计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
2025,34(11):127−138 [doi: 10.15888/j.cnki.csa.009996] [CSTR: 32024.14.csa.009996]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于双流门控视听融合的多模态语音增强①

彭敏轩, 梁　艳

(华南师范大学 人工智能学院, 佛山 528225)
通信作者: 梁　艳, E-mail: liangyan@m.scnu.edu.cn

摘　要: 针对现有音视频语音增强方法在复杂场景下存在的鲁棒性不足、多模态信息融合效率低下、计算复杂度高等问题, 本文提出一种双流门控视听融合 (dual-stream gated audio-visual fusion, DS-GAVF) 架构. 该架构通过静态与动态视觉协同建模, 以及细粒度的跨模态交互, 实现语音增强性能的提升. 在特征提取阶段, 采用 U-Net 编码音频时频特性, 同时设计双流视觉网络, 采用 ResNet-18 提取单帧面部静态特征, 使用时空图卷积网络捕捉连续面部标志点的动态运动模式. 为解决视听模态时序差异, 提出动态时间插值对齐方法, 并设计了一种门控交叉注意力机制, 通过门控单元动态调节特征融合权重, 在视觉可信度低时抑制噪声干扰. 在解码阶段, 通过跨模态跳跃连接, 将多尺度视觉线索注入 U-Net 解码层, 最终输出目标语音时频掩码. 实验结果表明, DS-GAVF 在 3 个数据集上的混合噪声环境下均表现出优异性能. 与现有方法相比, DS-GAVF 在保持较低计算复杂度的同时, 实现了语音质量、可懂度与鲁棒性的协同提升.

关键词: 语音增强; 视听融合; 门控交叉注意力; U-Net

引用格式: 彭敏轩,梁艳.基于双流门控视听融合的多模态语音增强.计算机系统应用,2025,34(11):127–138. http://www.c-s-a.org.cn/1003-3254/9996.html

Multimodal Speech Enhancement Based on Dual-stream Gated Audio-visual Fusion

PENG Min-Xuan, LIANG Yan

(School of Artificial Intelligence, South China Normal University, Foshan 528225, China)

Abstract: Given the problems of insufficient robustness, low efficiency of multimodal information fusion, and high computational complexity in existing audio-visual speech enhancement methods under complex scenarios, a dual-stream gated audio-visual fusion (DS-GAVF) architecture is proposed. This architecture achieves performance improvement in speech enhancement via static-dynamic visual collaborative modeling and fine-grained cross-modal interaction. In the feature extraction stage, U-Net is employed to encode the audio's time-frequency characteristics, and meanwhile a dual-stream visual network is designed. ResNet-18 is adopted to extract static features of single-frame facial images, and a spatiotemporal graph convolutional network is utilized to capture dynamic motion patterns of continuous facial landmarks. Additionally, a dynamic temporal interpolation alignment method is proposed to solve the time series difference of audio-visual modalities, and a novel gated cross-attention mechanism is designed, which dynamically adjusts the feature fusion weights via a gated unit and suppresses noise interference when visual reliability is low. In the decoding stage, cross-modal jump connections are leveraged to inject multi-scale visual cues into the U-Net decoding layers, finally outputting the time-frequency mask of the target speech. Experimental results show that DS-GAVF yields excellent performance under mixed noise environments on three datasets. Compared with existing methods, DS-GAVF achieves a synergistic improvement in speech quality, intelligibility, and robustness while maintaining low computational complexity.

① 基金项目: 广东省自然科学基金面上项目 (2025A1515011526)
收稿时间: 2025-03-24; 修改时间: 2025-05-07; 采用时间: 2025-05-26; csa 在线出版时间: 2025-09-30
CNKI 网络首发时间: 2025-10-09

Key words: speech enhancement; audio-visual fusion; gated cross-attention; U-Net

1 引言

语音增强是音频信号处理领域的重要研究方向和长期研究热点, 其核心目标是有效去除背景噪声, 最大程度保留语音信号的清晰度和可懂度, 以显著提升语音信号质量[1]. 该技术在诸多实际应用场景中发挥关键作用, 如保障嘈杂环境中语音通信的清晰度、提升助听设备的语音感知体验、增强语音识别系统在复杂环境下的抗噪声能力与鲁棒性[2]. 随着智能音箱、语音助手等设备的普及, 其重要性愈发凸显, 成为推动语音交互技术发展的重要支撑.

近年来, 基于深度学习的语音增强方法取得了显著进展. Richter 等人[3]将基于扩散的生成模型应用于语音增强领域. Lin 等人[4]提出了一种基于两阶段掩码变换器和信息交互的新型生成对抗网络, 显著提升了语音分离的效果. 在实际应用中, 语音增强系统的设计需要综合考虑噪声抑制和语音失真之间的平衡. 尤昕源等人[5]提出一种用于复数频谱映射的门控膨胀卷积循环网络, 有效利用了全局语音进行语音增强. 该系统在 TIMIT 电话语音数据集上的语音质量感知评估 (perceptual evaluation of speech quality, PESQ) 为 1.55, 短时客观可懂度 (short-time objective intelligibility, STOI) 为 81.15%, 验证了其在噪声抑制和语音质量保持方面的有效性. 这些方法仅利用音频信息提升语音质量, 实现简单, 应用广泛. 然而, 在信噪比极低、噪声与语音信号高度相关等特定场景下, 其性能受到限制.

考虑到人类对语音的感知是多模态的, 与视觉信息关联性强, 因此, 近年来有学者尝试探索结合视觉信息的解决方案[6]. 例如, 利用面部特征或语音特征嵌入身份信息, 进行特定身份的语音分离[7,8]; 基于静态面部特征信息, 结合音视频分离乐器声音等[9]. 然而, 这些方法仅依赖静态面部特征线索, 忽略了人说话时的动态面部运动信息 (如嘴唇动作), 而这些动态信息对语音增强和分离任务有重要补充作用.

最近有研究进一步尝试将面部运动信息 (如目标说话者的嘴唇动作) 作为深度学习框架的输入之一, 结合静态与动态视觉线索进行语音增强. 这些方法充分利用了多模态信息的优势, 显著提升了语音增强和语音分离任务的性能, 为解决复杂场景下的语音处理问题提供了新思路. 例如, Sadeghi 等人[10]采用音视频变分自编码器 (variational autoencoder, VAE) 进行语音增强, 在 GRID 音视频句子语料库数据集, 信噪比 (signal-to-noise ratio, SNR) 为−5 dB 的音频上获得 0.32 的 PESQ 和 9.5 dB 的信号失真比 (signal-to-distortion ratio, SDR). Ayilo 等人[11]提出了一种无监督视听语音增强方法, 在 TCD-TIMIT 数据集上获得 2.48 的 PESQ.

尽管利用视频中的运动线索来分离背景噪声, 可以克服静态面部属性特征的局限性[12]. 然而, 当人脸运动信息不可靠时, 例如唇部被遮挡或人脸运动信息不明显, 模型性能可能会出现明显下降. 因此, 在语音分类任务中, 若仅依赖人脸运动信息中的嘴唇动作信息进行语音分类, 往往会导致模型忽略具有重要价值的面部静态特征信息, 如年龄、性别、脸型等生物特征, 它们能够提供关于说话人声音特征的重要线索.

