小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2025-0017
2026 年 2 月 第 2 期
Vol.47 No.2 2026

UGA-SR:融合 U-Net 与 GAN 的加速度计语音恢复模型

沈 澍,陈林浩,袁晓博

(南京邮电大学 计算机学院,南京 210003)
E-mail:shens@njupt.edu.cn

摘 要:本文提出了一种基于安卓设备内置加速度传感器的语音恢复模型 UGA-SR,展示了利用加速度传感器进行语音恢复的可行性.该模型结合生成对抗网络和改进的 U-Net 架构,引入多尺度卷积的空间注意力模块,优化特征提取和信号重构,将预处理的加速度信号转换为梅尔频谱图,以生成高质量语音信号.在自制 6 位数字验证码数据集和公开开放词汇数据集的实验中,UGA-SR 在限定词汇任务中的语音恢复清晰度指标达到 0.731,能够被语音识别系统准确识别;在开放词汇任务中,尽管难度较高,模型仍有效恢复了低频和高频特征,清晰度指标为 0.553.实验结果表明,UGA-SR 在语音恢复任务中具有显著优势,为从加速度信号到语音的跨模态转换提供了新思路.

关 键 词:加速度传感器;安卓系统;语音恢复;对抗神经网络;注意力机制
中图分类号:TP391 **文献标识码**:A **文 章 编 号**:1000-1220(2026)02-0309-09

UGA-SR:an Accelerometer-based Speech Recovery Model Integrating U-Net and GAN

SHEN Shu,CHEN Linhao,YUAN Xiaobo
(School of Computer,Nanjing University of Posts and Telecommunications,Nanjing 210003,China)

Abstract:This paper proposes UGA-SR,a speech recovery model that utilizes built-in accelerometer sensors in Android devices,demonstrating the feasibility of speech recovery using accelerometer data.The model combines a generative adversarial network (GAN) and an improved U-Net architecture,introducing a spatial attention module with multi-scale convolutions to optimize feature extraction and signal reconstruction.It converts preprocessed accelerometer signals into Mel-spectrograms to generate high-quality speech signals.Experiments conducted on a self-made six-digit verification code dataset and an open-vocabulary dataset show that UGA-SR achieves an STOI score of 0.731 on the limited-vocabulary task,enabling accurate recognition by speech recognition systems.In the open-vocabulary task,despite its higher difficulty,the model effectively recovers low and high-frequency features,with an STOI score of 0.553.The results indicate that UGA-SR has significant advantages in speech recovery tasks and provides a new approach for cross-modal conversion from accelerometer signals to speech.

Keywords:accelerometer;android system;speech recovery;generative adversarial network;attention mechanisms

0 引 言

近年来,随着智能设备的普及,科技在极大提升生活便利性的同时,也让隐私泄露问题日益突出.你是否曾有过这样的经历:和朋友讨论某款耳机,没过多久,购物软件上便开始推送与耳机相关的推荐内容.这一现象让人不禁思考,人们的隐私是否正在被“偷听”或“监视”.智能手机作为人们日常生活中最重要的设备之一,其安全性问题尤为关键.根据 Statista 的数据,安卓系统在 2024 年第 2 季度继续稳居全球领先的移动操作系统,市场份额约为 71%,而苹果 iOS 的市场份额为 28%[1],在智能手机广泛应用的背景下,本论文选择安卓手机作为研究对象以探索加速度计对电路板上扬声器声音信息的感知.手机内置麦克风作为最重要的信息获取组件必然是安全等级最高且系统权限等级最高,而现如今已经证明不需要获取麦克风的权限也可获得声音信息[2],智能手机有着丰富的传感器,其中加速度计和陀螺仪被认为是低风险的,它们通常被设置为零权限传感器,并且可以在没有警告智能手机用户的情况下访问.运动传感器对震动具有敏感性,当手机发出声音时会引起主板的震动,进而收集到一定的震动信号,这使运动传感器获取声音信息成为一种可能.类似于通过加速度传感器进行声音感知的方式称为侧信道攻击,关于语音侧信道攻击的方式多种多样,按照攻击的不同方式可以分为两种类型:主动激励类和被动监听类,二者的主要区别为是否主动施加信号.

主动激励类需要通过外部设备发出信号激励目标,利用这些信号与目标设备或介质的相互作用来间接获取语音信息.主动激励类攻击多种多样,Zhao 等人[3,4]通过 COTS 毫米波雷达采集 TIMIT[5]和 LJSpeech[6]语料库的射频数据,在安

收稿日期:2025-01-09 收修改稿日期:2025-02-27 基金项目:国家自然科学基金面上项目(52275535)资助;江苏省高等学校自然科学研究重大项目(22KJA520010)资助;江苏省研究生科研与实践创新计划项目(SJCX25_0348,SJCX24_0318)资助;浙江大学 CAD&CG 国家重点实验室开放课题项目(A2118)资助;南京邮电大学研究生教育教学改革课题项目(JGKT23_XJ07)资助. 作者简介:**沈 澍**,男,1982 年生,博士,副教授,CCF 高级会员,研究方向为人工智能物联网;**陈林浩**,男,1997 年生,硕士研究生,研究方向为人工智能与语音识别;**袁晓博**,男,2003 年生,硕士研究生,研究方向为计算机视觉和边缘计算.

静、噪音和隔音玻璃3种环境下完成了语音信息的恢复. SAMI等人[7]通过流行的商用机器人真空吸尘器中配备的激光雷达传感器,可以感应附近物体上引起的细微振动发出的声音. 基于主动激励的攻击具有一定的局限,因为其需要需要额外的信号源,容易被发现或干扰. 而被动监听类攻击利用现有信号泄露源或环境响应,被动窃取并分析其间接泄露的信息,具有隐蔽性强的特点. Kwong 等人[8]利用硬盘驱动器的 pes 信号感知来提取和解析人类的语音. Zhang 等人[9]采用光电探测器来捕捉由声波振动引起的环境光偏转,通过将这些光学测量值转换为音频信号.

本研究领域基于加速度计的声音感知技术属于被动监听类攻击,其原理在于通过捕获扬声器振动传导至 PCB 产生的加速度信号来重构语音信息. 其中,Gyrophone[10]作为该领域的开创性研究,采用机器学习的方式对 11 个数字单词进行识别,但其分类准确率仅达到 26%. 随后,Spearphone[11]进一步扩展了应用场景,在自建数据集上实现了性别分类和说话人识别,同时针对语音助手与通话场景的 58 个独立词语进行重建,但其词汇覆盖范围仍局限于有限集合. StealthyIMU[12]将场景扩展到语音用户界面,可以从 23 种常用语音命令中窃取私人信息,以高精度获取联系人,搜索历史记录,日历,家庭地址,甚至 GPS 跟踪. AccelEve[13]首次尝试通过神经网络技术重建 10 个数字与 26 个英文字母的发音特征,但其输出仅限于单个词语的恢复. 相比之下,iSpyU[14]采用 BLSTM 编码器-解码器架构,构建包含 9950 个单词的语音字典,在连续语音恢复任务中取得了显著进展. 然而需要指出的是,该研究仍受限于预设词汇库的约束条件,未能实现开放域语音的完整重构.

