

其中，a为固定常数.

计算对抗样本的第i帧特征时，通过DFT计算频率分量：

$$\boldsymbol{X}_i(k)=\sum_{n=0}^{N+\mu N-1}\boldsymbol{x}^{\boldsymbol{w}}_{i,\,[0,N+\mu N]}(n)\exp\left(\frac{-2\pi nk}{N+\mu N-1}\right). \tag{6}$$

根据DFT的线性性质$\boldsymbol{X}_i(k)=\boldsymbol{X}^{\boldsymbol{s}}_i(k)+\boldsymbol{X}^{\boldsymbol{\delta}}_i(k)$，在对抗样本优化过程中，$\boldsymbol{s}(n)$保持恒定，$\boldsymbol{\delta}(n)$根据梯度信息迭代优化. 因此，$\boldsymbol{X}_i(k)$主要受$\boldsymbol{X}^{\boldsymbol{\delta}}_i(k)$的影响而发生变化：

$$\boldsymbol{X}^{\boldsymbol{\delta}}_i(k)=\sum_{n=0}^{N+\mu N-1}\begin{cases}\boldsymbol{\delta}^{\mathrm{p}}_i(n)\boldsymbol{w}(n)\exp\left(\dfrac{-2\pi nk}{N+\mu N-1}\right),\ n\in[0,N],\\ \boldsymbol{\delta}^{\mathrm{p}}_{i+1}(n)\boldsymbol{w}(n)\exp\left(\dfrac{-2\pi nk}{N+\mu N-1}\right),\ n\in[N,N+\mu N],\end{cases} \tag{7}$$

其中重叠部分$\boldsymbol{\delta}^{\mathrm{p}}_{i+1,\,[0,\,\mu N]}(n)$可看作周期为$N+\mu N$的信号$\boldsymbol{\delta}^{\mathrm{p}}_{i,\,[0,\,\mu N]}(n)$右移$N$位，即$\boldsymbol{\delta}^{\mathrm{p}}_{i+1,\,[0,\,\mu N]}(n)=\boldsymbol{\delta}^{\mathrm{p}}_{i,\,[0,\,\mu N]}(n+N)$. 根据DFT的移位性质$\boldsymbol{\delta}(n+N)\overset{\mathrm{DFT}}{\longleftrightarrow}\boldsymbol{X}(k)\exp\left(\dfrac{-2\pi Nk}{N+\mu N-1}\right)$，且$\dfrac{-2\pi k}{1+\mu}<0$，可知重叠区间扰动对第$i$帧DFT计算结果的影响小于位于左侧的非重叠区间.

为了直观地展示帧内的位置权重差异，我们绘制了权重因子$\boldsymbol{w}_{[0,N+\mu N]}(n)\exp\left(\dfrac{-2\pi Nk}{N+\mu N-1}\right)$关于采样点位置$n$的函数图像，当窗口函数$\boldsymbol{w}$中的固定常数$a$分别设置为0.54(Hamming window)和0.5(Hann window)时，权重因子在帧内的变化趋势如图7所示.

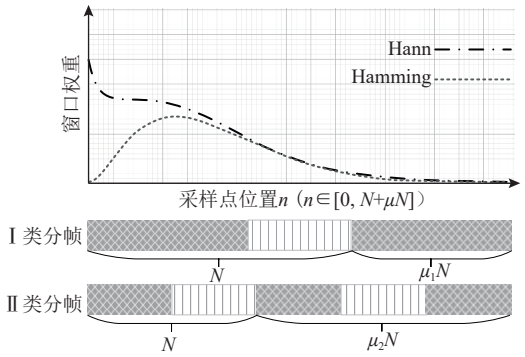

Fig. 7 Variation trend of weight with position in single frame

图7 单帧中权重随位置的变化趋势

在计算所有帧的权重时，扰动中的重叠部分$\boldsymbol{\delta}^{\mathrm{p}}_{i+1,\,[0,\,\mu N]}(n)$用于计算第$i$帧的DFT时，所乘权重在$\left[\dfrac{N}{N+\mu N},\ 1\right]$区间，权重较小；而用于计算第$i+1$帧的DFT时，所乘权重在$\left[0,\dfrac{\mu N}{N+\mu N}\right]$区间，权重较大.

由上述分析可知，耦合作用和窗口权重对同一采样点的影响作用是相反的，它们直接影响DFT特征计算，并扩展至MFCC或FBank等定义特征，这些定义特征作为神经网络的输入特征被进行分类. 在特征提取算法和参数固定后，语音序列中每个采样点对语音特征的贡献将由其位置决定，语音识别系统从中提取主要信息，但对于更精细的对抗性扰动来说，对由位置差异导致的变化更加敏感，根据噪声与定义特征的对应关系，我们划分出重点区域的可能分布. 又因为语音识别神经网络具有非线性及维度高的特点[32]，从输入特征到分类结果的对应关系无法被解析，本文通过实验确定上述因素对求解对抗样本的综合影响.

4 扰动区域评价方法设计

4.1 扰动范围设计

目标ASR模型的网络参数是通过对规模数据执行标准的特征提取流程后，对这些特征训练得到的，在求解对抗样本时，网络参数不再发生变化. 攻击过程中，只有保持和目标ASR相同的特征提取方法和参数，才能保证所求解对抗样本的有效性. 该条件限制了本文在实验设计方面的灵活性，不能通过定制

特征提取过程中的参数[37]来正向验证扰动效果，而只能通过划分不同位置的区间，根据每类位置上对抗样本的潜在求解空间的大小来验证扰动重点区域的分布. 因此，本文设置每步迭代的 DFT 特征计算过程和 ASR 模型中保持一致，通过调整扰动范围来探索影响因素的重要性. 在这种情况下，3 类影响对抗样本空间大小的因素为：

1）耦合作用. 在重叠区间上添加扰动，缩小了对抗样本求解空间.

2）位置权重差异. 在权重较大的区间上添加扰动对 DFT 特征具有更大的能动性，使神经网络的输入有更大的可选择空间.

3）区间长度差异. 当重叠比例μ偏离 0.5 时，甲、乙 2 类区间的长度不相等，在较长的扰动区间上生成对抗样本，对抗样本空间更大.

为了验证上述 3 种影响因素对对抗样本空间的影响，本文通过控制变量分别在 2 类分帧方式上对扰动范围限制设计了定性分析实验，扰动范围如图 8 所示.

