参赛学生姓名：曾一轩

中学：武汉英中高级中学

省份：湖北省

国家/地区：中 国

指导老师姓名：梁联晖

指导老师单位：广西大学

论文题目：基于多注意力融合的多语言视听语音识别方法研究

摘 要

目前全球有超过 4.3 亿人受到严重听力障碍的影响，而谷歌翻译眼镜是一款外观近似于普通眼镜的智能穿戴设备，其搭载的智能语音交互系统能够将语音转换为对应文字进行显示，一定程度上可以协助听障人士识别语音交互内容。然而，谷歌翻译眼镜依赖于音频信息进行语音识别，在嘈杂噪声的复杂环境中，其性能会大幅度下降。相比之下，视觉语音识别通过视频序列捕捉说话者的唇部运动及相关面部区域的动态特征，完全不受任何声学噪音的干扰，这使得它在嘈杂的声学环境中展现出无可比拟的鲁棒性。但是视觉语音识别对环境光线、遮挡物和拍摄角度极为敏感，这严重限制了其在实际应用场景中的适用性与可靠性。因此，探索将视觉与音频特征作为互补信息进行融合的视听语音识别技术，对提升听障人群的日常交流能力具有重要意义。视听语音识别算法主要是基于深度学习，例如，基于卷积神经网络的方法可以有效提取局部时空特征，并通过深层结构捕捉音视频的关键模式；基于 Transformer 的方法可以对音频和视频特征的远距离依赖关系进行建模。然而，现有的方法在在不同语言与多口音环境、嘈杂或远场语音以及多说话者干扰等复杂场景下的特征提取能力依然不足，易引入大量冗余特征，导致难以捕获有效信息，并且很难实现多模态之间的有效交互。

基于上述挑战，本文提出一种新型的基于多注意力融合的多语言视听语音识别模型，用来提升模型在复杂场景下的鲁棒性和识别效果。具体来说，本文提出头部混合注意力作为编码器分别处理音频信息和视觉信息，它通过将每个注意力头作为一个专家，每个专家头独立处理最相关的特征，所有专家头协同配合不断优化整体特征表示，以增强模型的远距离建模能力；提出多尺度稀疏交叉注意力融合模块增强多模态特征的融合，使用稀疏策略增强对关键特征的提取，使用交叉注意力实现音频模态特征和视觉模态特征的深层次交互，进一步提升语音识别效果。在实验部分，本文使用 ViSpeR 数据集，这是一个包含了四种语言（中文，阿拉伯语，西班牙语，法语）的较为复杂的多语言数据集。本文将所提出的方法与当前最先进的方法进行比较，实验结果显示了所提模型实现了最低的错误率，证明了其强大的语音识别能力。

关键词：视听语音识别；多语言；头部混合注意力；多尺度稀疏交叉注意力

目 录

1. 引言

1.1 研究背景与意义

根据世界卫生组织的数据显示，截至目前全球已有超过 4.3 亿人受到严重听力障碍的影响，该数字预计将在 2050 年增加至 7 亿人或超过总人口的 10%。然而，目前针对该疾病的有效临床治疗手段仍然匮乏。根据《2024 年中国老年人听力损失与助听器应用调研报告》，我国老年群体听力损失规模达 1.2 亿人，其中 7000 万人需要专业设备干预。目前，诸如谷歌翻译眼镜之类的辅助技术，能够通过语音转文本功能，有效帮助听障人士识别语音交互内容。然而，尽管这些设备对提升听障人群的言语理解能力和日常交流效果方面具有重要作用，但在嘈杂噪声环境中，其性能依然受到明显限制。这一局限给听障人士的日常沟通带来了困扰，因此，开发一种具有更强语音增强能力的新型语音识别算法显得尤为重要。

语音识别是人工智能和人机交互领域的重要研究方向，其核心目标在于将自然语音信号准确转化为对应的文本信息，并在辅助听力障碍人士的沟通方面具有重要应用价值[1]。目前常见的语音识别技术大体可分为三类：自动语音识别（Automatic Speech Recognition，ASR）[2]、视觉语音识别（Visual Speech Recognition，VSR）[3]以及视听语音识别（Audio-Visual Speech Recognition，AVSR）[4]。ASR 通过分析语音的波形将其转化为文本数据，但其在噪声环境下的性能会显著下降。VSR 则依赖视频数据捕捉说话者的唇部及面部动态信息进行识别，然而其易受环境光线、遮挡物和拍摄角度等环境因素影响，这意味着纯视觉方式难以完全替代音频信号。相比之下，AVSR 综合利用音频与视觉信息，实现跨模态信息的互补与融合[5]，在上述复杂环境下可展现出更强的语音识别能力。