最近有几项研究结合了静态和动态视觉线索来提升语音分离的性能. 例如, Gao 等人[13]通过声音和面部特征信息嵌入, 联合训练音视频说话人分离网络, 为分离过程提供先验帮助, 在 VoxCeleb2 数据集获得的 SDR 为 10.2 dB, PESQ 为 2.83. 此外, 有学者尝试在唇动和目标说话人声音嵌入的基础上对分离网络进行调节, 以提高对视觉遮挡的鲁棒性[14]. Rahimi 等人[15]提出了一个基于同步或异步线索的多模态语音分离与增强统一框架, 在 LRS3-TED 数据集上的 SDR 为 15.5 dB, PESQ 为 2.63, STOI 为 93.5%.

在语音增强领域, 多模态信息融合已成为提升语音分离和增强性能的重要研究方向. 然而, 基于音视频的多模态语音增强研究仍然存在诸多亟待解决的问题. 尽管现有模型通过整合静态和动态的视觉线索 (如唇部运动、面部表情等) 在一定程度上改善了语音分离的性能, 但多模态信息融合效率低的问题仍然突出. 例如, 文献[13–15]中所采用的融合方法容易引入冗余噪声, 且缺乏跨模态的复杂交互能力. 这种低效的融合方式限制了模型对多模态信息的充分利用, 限制了性能的提升.

此外, 许多已有研究虽然在实验环境中取得了良好的性能表现, 但其计算复杂度高、使用性低的特性使其难以在实际场景中得到广泛应用. 例如, 文献[16]

提出的多模态模型虽然在性能上表现优异, 但其庞大的模型参数量为前期训练和后期部署带来了巨大挑战, 不仅增加了对硬件资源的需求, 还可能导致模型在实际应用中出现响应延迟等问题, 影响用户体验.

综上所述, 尽管基于音视频的多模态语音增强技术在理论研究上已取得一定进展, 但在实际应用中仍面临多模态信息融合效率低、计算复杂度高、使用性低等关键问题. 这些问题不仅限制了现有方法在复杂场景中的性能表现, 也阻碍了多模态语音增强技术在智能语音助手、会议系统、语音识别等实际应用场景中的广泛应用. 因此, 探索高效、轻量化的多模态信息融合方法, 以实现性能与实际应用需求的平衡, 仍是一个具有重要研究价值和现实意义的方向.

针对上述问题, 本文对基于音视频的多模态语音增强算法展开研究, 提出一种视听融合架构对语音增强算法进行优化提升, 利用针对语音增强和语音分离任务设计的数据处理模块优化模型的训练以及预处理效果, 结合门控交叉注意力机制的细粒度融合、多尺度融合机制, 实现视觉模态与听觉模态信息的高效融合, 在降低整体计算复杂度的同时, 增强模型在复杂噪声环境下获取有效信息的能力, 提升模型的鲁棒性.

2 模型结构

为了解决基于音视频的语音增强方法鲁棒性不足、多模态信息融合效率低、计算复杂度高等问题, 本文提出了一种双流门控视听融合 (dual-stream gated audio-visual fusion, DS-GAVF) 架构, 该架构由双流视听特征提取模块、门控视听特征融合模块和视听特征解码模块构成, 如图 1 所示.

DS-GAVF 采用双流视听特征提取模块分别从音频和视频中提取特征. 音频特征通过短时傅里叶变换 (short-time Fourier transform, STFT) 获取频谱图, 并利用 U-Net 进行编码; 视觉特征则是通过静态流 (基于 ResNet-18 提取全局语义特征) 和动态流 (基于 ST-GCN 提取时序动作特征) 进行建模, 再通过特征对齐与融合形成视觉特征向量. 在门控视听特征融合模块中, 我们采用动态时间插值对齐视听模态的时间维度, 并引入门控交叉注意力机制 (gated cross-attention) 实现细粒度的特征交互, 得到融合特征. 最后, 通过 U-Net 解码层和跳跃连接进行特征解码, 输出目标语音的频谱图掩码, 并结合逆短时傅里叶变换 (inverse short-time Fourier transform, ISTFT) 重建出目标语音的时域波形.

2.1 双流视听特征提取

现有方法在利用面部信息提升语音增强模型性能时, 主要存在以下两种局限: (1) 仅截取唇部信息作为面部运动信息的来源, 忽视了人在说话时面部肌肉的全局运动信息及人脸本身自带的面部特征信息. (2) 对整个面部运动图像进行编码, 虽然能够涵盖所有面部运动信息, 但其计算成本巨大, 导致在训练和应用过程中, 特别是在计算资源受限的场景下, 需要较长时间完成, 甚至无法直接使用. 针对这些问题, 本文提出了一种轻量级的音视频特征提取网络, 充分利用视频帧面部运动信息和面部属性特征信息, 辅助语音增强.

2.1.1 音频信息提取

在语音增强任务中, 如何选择合适的方法高效提取音频信息, 并使其能够更好地应用于后续的多模态融合, 是本研究的关注点. 为此, 原始音频经过短时傅里叶变换及下采样, 得到频谱图, 并结合音频编码器来生成嵌入表示. 这一过程不仅能够有效提取音频特征, 还保留了时间分辨率, 避免了关键时频信息的丢失. 随后, 我们将处理后的音频数据输入到一个专门设计的 U-Net[17]风格网络中, 用于进一步的处理和特征提取. 该网络采用了经典的编码器-解码器结构, 其中编码器和解码器有 L 层, 每层由一维卷积神经网络 (1D CNN)、线性整流激活函数层 (ReLU)、卷积层和门控线性单元激活层 (GLU) 组成, 以增强网络的非线性建模能力.

设原始含有噪声的音频信号为 X_{audio}, 通过 STFT 获得一个维度为 $N_s \times D_s \times F_s \times T_s$ 的频谱图 S, 其中 N_s、D_s、F_s、T_s 分别代表频谱图的采样维数、通道维数、频率维数和时间维度. 每个时频点包含了频谱图值的实部和虚部. 随后, 频谱图 S 被输入到 U-Net 编码器进行特征编码. 在编码器中, 输入数据经过一系列卷积层处理, 并在卷积层之间插入频率池化层, 以在保持时间维度的同时降低频率维度的复杂性. 最终, 编码器输出一个维度为 $N_m \times D_{fa} \times F_{fa} \times T_{fa}$ 的音频特征映射 F_{audio}, 其中 N_m、D_{fa}、F_{fa}、T_{fa} 分别代表音频特征映射 F_{audio} 的采样维数、通道维数、频率维数和时间维度.

2.1.2 面部运动信息提取

面部肌肉的运动信息对指导语音增强具有重要意义. 然而如何以较低成本提取出对语音增强有指导作用的视频信息是关键问题. 针对这一问题, 本文结合面部标志点和时空图卷积网络, 提取面部运动信息.