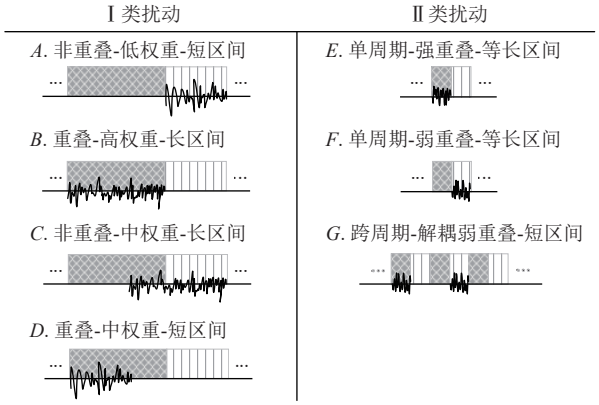

Fig. 8 Perturbation range design for two frame intervals

图 8 对 2 类分帧区间的扰动范围设计

以帧移为基本单位，我们设计了影响因素差异最大的扰动区域. 其中，*A*~*D* 为Ⅰ类分帧下的扰动范围，*A*, *B* 分别代表仅在每个帧移的乙、甲区间内添加扰动；*C*, *D* 分别为组合权重和长度差异的对照试验. *E*~*G* 为Ⅱ类分帧下的扰动范围，*E*, *F* 长度相等，由于单帧内存在多个帧移单位，且甲、乙区间交替重复出现，我们忽略 *E*, *F* 区间的权重大小差异，它们的主要差异在于受耦合作用影响的程度，为了进一步降低耦合作用的影响，我们设计了 *G* 组区域限制实验，如图 9 所示.

通过将扰动范围限制在跨帧移周期上，在单个甲、乙区间上交替添加扰动，由于图 9 中虚线部分所示的位置权重差异，当扰动位于某帧的后半部分时，

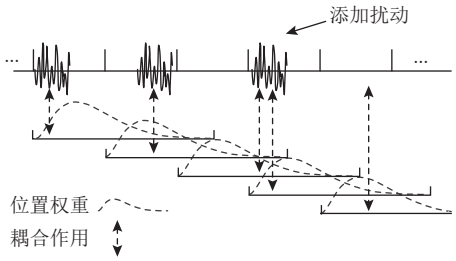

Fig. 9 Decoupling overlapping intervals by reducing the disturbance area

图 9 通过减少扰动区域对重叠区间解耦合

对该帧的影响几乎可以忽略不计，因此每处扰动可被视为只对单帧产生影响，耦合作用比仅在乙区间上添加更低. 但负面影响是每帧中包含的扰动区域降低为 1，扰动区间长度等效缩短.

这些扰动范围以帧移为单位重复，扩展至整个音频，在对抗样本求解过程中，只在图中扰动波形部分更新扰动，其余区域扰动大小固定置 0，表 2 对比了各类扰动范围上的影响因素分布.

Table 2 Experiment Design of Perturbation Range Constraints

表 2 扰动范围限制实验设计

影响因素	Ⅰ类分帧				Ⅱ类分帧		
	A	*B*	*C*	*D*	*E*	*F*	*G*
耦合作用	/	◎	/	◎	●	◎	/
位置权重	●	○	◎	○	○	/	◎
区间长度	◎	○	○	◎	/	/	◎

注：“○,◎,●”分别表示各因素在同类分帧方式中比较时对对抗样本空间的提升、缩减、大幅度缩减作用；“/” 表示不影响.

4.2 扰动生成

为了充分探索不同扰动区域上蕴含的潜在对抗样本空间大小，本文在白盒攻击场景下求解对抗样本. 为了模拟每条原始语音生成对抗样本的平均能力，我们为每条语音随机选取转录目标进行攻击. 针对 ASR 的目标攻击要使得所有帧的分类结果解码后满足目标语句，需要多次迭代计算梯度，每次迭代时通过梯度下降和反向传播算法更新对抗性扰动. 传统攻击中，梯度下降的优化目标[38]通常设置为

$$\min \ell_{\text{metric}}(\boldsymbol{\delta}) + c\ell_{\text{model}}(\boldsymbol{x}+\boldsymbol{\delta}, t), \tag{8}$$

其中$\ell_{\text{model}}(\cdot)$是目标 ASR 模型的损失函数，$\ell_{\text{metric}}(\cdot)$度量并限制对抗样本和原始语音之间的差异. 目前语音对抗样本领域对$\ell_{\text{metric}}(\cdot)$计算方法进行了各种探索：Carlini 等人[7]采用失真分贝$dB_x(\boldsymbol{\delta})$来描述$\boldsymbol{\delta}$的扰动水平，并将其添加到损失函数中作为优化目标，以降低对抗性扰动$\boldsymbol{\delta}$引起的失真；Liu 等人[9]分别计算了基

于全变分降噪(total variation denoising, TVD)正则化等3种扰动度量方法，并比较其对信噪比、$dB_x(\boldsymbol{\delta})$和攻击成功率等指标的影响. 这些正则化项均在成功执行攻击之外引入了额外的优化目标，以降低扰动大小，而本文的主要研究目标在于探索具有天然优势的扰动范围，以此为基础减少扰动点的个数. 这种情况下设置额外优化目标进行求解，将不能客观反映出限制扰动范围对对抗样本空间的影响.

为了探索耦合作用、权重因子对对抗样本空间的叠加影响，本文不设置$\ell_{\text{metric}}(\,)$，如式(9)所示，优化目标仅设置为当前语音到目标语句的损失值：

$$\min \ell\left(\boldsymbol{x}+\beta^k\boldsymbol{\delta}_n,\ t\right), \tag{9}$$

其中$\ell(\cdot)$为目标模型采用的损失函数，即$\ell_{\text{model}}(\,)$；$\beta^k\boldsymbol{\delta}_n$即第$n$次迭代的对抗性扰动，由根据梯度更新的扰动$\boldsymbol{\delta}_n$和衰减系数$\beta^k$构成，常数$\beta$满足$\beta\in(0,1)$，$k$即当前已成功攻击的次数；$\boldsymbol{\delta}$的更新规则为

$$\boldsymbol{\delta}_{n+1}=\boldsymbol{\delta}_n+\varepsilon\text{sgn}\left(\nabla_{\delta}\ell\left(\boldsymbol{x}+\beta^k\boldsymbol{\delta}_n,t\right)\right), \tag{10}$$

满足$\boldsymbol{\delta}_0=0$且$\boldsymbol{\delta}_n\in[-M,M]$. ε表示由攻击者指定的超参数，攻击者依据ε调整从梯度中计算的扰动大小，从而改变对抗样本解的搜索效率. 给定原始语音$\boldsymbol{x}$、目标语句t和最大迭代次数$iter$，在限制范围上添加对抗性扰动可分为3个步骤：

1) 在每步迭代中，首先根据当前样本到目标t的梯度确定样本更新的方向，然后以合适的步长ε更新样本，更新时扰动大小需满足$\boldsymbol{\delta}_n\in[-M,M]$.

2) 每次更新样本后即向目标模型查询，检查是否完成攻击，若ASR将当前样本转录为目标语句，则以β倍率对当前扰动$\beta^{k-1}\boldsymbol{\delta}_n$进行衰减；若没有完成攻击，则继续在当前扰动水平上进行迭代优化.

3) 如果发生衰减，衰减后的样本$\beta^k\boldsymbol{\delta}_n$通常失去目标攻击能力，样本将在更低的扰动水平上继续进行迭代优化，扰动大小满足$\beta^k\boldsymbol{\delta}_n\in[-\beta^kM,\beta^kM]$，以搜索更小的对抗性扰动.