上述相关研究表明,尽管现有的研究在加速度计与声音重建领域取得了一些进展,但当前研究主要集中于分类任务或限定词汇库的语音恢复,而对于开放词汇的语音恢复,研究仍较为稀缺. 开放词汇语音恢复需要处理更复杂和多样化的语言内容,这使得模型的鲁棒性和泛化能力面临更大的挑战. 另外,许多现有的研究采用的实验设备采样频率远高于加速度计的最大采样频率(500Hz)[15],高采样率设备能够捕获更多的高频细节信息. 相比之下,加速度计的较低采样频率无法提供足够的高频信息,这对高精度语音重建产生了限制. 除此之外,加速度计采集的信号包含硬件本身的大量噪声和由人类活动的影响引起的噪声,如何消除噪音的影响也是一大挑战. 针对这些挑战,本文提出提出了一个用于低采样率加速度计的语音恢复模型 UGA-SR,该模型基于改进的 U-Net 架构和生成对抗网络(GAN)设计,力求在低采样率和噪声环境下恢复清晰的语音信号. 本文的主要贡献包括:

1)本文优化了 U-Net[16]的编码器和解码器结构,以更好地适应开放词汇的语音恢复任务. 在编码器部分,采用卷积层替代传统池化层进行下采样,保留更多频谱细节;在解码器部分,引入 PixelShuffle[17]进行上采样,PixelShuffle 常用于超分任务,避免了反卷积层常见的伪影问题. 这些改进非常适合由低分辨率加速度信号还原语音信号的任务.

2)为了减少噪声干扰,本文结合了生成对抗网络[18]以增强 UGA-SR 模型的生成能力. 通过引入判别器和生成器的对抗训练,模型能够有效降低加速度计低采样率和环境噪声带来的干扰,优化语音信号的恢复质量. 尤其在复杂信号和开放词汇语音恢复任务中,生成对抗网络提升了模型的鲁棒性,有效减少了合成噪声.

3)为了提升 UGA-SR 模型在低采样率加速度计信号中的高频信息恢复能力,本文引入了 LSK 模块[19]. 该模块通过空间注意力机制和多尺度深度可分离卷积,能够有效捕捉语音信号的空间结构信息. LSK 模块利用空洞卷积扩大感受野,并通过池化操作生成自适应的空间注意力权重,从而帮助模型提取更丰富的语音特征,有效提高开放词汇语音的重建质量.

1 基本原理

1.1 加速度计采集的工作原理

加速度计是一种基于惯性传感技术的传感器,用于测量物体的加速度. 其核心组件通常为 MEMS(微机电系统)结构,能够将物体的三轴加速度转化为电信号. 本文使用了华为 P30 Pro 智能手机作为采集设备,其内置加速度计的最大采样频率为 500 Hz,根据奈奎斯特采样定理[20],能够有效捕获人类声音的基频信息 85 ~ 250Hz. 同时,当采样频率低于信号最高频率的两倍时,会发生频率混叠现象,使得高频信号折叠到低频范围. 这表明,尽管采样设备的分辨率有限,高频信息并未完全丢失. 该特性为 UGA-SR 模型从低采样率数据中恢复声音信号提供可能. 本研究将智能手机平放在桌面上,当播放自制的 6 位数字验证码“0253944”时,加速度计的三轴响应如图 1 所示. 从中可以观察到,Z 轴的响应最为强烈,其时域波形与原始音频波形存在显著的相关性. 这表明,加速度计能够有效捕捉语音信息.

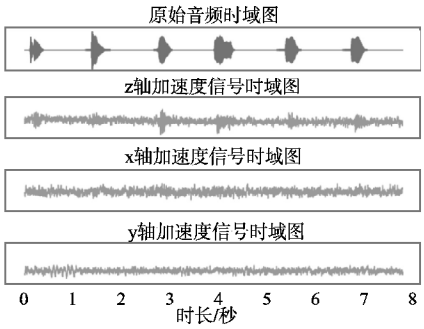


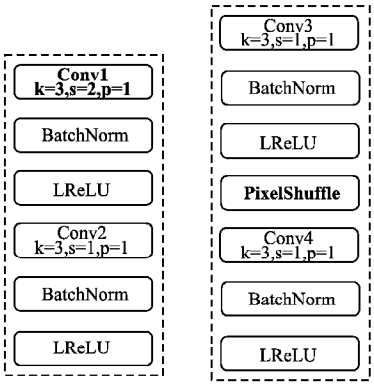
图 1 原始音频与三轴加速度信号的时域波形比较

Fig. 1 Comparison of time-domain waveforms between the original audio and the accelerometer signals

1.2 改进的 U-Net 网络

原始的 U-Net 网络存在一些局限性,首先是输入和输出之间存在尺寸不一致的问题,原论文[16]中输入尺寸为 572 × 572,输出尺寸为 388 × 388. 经过下采样和上采样操作后,输出的尺寸通常与输入不匹配,这会导致后处理困难,无法建立像素级的对应关系. 为了避免这种问题,本论文采用了适当的填充方式,确保在每层的卷积操作中,输入和输出尺寸始终保持一致. 其次,原始的 U-Net 网络下采样通常通过池化层来实现,会丢失一些高频细节,这对语音恢复等精细化任务具有负面影响. 为了更好地保留底层特征,本文用卷积层代替了池化

层,通过设置卷积核步长为 2 来实现下采样. 这种方式能够在降低特征分辨率的同时,最大限度地保留原始图像的细节信息. 最后,采用 PixelShuffle[17] 技术替代反卷积进行上采样操作,有效避免了伪影问题,产生更加平滑自然的分割结果,PixelShuffle 通过重排低分辨率特征图的通道数据,将 $r \times r \times C$ 个输入特征图通过渠道重排为 C 个 $r \times r$ 的高分辨率特征图,这种方式相对反卷积来说更加高效且避免了伪影问题,产生更加平滑自然的分割结果. 下采样和上采样模块如图 2 所示.