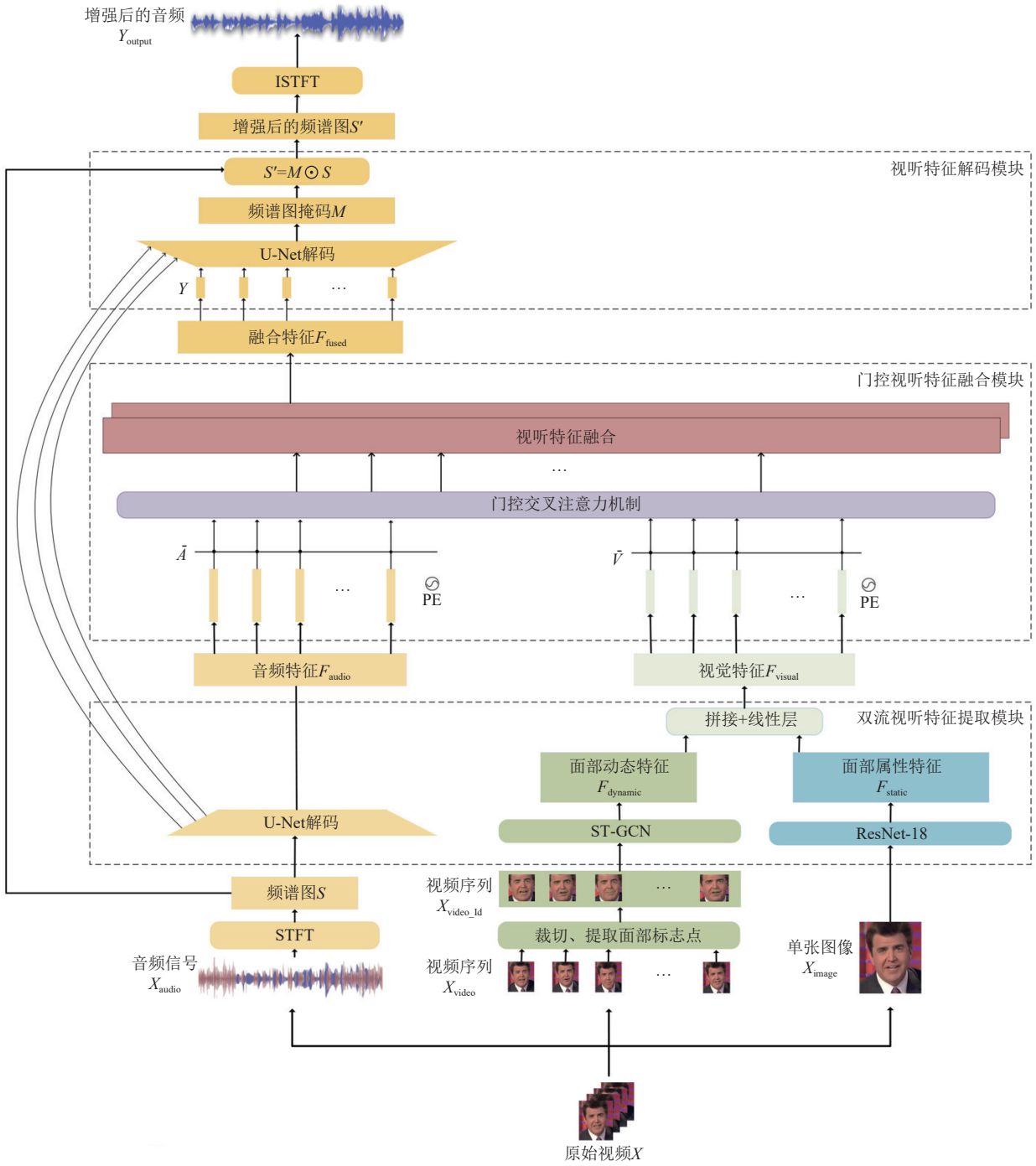


图 1 双流门控视听融合架构

将输入的视频序列 X_{video} 经过数据预处理, 得到一组面部标志点 X_{video_Id}. 随后, 把 X_{video_Id} 输入 ST-GCN 网络, 提取视觉信息中的运动信息表示 $F_{dynamic}$. 本文设计的 ST-GCN 网络参考文献[18], 该网络由一系列时空卷积块组成, 每个卷积块先对每一帧的图进行空间图卷积, 提取空间特征; 然后在时间维度上对这些空间特征进行卷积操作, 提取时间特征. 这种方法显著减少了需要处理和存储的数据量, 从每帧约 30 000 个值减少到每帧约 100 个值. 在处理大型视听数据集时, 这种数据量的减少不仅大幅降低了存储需求, 还提高了训练速度, 同时保持了模型性能.

2.1.3 面部属性特征信息提取

面部属性特征信息在语音增强任务中具有重要价值. 静态面部属性特征 (如性别、年龄等) 与语音特征

之间存在显著的相关性, 例如不同性别和年龄段的语音基频通常表现出显著差异. 这些静态属性特征可以作为可靠的视觉线索, 辅助语音分离任务的进行.

对于面部属性特征提取网络, 我们参考 ResNet-18[19] 设计了网络架构. 具体而言, 面部属性特征提取网络从输入的视频序列 X_{video} 中采样最具代表性的单个面部图像 X_{image}, 经过人脸区域检测、裁切以及归一化处理后, 输入到 ResNet-18 网络进行特征提取. 将网络提取出的面部属性特征在运动信息表示 V_f 的时间维度 T_f 上进行复制, 得到面部属性特征 F_{static}.

随后, 将面部属性特征在运动信息表示 F_{dynamic} 的时间维度 T_{fd} 上进行复制, 使其与提取出的面部运动表示 F_{dynamic} 在特征空间中进行连接. 连接后的特征经过线性层的映射, 最终得到视觉特征 F_{visual}:

$$F_{\text{visual}} = F_{\text{dynamic}} + F_{\text{static}} \tag{1}$$

其中, F_{visual} 维度为 $N_{fv} \times T_{fv} \times D_{fv}$, N_{fv} 表示输入维数, T_{fv} 表示时间维度, D_{fv} 表示通道维数.

在这一设计中, 面部属性特征被用作身份编码, 其主要目的是识别说话人声音的频率特性或其他音频属性的空间分布. 这有助于增强模型对目标说话人身份和发音习惯的建模能力. 与此同时, 说话时的面部运动特征则用于从噪声环境中识别并分离出目标语音信息. 通过将静态面部属性特征与动态面部运动特征相结合, 我们能够获取互补的视觉线索, 从而为语音分离任务提供更全面的指导. 这种设计不仅充分利用了静态面部属性与动态面部运动特征的互补性, 还显著提升了模型在复杂噪声环境下的语音分离性能.

2.2 多模态特征融合

音频特征和视觉特征的融合方式是多模态语音增强研究的关键问题. 目前, 主流的特征融合方案主要包括相加 (add) 和拼接 (concat) 两种方式. 虽然这些方法简单易行, 但它们可能无法充分利用两种模态的信息. 在多模态语音增强任务中, 神经网络往往倾向于依赖更可靠的音频模态, 而忽视视频模态提供的有用信息. 这种过度依赖可能导致模态失衡, 影响模型整体性能[20].

为了解决这一问题, 本文设计了一种双流门控视听融合的多模态特征融合结构. 该结构通过基于 Transformer 的架构添加位置编码, 利用多头注意力机制捕捉长时依赖关系, 并引入门控交叉注意力机制, 有效控制音频特征和视觉特征的信息流量, 确保网络能够正确学习视觉动态信息与音频信号之间的关系, 并在适当的时候使用相应的模态信息.

