◆自然语言理解◆

大词汇量环境噪声下的多模态视听语音识别方法*

吴　兰**,杨　攀,李斌全,王　涵
(河南工业大学电气工程学院,河南郑州　450001)

摘要:视听语音识别(Audio-Visual Speech Recognition,AVSR)技术利用唇读和语音识别(Audio-Visual Speech Recognition,AVSR)的关联性和互补性可有效提高字符识别准确率。针对唇读的识别率远低于语音识别、语音信号易受噪声破坏、现有的视听语音识别方法在大词汇量环境噪声中的识别率大幅降低等问题,本文提出一种多模态视听语音识别(Multi-modality Audio-Visual Speech Recognition,MAVSR)方法。该方法基于自注意力机制构建双流前端编码模型,引入模态控制器解决环境噪声下音频模态占据主导地位而导致的各模态识别性能不均衡问题,提高识别稳定性与鲁棒性,构建基于一维卷积的多模态特征融合网络,解决音视频数据异构问题,提升音视频模态间的关联性与互补性。与现有主流方法对比,在仅音频、仅视频、音视频融合 3 种任务下,该方法的识别准确率提升 7.58%以上。

关键词:注意力机制;多模态;视听语音识别;唇读;语音识别

中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1005-9164(2023)01-0052-09
DOI:10.13656/j.cnki.gxkx.20230308.006

从单模态机器学习发展到多模态机器学习对提升机器感知能力至关重要,其中,音频和视频两种模态代表了日常生活中最重要的两种感知方式。视听语音识别(Audio-Visual Speech Recognition,AVSR)技术利用基于视频模态的唇读技术和音频模态的语音识别(Automatic Speech Recognition,ASR)技术实现唇读和语音识别技术的融合,可更好地克服单模态感知任务的局限性。

在语音识别中,相比于传统的混合高斯模型(Gaussian Mixture Model,GMM)、隐马尔科夫模型(Hidden Markov Model,HMM)语音识别方法,基于神经网络的语音识别方法能更准确地描述语音模型内部的复杂结构,表现出更强的表征和建模能力,从而使语音识别的准确率取得重大突破。当前,最先进的 ASR 系统能够达到 95%以上的转换准确率,但在噪声干扰下 ASR 的识别准确率却大幅下降,这是因为音频模态容易受到噪声破坏。然而,视频模态却有着不受声学噪声干扰的显著特性,此时唇读的优势更

收稿日期:2022-11-14　　修回日期:2022-11-18
*国家自然科学基金项目(61973103),河南省自然科学基金项目(222300420039)和郑州市科技局自然科学项目(21ZZXTCX01)资助。
【第一作者简介】
吴　兰(1981-),女,教授,主要从事智能感知与智能控制、多模态机器学习、视听语音识别研究,E-mail:wulan@haut.edu.cn。
【**通信作者】
【引用本文】
吴兰,杨攀,李斌全,等.大词汇量环境噪声下的多模态视听语音识别方法[J].广西科学,2023,30(1):52-60.
WU L,YANG P,LI B Q,et al. A Multi-modality Audio-Visual Speech Recognition Method under Large Vocabulary Environmental Noise [J]. Guangxi Sciences,2023,30(1):52-60.

为明显[1]。

唇读又叫视觉语音识别(Visual Speech Recognition,VSR)。关于唇读的研究,早期主要采用以HMM为代表的非深度学习方法[2],而现有研究大都使用卷积神经网络(Convolutional Neural Network,CNN)、长短期记忆网络(Long Short-Term Memory,LSTM)的深度学习方法[3-6]。Noda等[3]使用CNN提取高度泛化的视频特征,这种"深度"特征远胜于传统的手工设计特征,并取得不错的识别准确率。Assael等[7]提出的端到端句子级唇读方法,可同时学习时空视频特征,实现句子级序列预测。目前唇读技术在小词汇量数据集中取得不错的识别准确率,但在大词汇量的数据集任务中,因为说话人数多、词汇量大、不同的音素对应相同的发音嘴形等原因,使得一些单词几乎无法单独依靠视觉系统来区分,导致唇读识别准确率大幅下降。视听语音识别技术突破了唇读和语音识别的限制,结合唇读不受噪声干扰、语音识别准确率高的优势,有效利用二者的关联性和互补性。

在视听语音识别技术研究中,Sterpu等[8]提出一种音视频对齐机制,并尝试在帧级对齐声学和视觉表示,在TCD-TIMIT[9]数据集上获得不错的识别准确率,但在更具挑战性的大词汇量数据集LRS2[10]上表现不佳,存在音频模态、视频模态融合情况下的识别准确率低于仅音频模态下的识别准确率,音频模态占据主导性等问题。此外,Stafylakis等[6]发现AVSR过度依赖音频模态,Chung等[11]也观察到在AVSR中音频信号占据主导地位。在环境噪声下,由于音频模态容易受到噪声破坏,因此音频模态的主导地位会给AVSR带来更多负面影响。此外,当各模态的性能相差很大时,AVSR准确率也会受低性能模态识别率的影响,造成整体性能的损失,给多模态视听语音识别融合音频模态和视频模态带来更多挑战。