4.3 对抗样本空间度量

一条对抗样本在某些采样点上随机多次+1或−1，仍能够对目标模型造成目标攻击，但神经网络输出层***logits***几乎没有变化. 因此对抗样本空间可被视为由很多高维子空间组成，扰动差异较小且具有相近***logits***分布的对抗样本视为位于同一子空间. 我们用对抗样本空间大小来描述一条语音在一个具体模型上的潜在可求解对抗样本的质量，对抗样本空间越大，对抗性扰动的幅值越小，可求解的对抗样本质量越好. 不同的原始语音和目标转录设置之间的对抗样本空间不具有可比性，同一组源语音和转录目标设置下，不同扰动区间上的对抗样本空间才能进行比较.

在目标模型和网络参数已知的条件下，一条语音到目标语句对抗样本解的空间是固定的. 而限制扰动范围会导致某些从原始语音到对抗样本的路径不可达，我们用对抗样本空间的缩减来描述这一现象. 同时，由于对抗样本空间是不可测量的，我们用有限次迭代下的成功攻击次数k来描述对抗样本空间的大小，根据不同区间上求解对抗样本的k值比较耦合作用、位置权重和区间长度对对抗样本空间的综合影响.

在对抗样本求解过程中，本文攻击方案主要解决在求得对抗样本后存在局部最优解的问题. 在当前扰动水平上求得对抗样本后，如果不衰减继续执行迭代，*loss*值仍可以被进一步降低，经过一定次数的迭代后，求解算法将会在局部最优解[39]附近震荡，但此时的迭代对于度量对抗样本空间是没有意义的，只探索了某子空间中的附近区域.

所提出的衰减系数β^k的主要作用包含：1)降低扰动大小；2)跳出当前局部最优解继续进行优化. 在我们的方案中，每次执行衰减，***logits***输出都发生了较大改变，意味着其在对抗样本空间也发生了较大程度的转移，因此可以用衰减次数k度量对抗样本空间. 优化算法示意图如图10所示.

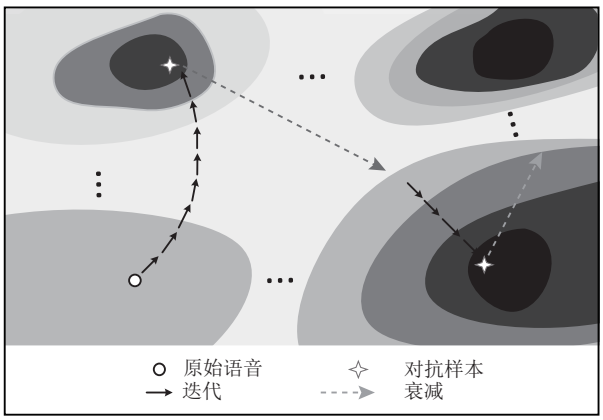

Fig. 10 Exploring adversarial example space through perturbation decay

图10 通过扰动衰减探索对抗样本空间

每步迭代添加的扰动都会使*loss*减小，实现目标攻击时的*loss*并不一定是局部最小值. 每次实现目标攻击后，本文优化算法不继续降低*loss*值，而是执行扰动衰减，以降低扰动水平并跳出当前局部最优解范围.

5 实验设置

5.1 数据集和目标模型

为了探究在固定迭代次数下在语音上限制不同范围对生成对抗样本的影响，针对2类分帧方式，本文选取了4种语音识别模型中的6个模型作为目标模型：DeepSpeech v0.9.3, DeepSpeech v0.4.1, DeepSpeech v0.1.1, Lingvo, SpeechBrain-Transducer, Whisper进行交叉验证，它们的信息介绍如表3所示.

Table 3 Target Model Configuration Information

表3 目标模型配置信息

语音识别模型	分帧方式	特征提取	窗口函数
DeepSpeech v0.9.3	Ⅰ类	MFCC	Hamming
DeepSpeech v0.4.1	Ⅰ类	MFCC	Hann
DeepSpeech v0.1.1	Ⅱ类	MFCC	不加窗
Lingvo	Ⅱ类	Mel谱	Hann
SpeechBrain-Transducer	Ⅱ类	FBank	Hamming
Whisper	Ⅱ类	FBank	Hann

1）DeepSpeech. 是由百度公司在2014年发布的端到端语音识别模型，各个版本之间网络结构无变化，新版本比旧版本采用了更多训练数据，特征提取方式也存在部分差异. 其中v0.1.1属于Ⅱ类分帧方式，帧长为400，帧移为160；v0.4.1和v0.9.3属于Ⅰ类分帧方式，帧长为512，帧移为320.

2）Lingvo. 是由谷歌公司在2019年开源的语言相关任务序列模型. Lingvo模型采用了金字塔式特征提取，同一帧的特征比DeepSpeech分布在更多的原始语音区间内. 提取语音的Mel谱图特征，帧长为400，帧移为160.

3）SpeechBrain-Transducer. 是由Mila研究所等在2020年主导的开源一体化语音工具包. 我们选取了其中的Transducer网络作为目标模型. 其预训练模型提取语音的FBank特征，帧长400，帧移160.

4）Whisper. 是由OpenAI公司在2022年发布的通用语音识别模型，采用自注意力机制的MLP（multilayer perceptron）作为Transformer的编解码器，提取语音的FBank特征，帧长400，帧移160.

其中2类分帧方式的对比验证了本文分析规律的普遍性；DeepSpeech v0.9.3和v0.4.1对比，验证能够兼容窗口函数中不同的权重分布；DeepSpeech v0.1.1, Lingvo, SpeechBrain-Transducer, Whisper对比验证能够兼容多种特征提取方法. 对抗攻击不涉及模型的训练过程，本文针对训练完成的ASR模型生成对抗样本.

我们使用LibriSpeech数据集进行规模测试. LibriSpeech数据集来源于LibriVox项目，由采样率为16 kHz的英语音频数据组成，发音较清晰，不会因为数据质量问题影响实验结果. 为了探索对抗样本空间的分布差异，所求解对抗样本应有一定难度，过短的原始语音和目标语句设置会导致对抗样本的求解简单，甚至在黑盒攻击条件下也能成功，因此本文过于在test-clean分支上随机选取600条平均时长为5 s的原始语音组成数据集进行实验，其中300条作为原始语音，另外300条的转录作为目标语句.

5.2 攻击参数

随着攻击成功次数的增大，扰动幅度呈指数级减小，本文采用Adam优化器来适应扰动幅度的改变. 学习率设置为100，初始扰动幅值阈值M设置为2000，衰减系数β设置为0.8.

迭代次数$iter$即停止优化的条件，一步迭代包含完整的梯度下降和反向传播流程. 通过在不限制扰动范围的条件下生成对抗样本进行实验测试，本文攻击方法能够使99%以上的语音在500步之内完成对抗样本的优化. 限制扰动范围会增大对抗样本的求解难度，但大部分样本仍在500步之内找到局部最优解，为了统一条件，本文设置除Whisper之外的其他模型上迭代次数$iter=500$，Whisper模型上迭代次数$iter=2000$.