(a)下采样模块　　(b)上采样模块

图 2　下采样模块与上采样模块

Fig. 2　Down block and up block

1.3　生成对抗网络的基本原理

生成对抗网络(GAN)是一种由生成器和判别器组成的深度学习框架,其基本思想是通过两者的对抗性训练,实现生成数据的分布与真实数据分布的匹配. 在训练过程中,生成器的目标是生成尽可能真实的语音梅尔谱图来欺骗判别器,而判别器的目标是尽可能准确地区分真实语音的梅尔谱图与生成的梅尔谱图. 这种训练机制构成了一个动态的博弈过程. 在最理想的情况下,生成器可以生成难以区分的语音梅尔谱图. 对于判别器来说,它难以判断合成的梅尔谱图是生成的还是来自真实数据集,因此判别器对生成样本的判定概率接近 0.5,即判别器的输出接近随机猜测. 在数学上,GAN 的目标函数可以表示为:

$$\min_G \max_D V(G,D) = \mathbb{E}_{x\sim p_{data}}[\log D(x)] + \mathbb{E}_{z\sim p_z}[\log(1-D(G(z)))] \tag{1}$$

其中,G 是生成器,D 是判别器. $D(x)$ 是判别器对真实样本 x 的判定概率,$x \sim p_{data}$ 表示从真实数据分布采样的样本. $D(G(z))$ 是判别器对生成样本 $G(z)$ 的判定概率. $z \sim p_z$ 表示生成器的输入分布. 公式中的 $V(G,D)$ 表示真实语音梅尔谱图和生成语音梅尔谱图之间的差异度.

1.4　LSK 模块

为了更好地处理低采样率加速度计信号中的空间特征,本文引入了 LSK 大核选择模块[19]. 其结构如图 3 所示,主要包含多尺度特征提取和空间注意力两个关键组件.

多尺度特征提取部分采用双分支并行结构. 给定输入特征图 $X \in \mathbb{R}^{C\times H\times W}$,模块通过两个不同配置的深度可分离卷积分支实现多尺度感知. 第一分支使用标准 5×5 深度可分离卷积,专注于捕获局部精细特征. 这种设计在保持计算效率的同时,能够有效提取信号中的细节信息. 第二分支采用具有膨胀率为 3 的 7×7 空洞卷积,显著扩大了感受野范围,使模块能够获取更广泛的上下文信息. 这种双尺度特征提取机制特别适合处理低采样率信号中的多尺度时空特征. 为了降低计算复杂度并促进特征融合,模块随后通过 1×1 卷积对两个分支的特征进行降维处理. 这一操作不仅压缩了特征维度,还能够整合不同尺度的特征信息.

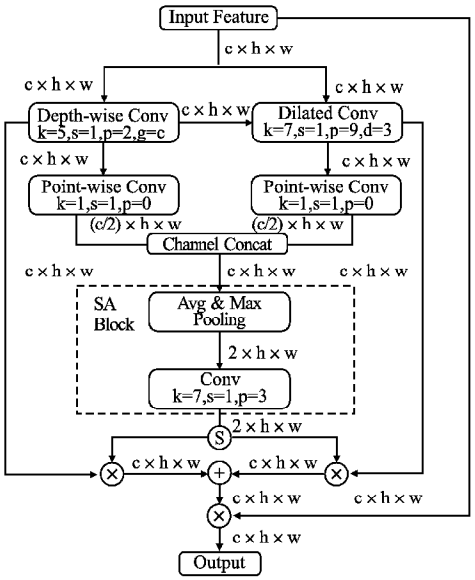


图 3　LSK 模块结构图[19]

Fig. 3　Structure of LSK block

在特征提取之后,LSK 模块引入了空间注意力机制来动态调整特征的重要性. 首先将降维后的双分支特征进行拼接,随后分别通过平均池化和最大池化两个操作获取特征的全局信息. 其中,平均池化能够提取特征图的整体统计特征,最大池化则突出了特征图中的显著响应. 这两种互补的池化操作共同构建了一个全面的特征表示. 注意力权重的生成过程可表示为:

$$W = \sigma(\mathrm{Conv}_{7\times 7}([\mathrm{AvgPool}(F), \mathrm{MaxPool}(F)])) \tag{2}$$

其中 σ 为 Sigmoid 激活函数,$W \in \mathbb{R}^{2\times H\times W}$ 表示双分支的加权系数,F 为融合后的多尺度特征. 最后,LSK 模块通过加权融合完成特征整合,得到输出特征:

$$Y = X \cdot \mathrm{Conv}_{1\times 1}(F_1' \cdot W_1 + F_2' \cdot W_2) \tag{3}$$

其中 W_1 和 W_2 分别为两分支的注意力权重,F_1'、F_2' 分别为多尺度特征提取的降维结果. 输入特征 X 与加权融合结果逐元素相乘,进一步增强了对输入信号关键区域的表征能力.

2　方法设计

2.1　整体框架

整体系统框架如图 4 所示,包括语音采集模块、预处理模块和语音还原模块. 在语音采集模块中,本文开发了一款名为 AccCapture 的数据采集应用程序. 与以往的方法相比,传统的数据集采集方式需要先播放所有音频数据集文件,再收集相应的加速度计数据生成单一文件,随后通过人工切割和对齐的方式完成数据集的整理. 这种方式在处理大型数据集(如 CSMSC[21])时,既耗时又费力. 而 AccCapture 实现了全自动

化的数据采集,在每个音频文件播放结束后会自动生成对应的加速度数据文件,无需人工对齐,大幅提升了数据采集的效率.

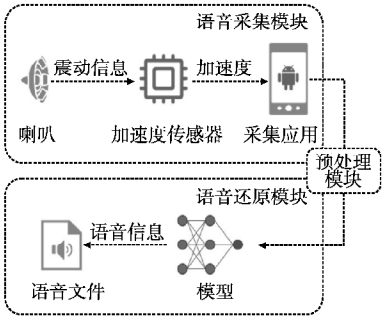


图 4 系统架构图

Fig. 4 System architecture diagram

采集的原始加速度数据经过预处理模块处理后生成加速度的频谱图,作为输入传递给语音还原模块中的 UGA-SR 模型,模型的输出为合成语音的梅尔频谱图. 梅尔频谱图是语音信号的重要频率特征表示,能够有效捕捉人类听觉系统感知的频谱信息,便于神经网络对语音特征的建模和重建. 在将梅尔频谱图转换为语音文件时,本文没有采用传统的 Griffin-Lim 声码器,而是选择了基于神经网络的语音合成器 Parallel WaveGAN[22]. 传统 Griffin-Lim 算法虽然能够将梅尔频谱图转换为语音信号,但其重建过程基于迭代优化,容易引入伪影,且生成的语音质量和自然度有限. 而 Parallel WaveGAN 通过端到端的神经网络架构,利用卷积生成器生成高质量的波形,能够将估计的语音信号梅尔频谱图转换为自然的语音波形. 相比其他神经语音合成器,本文选择 Parallel WaveGAN 的原因在于其兼具生成速度快和资源占用少的优势. 这种高效的并行波形生成器在保证生成语音自然度的同时,大幅提升了系统的实时性和适用性,使得恢复的语音信号更加接近真实语音.