针对上述问题,本文提出一种多模态视听语音识别(Multi-modality Audio-Visual Speech Recognition,MAVSR)方法。针对大词汇量任务中,VSR识别率远低于ASR、各模态识别性能不均衡的问题,本文提出一种基于自注意力机制与一维卷积相结合的特征级多模态融合方法,从音频和视频编码特征中获取具有高度局部相关性的音视频联合编码特征,提升音频模态和视频模态的关联性和互补性。针对环境噪声下音频易遭受破坏,且在AVSR任务中音频模态又占据主导性的问题,本文引入模态控制器,重新定义音频数据权重,发挥视频模态不受噪声影响的优势,提升MAVSR方法在噪声数据中的稳定性和鲁棒性。最后,在MAVSR解码推理中嵌入语言模型(Language Model,LM)并联合transformer解码器实现字符级预测,实验表明嵌入LM联合解码可在transformer解码输出的基础上进一步提升MAVSR方法的识别准确率。该方法可实现在仅音频、仅视频和音视频融合3种模态下的识别任务。

1 相关工作

1.1 唇读

唇读主要使用图像处理技术提取说话人唇部连续运动图像,利用神经网络提取说话人的唇部运动特征,并在此基础上实现视觉语音识别。提取有效的视频特征对模型推理至关重要,在深度学习出现之前,唇读大多采用手工设计的特征,需要对视频帧进行大量预处理工作,非常耗时且提取的特征比较简单。近期的研究都是基于深度学习的方法提取更深、更为抽象的"深度"特征。Assael等[7]基于时空卷积、循环网络和连接主义时间分类损失等方法,提出一种端到端唇读模型LipNet,直接以图像序列作为输入便可输出字符序列预测概率,实现完全端到端的训练,把特征提取融入网络模型,让整个模型更为简洁。最近,transformer[12]和时间卷积网络(Temporal Convolutional Network,TCN)[13]正逐步取代循环神经网络(Recurrent Neural Network,RNN),transformer能更好地执行并行计算和学习长输入的序列关系,并缩短训练时间。TCN采用因果卷积和空洞卷积加残差连接的结构,在处理时间序列建模任务中比LSTM等递归架构表现更佳,而transformer在处理可变长序列、推理超过训练数据最大长度序列方面更具优势。因此,本文使用基于3D卷积和2D ResNet网络提取视频特征,基于transformer自注意力机制实现双流通道特征编码。

1.2 视听语音识别

AVSR与唇读是密切相关的,融合音频和视频两种模态往往能够为AVSR提供更多的先验信息[14]。在最近的研究中,端到端架构[15]在AVSR中得到广泛的应用,这些研究一方面利用全连接层和LSTM来提取特征并对时间信息建模[16,17],另一方面使用3D卷积层结合CNN、LSTM或者使用CNN的变体构建模型[18,19]。此外,在AVSR中一个有效

的视听融合策略非常重要,Paraskevopoulos 等[20]提出一种特征级融合方法,可直接在编码层提取音频特征,在跨模态多头注意力层[21]融合音频、视频特征,这种融合方法已经被证明是有效的。音频流和视频流的融合应该确保融合后的系统性能比两个单独模态的识别性能更优秀,为解决这个问题,本文提出音视频双流前端模型和双通道多头注意力编码模型,基于卷积网络实现多模态特征融合策略,采用端到端架构实现本文提出的 MAVSR 方法。

2 视听语音识别方法

在这一节,主要叙述多模态视听语音识别方法(MAVSR)。针对大词汇量视听语音识别任务中 VSR 性能远低于 ASR、音频易受噪声破坏等问题,利用唇读在噪声中的鲁棒性,设计多模态双流编码模型和多模态融合网络,解决音频和视频网络融合问题,提升二者互补性。在联合解码推理中,嵌入 LM 联合解码器实现字符级预测,并在集束搜索的每一步结合 LM 执行解码推理。MAVSR 方法既可以实现音频、视频两种模态融合的识别任务,又可以完成单模态下的 ASR 和 VSR 任务。

2.1 总体架构

本文提出的 MAVSR 方法主要由 3 个部分组成:音频、视频前端双流编码模型,多模态特征融合网络和联合解码。MAVSR 网络架构见图 1。该方法有音频和视频两个模态流分别处理音频模态和视频模态数据,两个模态流既可以独立工作,又可以联合工作。