5.3 评价指标

1) 成功攻击次数k. k值能够反映所求解对抗性扰动的幅值大小，第k次攻击成功后，对抗性扰动的值域为$[-0.8^k\times 2\,000, 0.8^k\times 2\,000]$，$k$值越大，最终求解对抗性扰动越小. 同时$k$值每次增长所需的迭代次数也能反映出对抗样本的求解难度：$k+1$所需要的迭代次数越多，当前扰动水平下对抗样本的求解难度越大.

2) 功率信噪比（SNR）. k值反映了对抗性扰动幅值的极值水平. 语音信号作为1维序列数据，其整体扰动水平应在全序列上计算. 求解难度增大意味着最后求解出的全局扰动水平较高，本文采用功率信噪比来量化评价扰动水平，计算方法如式(11)所示，功率信噪比越小，意味着噪声能量相对越大，

$$SNR=10\lg\frac{\sum_{n=0}^{tN}s^2(n)}{\sum_{n=0}^{tN}\delta^2(n)}. \tag{11}$$

3) 攻击成功率（SR）. 如果1条语音在500次迭代

内没有求解出符合条件的对抗样本,则认为攻击失败.本文采用在300条语音上测试的整体攻击成功率来检验限制扰动范围对攻击可用性的影响.

4) 对抗样本空间.在数据集实验层面,如果k值平均值显著降低,则意味着该扰动范围缩减了对抗样本空间.

5) 对抗样本的求解难度.随着k值增大,攻击方法将限制在更小的扰动幅值内求解对抗样本,为了比较限制不同范围对求解难度的影响,本文采用k值增加所需的迭代次数来反映求解难度.在扰动区间固定的条件下,2次成功攻击之间所需的迭代次数越多,意味着对抗样本求解难度越大.

5.4 实验验证

在已完成训练的6个模型上,我们以固定排列的原始语音和目标句子进行对抗样本生成实验.统计300条语音在500次迭代下的平均成功攻击次数(k值)与不限制扰动范围下的平均k值.统计信息如表4和表5所示,同一模型的不同实验间(横向比较)唯一变量是扰动范围.

Table 4 Perturbation Range Constraints Experiments of Class-I Framing

表4 Ⅰ类分帧中限制扰动区间实验

扰动区间	DeepSpeech v0.9.3			DeepSpeech v0.4.1			扰动比例
	SNR/dB	*k*	*SR*/%	*SNR*/dB	*k*	*SR*/%	
A	22.44	9.08	93.7	21.09	8.36	97.3	0.4
B	18.11	7.71	88.7	17.19	7.24	95	0.6
C	**23.44**	**10.54**	**94.3**	**22.25**	**9.91**	**99**	0.6
D	15.88	5.60	83	14.32	4.77	83.7	0.4
不限	**25.12**	**12.58**	**96.3**	**24.13**	**12.1**	**100**	1.0

在包含非重叠区间的Ⅰ类分帧方式中,DeepSpeech v0.9.3和v0.4.1这2组实验数据表现出相同规律:区间A和区间B相比,A中扰动范围的长度和位置权重影响因素均比B差,只有耦合作用影响因素优于B,但仍取得了较大的k值,这说明在非耦合区间上求解对抗样本,其对抗样本空间更大.同时,区间A对抗样本的信噪比也显著优于区间B,在更小的扰动范围上获得了更小的噪声能量.

在区间C上添加扰动,取得了4个区间中最佳的评价结果.区间C包含的重叠部分同样包含在区间B中,其扰动比例也与区间B相同.但是,攻击效果显著优于区间B,验证了在帧内各区间上生成对抗样本时,对抗样本解的空间分布是非均衡的.其次,区间C包含完整的区间A和部分重叠帧,可被近似视为不受耦合作用影响.在扰动区间长度增大且权重小幅度提升的条件下,区间C上的扰动幅值减小,功率信噪比小幅度提升.

区间B包含区间D和部分重叠帧.在均受耦合作用影响的条件下,区间D中扰动范围的长度和位置权重影响因素均比B差,所求解对抗样本的扰动幅值大幅度增大,功率信噪比大幅度降低.

因此,根据Ⅰ类分帧方式的4组实验和不限制扰动范围的对照试验结果,我们总结规律有4点:

1) 不限制扰动范围的对抗样本空间最大;

2) 对抗样本空间缩减主要由耦合作用导致;

3) 扰动范围由非重叠区间扩展加入部分重叠区间时,对抗样本空间增大,功率信噪比小幅度提升;

4) 扰动范围由重叠区间进行截断时,对抗样本空间大幅度缩减,求解到的对抗样本功率信噪比大幅度缩减.

我们在所有区间都属于重叠帧的Ⅱ类分帧方式上验证上述规律,扰动范围限制试验结果如表5所示.

在所有区间都属于重叠帧的Ⅱ类分帧方式中,DeepSpeech v0.1.1, Lingvo, SpeechBrain-Transducer的表现相同,当把扰动范围完全限制在强重叠区间或重叠区间上时,对抗样本求解空间均大幅度缩减,且强重叠区间E上的对抗样本空间比弱重叠区间F更小;而在通过减少扰动区域对重叠区间解耦合的G区间上,以更小的位置权重和更小的扰动范围反而取得了更大的k值,在3组限制范围实验中实验效果最佳.结合图9分析,增大扰动区间的间隔后,重叠

Table 5 Perturbation Range Constraints Experiments of Class-Ⅱ Framing

表5 Ⅱ类分帧中限制扰动区间实验

扰动区间	DeepSpeech v0.1.1			Lingvo			SpeechBrain-Transducer			Whisper			扰动比例
	SNR/dB	*k*	*SR*/%	*SNR*/dB	*k*	*SR*/%	*SNR*/dB	*k*	*SR*/%	*SNR*/dB	*k*	*SR*/%	
E	0	0	0	4.12	1.82	39	34.43	4.02	100	16.77	11.51	80.3	0.5
F	0	0	0	5.33	2.34	48.7	35.09	5.6	100	**16.82**	**11.62**	**81.5**	0.5
G	**19.28**	**7.56**	**98**	**21.85**	**8.22**	**95.3**	**45.07**	**12.97**	**100**	11.12	9.04	73.3	0.33
不限	25.01	12.45	100	27.96	14.03	97.3	52.55	18.76	100	25.18	15.46	86.8	1.0

部分更容易分布在权重较低的位置，每区间添加的扰动可被近似视为对单帧起作用，其评价指标结果也类似于Ⅰ类分帧方式中在区间 A 上添加扰动.