2.2 预处理

AccCapture 收集后的原始三轴加速度数据,经过一系列预处理步骤,转化为适合后续语音恢复模型 UGA-SR 使用的频谱表示. 预处理过程主要包括以下 4 个关键步骤:三次样条插值、高通滤波、频谱图提取和数据归一化.

由于加速度计在语音采集过程中存在采样频率限制,最高只能达到 500 Hz,需要通过插值方法提高采样密度,从而更精确地捕获语音信号的细节变化. 本研究采用三次样条插值方法,通过拟合每一段数据点之间的平滑曲线来生成高分辨率的插值数据,将原始采样频率提升至 1000 Hz. 对于每个区间$[t_i,t_{i+1}]$,三次样条函数是一个三次多项式:

$$S_i(t)=a_i(t-t_i)^3+b_i(t-t_i)^2+c_i(t-t_i)+d_i \quad (4)$$

其中 t 是时间,a_i,b_i,c_i,d_i是通过边界条件待求的系数.

在加速度计相关噪声中,空气中声音振动对加速度计的影响较小,因此可以忽略. 然而,由于加速度计与其他电子元件共享同一电路板,并且智能手机在使用过程中不断移动,采集到的加速度信号容易受到设备噪声和外部人类活动的干扰. 人体运动通常发生在 30 Hz 以下的低频范围,设备噪声通常也是低频的,这些低频噪声对加速度数据产生较大影响. 此外,设备还可能产生直流分量噪声,如重力效应或电路噪声,进一步干扰加速度数据. 为了有效抑制这些低频噪声,本文在数据预处理过程中应用了 8 阶巴特沃斯高通滤波器,设置截止频率为 40 Hz. 通过这一处理,可以去除无关的低频成分,保留与语音信号相关的高频特征,为后续的语音恢复提供更加干净的输入信号.

接下来,为了提取加速度信号的时频特征,本文使用离散时间短时傅里叶变换 DTSTFT 对加速度信号进行频谱图提取. DTSTFT 的基本原理是通过将信号分成小的时间段并对每个时间段进行傅里叶变换得到频谱图,DTSTFT 的公式为:

$$X(t,f)=\sum_{n=-\infty}^{\infty}x[n]w[n-t]e^{-j2\pi fn} \quad (5)$$

其中,$x[n]$是原始加速度信号,$w[n]$是窗函数,t 为时间位置,f为频率. 由于加速度计的采样率较低(500 Hz),相比于传统的语音信号,梅尔频谱图并不适用于加速度信号的频谱提取. 在低采样率下,梅尔频谱图的频率分辨率较低,难以有效表示加速度信号的频率特征. 因此,DTSTFT 更适合作为加速度计数据的频谱提取方法.

为了使加速度的频谱图和语音的梅尔谱图具有一致的尺度,本文对其进行了 0-1 归一化处理,将频谱图的值映射到[0,1]之间,从而避免不同数据源间的量纲差异对后续模型训练的影响. 与零均值单位方差归一化不同,0-1 归一化能更好地适应加速度信号和梅尔谱图的特征,因为加速度信号的幅度范围较小且波动较大,标准化可能会压缩或丢失关键特征. 0-1 归一化能有效保留信号特征,避免不必要的尺度转换,从而更适合 UGA-SR 模型的训练.

以 CSMSC 数据集中的“000013. wav”处理为例,经过预处理后,原音频信号与加速度预处理后的时频特征对比如图 5 所示.

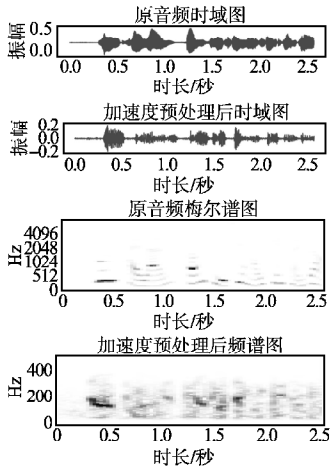


图 5 原音频信号与加速度预处理后的时频特征对比

Fig. 5 Comparison of time-frequency features between original audio and preprocessed acceleration signal

从图 5 中可以看出,加速度信号在预处理过程中通过插值、滤波和归一化处理后,去除了由设备噪声和人类活动引起的低频干扰,使得加速度信号更加专注于与语音信号相关的高频部分. 此外,加速度信号的时域波形与原音频信号的时域波形在形态上具有一定的相似性.

2.3　模型设计

本文提出了一种结合改进 U-Net 和生成对抗网络的语音恢复模型 UGA-SR. 在生成器的解码器部分,引入了多尺度特征融合和空间注意力机制的 LSK 模块,以提升语音恢复的精度与质量. 该模型的结构图如图 6 所示.

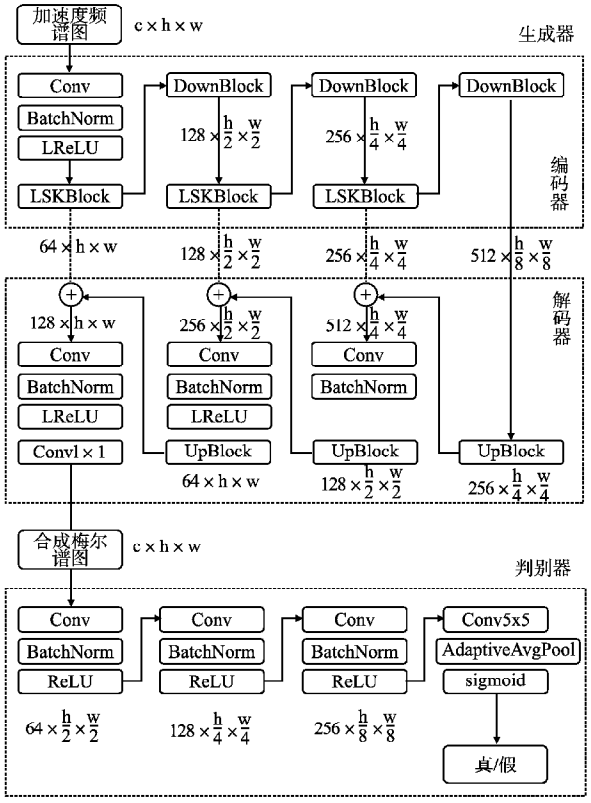


图 6　UGA-SR 模型结构图

Fig. 6　Structure of UGA-SR

生成器是 UGA-SR 模型的核心部分,负责从低采样率的加速度计信号中恢复语音. 生成器主要由编码器和解码器组成. 编码器通过多个 LSK 模块和 Downsampling 模块进行多层特征提取. 每个 Down 模块执行卷积操作,逐步减小特征图的尺寸,同时增加通道数,捕捉更多的上下文信息. 这一部分通过捕捉多尺度的低频特征来增强语音恢复的准确性. 解码器部分通过多个 Up 模块进行上采样,降低通道数,逐步恢复特征图的空间分辨率. 每个 Up 模块首先通过卷积和 PixelShuffle 操作进行上采样,然后通过跳跃连接(图 6 中虚线)将对应层的编码器特征与解码器的输出进行融合,帮助模型恢复更多细节. 跳跃连接有效地避免了特征丢失,从而提升了恢复语音的质量.