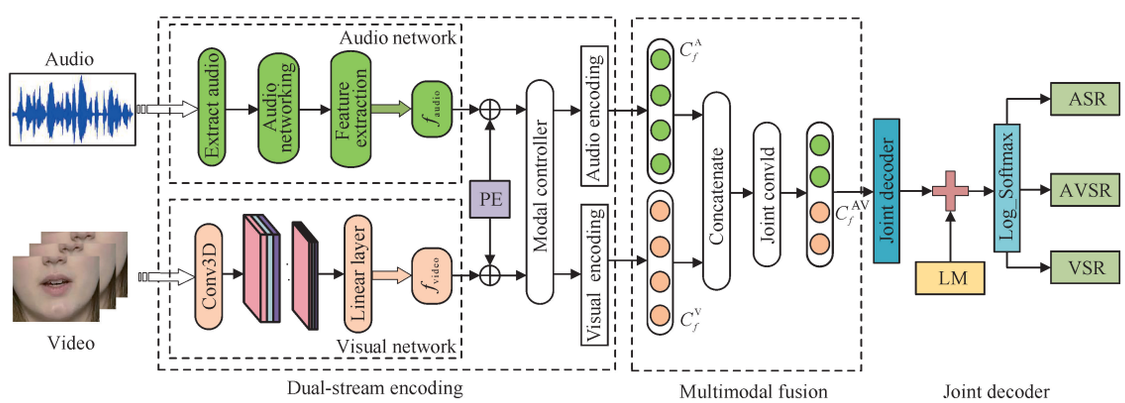

图 1 MAVSR 网络架构

Fig. 1 MAVSR network architecture

位置编码(Positional Encoding,PE)用于在输入序列中嵌入位置信息,本文采用正弦和余弦函数位置编码,具体方法如下:

$$\mathrm{PE}(\mathrm{pos},2i)=\sin\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{d_{\mathrm{model}}}}}\right), \tag{1}$$

$$\mathrm{PEB}(\mathrm{pos},2i+1)=\cos\left(\frac{\mathrm{pos}}{10000^{\frac{2i}{d_{\mathrm{model}}}}}\right), \tag{2}$$

其中,pos 表示单词在句子中的绝对位置,i 表示词向量中的第几维,本文中 $d_{\mathrm{model}}=512$,故 $i=0,1,\cdots,255$。模型输入的图像和音频都是可变长的序列,使用正弦、余弦函数位置编码,可根据序列长度变化进行调整。其中,位置编码向量维度和音视频特征向量维度一致。

2.2 音视频前端模型

音视频前端模型分为音频流和视频流两个部分,分别对应音频前端模型和视频前端模型。音频前端模型用于提取音频特征,并生成噪声数据。噪声数据用于在不同信噪比(Signal-to-Noise Ratio,SNR)的环境中训练、验证本文提出的方法。音频流的输入数据是原始的单通道音频信号,经过声学网络提取对应的音频特征 $f_{\mathrm{audio}}=[a_1,a_2,a_3,\cdots,a_m]$,$a\in R^d$。视频前端模型从连续视频帧中提取出唇部感兴趣区域(Region of Interest,ROI),经过 ResNet18 构成的视频特征提取器,将连续的视频帧转化为说话人唇部区域视频特征 $f_{\mathrm{video}}=[v_1,v_2,v_3,\cdots,v_n]$,$v\in R^d$,其中音频流帧率是 100 f/s,视频流帧率是 25 f/s。为了使音视频特征对齐,本文将 4 个连续的音频特征向量和 1 个视频特征向量对齐。若两个模态流的特征长度不齐,则用零向量填充音频特征和视频特征。

2.3 多模态特征融合网络

多模态融合特征可以有效利用音频模态和视频

模态特征间的互补性和关联性，如图 2 所示。本文采用双流编码通道分别将前端模型提取的音频特征 f_{audio} 和视频特征 f_{video} 送入特征编码器。