但 Whisper 表现出不同的规律：强重叠区间上的对抗样本空间大小和弱重叠区间上的几乎相等；G 区间的主动解耦合操作减小了整体的可扰动范围，缩减了对抗样本空间，起到了和其余 3 个模型完全相反的作用. 由于帧长和帧移参数决定了语音帧内的采样点贡献的不均衡分布，相同的帧长和帧移意味着同等幅度的扰动对 DFT 特征有相同的控制能力，已知由 DFT 特征计算的 MFCC、Mel 频率谱、FBank 特征表现出相同的规律，且 Whisper 采用的 Log-Mel 特征由在 Mel 频率谱的基础上取对数得到，因此我们更倾向于认为这种规律差异是由于模型结构造成的，可能的原因为：DeepSpeech 是 CTC 结构的模型，Lingvo 和 SpeechBrain 是 Transducer 结构的模型，它们都是逐帧解码的模型结构，语音结束则解码过程结束；而 Whisper 是一种基于 seq2seq 结构的模型，特点是逐词解码，直到解码出<EOS>标记，解码过程结束. 在多次解码的过程中，帧与帧之间的位置划分发生相对变化，会导致强重叠与弱重叠结构的相互转化，从而对抗样本空间只由扰动区间长度决定，且和区间长度正相关.

本文只从输入特征的扰动能力差异分析了对抗样本空间受影响的规律，把模型对特征的处理作为黑盒，不考虑模型处理机制对对抗样本空间的影响，因此我们暂时把本文规律的适用范围限制在 CTC 及 Transducer 结构的模型上.

总的来说，重叠程度较弱的区间上更容易求解对抗样本，而为了利用该结论限制目标攻击的扰动范围，要付出的代价有所差异：如果特征提取过程中天然存在非重叠区间，直接将扰动范围限制在重叠区间上，即可有效降低扰动范围；如果特征提取流程中不存在非重叠区间，若限制扰动范围到弱重叠区间不能有效降低，则以增大扰动区间间隔的方式对扰动区间解耦合；若需进一步提升语音质量，从扰动范围的左侧（权重更大的地方）进行扩充能够取得更好的扰动效果.

上述实验结果展示了限制扰动范围对对抗样本求解结果的影响，是一种静态结果，代表了对抗样本空间的缩减程度. 为了理解对抗样本的求解难度随扰动范围的变化，我们绘制了本节实验中不同区间的平均 k 值随迭代次数的增长趋势，如图 11 所示，该图中所示 k 值为每个模型上 300 条对抗样本的求解过程的平均值.

在所有模型上，未限制扰动范围的 k 值变化最快，对抗样本求解过程最活跃，每次缩减后，仅需较少次迭代，即可求得更小扰动的对抗样本解. 除 Whisper 模型外，所有子图中耦合作用更弱的区间的平均 k 值均处于较高的水平，持续大于等于比自己扰动范围更大的限制区间. 不同区间的求解规律和对抗样本的空间缩减特性一致：在限制扰动区间增大了对抗样本求解难度的条件下，非重叠帧上的扰动范围越多，可扰动区间越大，越容易求解对抗样本.

6 讨　论

6.1 无耦合作用下区间权重及长度的影响

本文对 ASR 的数据预处理过程进行分析，根据扰动作用随帧内权重和复用程度随采样点位置的变化，提出了 3 种对抗样本空间的影响因素，并对其影响大小进行了实验分析. 但考虑到神经网络具有非线性特点，位置权重和区间长度优势对对抗样本空间的影响不能确定. 针对该问题，本节屏蔽耦合作用的影响，以攻击单字符为目标进行对抗样本生成实验.

当以一句话中的单个字符为目标进行攻击时，如使目标模型将原始语音“I think so”转录为“I thank so”，如果对抗样本空间较大，在单帧（转录结果为“i”的对应帧）上添加扰动就能使 ASR 的识别结果发生改变. 但当对抗样本空间较小时，需要在左右相邻帧（“i”的邻近帧，可能为“h-i--”等，其中“-”表示空白伪字符）上添加扰动，才能将该帧的转录结果误导为目标字符. 这种情况下，所有扰动的优化目标只有 1 个，而非 5.4 节实验中每帧都有对应的优化目标，因此不存在耦合作用.

在实际实验中，我们发现将一个字符的转录结果指定为不存在的单词时（如“think”攻击为“thgnk”）难以攻击成功，这是由于训练集中没有该单词，模型网络参数中也没有对应的模式. 对抗样本研究中，添加的扰动只能使模型将数据判断为错误的已有类别，而不能新增类别. 因此，本文随机选取了 10 条语音，只改变其中的 1 个字符进行目标攻击，测量无耦合作用条件下区间 A 和区间 B 的对抗样本空间大小，来比较权重和长度优势对对抗样本空间的影响. 我们仍采用 5.4 节攻击方式和评价指标 k 进行测量. 其中原始语音及其目标设置如表 6 所示.