判别器的设计基于经典的卷积神经网络架构,通过多个卷积层提取特征,并最终输出判别结果. 首先 3 个卷积模块,分别使用不同的通道数和卷积核大小,每个卷积块包括卷积层、批量归一化层和 ReLU 激活函数. 判别器在最后一层使用一个 5×5 的卷积层,对输入信号进行局部区域的真假判别. 随后,通过自适应平均池化将输出尺寸降至 1×1,并使用 Sigmoid 激活函数将结果映射到 0 到 1,从而生成整体的判别结果. 其中生成器的激活函数采用 LeakyReLU,适用于生成任务;而判别器采用 Relu,适用于分类任务.

3　实验设计与结果分析

3.1　数据来源

由于基于加速度信号重建声音的研究方向较为小众且具有一定敏感性,目前尚无公开的相关数据集可供使用. 因此,本文实验数据集均为自制,包含成对的加速度信号频谱图和对应的语音信号梅尔谱图. 其中加速度信号频谱图的采集来源主要主要基于两个音频数据集:自制的 6 位验证码音频数据集 6Nums 和公开的音频数据集 CSMSC[21],数据集 6Nums 包含大量 6 位数字组合语音样本,旨在模拟实际场景中具有结构性和有限音素的语音数据. 数据集 CSMSC 包含标准普通话录音样本,主要用于消融实验以及评估模型在更复杂语音样本上的泛化能力,两个音频数据集的采样率均为 16000Hz. 具体信息如表 1 所示.

表 1　数据集

Table 1　Datasets

数据集	开放词汇	语言	文件数量	总时长(时)
6Nums	否	英语	1000	11.86
CSMSC	是	普通话	10000	2.21

以 6Nums 数据集的收集过程为例,说明实验数据集的制作流程:现代软件验证机制中,除了短信验证码外,还广泛支持电话验证码的方式. 如果验证信息一旦泄露,将会造成危险. 所以本文自制了 6 位验证码数据集 6Nums 来模拟这一场景进行实验. 首先利用 Pyttsx3 语音合成库生成 6 位数字的验证码语音文件. 通过随机生成的方式,制作了 1000 个音频样本,每个样本的时长约为 8 秒. 然后将智能手机水平放置在桌面上,通过 AccCapture 应用设备同步采集语音对应的加速度信号. 因为 z 轴反应最强烈,所以将采集到的 z 轴加速度信号进行预处理,得到加速度信号频谱图,并与语音信号的梅尔谱图配对,构成模型训练所需的成对数据集. 最后,将处理后的数据集按照 8∶2 的比例划分为训练集和测试集,用于模型的训练与验证. CSMSC 数据集的采集过程与 6Nums 数据集一致,经过同样的步骤构建成对的数据.

3.2　实验设置与训练

本实验在 Windows 10 64 位操作系统上进行,硬件配置包括 NVIDIA GeForce RTX 4070(12GB 显存)、32GB 内存,以及 Intel Core i5-12600KF 处理器. 深度学习框架使用 PyTorch 2.0,运行环境为 Python 3.10,结合 PyCharm 和 Anaconda 管理环境依赖. 实验中使用了 librosa 库进行语音信号的特征提取和处理. 加速度数据集采集设备为华为 P30 Pro 智能手机,系统为 Android 10,将加速度计开启_SENSOR_DELAY_FASTEST_模式,加速度计的采样率可以达到最高的 500hz.

为了生成适合模型输入的频谱图,本文针对不同的数据集采用了不同的超参数配置. 表 2 详细列出了 6Nums 数据集和 CSMSC 数据集的具体参数设置. 为适应不同数据集的特性,本文分别调整了频谱图尺寸、采样率、傅里叶变换点数、窗口类型与步长等参数. 加速度信号的采样率较低,而音频信号则具有更高的采样率,所以音频信号采用梅尔频谱图提取特征,其梅尔频带数分别为 256 和 80. 所有信号均使用 Hann 窗口平滑频谱图边界.

UGA-SR 模型训练如下,在 UGA-SR 模型中,生成器以加速度频谱图作为输入,输出合成的语音梅尔谱图;判别器以语音梅尔谱图和加速度频谱图作为联合输入,判断语音梅尔谱图的真实性. 当输入为生成器生成的语音梅尔谱图时,理想情况下判别器输出应该为 0(假的);当输入为真实语音的梅尔谱图时,判别器输出应该为 1(真实的). 生成器的目标是生成足够逼真的语音梅尔谱图,使判别器无法区分真假,从而实现对语音信号的重建. UGA-SR 的训练分为以下两个阶段交替进行:

表 2 超参数设置

Table 2 Hyperparameter settings

参数	6Nums 数据集		CSMSC 数据集	
	加速度	音频	加速度	音频
频谱图尺寸	256×256	256×256	80×80	80×80
预处理后采样率	1000	16000	1000	16000
傅里叶变换点数	510	1024	158	1024
窗口长度	510	1024	158	1024
窗口类型	hann	hann	hann	hann
窗口步长	32	512	32	512
梅尔频带数	—	256	—	80
截取时间	8 秒	8 秒	2.5 秒	2.5 秒

1)固定生成器 G,训练判别器 D:

判别器通过最大化交叉熵损失来优化,其目标是区分真实语音和生成语音,使得 $V(G,D)$ 值最大化,通过优化以下损失函数实现:

$$L_D = L_D^{\text{real}} + L_D^{\text{fake}} \tag{6}$$

其中,为了使判别器对真实语音的判别值 $D(x)$ 尽可能接近 1,真实数据损失:

$$L_D^{\text{real}} = -\mathbb{E}_{x\sim P_{\text{data}}(x)}[\log D(x)] \tag{7}$$

为了使判别器对生成语音的判别值尽可能接近 0,生成数据损失:

$$L_D^{\text{fake}} = -\mathbb{E}_{z\sim p_z(z)}[\log(1-D(G(z)))] \tag{8}$$

2)固定判别器 D,训练生成器 G:

生成器的目标是生成尽可能接近真实语音的梅尔谱图,使判别器误判其为真实数据. 生成器的损失函数包括以下两部分:

$$L_G = L_G^{\text{bce}} + \lambda L_G^{l1} \tag{9}$$

其中,λ 是控制 $l1$ 损失权重的超参数,本实验 $\lambda = 100$. 其中对抗性损失:

$$L_G^{\text{bce}} = -\mathbb{E}_{z\sim p_z(z)}[\log D(G(z))] \tag{10}$$

$l1$ 重构损失:

$$L_G^{l1} = \| G(z) - x \|_1 \tag{11}$$

通过计算生成语音梅尔谱图与真实语音梅尔谱图之间的 $l1$ 距离,提升语音重建的精度.