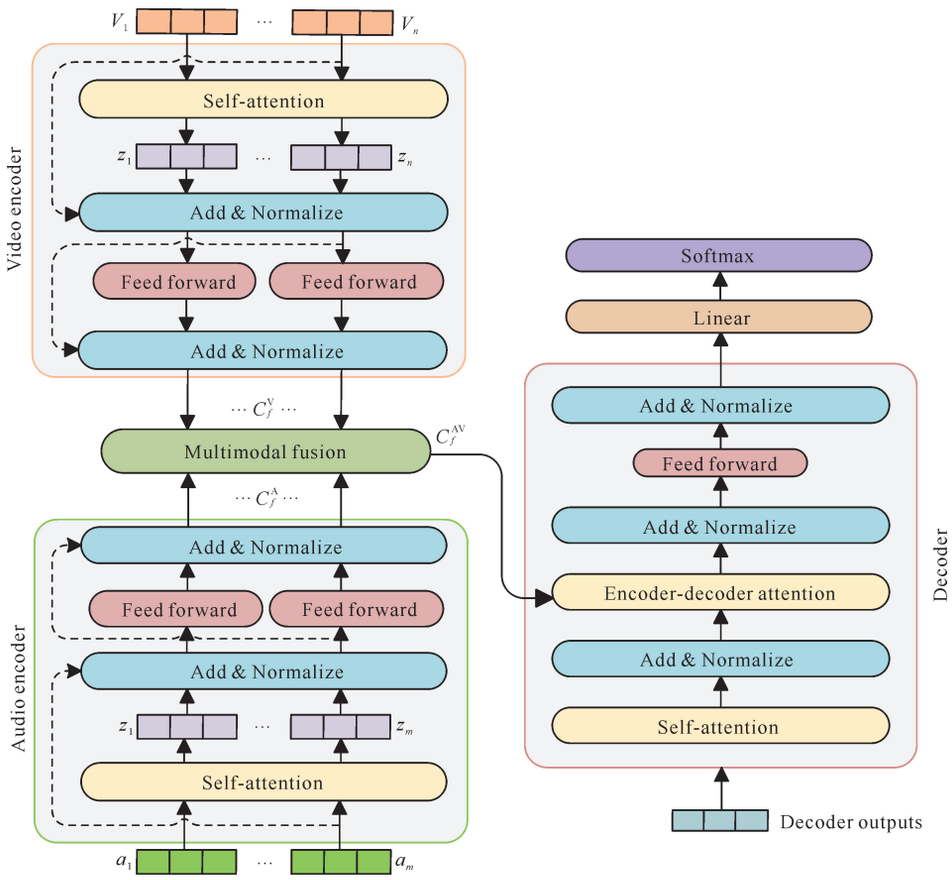

图 2　双流多模态编解码网络

Fig. 2　Dual-stream multi-modality codec network

首先，音频编码和视频编码的结构相同，均使用 transformer 自注意力编码层，分别对音频、视频流两个通道的数据单独编码，获得音频流编码特征 $C_f^{\text{A}}=[C_1^{\text{A}},C_2^{\text{A}},C_3^{\text{A}},\cdots,C_m^{\text{A}}],C_m^{\text{A}}\in R^d$，视频流编码特征 $C_f^{\text{V}}=[C_1^{\text{V}},C_2^{\text{V}},C_3^{\text{V}},\cdots,C_n^{\text{V}}],C_n^{\text{V}}\in R^d$，音频、视频流的编码特征向量的维度均为 512。然后，通过多模态特征融合网络输出音视频联合特征。具体公式如下：

$$C_f^{\text{A}}=\text{Audio}_{\text{attention}}(Q_a,K_a^{\text{T}},V_a)=\text{softmax}(\frac{Q_a\cdot K_a^{\text{T}}}{\sqrt{d_k}})V_a, \tag{3}$$

$$C_f^{\text{V}}=\text{Video}_{\text{attention}}(Q_v,K_v^{\text{T}},V_v)=\text{softmax}(\frac{Q_v\cdot K_v^{\text{T}}}{\sqrt{d_k}})V_v, \tag{4}$$

$$C_f^{\text{AV}}=\text{Joint}\{\text{cat}(C_f^{\text{A}},C_f^{\text{V}})\}, \tag{5}$$

将音频编码特征 C_f^{A} 和视频编码特征 C_f^{V} 送入多模态特征融合网络，经特征拼接后送入联合卷积网络，最终输出多模态联合特征 $C_f^{\text{AV}}\in R^d$，联合特征的维度与前面音频编码特征和视频编码特征维度相同。最后，将音频、视频联合编码特征送入解码单元。在双流多模态编码网络中，采用两个编码网络分别处理音频特征和视频特征得到对应的编码特征，每个数据流的编码网络由 6 个编码层组成。经特征融合网络学习音频模态、视频模态特征相关性获取增强型联合特征 $C_f^{\text{AV}}\in R^d$，实现音频、视频两种模态在帧级的增强型特征表示。该网络可以根据不同任务的需要在仅音频、仅视频、音视频融合 3 种模态任务下工作，在仅音频模态数据下可以完成 ASR 任务，在仅视频模态数据下可以完成 VSR 任务，在音视频多种模态数据下，可以有效发挥 VSR 不受噪声干扰和 ASR 识别率高的优势执行 AVSR 任务并提升识别准确率及模型稳定性。

2.4　联合解码

该方法在 transformer 解码器上嵌入 LM 执行联合解码推理，LM 由多层堆叠的 LSTM 构建，使用训练集数据对其进行训练。Transformer 解码器与

前面编码器一样也由 6 层组成。在 AVSR 中,原有解码方法由当前 t 时刻的联合编码特征向量 $C_{f_t}^{\mathrm{AV}}$ 和前一时刻预测 y_{t-1} 通过 softmax 层计算概率分布。本文在原有的方法中嵌入 LM,结合 transformer 解码器执行联合解码,方法如下:

$$y=\lg p^{\mathrm{dec}}(y_t \mid y_{t-1},C_{f_t}^{\mathrm{AV}})+\beta\lg p^{\mathrm{LM}}(y_t \mid y_{t-1}), \tag{6}$$