如 4.3 节所述，对抗样本空间和神经网络参数、当前语音、目标字符有关，我们首先在攻击目标的

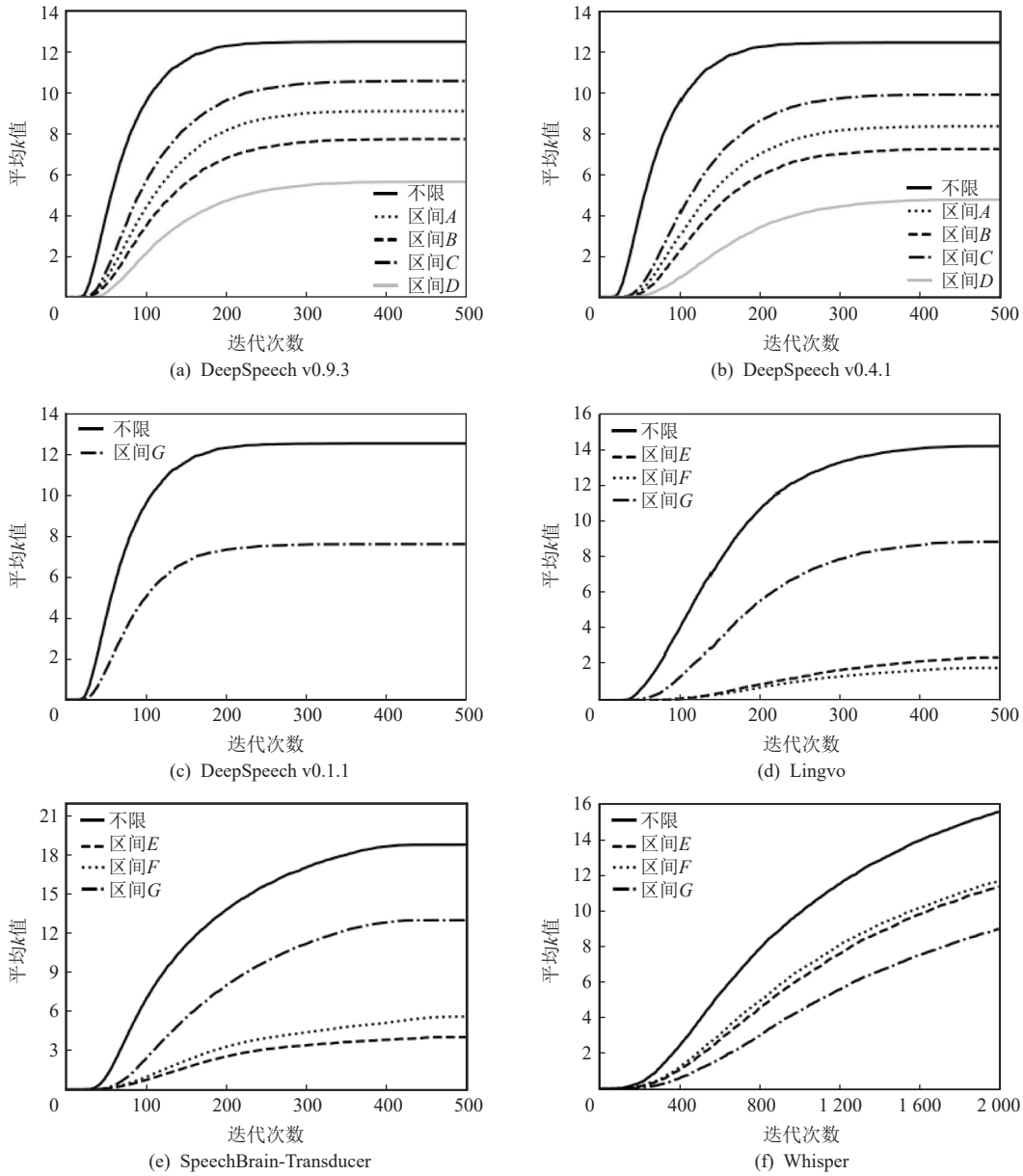

Fig. 11 Variation of the average k values in ASR

图 11 语音识别模型中的平均 k 值变化

Table 6 Original Speech and Attack Target Setting for Single Frame Attack

表 6 针对单帧攻击的原始语音及攻击目标设置

文件序号	原始语音转录	攻击目标
1188-133604-0014	“Do not therefore think that the gothic school is an easy one”	think → thank
61-70970-0009	“This late and i go myself within a short space”	short → shirt
6829-68769-0051	“There was a grim smile of amusement on his shrewd face”	on → of
7176-92135-0009	“Nd i should begin with a short homily on soliloquy”	on → in
7127-75946-0004	“Certainly sire but i must have money to do that what”	have → hate
7176-92135-0020	“Double nine two three elsinore double nine yes hello is that you horatio hamlet speaking”	nine → nice
7176-92135-0003	“Your play must be not merely a good play but a successful one”	good → food
8555-292519-0012	“Through the black night rain he sang to her window bars”	night → right
5683-32866-0027	“A cold bright moon was shining with clear sharp lights and shadows”	clear → clean
61-70970-0012	“Yet he will teach you a few tricks when morning is come”	teach → beach

第 t 帧上选取扰动范围，如果没有攻击成功，向左右扩展 1 帧再次尝试攻击，重复这个流程，直到攻击成功，实验结果如表 7 所示，k_A, k_B 分别表示在区域 A, B 上添加扰动时的 k 值.

Table 7 Experimental Results of Single Frame Attack

表 7 单帧攻击实验结果

序号	扰动帧	k_A	k_B
1188-133604-0014	[t]	5	11
61-70970-0009	[t−2, t+2]	5	6
6829-68769-0051	[t−2, t+2]	8	9
7176-92135-0009	[t−2, t+2]	3	4
7127-75946-0004	[t−1, t+1]	4	9
7176-92135-0020	[t−1, t+1]	5	7
7176-92135-0003	[t−2, t+2]	0	2
8555-292519-0012	[t−2, t+2]	7	6
5683-32866-0027	[t−2, t+2]	2	7
61-70970-0012	[t−3, t+3]	4	5

在重叠区间 B 上添加扰动时，对抗样本求解结果普遍优于区间 A，表明在不受耦合作用影响时，具有权重和长度优势的区间上具有更大的对抗样本空间. 这些优势发生在对输入数据的预处理阶段，经过特征提取和神经网络的分类，仍能作用于对抗样本空间.

6.2 度量方法客观性讨论

本文所提出攻击方法的特点在于设置阶段性的优化目标. 随着成功攻击次数的增大，求解到的对抗样本扰动减小，即能够以更精细的扰动实现攻击. 因此攻击成功次数可以作为衡量对抗样本空间大小的指标. 攻击方法包含梯度下降和反向传播 2 个阶段，在梯度下降过程中，目标函数关于参数的梯度是在完整语音上进行计算的，而在反向转播更新对抗性扰动时，扰动范围的限制使得只有部分梯度信息被用来更新扰动，选用带有动量的优化算法更有利于实现优化目标.

5.4 节实验中采用 Adam 优化器，每次迭代的优化方向和步长由原始语音、攻击目标、历史扰动决定，当陷入局部最优解时，无法求解到更小的对抗性扰动，可能存在探索对抗样本空间不充分的问题. PGD 攻击[40] 采用随机重启策略解决这一问题，本文借鉴该方案，在攻击过程中每迭代固定间隔次数，即在对抗性扰动上添加随机噪声，以微调优化方向，从而增加跳出局部最优解的机会以继续进行优化. 本节在 DeepSpeech v0.4.1 模型上进行噪声扰动实验，在迭代过程中，每隔 10 次迭代添加 1 次噪声，该噪声采样数和原始语音保持一致，每个采样点噪声服从 $N(0,9)$正态分布，其余设置和 5.4 节保持一致. 表 8 测试了 A, B, C, D 这 4 个区域在添加随机噪声扰动的条件下的对抗样本空间大小.

Table 8 Experimental Results of Noise Attack

表 8 噪声攻击实验结果

扰动区间	SNR/dB	k	SR/%
A	21.75	8.4	97.3
B	17.15	7.21	95
C	22.45	9.99	99
D	14.51	4.84	83.7
不限	23.94	11.92	100

添加噪声的扰动实验与无噪声扰动实验表现出相同的规律，即不限制扰动范围时对抗样本空间最大，其次是非耦合帧占主体的区间 C 和区间 A. 另外，与表 4 相比，表 8 中各区间的 SNR 值和 k 值没有增大. 我们对比分析了原始实验和噪声实验中的个体差异，发现确实存在部分语音和目标转录在添加随机噪声后能够求解出更小扰动的对抗样本，但是，还有一部分样本数据添加随机噪声后 k 值减小，即比无噪声更早地陷入了局部最优解. 因此，在数据集规模上，添加随机噪声不能更客观地探索对抗样本空间，我们不建议在探索对抗样本空间时添加随机扰动.