在模型训练中,为了优化生成器和判别器的参数,本文选择了 AdamW 优化器配合 Warmup 学习率调度策略. AdamW 优化器是一种改进的自适应学习率优化方法,能够在训练过程中通过权重衰减有效抑制过拟合. Warmup 调度器则在初始阶段逐步增加学习率,以避免训练初期因学习率过高导致的不稳定性,随后线性降低学习率,确保模型在后期训练中的收敛性. 针对 CSMSC 数据集,其频谱尺寸较小(80×80)但数据量较大,所以训练参数 batch size 设置为 32,epochs 为 800,生成器和判别器的学习率分别为 1e-4 和 4e-5. 而对于 6Nums 数据集,其频谱尺寸较大(256×256)J 但数据量较小,因此训练参数调整为 batch size 为 8,epochs 为 600,生成器和判别器的学习率分别为 1e-4 和 2e-4.

3.3 评估指标

由加速度信号恢复语音信号是一个跨模态任务,评估指标需要覆盖从时域到频域、从全局到局部的多个层面,为了全面评估本文提出的 UGA-SR 模型在语音恢复任务中的性能,采用以下 5 种评估指标:

短时客观可懂度(STOI)是一种基于时间频率分析的指标,其核心是通过频带相关性来度量语音的可懂度,并将短时间内的语音信号理解度量化为数值. STOI 分数与主观语音可懂度密切相关,通常分值范围在 0~1 之间,且分值大于 0.7 时一般认为恢复语音质量较高.

归一化协方差度量(NCM)是一种评估恢复语音与目标语音在带通频带内相关性的指标,反映了恢复语音与目标语音在能量分布上的一致性. 该指标基于对语音信号的带通滤波和包络计算,量化恢复语音与目标语音在多频带上的相似性. 分值越高,表示恢复语音与目标语音的相似度越高.

分段信噪比(SNGSEG)是一种信噪比相关的指标,用于衡量语音恢复过程中信号增益的质量. 其核心思想是通过逐帧计算目标语音和恢复语音的信噪比,综合分析模型对语音信号的增强效果. 分值越高,表示恢复语音的质量越好,噪声成分越少,语音信号与目标语音更为接近.

对数谱失真(LSD)衡量恢复语音与目标语音之间的频谱差异,主要用于评估语音频谱的精确性. LSD 的单位为分贝,分值越低,表示恢复语音的频谱结构与目标语音越接近,LSD 的设计特性使其在语音恢复任务中非常适用于评估语音频谱的保真性和一致性.

梅尔倒谱失真(MCD)通过计算梅尔倒谱系数的差异评估语音失真程度,是语音合成和语音转换领域的常用指标. 具体通过动态时间规整 DTW 方法对参考信号和目标信号的特征进行对齐,从而计算失真值,对齐后的特征序列逐帧计算欧氏距离,并取平均值作为最终的 MCD 指标. 单位为分贝,分值越低,表示恢复语音与目标语音在梅尔频率倒谱上的差异越小,恢复效果越佳.

3.4 消融实验与模块有效性分析

为了验证不同模块对语音恢复性能的影响,并找到最优模型,本文在开放词汇数据集 CSMSC 和有限词汇数据集 6Nums 上进行了消融实验. 实验结果分别列于表 3 和表 4. UGA-SR 模型的架构设计受到 Radio2Speech[3] 的启发,该研究提出了基于毫米波雷达信号的语音恢复方法,并验证 UNet 架构在该类非传统语音信号处理中的有效性. 针对加速度传感器语音恢复这一领域,本研究采用三阶段消融实验框架:1)基础架构验证:复现标准 UNet 作为性能基准;2)模块增量分析:通过控制变量法逐步集成改进的 Up/Down 采样、注意力机制等组件;3)架构融合实验:探究 GAN 对抗训练与 UNet-LSK 组合的协同效应. 表 3 展示了开放词汇数据集 CSMSC 数据集上的层次化消融实验结果,实验设计严格遵循单一变量原则. 实

验组在保持主干网络一致性的前提下,依次引入 Up(基于 PixelShuffle 的改进上采样模块)、Down(以卷积替代池化的改进下采样模块)、FTL(频率变换块)、SE(通道注意力模块)以及 LSK(多尺度特征融合与空间注意力模块).

表 3　CSMSC 数据集评测结果

Table 3　Evaluation results on the CSMSC dataset

模　型	STOI ↑	NCM ↑	SNGSEG ↓	LSD ↓	MCD ↓	参数量
白噪音	0.224	0.007	-10	7.302	54.041	-
Unet(base)	0.429	0.266	-2.487	2.140	2.527	**7.69m**
Unet + Up[17]	0.473	0.320	-2.008	2.232	2.107	12.436m
Unet + Down	0.477	0.321	-2.057	1.766	2.161	7.7m
Unet + Down + Up	0.491	0.339	-1.888	1.635	2.010	12.437m
Unet + FTL[3]	0.453	0.289	-2.241	1.993	2.318	11.58m
Unet + SE[23]	0.482	0.325	-1.412	1.490	1.868	7.72m
Unet + LSK[19]	0.502	0.332	-1.761	1.605	1.812	7.86m
Unet + Down + Up + LSK	0.515	0.357	-1.632	1.451	1.732	12.6m
GAN	0.493	0.344	-0.976	1.536	1.756	9.02m
GAN + LSK	0.521	0.359	-0.887	1.443	1.610	9.19m
UGA-SR	**0.553**	**0.367**	**-0.695**	**1.411**	**1.563**	13.93m

为建立可靠的性能基准,首先复现标准 UNet 架构作为基础模型:由表 3 可以看出,基础 U-Net 模型在语音恢复任务中相较于白噪声表现显著提升,STOI 提升至 0.429,NCM 达到 0.266,LSD 和 MCD 分别下降至 2.140 和 2.527,表明其在改善语音信号的可懂度和频谱恢复方面已具备一定能力. 在控制主干网络一致性的前提下,通过渐进式模块集成策略评估各创新组件的边际贡献:单独添加 Up 或 Down 模块时,STOI 分别提升至 0.477 和 0.473,且 MCD 分别降低至 2.107 和 2.161. 两者联合时,模型性能进一步提升,STOI 达到 0.491,相较于基准提升了 14.5%,且 NCM 达到 0.339,表明改进的上下采样模块有助于捕获更多的特征信息. FTL 模块对模型性能有一定提升,但效果有限. 加入 FTL 模块后,STOI 从 Unet 的 0.429 提升至 0.453,仅提高约 5.6%;LSD 从 2.140 降低至 1.993,减少约 6.9%. 这表明频率变换块虽能增强频域特征提取能力,但其贡献相对较小. 相比之下,注意力机制的引入带来了更显著的提升. SE 通道注意力模块的加入使 STOI 从 Unet 的 0.429 提升至 0.482,提高约 12.4%;LSD 从 2.140 降低至 1.490,减少约 30.3%,证明了通道注意力在提升关键特征关注方面的有效性. 然而,LSK 空间注意力模块的效果更为显著,STOI 从 Unet 的 0.429 提升至 0.502,提高约 17%;LSD 从 2.140 降至 1.605,减少约 25%,表明多尺度特征与空间注意力的结合显著增强了模型的特征表达能力.