具体来说, 首先对音频信息和视觉信息进行特征对齐, 其中视觉信息包含面部运动信息以及面部属性特征信息. 对于音频特征 F_{audio}, 将其展平为时序特征; 对于视觉特征 F_{visual}, 由于音频采样率与视频采样率不同, 使用线性插值使得 F_{audio} 和 F_{visual} 对齐时间步, 生成在时间维度上匹配的音频信息表示 A 和视觉信息表示 V. 随后, 为了让模型了解其输入的时间顺序, 即视频和音频特征的信号时间戳, 本文添加了位置编码 $PE_{\{a,v\}} \in \mathbb{R}^{t_{\{a,v\}} \times c}$, 其中, $PE_{\{a,v\}}$ 为正弦向量, 从而得到了包含时序信息的音频信息表示$\bar{A}$和视觉信息表示$\bar{V}$. 每个模态表征的计算如下:

$$\bar{A} = A + PE_a \tag{2}$$

$$\bar{V} = V + PE_v \tag{3}$$

沿着时间维度把这些特征串联起来, 获得音视频特征:

$$Z = (\bar{A}; \bar{V}) \in \mathbb{R}^{(t_a + t_v) \times c} \tag{4}$$

其中, t_a 和 t_v 分别表示音频特征和视觉特征的时间步数, c 为音频和视觉特征向量的维度.

接着, 我们为其添加跨模态门控交叉注意力机制. 将 Query 设置为音频特征, Key/Value 设置为视觉特征, 生成对应的注意力权重矩阵, 该矩阵用于表示每个音频时间步对视觉时间步的关注程度, 并根据注意力矩阵生成加权后的 Value 输出. 然后, 将原始音频特征与注意力输出拼接, 通过全连接层和 Sigmoid 生成门控权重矩阵, 最终得到融合特征. 融合特征计算如下:

$$F_{\text{fused}} = A + G \odot E \tag{5}$$

其中, A 表示音频特征, G 表示门控权重矩阵, E 表示注意力权重矩阵, $\odot$表示元素级的乘积.

2.3 视听特征解码

在视听特征解码模块中, 我们采用与编码器相反的处理流程, 通过上采样和重建操作, 实现 U-Net 解码, 将编码器生成的低维特征映射逐步恢复为高分辨率的时频表示. 具体而言, 解码器由 L 层组成, 每层包含反卷积层、ReLU 激活函数和 GLU 激活函数, 以增强网络的非线性建模能力. 在解码过程中, 首先将融合后的特征映射 F_{fused} 输入到解码器中, 通过反卷积操作逐步恢复频率维度的分辨率, 同时保持时间维度的连贯性.

最终, 解码器输出一个维度为 $N_m \times D_m \times F_m \times T_m$ 的掩码表示 M, 其中 N_m、D_m、F_m、T_m 与编码器输入的频谱图维数大小一致. 随后, 我们将掩码表示与原始频谱图相乘, 得到增强后的频谱图:

$$S' = M \odot S \tag{6}$$

其中, M 为掩码表示, S 为原始频谱图.

最后, 增强后的频谱图通过 ISTFT 处理, 转换回时域信号, 从而获得高质量语音增强音频 Y_{output}.

模型的目标是通过最小化预测值和真实值间的 $Loss$ 来预测复数掩码 $\widehat{M}$:

$$Loss = \|G \odot (M - \widehat{M})\|^2 \tag{7}$$

其中, M 和 $\widehat{M}$ 分别是真实值的有界掩码和模型预测的有界掩码, $\|\cdot\|$ 是 L2 范数, G 是梯度罚项, 这个罚项根据混合谱图 X 中类似点的能量来加权掩码的时频点, 从而进一步优化掩码的估计. G 的计算公式如下:

$$G(ft;X) = \max(\min(\log(1 + \|X(ft)\|), 10), 10^{-3}) \tag{8}$$

3 实验

3.1 实验数据集

本文采用 VoxCeleb2、LRS2、LRS3-TED 和 AudioSet 这 4 个公开数据集对模型进行测试.

VoxCeleb2[21]是语音增强领域中广泛应用的重要数据集, 包含超过 100 万段音频片段. 这些片段来源于 YouTube 上的名人视频, 每个片段时长约 2–5 s. 数据集涵盖了来自 6 112 个不同身份的音频, 包括多种语言、背景环境、音质情况以及不同光照条件下的面部图像, 确保了数据的高度多样性.

LRS2 (lip reading sentences 2)[22]和 LRS3-TED[23]是两个专门用于唇读研究的公开数据集, 提供了大量的音视频样本, 用于训练和评估自动唇读系统. LRS2 数据集的样本主要来源于英国电视台的广播片段, 包括新闻、访谈和其他节目类型, 包含数千个短片段, 每个片段展示一个说话人的面部和对应的音频, 时长通常为几秒, 包含连续的讲话. LRS3-TED 数据集则来自从 YouTube 下载的 TED 和 TEDx 演讲视频, 同样包含大量的小片段, 每个片段时长约几秒的说话人的面部和音频, 提供了多样的语言和说话风格.

AudioSet[24]是由谷歌团队开发的大规模、手工标注的音频事件数据集. 该数据集涵盖了 600 种不同的音频事件类别, 从自然界的声响到人造声音, 范围广泛. 本文从 AudioSet 数据集中选择了日常生活中常见的声音类型, 如笑声、哭声、音乐声、引擎声、风声等, 保留了约 40 h 的音频, 用于制作含噪声的训练数据.

本文采用 VoxCeleb2 数据集中的音视频样本作为高质量的干净数据样本, 并从 AudioSet 中选取适当数量的音频数据作为噪声源, 与干净样本进行混合, 模拟真实世界中的噪声环境. 混合后的音频数据被用作训练数据. 同时, 使用部分 LRS2 和 LRS3-TED 中的数据样本验证本文模型的泛化性能.

在对 VoxCeleb2、LRS2 和 LRS3-TED 数据集进行检查时, 我们发现部分样本中存在两个说话者的语音或包含背景噪声 (如掌声、背景音乐声或人群的笑声). 为了保证数据集的质量, 我们对这些样本进行了筛选, 仅保留包含单一说话者语音且无其他背景噪声干扰的样本. 最终, VoxCeleb2 保留了约 500 h 的数据量, LRS2 保留了约 60 h 的数据量, LRS3-TED 保留了约 320 h 的数据量.

3.2 数据增强

为了确保训练数据的多样性和高质量, 本文采用系统化的视频数据处理方法. 首先, 通过二元化和背景噪声检测方法的组合, 对原始数据进行噪声检测和去除, 确保数据的纯净性. 接着, 对视频和音频数据进行归一化处理, 将长视频分割为多个片段, 并去除首尾帧, 以提高数据处理的效率和准确性.

在数据混合策略方面, 本文设计了一种多源干扰混合与多类型噪声数据增强 (multi-source interference mixing & multi-type noise data augmentation, MI-MNA) 方法对数据进行处理, 具体如下.