6.3 应 用

除在度量对抗样本空间大小时访问了模型梯度外，本文在更严格的条件下设置了攻击目标和条件，以探索对抗样本重点区域的真实分布. 所设计的扰动范围限制实验以帧为单位在整条语音的部分区间上添加扰动. 根据实验过程中的人耳监听，对于某些天然难求解对抗样本的原始语音，在限制扰动范围后 k 值更小，所求解出的对抗性扰动的幅值也普遍较大，均匀分布在整条语音上时将产生啁啾噪声(Chirp)，不能完全用于实际对抗攻击. 同时，我们也在采用基于心理声学掩蔽[12, 16] 的对抗样本生成方法上进行了测试，以这些语音为原始语音生成的目标攻击对抗样本能感觉到底噪的存在. Vadillo 等人[12] 也认为语音对抗样本研究中的评价指标只是定量描述了添加的扰动量，不能客观反映出对人耳的影响，考虑将底噪转化为噪点[41] 是降低扰动感知度研究中更具潜力的研究方向.

本文所证明的对抗样本重点区域分布规律，为语音对抗攻击和防御提供了新的思路：对于攻击方，

如果要执行特定短语的目标攻击，以弱重叠区间或向左侧扩展的扰动范围能最大程度保持信噪比，进一步探索出序列模型中帧与帧识别结果相互影响的规律并予以规避后，有希望实现针对语音识别的最小范围攻击甚至每帧单采样点扰动攻击；如果要执行扰乱原始语音识别结果的无目标攻击，则特征耦合作用与模型识别结果的帧间相互影响则转变为优势，将扰动添加在重叠区域上即可实现高信噪比的无目标攻击. 对于防御方，利用对抗样本比正常语音鲁棒性差的特点，在重叠区间上添加随机干扰噪声，能够破坏对抗样本而尽可能降低对正常业务的影响.

对于希望在语音中添加对抗性扰动以保护日常对话隐私免受广告服务商窃取的防御者[42]来说，针对离线语音文件防识别的应用需求，普遍做法是在文件传输至互联网前添加通用扰动[43]，由于通用扰动的生成不依赖于具体的语音文件，采用心理声学降低扰动感知度的方法将不再适用，本文方法同样不依赖具体的语音文件，能更好地和通用扰动结合，降低扰动感知度；针对实时添加扰动干扰任意语音识别结果的应用需求[33]，也可以结合本文规律在重叠区间上添加噪声.

7 总　　结

本文从帧的结构对求解对抗样本的影响展开分析，证明了在不考虑模型对特征处理机制差异的条件下，分帧过程中存在的耦合作用是对抗样本空间缩减的主要原因，并给出了在限制扰动范围时最应该保留的扰动区间. 在研究过程中，本文采用交叉试验方法，将复合因素叠加分析问题转变为对抗样本空间求解问题，并设计了针对序列到序列模型的对抗样本空间度量方法和评价指标，解决了固定结构的耦合作用、位置权重、区间长度影响难以在同一尺度下进行比较的问题. 经检验，该度量方法能够在数据集规模上客观地度量对抗样本空间. 最后，我们提出了应用该一般规律的应用场景，为语音识别攻击与防御提供新的思路.

作者贡献声明：韩松莘提出论文选题，设计实验并编写代码进行测试，完成论文初稿撰写；郭松辉对现象进行理论分析，指导实验的总体设计；徐开勇指导从理论到现象之间的总结，完善规律的应用范围；杨博完善论文中前后逻辑，对设计思路和分析部分做出重要修改；于淼参与多次实验，验证规律.

参　考　文　献

[1] Li Jinyu. Recent advances in end-to-end automatic speech recognition[J]. APSIPA Transactions on Signal and Information Processing. 2022, 11(1): e8

[2] Pan Shanrong. Design of intelligent robot control system based on human–computer interaction[J]. International Journal of System Assurance Engineering and Management, 2023, 14: 558−567

[3] Wei Chunyu, Sun Meng, Zou Xia, et al. Reviews on the attack and defense methods of voice adversarial examples[J]. Journal of Cyber Security, 2022, 7(1): 100−113 (in Chinese)
(魏春雨，孙蒙，邹霞，等. 语音对抗样本的攻击与防御综述 [J]. 信息安全学报，2022，7(1)：100−113)

[4] Wang Donghua, Wang Rangding, Dong Li, et al. Adversarial examples attack and countermeasure for speech recognition system: A survey[C] //Proc of the 1st Security and Privacy in Digital Economy. Berlin: Springer, 2020: 443−468

[5] Li Zhouhang, Wu Yi, Liu Jian, et al. Advpulse: Universal, synchronization-free, and targeted audio adversarial attacks via subsecond perturbations[C] //Proc of the 20th ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2020: 1121−1134

[6] Zheng Baolin, Jiang Peipei, Wang Qian, et al. Black-box adversarial attacks on commercial speech platforms with minimal information[C] //Proc of the 21st ACM SIGSAC Conf on Computer and Communications Security. New York: ACM, 2021: 86−107

[7] Carlini N, Wagner D. Audio adversarial examples: Targeted attacks on speech-to-text[C] //Proc of the 39th IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2018: 1−7

[8] Taori R, Kamsetty A, Chu B, et al. Targeted adversarial examples for black box audio systems[C] //Proc of the 40th IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2019: 15−20

[9] Liu Xiaolei, Wan Kun, Ding Yufei, et al. Weighted-sampling audio adversarial example attack[C] //Proc of the 34th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2020: 4908−4915

[10] Tay K Y, Ng L, Chua W H, et al. Audio adversarial examples: Attacks using vocal masks[J]. arXiv preprint, arXiv: 2102. 02417, 2021

[11] Qin Yao, Carlini N, Cottrell G, et al. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition[C] //Proc of the 36th Machine Learning Research. New York: PLMR 2019: 5231−5240

[12] Vadillo J, Santana R. On the human evaluation of universal audio adversarial perturbations[J]. Computers & Security, 2022, 112: 102495

[13] Xie Yi, Li Zhuohang, Shi Cong, et al. Real-time, robust and adaptive universal adversarial attacks against speaker recognition systems[J]. Journal of Signal Processing Systems, 2021, 93(10): 1187−1200

[14] Xie Yi, Li Zhuohang, Shi Cong, et al. Enabling fast and universal audio adversarial attack using generative model[C] //Proc of the 35th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021: 14129−14137

[15] Eisenhofer T, Schönherr L, Frank J, et al. Dompteur: Taming audio adversarial examples[J]. arXiv preprint, arXiv: 2102. 05431, 2021

[16] Schönherr L, Kohls K, Zeiler S, et al. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding[J]. arXiv preprint, arXiv: 1808.05665, 2018

[17] Malik M, Malik M K, Mehmood K, et al. Automatic speech recognition: A survey[J]. Multimedia Tools and Applications, 2021, 80(6): 9411−9457

[18] Gupta D, Bansal P, Choudhary K. The state of the art of feature extraction techniques in speech recognition[C] //Proc of Speech and Language Processing for Human-Machine Communications. Berlin: Springer, 2018: 195−207