AVSR 的核心思想是在传统声学建模的基础上引入视觉模态，以实现跨模态信息的融合，从而增强语音转文本能力。通过融合音频与视觉信息，AVSR 能够有效缓解噪声环境下音频信号的不完整性问题，提高语音识别的鲁棒性和准确性。由于深度学习能够从原始数据中自主学习特征表示，并具备处理大规模数据的能力，基于深度学习的方法逐渐成为 AVSR 领域的主流方法。尽管已有大量基于深度学习的 AVSR 方法，但现有方法在不同语言与多口音环境、嘈杂或远场语音以及多说话者干扰等复杂场景下仍存在识别性能下降的问题。此外，多模态信息在时间和空间尺度上的关联特征仍未被充分挖掘，这限制了 AVSR 在实际应用中的鲁棒性与泛化能力。因此，深入研究并开发一种新型深度学习模型，对于提升 AVSR 在复杂环境下的识别性能，以及辅助听力障碍人士的日常沟通和语言理解，具有重要的理论意义和应用价值。

1.2 国内外研究现状

1.2.1 自动语音识别

ASR 是语音识别技术发展的基础形态，也是目前应并用最为广泛的语音转文字方法[6]。该系统通过对声音信号进行采集、预处理和特征提取，结合语言模型、声学模型和深度神

经网络，从而实现语音内容的精准识别与语义理解。其技术发展历经模式匹配、统计建模到端到端深度学习等多个阶段，持续推动着人机交互技术的演进与革新。

早期语音识别多依靠滤波器结合基础的模式匹配方法实现。1952 年，贝尔实验室研发出了首个 ASR 系统 Audrey，能够实现对十个英文数字的识别[7]。1959 年，林肯实验室研制出一套可识别十种元音的系统，并支持在不同说话者之间进行区分[8]。此类系统多基于对语音信号的频谱分析，通常借助滤波器提取特征，并利用模式匹配完成识别[9]。然而，这一时期的语音识别系统在处理发音差异、口音以及语速变化等复杂情况时表现有限，识别准确率仍不理想。

计算机科学和概率统计模型的发展为语音识别提供了新范式。1967 年，Leonard Baum 提出了隐马尔可夫模型（Hidden Markov Model，HMM）的理论，该方法能够同时对语音特征与语言规律进行统计建模，并借助马尔可夫链来刻画说话者差异和音调变化等动态特性，可以成功识别未知语音信号[10]。20 世纪 80 年代，HMM 与高斯混合模型（Gaussian Mixture Model，GMM）的结合，实现了对连续语音和大词汇量任务的支持[11]。在深度学习兴起之前，语音识别主要依赖于 HMM-GMM 框架。虽然这些方法提高了识别准确率，但是在应对日常交流时，它的表现依旧存在明显的不足。

在深度学习推动下，ASR 取得了显著进展[12]。深度学习不仅可以直接从原始数据中学习特征，还具备处理大规模数据的能力。这一时期，语音识别技术分为两个主要方向：其一是将 HMM 与深度神经网络（Deep Neural Network，DNN）结合，其二是端到端模型。2011 年，Dong 等人提出了一种新型 CD-DNN-HMM 模型，通过将后验概率转化为似然函数，并利用 DNN 连接相邻特征以学习上下文信息[13]。虽然 DNN-HMM 在大词汇任务中较 HMM-GMM 更优，但其结构通常较复杂，计算效率也相对较低。相比之下，端到端模型因能够省略声学模型设计，直接将声学信号映射为标签序列而受到广泛关注。常见的端到端方案包括两类：一类是联结主义时间分类（Connectionist Temporal Classification, CTC）[14]，另一类是序列到序列的模型[15]。

CTC 的动态规划算法在损失计算中具有较高的效率，因而特别适用于大规模数据集的训练。Florian Eyben 等人首次将 CTC 应用于语音识别任务，并结合长短时记忆网络（Long Short-Term Memory，LSTM），取得了显著效果[16]。随后，Yao 等人提出了一种正则化 CTC 方法，通过引入掩码预测增强上下文表示，从而缓解过拟合并提升模型的泛化能力[17]。