其中,p^{dec} 由 transformer 解码器提供,p^{LM} 由 LM 提供,y_t 是 t 时刻的预测输出,β 为超参数。该方法既能实现音频、视频融合的解码任务,又可以完成单模态下仅音频或仅视频的解码任务,得到最终预测输出。

3 实验结果与分析

3.1 数据集

本文所有实验都是在大词汇量数据集 LRS2 上进行,该数据集是最大公开可获得的大规模连续语音识别数据集之一。LRS2 数据集设置及其与其他公开主流数据集对比情况如表 1 所示。从表 1 中可以看出 LRS2 共包含 14.4 万多条来自 BBC 的英国电视口语句子,时长约 225 h,具有说话人数更多、词汇量大等特点。

表 1 LRS2 和其他公开数据集

Table 1 LRS2 and other publicly available datasets

数据集 Datasets	子集 Subsets	人数 Number	句子 Sentence	词汇量 Vocabulary	时长(h) Times (h)
TCD-TIMIT	Train/test	62	6 913	-	-
LRW	Train	1 000	514 000	500	165.0
	Test	1 000	25 000	500	8.0
LRS2	Pre-train	1 000	96 318	41 427	195.0
	Train	1 000	45 839	17 660	29.0
	Validation	1 000	1 082	1 984	0.7
	Test	1 000	1 243	1 698	0.6

Note:"-" indicates no data

同时,针对环境噪声影响的问题,分析在不同环境噪声下 MAVSR 方法的性能,本文在 LRS2 数据集中添加嘈杂人声环境噪声,分别选择 10 dB、0 dB、-10 dB 3 种不同信噪比的噪声数据用于训练、验证和测试 MAVSR 方法。

3.2 数据预处理

3.2.1 音频

本文先采用 ffmpeg 工具包从视频中提取音频数据,并存储到对应的文件路径下,音频采样率为 16 kHz,单声道。使用窗口长度为 25 ms、帧移为 10 ms 的汉明窗,将连续音频信号转换成音频帧,并进行快速傅里叶变换获得频谱特征,具体方法如下:

$$\mathrm{STFT}\{x(t)\}(\tau,\omega)=\int_{-\infty}^{+\infty}x(t)\omega(t-\tau)e^{-j\omega t}\mathrm{d}t, \tag{7}$$

其中,$x(t)$ 是输入音频信号,τ 是帧移,$\omega(t)$ 是窗函数,$e^{-j\omega t}$ 频域变换。$\omega(t)$ 计算方法如下:

$$\omega(t)=\begin{cases}0.54-0.46\cos(\dfrac{2\pi t}{L-1}),0\leqslant t\leqslant L-1\\0,\text{其他}\end{cases}。 \tag{8}$$

3.2.2 视频

在视频处理部分,视频的帧率为 25 f/s。首先从每段视频中提取视频帧,将提取的每一帧图像转换为灰度图,并存储到对应文件目录下;然后对每一帧图像进行裁剪,只保留人脸区域。本文只采用说话人唇部连续运动的图像帧作为视频特征,剔除其他区域冗余信息,可提高数据质量,减少计算量和内存消耗。

3.3 实验设置与评价标准

本文所有实验都是在搭载 RTX2080Ti、显存 11 M 的 GPU 工作站上进行,采用 PyTorch 机器学习框架。使用连接时间分类损失函数(CTC Loss);采用 Adam 优化器,平滑参数 $\beta_1=0.9$,$\beta_2=0.999$;初始学习率均为 0.001,衰减步设置为 20。

本文实验的评价标准是以 LRS2 测试集上报告的字符错误率(Character Error Rate,CER)来衡量,其公式定义如下:

$$\mathrm{CER}=\frac{N_{\mathrm{Del}}+N_{\mathrm{Sub}}+N_{\mathrm{Ins}}}{N}, \tag{9}$$

其中，N_{Del} 表示识别结果相对于实际标注发生删除错误的字符数量，N_{Sub} 代表发生替换错误的字符数量，而 N_{Ins} 代表发生插入错误的字符数量，N 表示测试集上所有的字符数量。最终的结果用百分比表示，字符错误率数值越小表明实验结果性能越好。