我们将音频信号按照目标音频、干扰音频、环境噪声的一定比例进行混合, 用于制作训练集. 其中目标音频来自干净数据集中的音频样本, 干扰音频来自干净数据集中另一个说话者的音频样本, 环境噪声来自含有环境噪声音频数据集中的样本. 本实验中干净数据集为 VoxCeleb2 数据集中的音频样本, 含有环境噪声音频的数据集为 AudioSet 数据集中的样本.

我们通过把每个目标音频分别与以下 4 种类型音频信号混合来制作干扰音频: 与其相同性别和相同场景的人声、与其相同性别但不同场景的人声、与其不同性别且不同场景的人声、与其不同性别和相同场景的人声.

在上述音频混合方法的基础上, 将生成的音频样本分成两份, 给其中一份添加不同分贝的环境噪声, 使得添加噪声音频样本的 SNR 在−5 dB 到 15 dB 之间, 保证了数据样本的多样性, 其中环境噪声音频信号类型包括音乐、笑声、哭声、发动机声、风声等.

3.3 实验细节

实验在 Linux 系统上进行, 使用 RTX 3090 显卡作为 GPU 加速设备. 模型基于 PyTorch 深度学习框架进行训练, 并在 PyCharm 上进行编译和测试. U-Net 编码器采用了 4 层卷积结构, 通道数分别为 64、128、256 和 512, 卷积核尺寸为 3×3. Transformer 层使用 4 头注意力机制, 隐藏层维度为 512, 前馈网络维度为 2 048. 门控机制采用了 Sigmoid 激活函数.

在训练过程中, 批处理大小设置为 16. 优化器选择 Adam, 学习率设为 0.001, 动量设为 0.9. 为了防止过拟合, 采用早停法, 当模型收敛且性能不再提升时, 训练过程自动停止.

在训练数据处理方面, 使用 VoxCeleb2 数据集作为训练集, 同时, LRS2 和 LRS3-TED 数据集中的样本被用作补充测试集.

在视觉特征处理中, 首先对长视频进行分段并去除首尾帧, 然后通过 3D 人脸模型对齐面部标记点, 提取 68 个标志点并裁剪出 112×112 的人脸区域, 选取最能代表面部特征的视频帧, 经标准化和中心裁剪处理后调整为 224×224 像素. 在听觉特征处理中, 音频采样率设为 16 kHz, 采用 STFT 生成 201 维幅度谱, 并进行对数压缩以增强频谱动态范围.

3.4 评价指标

在语音分离任务中, 我们使用一系列标准度量来评估分离结果的质量, SDR[25]是衡量分离语音信号与原始语音信号之间相似度的指标, 单位是分贝 (dB), 它反映了分离信号中保留的原始信号成分与引入的失真成分之间的比例. SDR 值越高, 表示分离信号的质量越好, 失真越小. PESQ[26]是一种基于人类听觉感知的语音质量评估方法, 通过比较参考语音信号和经过处理后的测试语音信号, 综合考虑语音的失真和噪声等因素, 得到一个介于−0.5 到 4.5 之间的分数. PESQ 分数越高, 表示分离语音的感知质量越好. STOI[27]是一种用于评估语音信号可懂度的客观指标. 它通过分析语音信号的短时频谱特征, 计算分离语音与原始语音之间的相关性, 从而评估语音的可懂度. STOI 的取值范围在 0–1 之间, STOI 值越接近 1, 表示语音的可懂度越高; 反之, 则表示可懂度越低.

3.5 消融实验

3.5.1 数据增强方法比较

为了验证训练样本的多样性对模型收敛速度、模型性能以及泛化性的影响, 本文设计了如下实验.

我们使用 MI-MNA 方法处理 VoxCeleb2 数据得到增强后的数据. 作为对照, 在数据集中随机取人声信号进行混合, 并且不添加环境噪声生成对照训练集. 这种数据处理方法后续称为 Random 方法. 我们把 MI-MNA 和 Random 方法获得的数据分别用于网络训练. 实验中确保所有模型在相同的数据量和相同训练轮次下进行训练和评估, 以排除其他变量的影响.

在 VoxCeleb2 数据集进行实验, 结果见图 2. 从图中结果可以看出, 模型使用 MI-MNA 方法增强后的数据训练时, 表现出更快的收敛速度, 同时模型性能也有所提升. 具体来说, 使用 MI-MNA 方法的训练损失比 Random 方法提升了 0.98, MI-MNA 方法和 Random 方法的 SDR 分别为 15.00 dB 和 11.03 dB, 证明 MI-MNA 方法有助于语音增强性能的提升. 同时, MI-MNA 方法在 PESQ 和 STOI 指标上也表现出显著优势, MI-MNA 方法和 Random 方法的 PESQ 分别为 2.81 和 2.48, STOI 分别为 82.03% 和 79.22%, 分别提升了 0.33% 和 2.81%, 进一步验证了其在语音质量改善方面的有效性. 实验结果表明本文 MI-MNA 方法能生成多样化的训练数据, 可以帮助模型更快地学习到对语音增强任务的有效特征, 提升模型综合性能.

随后, 在 LRS2、LRS3-TED 数据集进行测试, 结果如表 1 所示. 从表 1 数据可以看出, 无论是在 LRS2 数据集还是 LRS3-TED 数据集上, MI-MNA 数据增强方法训练的模型在各项指标上均优于 Random 方法训练的模型. 具体来说, 在 LRS2 数据集上, MI-MNA 方法获得了 15.57 dB 的 SDR, 比 Random 方法提升了 4.47 dB, 而在 LRS3-TED 数据集, MI-MNA 方法的 SDR 比 Random 方法提升了 3.80 dB, 表明 MI-MNA 方法有利于减少失真. MI-MNA 方法在 LRS2 和 LRS3-TED 上的 PESQ 分别比 Random 方法提升了 0.23 和 0.23, STOI 分别比 Random 方法提升了 3.53% 和 2.97%. 数据表明, MI-MNA 方法训练的模型语音质量更好、可懂度更高, 本文设计的多样化的训练数据能够帮助模型更快地学习通用特征, 减少过拟合风险.