[19] Shen Yijie, Li Liangcheng, Liu Ziwei, et al. Stealthy attack towards speaker recognition based on one-"audio pixel" perturbation[J]. Journal of Computer Research and Development, 2021, 58(11): 2350−2363 (in Chinese)
(沈铁杰, 李良澄, 刘子威, 等. 基于单"音频像素"扰动的说话人识别隐蔽攻击[J]. 计算机研究与发展, 2021, 58(11): 2350−2363)

[20] Sood M, Jain S. Speech recognition employing MFCC and dynamic time warping algorithm[C] //Proc of Innovations in Information and Communication Technologies. Berlin: Springer, 2021: 235−242

[21] Pardede H F, Zilvan V, Krisnandi D, et al. Generalized filter-bank features for robust speech recognition against reverberation[C] //Proc of the 7th Int Conf on Computer, Control, Informatics and Its Applications. Piscataway, NJ: IEEE, 2019: 19−24

[22] Keshishian M, Norman-Haignere S, Mesgarani N. Understanding adaptive, multiscale temporal integration in deep speech recognition systems[C] //Proc of the 35th Advances in Neural Information Processing Systems. Cambridge, MA: MIT, 2021: 24455−24467

[23] Ravanelli M, Parcollet T, Bengio Y. The pytorch-Kaldi speech recognition toolkit[C] //Proc of the 44th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2019: 6465−6469

[24] Shen J, Nguyen P, Wu Yonghui, et al. Lingvo: A modular and scalable framework for sequence-to-sequence modeling[J]. arXiv preprint, arXiv: 1902.08295, 2019

[25] Hong Qingyang, Li Lin. Principle and Application of Speech Recognition[M]. 2nd ed. Beijing: Publishing House of Electronics Industry, 2023(in Chinese)
(洪青阳, 李琳. 语音识别: 原理与应用[M]. 第2版. 北京: 电子工业出版社, 2023)

[26] Panayotov V, Chen Guoguo, Povey D, et al. LibriSpeech: An ASR corpus based on public domain audio books[C] //Proc of the 40th IEEE Int Conf on Acoustics, Speech and Signal Processing. Piscataway, NJ: IEEE, 2015: 5206−5210

[27] Breithaupt C, Martin R. Statistical analysis and performance of DFT domain noise reduction filters for robust speech recognition[C/OL] //Proc of the 9th Int Conf on Spoken Language Processing. ISCA, 2006: 365−368. [2022-12-01]. https://www.isca-speech.org/archive/pdfs/interspeech_2006/breithaupt06_interspeech.pdf

[28] Ravindran S, Anderson D V, Slaney M. Improving the noise-robustness of mel-frequency cepstral coefficients for speech processing[J]. Reconstruction, 2006, 12(S14): 48−52

[29] Schuster G, Ansorge R. WOLA noise cancelling performance[C] //Proc of the 16th European Signal Processing Conf. Piscataway, NJ: IEEE, 2008: 1−5

[30] Zhang Weiqiang, Yang Dengzhou, Liu Jia, et al. Perturbation analysis of mel-frequency cepstrum coefficients[C] //Proc of the 2nd Int Conf on Audio, Language and Image Processing. Piscataway, NJ: IEEE, 2010: 715−718

[31] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks[J]. arXiv preprint, arXiv: 1312. 6199, 2013

[32] Chen Yuxuan, Zhang Jiangshan, Yuan Xuejing, et al. Sok: A modularized approach to study the security of automatic speech recognition systems[J]. ACM Transactions on Privacy and Security, 2022, 25(3): 1−31

[33] Xu Zirui, Yu Fuxun, Liu Chenchen, et al. HAMPER: High-performance adaptive mobile security enhancement against malicious speech and image recognition[C]// Proc of the 24th Asia and South Pacific Design Automation Conf. New York: ACM, 2019: 512−517

[34] Abdullah H, Rahman M S, Garcia W, et al. Hear" no evil", see" kenansville": Efficient and transferable black-box attacks on speech recognition and voice identification systems[C] //Proc of the 42nd Symp on Security and Privacy. Piscataway, NJ: IEEE, 2021: 712−729

[35] Wang Qian, Zheng Baolin, Li Qi, et al. Towards query-efficient adversarial attacks against automatic speech recognition systems[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 896−908

[36] Xie Yi, Li Zhuohang, Shi Cong, et al. Enabling fast and universal audio adversarial attack using generative model[C]// Proc of the 35th Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2021, 35(16): 14129−14137

[37] Zhang Wanli, Chen Yue, Yang Kuiwu, et al. An adversarial example generation method for locally occluded face recognition[J]. Journal of Computer Research and Development, 2023, 60(9): 2067−2079(in Chinese)
(张万里, 陈越, 杨奎武, 等. 一种局部遮挡人脸识别的对抗样本生成方法[J]. 计算机研究与发展, 2023, 60(9): 2067−2079)

[38] Abdullah H, Warren K, Bindschaedler V, et al. SoK: The faults in our ASRs: An overview of attacks against automatic speech recognition and speaker identification systems[C] //Proc of the 42nd Symp on Security and Privacy. Piscataway, NJ: IEEE, 2021: 730−747

[39] Zong Wei, Chow Y W, Susilo W. Towards visualizing and detecting audio adversarial examples for automatic speech recognition[C] //Proc of the 26th Symp Information Security and Privacy. Berlin: Springer, 2021: 531−549

[40] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv preprint, arXiv: 1706.06083, 2017

[41] Wu Xiaoliang, Rajan A. Catch me if you can: Blackbox adversarial attacks on automatic speech recognition using frequency masking[C] //Proc of 29th Asia-Pacific Software Engineering Conf. Piscataway,

NJ: IEEE, 2022: 169−178

[42] Kwon H, Kim Y, Yoon H, et al. Selective audio adversarial example in evasion attack on speech recognition system[J]. IEEE Transactions on Information Forensics and Security, 2020, 15: 526−538

[43] Mathov Y, Ben Senior T, Shabtai A, et al. Stop bugging me! Evading modern-day wiretapping using adversarial perturbations[J]. Computers & Security, 2022, 121: 102841

Han Songshen, born in 1999. Master. His main research interests include artificial intelligence security and cloud computing security.

韩松莘, 1999 年生. 硕士. 主要研究方向为人工智能安全、云计算安全.

Guo Songhui, born in 1979. PhD, professor. His main research interests include 5G security and cloud computing security.

郭松辉, 1979 年生. 博士，研究员. 主要研究方向为 5G 安全、云计算安全.

Xu Kaiyong, born in 1963. PhD, professor. His main research interests include information security and trusted computing.

徐开勇, 1963 年生. 博士，研究员. 主要研究方向为信息安全、可信计算.

Yang Bo, born in 1993. PhD candidate. His main research interests include deep learning, and intelligent system security testing and evaluation.

杨　博, 1993 年生. 博士研究生. 主要研究方向为深度学习、智能系统安全测试与评估.

Yu Miao, born in 1987. Master candidate. His main research interests include artificial intelligence security and natural language processing.

于　淼, 1987 年生. 硕士研究生. 主要研究方向为人工智能安全、自然语言处理.