3.4 结果与分析

本节中 A、V 和 AV 分别表示仅音频模态、仅视频模态以及音视频融合模态任务。

3.4.1 MAVSR 方法在不同模态下的结果对比分析

本文所提 MAVSR 方法在 A、V、AV 3 种模态数据和 4 种音频环境(清晰、10 dB、0 dB、－5 dB)中的对比实验结果，如图 3 所示。

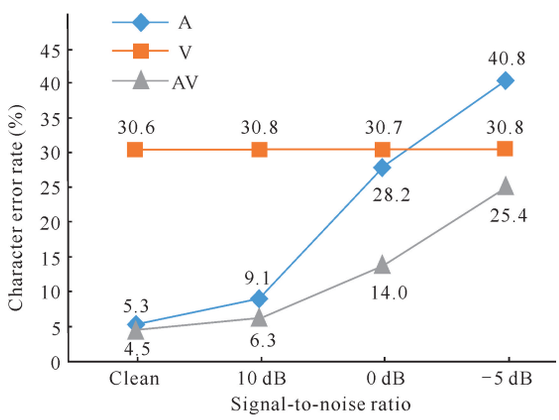

图 3 MAVSR 在 LRS2 数据集上的 CER

Fig. 3 MAVSR CER on the LRS2 dataset

随着噪声的加剧，MAVSR 方法在仅视频模态任务下的识别准确率依然保持稳定，而音频的准确率急剧下降。这表明视觉语音识别相对于音频语音识别在环境噪声中具不受噪声干扰的优势。同时在不同环境噪声中，AV 模态的识别准确率相较于音频模态有明显的提升，并且这种提升效果随着噪声的加剧更显著。具体表现为从清晰音频条件下 0.8% CER 的下降，到信噪比为－5 dB 环境噪声下 15.4% CER 的下降(图 3)，表明本文融合视频模态的方法可以有效提升 MAVSR 方法在环境噪声下的识别准确率。

3.4.2 MAVSR 在 LRS2 数据集上的消融实验

与基线模型对比 MAVSR 方法各个模块改进带来的性能改善效果，结果如表 2 所示。

表 2 MAVSR 消融实验

Table 2 MAVSR ablation experiments

方法 Method	CER (%)
Baseline	15.57
ASR + VSR (E2E)	13.60
E2E + Pre-training (Ⅰ)	9.80
Ⅰ + Transformer encoder (Ⅱ)	6.10
Ⅱ + modality controller	5.00
Ⅲ + LSTM LM	4.50

首先，本文采用音频与视频双流模型并以端到端的方式训练模型，对比基线方法实现 1.97%的改进。其次，将 LRS2 预训练数据集上训练的模型对前端模型初始化，观察到 CER 进一步降低 3.80%。再次，使用 transformer 编码器并加入卷积融合网络，性能提升 3.70%。最后，加入模态选择器和基于 LSTM 的语言模型实现了 1.60%的性能改善，以及 4.50%的字符错误率(表 2)。

3.4.3 MAVSR 方法与其他主流方法对比分析

MAVSR 方法与其他主流方法实验结果对比分析，以及每种方法在不同模态下的实验结果，如表 3 所示。

从表 3 可以看出，MAVSR 方法的性能优于其他几种方法，与 AV-transformer 方法相比，MAVSR 方法在清晰语音条件下的字符错误率降低 7.6%，与 BLSTM-DFN 相比降低 3.3%。在信噪比为 10 dB、0 dB、－5 dB 的环境噪声中，与其他方法相比，MAVSR 方法在 AV 模态比在仅音频模态中的 CER 下降更为明显。这表明本文提出的方法在嘈杂的环境噪声中能更好地利用视频模态不受声学噪声干扰的优势，提升视频模态与音频模态的互补性。

表 3 MAVSR 与其他主流方法的 CER (%)对比

Table 3 Comparison of MAVSR and other mainstream methods with CER (%)

方法 Method	模态 Modality	清晰 Clean	10 dB	0 dB	－5 dB
AV- transformer	A	13.7	18.8	31.2	43.7
	AV	13.1	17.8	30.6	43.3
	AV + AU	12.1	14.8	23.5	31.7

续表

Continued table

方法 Method	模态 Modality	清晰 Clean	10 dB	0 dB	−5 dB
AV-Align	A	16.4	21.9	36.3	49.1
	AV	15.8	18.3	26.6	34.0
BLSTM-DFN	AV	7.8	10.8	16.4	23.1
MAVSR	A	5.3	9.1	28.2	40.8
	V	30.6	30.8	30.7	30.8
	AV	4.5	6.3	14.0	25.4