最终整合 GAN 对抗训练框架与 UNet-LSK 组合架构,实现跨模态优化:GAN 框架的引入显著改善了噪声抑制效果,SNGSEG 从 Unet 的 -2.487 提升至 -0.976,结合 LSK 后进一步提升至 -0.887. 最终模型 UGA-SR 综合性能最佳,STOI 达到 0.553,LSD 和 MCD 分别降至 1.411 和 1.563,相比基准模型 LSD 降低 34.0%,MCD 降低 38.2%,在信噪比增益和语音质量指标上实现了显著提升. 而本文提出的 UGA-SR 模型综合了 GAN 架构、Unet 以及改进的 Down、Up 和 LSK 模块,实现了整体性能的最大化. 与 Unet 相比,UGA-SR 的 STOI 从 0.429 提高至 0.553,提升约 28.9%;LSD 从 2.140 降至 1.411,减少约 34%;MCD 从 2.527 降至 1.563,减少约 38.2%. 这一显著提升验证了所提出模块设计的有效性,特别是 LSK 模块在模型性能优化中的关键作用. 尽管 UGA-SR 模型的参数量为 13.93M,是消融实验中模型中最大的,但实验结果表明,参数量的增加并不一定与性能提升呈正相关. 例如,在引入 FTL 模块的实验中,虽然参数量有所增加,但评测指标未达到最优.

表 4　6Nums 数据集评测结果

Table 4　Evaluation results on the 6Nums dataset

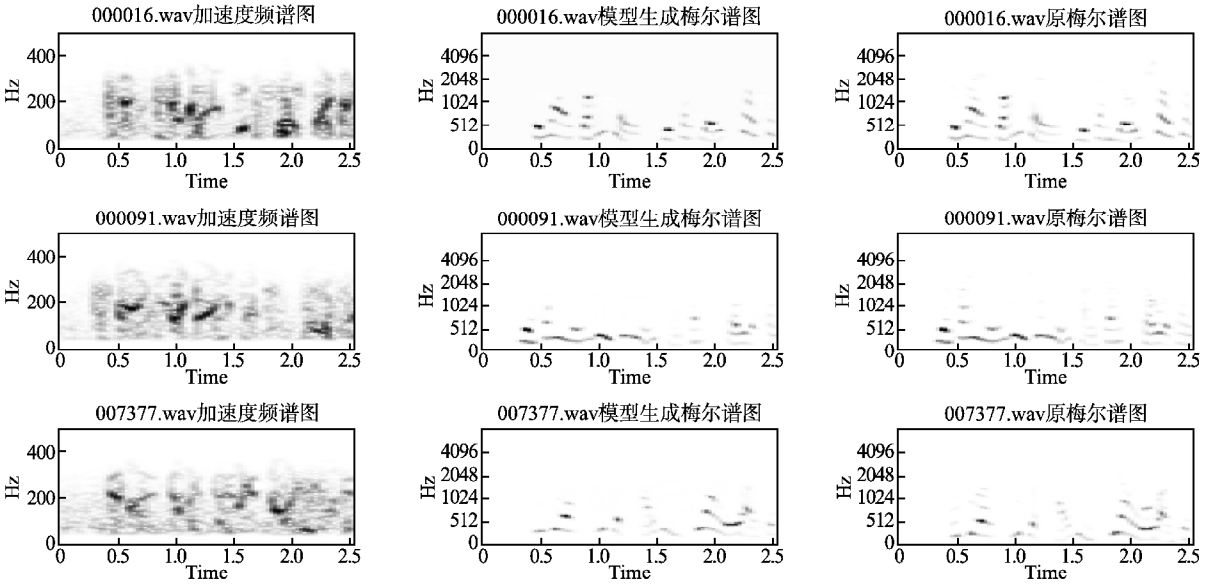
模　型	STOI ↑	NCM ↑	SNGSEG ↓	LSD ↓	MCD ↓	参数量
白噪音	0.295	0.001	-10	8.947	52.786	-
Unet(base)	0.692	0.584	-6.864	4.272	1.350	**7.69m**
Unet + LSK[19]	0.708	0.609	-6.800	3.750	1.333	7.86m
GAN	0.716	0.618	-6.291	3.565	1.267	9.02m
UGA-SR	**0.731**	**0.621**	**-6.129**	**3.518**	**1.235**	13.93

在有限词汇数据集 6Nums 上进行了进一步实验,6Nums 数据集模拟电话语音验证码场景,具有特定词汇和有限音素的特点. 由于在开放词汇数据集 CSMSC 上已经验证了 UGA-SR 的有效性,且全面评估了各模块对语音恢复性能的影响,因此在 6Nums 数据集上没有对每个模块单独进行实验,而是聚焦于 LSK 和 GAN 模块的消融实验,以进一步分析其在特定任务中的作用. 由表 4 可以看出,各模型均在语音恢复任务上总体提升幅度有限,但取得了一定的性能提升. 在该实验

中,短时客观可懂度 STOI 成为衡量模型性能的重要指标. 值得注意的是,STOI 在多个模型中均接近或超过了 0.7,其中基础 U-Net 模型的 STOI 达到 0.692,表现出对有限词汇语音数据的较好恢复能力. 引入多尺度特征融合和空间注意力的 LSK 模块后,STOI 提升至 0.708,同时其他指标如 NCM 和 MCD 也有显著优化. 进一步引入 GAN 框架后,STOI 达到 0.716,语音信号的主观可懂度和频谱恢复效果进一步提升. 最终,采用 UGA-SR 模型后,STOI 达到 0.731,MCD 降至 1.235. 通常认为 STOI 分数超过 0.7 通常被认为语音质量已达到较高水平. 且实验还验证了生成的语音可以被谷歌 ASR 准确识别. 这一结果验证了 UGA-SR 模型在有限词汇语音数据恢复任务中的优越性能,同时能够生成更清晰、准确的语音梅尔谱图.