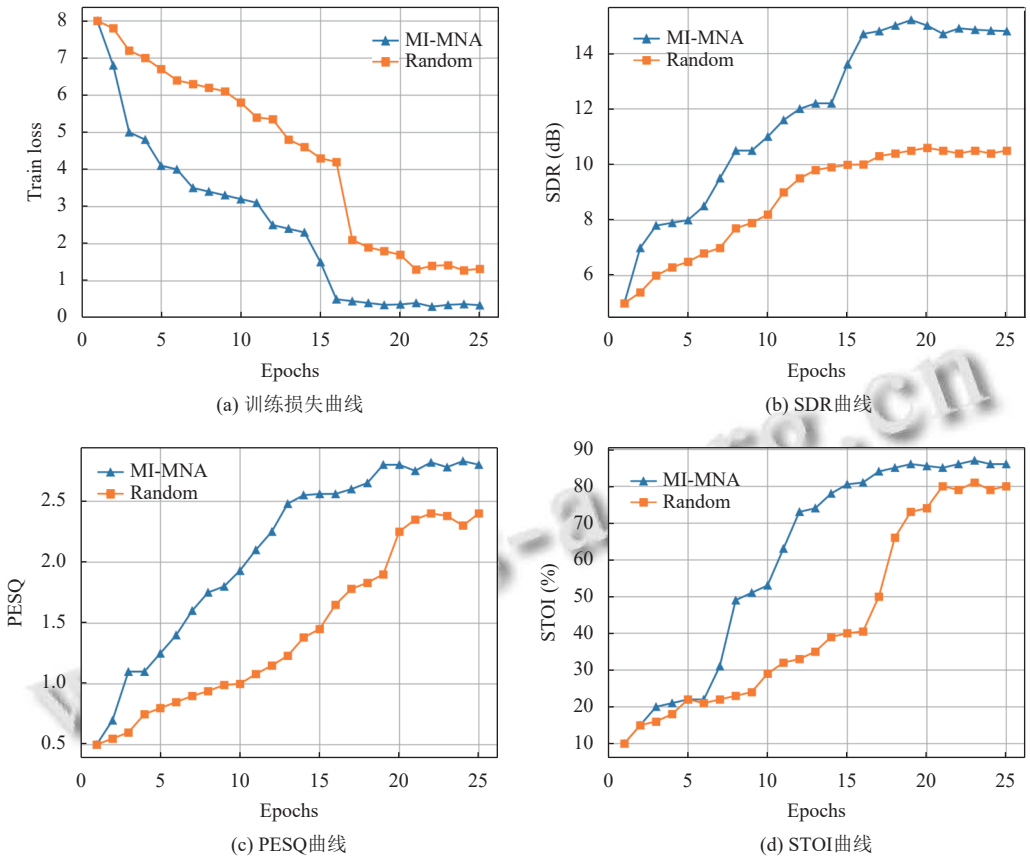


图 2 在 VoxCeleb2 数据集下使用不同数据增强方法训练模型的结果

表 1 不同数据增强方法在不同数据集下的结果对比

指标	LRS2		LRS3-TED	
	Random	MI-MNA	Random	MI-MNA
SDR↑ (dB)	11.10	**15.57**	12.30	**16.10**
PESQ↑	2.43	**2.66**	2.55	**2.78**
STOI↑ (%)	87.90	**91.43**	89.30	**92.27**

注: 最优结果加粗显示

3.5.2 多模态特征融合

本实验聚焦于多模态特征对语音增强模型性能与复杂度的影响, 对比了结合不同面部特征与音频特征的方法, 实验结果见表 2. 其中, 音频特征、面部属性特征以及整体面部运动特征点采用第 2.1 节中的双流视听特征提取方法获取; 唇部运动特征点为整体面部运动特征点中仅涉及唇部运动的部分特征点; 面部运动特征则是指使用 Visualvoice[13]方法对人脸裁切后的视频序列所提取的特征.

仅使用音频特征时, 模型的 SDR、PESQ 和 STOI 分别为 7.25 dB、1.86 和 80.25%, 参数量为 36.14M. 当结合音频特征与面部属性特征时, 模型的 SDR、PESQ 和 STOI 均有所提升, 分别增加至 7.81 dB、2.12 和 82.13%, 但参数量也增加到 47.23M, 表明加入面部属性特征在一定程度上提升了模型性能, 但也增加了模型复杂度.

表 2 多模态特征语音增强性能对比

音频特征	面部属性特征	唇部运动特征点	面部运动特征	整体面部运动特征点	SDR↑ (dB)	PESQ↑	STOI↑ (%)	参数量 (M)
√	—	—	—	—	7.25	1.86	80.25	**36.14**
√	√	—	—	—	7.81	2.12	82.13	47.23
√	—	√	—	—	8.91	2.73	83.54	50.42
√	—	—	—	√	10.37	2.78	84.72	57.17
√	—	—	√	—	11.95	2.80	86.38	77.51
√	√	—	—	√	**17.31**	**3.12**	**92.58**	68.36

注: 最优结果加粗显示

结合音频特征与唇部运动特征点时, 模型的 SDR 进一步提升至 8.91 dB、PESQ 达到 2.73、STOI 为 83.54%, 参数量为 50.42M, 说明唇部运动特征点对模型性能的提升效果优于面部属性特征. 当结合音频特征与整体面部运动特征点时, SDR 为 10.37 dB、PESQ 为 2.78、STOI 为 84.72%, 参数量为 57.17M, 性能较前两者有提升, 但参数量也相应增加. 结合音频特征与面

部运动特征时, SDR 达到 11.95 dB、PESQ 为 2.80、STOI 为 86.38%, 但参数量大幅增加到 77.51M. 当同时结合音频特征、面部属性特征以及整体面部运动特征时, 模型的性能得到显著提升, SDR 高达 17.31 dB、PESQ 为 3.12、STOI 为 92.58%, 且参数量为 68.36M, 相较于结合音频特征与面部运动特征点的情况, 在性能提升的同时参数量有所降低.

在表 2 的 6 种方法中, 使用音频信息、面部运动特征点和面部属性特征 (本文模型) 的方法效果最佳, 在 VoxCeleb2 数据集上的 SDR、PESQ 和 STOI 均获得了最优的结果, 同时参数量也处于较低水平. 这证明了本文提出的特征提取方法能够高效地提取语音增强所需的面部信息. 通过结合音频信息和多模态的面部特征, 本文提出的模型能够更好地捕捉与语音增强相关的面部运动特征, 从而在提升语音质量的同时, 实现对模型复杂度的有效控制.

3.5.3 门控注意力机制在复杂视觉条件下的语音增强

为验证门控机制在提升模型于复杂条件下鲁棒性方面的有效性, 本文进行了消融实验, 测试了在不同特征融合策略下, 不同模型在正常场景与遮挡场景中的性能表现. 其中遮挡场景的数据是指人的唇部被完全遮挡的情况. 实验在 VoxCeleb2 数据集上进行测试. 我们设计了模型对照组, 具体设置和实验结果见表 3.

表 3 多模态面部特征语音增强性能对比

融合策略			正常场景			遮挡场景		
视听特征拼接+全连接	交叉注意力	门控	SDR↑ (dB)	PESQ↑	STOI↑ (%)	SDR↑ (dB)	PESQ↑	STOI↑ (%)
√	—	—	9.25	1.91	84.12	8.25	1.71	71.02
—	√	—	12.82	2.27	85.93	11.80	2.17	72.97
—	√	√	**17.31**	**3.12**	**92.58**	**14.36**	**2.82**	**84.59**

注: 最优结果加粗显示

在正常场景下, 仅采用“视听特征拼接+全连接”策略时, 模型的 SDR 为 9.25 dB、PESQ 为 1.91、STOI 为 84.12%; 仅使用交叉注意力机制时, SDR、PESQ 和 STOI 均有所提升, 表明交叉注意力机制能在一定程度上优化模型性能; 而当同时使用交叉注意力和门控机制时, 这 3 个性能指标的值获得进一步提升, 相较于仅使用交叉注意力的情况, SDR 提升了 4.49 dB, PESQ 提升了 0.85, STOI 提升了 6.65%, 说明门控机制在正常场景下能够与交叉注意力机制协同工作, 进一步增强模型对视听特征的处理能力, 显著提升模型性能.