3.4.4 语言模型对 CER 的影响

在 ASR 中,LM 被证实可以提升 ASR 的识别准确率。LM 对 AVSR 识别准确率的提升效果如图 4 所示。

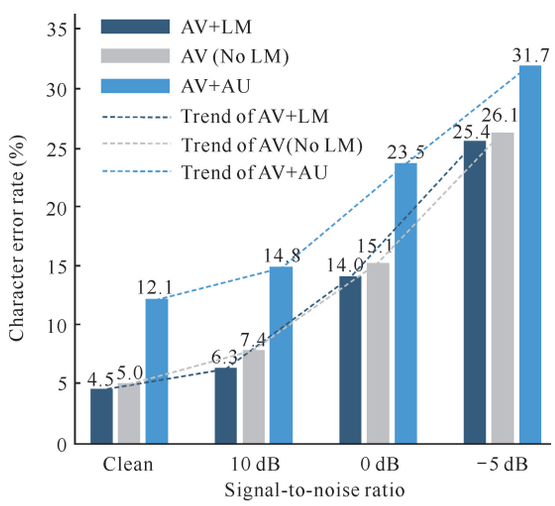

图 4 联合 LM 解码在不同 SNR 下的 CER

Fig. 4 Combining LM decodes CER under different SNR

在 MAVSR 方法中,通过选择是否联合 LM 解码,分别获得 AV + LM 和 AV (No LM)两种方法的 CER,并与其他方法进行对比,通过图 4 可以清楚看出 AV + LM 方法的 CER 最低可以达到 4.5%。在 MAVSR 方法中,联合外部语言模型解码(AV + LM)的 CER 明显低于 AV(No LM)解码的 CER,并且低于 AV-transformer (AV + AU)方法的 CER。这一实验结果有力地说明了外部语言模型不仅在传统语音识别模型中可以提升网络的性能,而且在 AVSR 中也能带来显著的 CER 改善。

3.4.5 音视频模态数据比例对系统性能的影响

为降低 MAVSR 对音频数据的依赖性,探索音视频模态数据比对 CER 的影响,本文设计模态选择器控制音频模态的数据权重,通过控制音频、视频两种模态在训练时的数据比,获得不同音频数据权重下的最优模型。采用同一测试集进行测试,获得不同模态数据比的 CER,验证不同比例的音视频数据与 MAVSR 方法识别准确率的关系。以下实验分别在清晰语音和 SNR 为 10 dB 噪声语音数据中进行,实验结果如图 5 所示。

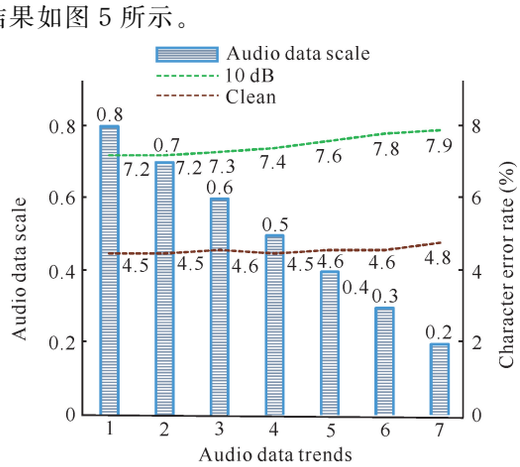

图 5 不同音频数据比下的 CER

Fig. 5 CER under different audio data ratios

由图 5 可以看出,随着音频模态数据逐步减少,MAVSR 方法的 CER 在清晰语音和 SNR 为 10 dB 的噪声语音数据中基本趋于稳定。该实验结果表明 MAVSR 方法能够有效发挥视频模态对音频的互补作用,降低音频模态数据占据主导性的影响,确保 MAVSR 方法在弱音频数据中的识别稳定性。同时考虑到在某些特定领域中难以获取足够的音频数据,采用本文方法依然可以在弱音频数据条件下取得稳定的识别准确率。

4 结论

本文提出一种多模态视听语音识别(MAVSR)方法,该方法构建基于自注意力机制双流编码模型,音视频模态联合卷积网络获取音视频模态特征联合

表示，可有效利用唇读与 ASR 的关联性和互补性；在前端编码模型中嵌入模态选择器，降低 MAVSR 方法对音频模态数据的依赖性；在解码推理中嵌入语言模型联合解码并进一步降低识别错误率。MAVSR 可完成在仅音频模态、仅视频模态和音视频模态融合 3 种任务下的字符识别。验证实验表明，本文提出的方法可有效利用视频模态改善 MAVSR 的性能，在大词汇量环境噪声下获得显著的准确率提升并明显优于其他主流方法，提升模型在弱音频数据中识别率的稳定性以及环境噪声下的鲁棒性。下一步，笔者考虑采用特征级与决策级的多级自适应融合机制来实现多模态的融合，研究弱标签或无标签视听语音识别任务。

参考文献

[1] CROSSE M J，DILIBERTO G M，LALOR E C. Eye can hear clearly now：inverse effectiveness in natural audiovisual speech processing relies on long-term cross modal temporal integration [J]. Journal of Neuroscience，2016，36(38)：9888-9895.

[2] BOZKURT E，ERDEM C，ERZIN E，et al. Comparison of phoneme and viseme based acoustic units for speech driven realistic lip animation [C]//Signal Processing and Communications Applications，2007. Piscataway，NJ：IEEE，2007.

[3] NODA K，YAMAGUCHI Y，NAKADAI K，et al. Lipreading using convolutional neural network [C]//Proceedings of the 15th Annual Conference of the International Speech Communication Association. Singapore：International Speech Communication Association，2014：1149-1153.

[4] 卢小春，胡维平，王修信. 基于人工神经网络的汉语数字语音识别系统[J]. 广西科学，2004，11(4)：320-322.

[5] 梁骁，黄文明，姚俊，等. 结合多注意力和条件变分自编码器的宋词生成模型[J]. 广西科学，2022，29(2)：308-315.