3.5 频谱图分析

图 7 展示了开放词汇数据集 CSMSC 在不同阶段的频谱图对比,包括图 7(a)加速度信号频谱图、图 7(b)UGA-SR 模型生成的梅尔谱图以及图 7(c)目标音频原始梅尔谱图. 从图 7(a)可以看出,加速度信号的频谱图包含多种叠加信号及噪声成分,频率主要集中在 500Hz 以下. 这些频谱特征与目标音频的原始梅尔谱图(图 7(c))在频率分布和结构上几乎没有直接相似性. 然而,通过 UGA-SR 模型生成的梅尔谱图(图 7(b))与原始音频的梅尔谱图相比,在低频和高频区域均表现出显著的还原效果,有效恢复了语音的关键频谱特征,并抑制了原始信号中的噪声干扰. 进一步将 UGA-SR 模型生成的梅尔谱图通过 Parallel WaveGAN 声码器进行还原,结合音频主观听觉评估,恢复的语音展现了较高的清晰度和可懂度. 这表明 UGA-SR 模型在开放词汇语音恢复任务中具有较强的适用性和稳定性.



(a)加速度信号的频谱图 (b)模型生成的梅尔谱图 (c)目标音频的原始梅尔谱图

图 7 数据集频谱对比图

Fig. 7 Spectrogram comparison of datasets

4 结束语

本文通过利用加速度计感知声音振动的特性,提出了一种跨模态的语音恢复模型 UGA-SR,实现了从加速度信号到语音信号的有效重建. UGA-SR 结合了生成对抗网络 GAN 与改进的 U-Net 架构,并引入了多尺度融合卷积的空间注意力模块,显著提升了语音恢复的准确性和质量. 在实验中,UGA-SR 分别在限定词汇数据集 6Nums 和开放词汇数据集 CSMSC 上进行了验证. 在限定词汇任务中,模型恢复的语音信号能够被谷歌 ASR 系统准确识别,展现了出色的性能. 在开放词汇任务中,虽然恢复效果相比限定词汇任务略逊一筹,但 UGA-SR 依然能够较好地还原原始音频的低频和高频特征,充分体现了模型在处理复杂语音任务中的鲁棒性和适应性. 未来的研究可以从生成语音的方式出发,进一步优化为流式实时生成,提升模型在实际应用场景中的响应速度. 同时,作为一个通用的跨模态语音恢复框架,本模型还具备广泛的扩展潜力,未来工作可尝试应用于其他模态场景,探索更多跨模态信息转化的可能性,为相关领域的发展提供新的思路和技术支持.

References:

[1] Global market share held by mobile operating systems from 2009 to 2024[EB/OL]. https://www.statista.com/statistics/272698/global-market-share-held-by-mobile-operating-systems-since-2009/, 2024.

[2] Michalevsky Y, Boneh D, Nakibly G, et al. Gyrophone: recognizing speech from gyroscope signals[C]//Proceedings of the 23rd USENIX Conference on Security Symposium (SEC'14), 2014: 1053-1067.

[3] Zhao R, Yu J T, Li T, et al. Radio2Speech: high qual-ity speech recovery from radio frequency signals[C]//Proceedings of Interspeech, 2022: 4666-4670.

[4] Zhao R, Yu J T, Zhao H, et al. Radio2Text: streaming speech recognition using mmWave radio signals[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2023, 7(3): 1-28.

[5] Zue V, Seneff S, Glass J. Speech database development at mit: timit and beyond[J]. Speech Communication, 1990, 9(4): 351-356.

[6] Ito K, Johnson L. The lj speech dataset[EB/OL]. https://keithito.com/LJ-Speech-Dataset, 2017.

[7] Sami S, Dai Y, Tan S R X, et al. Spying with your robot vacuum cleaner: eavesdropping via lidar sensors[C]//Proceedings of the 18th Conference on Embedded Networked Sensor Systems, 2020: 354-367.

[8] Kwong A, Xu W, Fu K, et al. Hard drive of hearing: disks that eavesdrop with a synthesized microphone[C]//IEEE Symposium on Security and Privacy (SP), 2019: 905-919.

[9] Zhang G, Fu H, Xiang Z, et al. Ambient light reflection-based eavesdropping enhanced with cGAN[J]. IEEE Transactions on Mobile Computing, 2025, 24(1): 72-85.

[10] Michalevsky Y, Boneh D, Nakibly G. Gyroph-one: recognizing speech from gyroscope signals[C]//23rd USENIX Security Symposium, 2014: 1053-1067.

[11] Anand S A, Wang C, Liu J, et al. Spearphone: a lightweight speech privacy exploit via accelerometer-sensed reverberations from smartphone loudspeakers[C]//Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks, 2021: 288-299.

[12] Sun K, Xia C, Xu S, et al. StealthyIMU: stealing permission-protected private information from smartphone voice assistant using zero-permission sensors[C]//Network and Distributed System Security (NDSS), 2023.

[13] Ba Z, Zheng T, Qin Z, et al. Accelerometer-based smartphone eavesdropping[C]//Proceedings of the 26th Annual International Conference on Mobile Computing and Networking, 2020: 1-18.

[14] Zhang S, Liu Y, Gawda M K. I spy you: eavesdropping continuous speech on smartphones via motion sensors[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2023, 6(4): 1-31.

[15] Google. Android api reference[EB/OL]. https://developer.android.com/reference/, 2024.

[16] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation[J]. arXiv eprints, 2015, doi: 10.48550/arXiv.1505.04597.

[17] Shi W Z, Caballero J, Huszár F, et al. Real-time single image and video super-resolution using an efficient subpixel convolutional neural network[J]. arXiv eprints, 2016, doi: 10.48550/arXiv.1609.05158.

[18] Goodfellow I J, Pouget Abadie J, Mirza M, et al. Generative adversarial networks[J]. arXiv eprints, 2014, doi: 10.48550/arXiv.1406.2661.

[19] Li Y X, Li X, Dai Y, et al. LSKNet: a foundation light-weight backbone for remote sensing[J]. International Journal of Computer Vision, 2024, doi: 10.1007/s11263-024-02247-9.

[20] Wikipedia. Nyquist-Shannon sampling theorem[EB/OL]. https://en.wikipedia.org/wiki/Nyquist-Shannon_sampling_theorem/, 2024.

[21] Databaker. TNtts dataset[EB/OL]. https://test.data-baker.com/data/index/TNtts/, 2024.

[22] Yamamoto R, Song E, Kim J M. Parallel WaveGAN: a fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram[J]. arXiv eprints, 2019, doi: 10.48550/arXiv.1910.11480.

[23] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 7132-7141.