在遮挡场景下也获得了类似的结果, 同时使用交叉注意力和门控机制时, SDR、PESQ 和 STOI 都获得了最优值, 表明门控机制在遮挡场景下同样能够有效提升模型的鲁棒性和性能, 帮助模型更好地应对遮挡带来的干扰.

值得注意的是, 模型在遮挡场景下的整体性能比正常场景下的性能要低, 这主要是因为遮挡场景下, 视觉特征的完整性和准确性受到显著影响, 导致模型难以充分利用视听信息进行有效的语音增强. 此外, 遮挡可能引入额外的噪声或干扰, 进一步增加了语音分离和恢复的难度, 从而导致性能下降.

综合来看, 门控机制在正常和遮挡这两种不同视觉条件下, 与交叉注意力机制相结合, 都能使模型在关键指标上取得更好表现, 对提升模型性能、增强模型的鲁棒性和适应性具有重要作用.

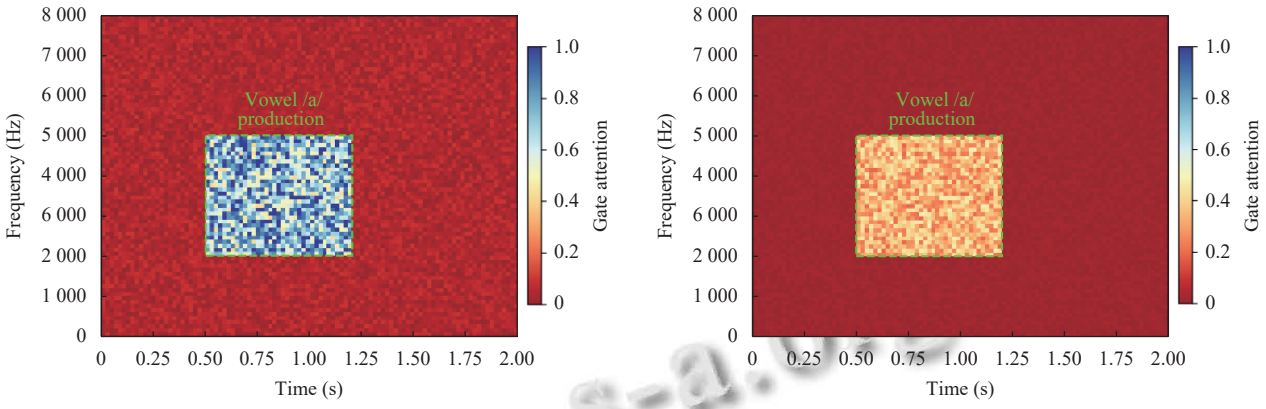
为验证门控机制在不同视觉条件下的鲁棒性, 本文进一步设计了门控机制在正常场景与遮挡场景下的对比实验, 如图 3 所示.

图 3 通过热力图形式展示了说话人发音“/a/”时, 门控权重矩阵的时空分布特性, 包含两个子图. 其中图 3(a) 代表唇部没有被遮挡的正常场景, 图 3(b) 代表唇部被完全遮挡的场景. 横轴表示时间维度, 纵轴表示频率维度; 采用 RdYlBu 色阶展示, 红色 (低值, 权重≈0) 至蓝色 (高值, 权重≈1) 的颜色变化表征门控激活强度; 绿色虚线框的横轴表示发音时段, 纵轴表示发音频段.

图 3(a) 中, 绿色虚线框区域中的蓝色 (权重>0.8) 表示在发音时段的基频与共振峰区域高度激活, 表明模型优先融合与唇动强相关的视觉信息; 绿色虚线框区域外的红色 (权重<0.2) 表示静音段与非语音频段自动抑制视觉输入. 图 3(b) 中, 整体色阶向红色偏移, 表示全局门控权重下降约 60% (对比红色区域峰值从 0.9→0.4); 绿色虚线框区域中的橙色 (权重≈0.4) 表示发音时段的残余激活, 反映视觉信息部分缺失时模型的退让性融合策略.

整体而言, 当视觉可信度高时, 在发音相关的时频

点形成局部高激活峰, 强化视听特征对齐; 视觉可信度低时, 全局激活水平压缩, 避免错误特征注入, 此时模型依赖音频自注意力维持基本性能, 表明模型能够有效调整特征融合策略. 通过绿色虚线框标记的目标发音段, 可以清晰地看到模型精准捕捉到了听觉与视觉同步区域, 进一步证明了门控机制在复杂场景下的有效性.



(a) 正常场景 (b) 遮挡场景

图 3 门控机制在正常场景与遮挡场景中的时间-频率分布图

3.5.4 跨数据集语音增强的泛化能力

从表 4 可以看出, 本文提出的 DS-GAVF 模型在 VoxCeleb2 数据集上训练后, 在 LRS2 和 LRS3-TED 数据集上展现了良好的跨数据集泛化能力, 并显著优于 Visualvoice[13]和 Reading[15]. 在 LRS2 数据集上, 本文模型在 SDR、PESQ 和 STOI 指标上分别达到了 15.57 dB、2.66 和 91.43%, 比模型 Visualvoice 分别提升了 4.77 dB、0.50 和 3.03%, STOI 仅比 Reading 下降了 0.27%. 在 LRS3-TED 数据集上, 本文模型在 SDR、PESQ 和 STOI 指标上分别达到了 16.10 dB、2.78 和 92.27%, 比模型 Visualvoice 分别提升了 4.4 dB、0.37 和 2.27%. 这些结果表明, 本文模型在 VoxCeleb2 数据集上训练后, 不仅在目标数据集上表现出色, 还具有优异的跨数据集泛化能力, 显著优于现有模型.

3.6 方法对比实验

为了进一步验证 DS-GAVF 模型在语音增强任务中的有效性和优越性, 本文把该模型与近年来的方法在 VoxCeleb2 数据集上进行对比, 结果如表 5 所示. 其中, 对比方法的结果来自文献.

表 4 DS-GAVF 在 LRS2 和 LRS3-TED 数据集上的性能对比

指标	LRS2			LRS3-TED		
	Visual-voice	Reading	DS-GAVF	Visual-voice	Reading	DS-GAVF
SDR↑ (dB)	10.80	14.20	**15.57**	11.70	15.50	**16.10**
PESQ↑	2.16	2.41	**2.66**	2.41	2.63	**2.78**
STOI↑ (%)	88.40	**91.70**	91.43	90.0	**93.50**	92.27

注: 最优结果加粗显示

从实验结果可以看出, 与现有方法相比, 本文方法在 SDR、PESQ 和 STOI 这 3 种评估指标上均取得了显著的优势. 具体而言, DS-GAVF 在 SDR 指标上达到了 17.31 dB, 较 VoViT 提升了 2.46 dB; 在 PESQ 和 STOI 指标上达到了 3.12 和 92.58%, 较 Optimizing 分别提升了 0.23 和 9.28%. 这些结果表明, 本文方法在语音质量、可懂度和感知质量方面均显著提升.