[6] STAFYLAKIS T，TZIMIROPOULOS G. Combining residual networks with LSTMs for lipreading [C]//Proceedings of the 18th Annual Conference of the International Speech Communication Association. Stockholm，Sweden：International Speech Communication Association，2017：3652-3656.

[7] ASSAEL Y M，SHILLINGFORD B，WHITESON S，et al. Lip net：end-to-end sentence-level lipreading [Z/OL]. (2016-11-05)[2022-11-14]. https://doi.org/10.48550/arXiv.1611.01599.

[8] STERPU G，SAAM C，HARTE N. How to teach DNNs to pay attention to the visual modality in speech recognition [J]. IEEE/ACM Transactions on Audio，Speech，and Language Processing，2020，28：1052-1064.

[9] HARTE N，GILLEN E. TCD-TIMIT：an audio-visual corpus of continuous speech [J]. IEEE Transactions on Multimedia，2015，17(5)：603-615.

[10] AFOURAS T，CHUNG J S，SENIOR A，et al. Deep audio-visual speech recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，2018，42(4)：722-737.

[11] CHUNG J S，SENIOR A，VINYALS O，et al. Lip reading sentences in the wild [C]//2017 Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Piscataway，NJ：IEEE，2017：6447-6456.

[12] KANNAN A，WU Y，NGUYEN P，et al. An analysis of incorporating an external language model into a sequence-to-sequence model [C]//2018 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). Piscataway，NJ：IEEE，2018.

[13] 崔海燕，李雅文，徐欣. 基于时间卷积网络的科技需求主题热度预测算法[J]. 广西科学，2022，29(4)：627-633.

[14] MARTINEZ B，MA P，PETRIDIS S，et al. Lipreading using temporal convolutional networks [C]//2020 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). Piscataway，NJ：IEEE，2020：6319-6323.

[15] PETRIDIS S，LI Z，PANTIC M. End-to-end visual speech recognition with LSTMs [C]//2017 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). Piscataway，NJ：IEEE，2017：2592-2596.

[16] 欧阳苏宇，邵莹侠，杜军平，等. 基于字词混合和 GRU 的科技文本知识抽取方法[J]. 广西科学，2022，29(4)：634-641.

[17] PETRIDIS S，STAFYLAKIS T，MA P，et al. End-to-end audiovisual speech recognition [C]//2018 IEEE International Conference on Acoustics，Speech and Signal Processing (ICASSP). Piscataway，NJ：IEEE，2018：6548-6552.

[18] TORFI A，IRANMANESH S M，NASRABADI N，et al. 3D convolutional neural networks for cross audio-visual matching recognition [J]. IEEE Access，2017，7(5)：22081-22091.

[19] WAND M，KOUTNÍK J，SCHMIDHUBER J. Lipreading with long short-term memory [C]//2016 IEEE In-

ternational Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway, NJ: IEEE, 2016: 6115-6119.

[20] PARASKEVOPOULOS G, PARTHASARATHY S, KHARE A, et al. Multiresolution and multimodal speech recognition with transformers [C]//2020 Proceedings of the Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 2381-2387.

[21] STERPU G, SAAM C, HARTE N. Should we hard-code the recurrence concept or learn it instead? Exploring the transformer architecture for audio-visual speech recognition [C]//2020 Proceedings of the Annual Conference of the International Speech Communication Association. [S. l.: s. n.], 2020: 3506-3509.

A Multi-modality Audio-Visual Speech Recognition Method under Large Vocabulary Environmental Noise

WU Lan**, YANG Pan, LI Binquan, WANG Han

(School of Electrical Engineering, Henan University of Technology, Zhengzhou, Henan, 450001, China)

Abstract: Audio-Visual Speech Recognition (AVSR) technology can effectively improve the accuracy of character recognition by using the relevance and complementarity of lip reading and speech recognition. In view of the problems that the recognition rate of lip reading is much lower than that of speech recognition, the speech signal is easily damaged by noise, and the recognition rate of existing Audio-Visual Speech Recognition (AVSR) methods in large vocabulary environment noise is greatly reduced, a Multi-modality Audio-Visual Speech Recognition (MAVSR) method is proposed. This method constructs a dual-stream front-end coding model based on the self-attention mechanism, and introduces a modal controller to solve the problem of unbalanced recognition performance of each mode caused by the dominance of audio modes in the environment noise, and improves the stability and robustness of recognition. A multi-modal feature fusion network based on one-dimensional convolution is constructed to solve the heterogeneous problem of audio and video data and improve the correlation and complementarity between audio and video modes. Compared with the existing mainstream methods, the recognition accuracy of this method is increased by more than 7.58% under the three tasks of audio-only, video-only, and audio-video fusion.

Key words: attention mechanisms; multi-modality; audio-visual speech recognition; lip reading; automatic speech recognition

责任编辑:陆媛峰