表 5 在 VoxCeleb2 数据集的对比结果

方法	SDR↑ (dB)	PESQ↑	STOI↑ (%)
Conversation[28] (2018)	9.25	1.91	84.00
Visualvoice[13] (2021)	9.73	2.16	88.40
VoViT[29] (2022)	14.85	—	—
Optimizing[30] (2024)	—	2.89	83.30
DS-GAVF (本文)	**17.31**	**3.12**	**92.58**

注: 最优结果加粗显示

4 结论与展望

本研究针对音视频语音增强在复杂场景下的挑战, 提出了 DS-GAVF 架构, 通过多模态信息协同建模与细粒度特征交互提升语音增强的性能. 实验结果表明, DS-GAVF 在 VoxCeleb2 数据集上取得了优异的性能, 获得 17.31 dB 的 SDR、3.12 的 PESQ 和 92.58% 的 STOI, 验证了其在混合噪声环境下的有效性. 消融实验进一步验证了双流视觉建模和门控机制的关键作用. 门控单元能够根据视觉可信度动态调节特征融合权重, 在视觉信息不可靠时有效抑制噪声干扰, 结合 MI-MNA

数据增强策略, 显著提升了模型的跨数据集泛化能力.

相较于现有方法, DS-GAVF 不仅保持了较低的计算复杂度, 还在语音质量、可懂度和鲁棒性方面实现了协同提升. 未来工作将进一步探索更多模态信息的融合, 优化模型结构以适应实时语音增强需求, 并在更多实际应用场景中验证其有效性.

参考文献

1 Jia SC, Zhang TL, Zuo RC, *et al*. Explaining cocktail party effect and McGurk effect with a spiking neural network improved by motif-topology. Frontiers in Neuroscience, 2023, 17: 1132269. [doi: 10.3389/fnins.2023.1132269]

2 Sheng CC, Kuang GY, Bai L, *et al*. Deep learning for visual speech analysis: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(9): 6001–6022. [doi: 10.1109/TPAMI.2024.3376710]

3 Richter J, Welker S, Lemercier JM, *et al*. Speech enhancement and dereverberation with diffusion-based generative models. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023, 31: 2351–2364. [doi: 10.1109/TASLP.2023.3285241]

4 Lin LX, Li YW, Wang HZ. TSMGAN-II: Generative adversarial network based on two-stage mask Transformer and information interaction for speech enhancemen. Proceedings of the 20th International Conference on Advanced Intelligent Computing Technology and Applications. Tianjin: Springer, 2024. 174–185. [doi: 10.1007/978-981-97-5591-2_15]

5 尤昕源, 王恒. 基于门控膨胀卷积循环网络的单声道语音增强. 计算机应用, 2024, 44(4): 1317–1324. [doi: 10.11772/j.issn.1001-9081.2023040452]

6 Michelsanti D, Tan ZH, Zhang SX, *et al*. An overview of deep-learning-based audio-visual speech enhancement and separation. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 1368–1396. [doi: 10.1109/TASLP.2021.3066303]

7 Roth J, Chaudhuri S, Klejch O, *et al*. Ava active speaker: An audio-visual dataset for active speaker detection. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 4492–4496.

8 Gu RZ, Zhang SX, Xu Y, *et al*. Multi-modal multi-channel target speech separation. IEEE Journal of Selected Topics in Signal Processing, 2020, 14(3): 530–541. [doi: 10.1109/JSTSP.2020.2980956]

9 Rouditchenko A, Zhao H, Gan C, *et al*. Self-supervised audio-visual co-segmentation. Proceedings of the 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Brighton: IEEE, 2019. 2357–2361.

10 Sadeghi M, Leglaive S, Alameda-Pineda X, *et al*. Audio-visual speech enhancement using conditional variational auto-encoders. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 1788–1800. [doi: 10.1109/TASLP.2020.3000593]

11 Ayilo JE, Sadeghi M, Serizel R, *et al*. Diffusion-based unsupervised audio-visual speech enhancement. Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Hyderabad: IEEE, 2025. 1–5.

12 Gogate M, Dashtipour K, Hussain A. Robust real-time audio-visual speech enhancement based on DNN and GAN. IEEE Transactions on Artificial Intelligence, 2024. [doi: 10.1109/TAI.2024.3366141]

13 Gao RH, Grauman K. Visualvoice: Audio-visual speech separation with cross-modal consistency. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Nashville: IEEE, 2021. 15490–15500.

14 Wang J, Luo YY, Yi WM, *et al*. Speaker-independent audio-visual speech separation based on Transformer in multi-talker environments. IEICE Transactions on Information and Systems, 2022, 105(4): 766–777.

15 Rahimi A, Afouras T, Zisserman A. Reading to listen at the cocktail party: Multi-modal speech separation. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans: IEEE, 2022. 10483–10492.

16 Wang W, Xing C, Wang D, *et al*. A robust audio-visual speech enhancement model. Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona: IEEE, 2020. 7529–7533.

17 Cao H, Wang YY, Chen J, *et al*. Swin-unet: Unet-like pure Transformer for medical image segmentation. Proceedings of the 2022 European Conference on Computer Vision. Cham: Springer, 2022. 205–218.

18 Sheng C, Zhu X, Xu H, *et al*. Adaptive semantic-spatio-temporal graph convolutional network for lip reading. IEEE Transactions on Multimedia, 2021, 24: 3545–3557.

19 He KM, Zhang XY, Ren SQ, *et al*. Deep residual learning for image recognition. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 770–778.

20 Gabbay A, Shamir A, Peleg S. Visual speech enhancement. arXiv:1711.08789, 2017.
21 Chung JS, Nagrani A, Zisserman A. VoxCeleb2: Deep speaker recognition. arXiv:1806.05622, 2018.
22 Afouras T, Chung J S, Senior A, *et al*. Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 44(12): 8717–8727.
23 Afouras T, Chung JS, Zisserman A. LRS3-TED: A large-scale dataset for visual speech recognition. arXiv:1809.00496, 2018.
24 Gemmeke JF, Ellis DPW, Freedman D, *et al*. Audio set: An ontology and human-labeled dataset for audio events. Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). New Orleans: IEEE, 2017. 776–780.
25 Raffel C, McFee B, Humphrey EJ, *et al*. MIR_EVAL: A transparent implementation of common mir metrics. Proceedings of the 15th International Society for Music Information Retrieval Conference. Taipei: ISMIR, 2014. 10.
26 Rix AW, Beerends JG, Hollier MP, *et al*. Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs. Proceedings of the 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Salt Lake City: IEEE, 2001. 749–752.
27 Taal CH, Hendriks RC, Heusdens R, *et al*. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. IEEE Transactions on Audio, Speech, and Language Processing, 2011, 19(7): 2125–2136. [doi: 10.1109/TASL.2011.2114881]
28 Afouras T, Chung JS, Zisserman A. The conversation: Deep audio-visual speech enhancement. Proceedings of the 2018 Interspeech. Hyderabad: Interspeech, 2018. 3244–3248.
29 Montesinos JF, Kadandale VS, Haro G. VoViT: Low latency graph-based audio-visual voice separation Transformer. Proceedings of the 17th European Conference on Computer Vision. Tel Avi: Springer, 2022. 310–326.
30 Chen H, Wang Q, Du J, *et al*. Optimizing audio-visual speech enhancement using multi-level distortion measures for audio-visual speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2024, 32: 2508–2521. [doi: 10.1109/TASLP.2024.3393732]

(校对责编: 王欣欣)