近年来，得益于注意力机制强大的上下文建模能力，ASR 得到了飞速的发展[18]。2020 年，Google 提出了一种针对大规模数据集的 Transformer 架构，它通过引入循环神经网络转换器并结合掩码机制进行上下文特征提取，在语音识别任务中表现出了优异的性能[19]。随后，Gulati 等人设计了 Conformer，该模型通过卷积神经网络（Convolutional Neural Network，CNN）模块捕获语音信号的局部特征，同时利用 Transformer 建模全局上下文依赖关系，从而实现短时与长时依赖的有效融合，显著提升了识别准确率[20]。Peng 等人提出了 Branchformer，通过并行分支结构使模型能够同时捕获局部与全局特征，进一步改善了识别

性能[21]。E-Branchformer 在 Branchformer 的基础上堆叠额外的逐点模块，以增强特征建模能力，并进一步提高了语音识别的整体效果[22]。

尽管上述基于纯音频模态进行识别的方法已经得到了广泛的应用，但在实际场景中仍面临诸多挑战，例如在噪声环境下性能易大幅下降、说话人存在口音时识别率显著降低、多人重复讲话时模型难以区分语音来源。这些问题既影响语音识别精度，也限制了语音识别在实际应用中的泛化能力。

1.2.2 视觉语音识别

VSR 又称唇读识别，是一种不受声学噪音干扰，依赖于视觉模态进行语音信息解码的技术。该方法通过视频序列捕捉说话者的唇部运动及相关面部区域的动态特征，并利用深度学习等模型对这些视觉信号进行建模，从而推断对应的语音内容。

早期的 VSR 方法主要是基于机器学习和手工提取的方法。Hwang 等人提出了一种结合主成分分析（Principal Component Analysis，PCA）与 Snake 模型的方法，在单词级别的唇语数据上显著提升了识别准确率[23]。由于支持向量机在处理高维特征方面表现优异，Wang 等人将其引入唇读识别，并取得了良好效果。然而这类方法受限于手工提取的特征，在面对复杂场景时的特征表达能力仍显不足[24]。

随着深度学习技术的发展，VSR 技术迈入了全新的发展阶段，特别是基于 CNN 和 Transformer 的方法。2016 年，Google 和牛津大学的研究人员提出了 LipNet 系统，首次在唇读识别中采用端到端结构，并实现了优于传统方法的性能[25]。Stafylakis 等人结合了残差网络和 LSTM，利用 CNN 提取细节特征，利用循环神经网络（RNN）对视频序列的时间动态进行建模，从而捕捉连续唇部动作的时序信息，成功预测了对应的语音内容[26]。Martinez 等人引入了一种新型的基于时间卷积的框架，有效捕捉视频序列的时序特征，从而提升唇语识别准确率[27]。Prajwal 等人设计了视觉 Transformer 作为编解码器，通过引入注意力机制增强了模型的全局上下文建模能力[28]。Ma 等人提出了 GhostNet-TSM 网络，该模型通过结合同类自知识蒸馏对特征进行解耦，在提高了计算效率的同时性能超越了多数非轻量化模型[29]。Baaloul 等人将 CNN 和视觉 Transformer 结合起来，从而有效实现局部特征提取和全局上下文的建模，有效增强了唇读识别的能力[30]。

尽管 VSR 已取得显著进展，但仍面临诸多挑战，例如说话者间的口型差异、光照和拍摄角度变化以及视频质量波动等问题。实际上，仅依靠视觉模态进行语音识别难以应对音频信号缺失的情况，因此将视觉信息与音频信号相结合成为提高识别精度和鲁棒性的有效途径。

1.2.3 视听语音识别

AVSR 是一种综合利用音频信号与视频信号进行语音内容识别的多模态技术，其核心思想是在传统声学建模的基础上引入视觉模态，从而实现音视频信息的互补与融合。与单一模态识别相比，AVSR 在嘈杂环境或音频信号缺失的情况下表现出显著优势。

早在 1985 年，Petajan 等人开发了全球首个唇读识别系统，并将其与音频信息结合，引入一种新的相似度计算算法，实现了语音信息的有效提取，这一工作标志着利用视觉信息增强语音识别能力的开端。2000 年，Dupont 等人将 HMM 应用于音频信号处理，同时设计了针对唇部运动的视觉特征提取方法，并探索了多模态信息融合策略对语音识别精度的影响[31]。该研究不仅验证了 AVSR 技术可行性，也为后续多模态语音识别方向的深入探索奠定了重要基础。

近年来，深度学习的快速发展推动了 AVSR 技术的持续提升。Ma 等人构建了一种结合 ResNet-18 和卷积增强的 Conformer 混合 CTC/注意力模型来提取单模态特征，并设计了一个多层感知机（Multilayer Perceptron，MLP）模块来融合不同模态的特征，有效实现了字符识别[32]。Maxime 等人在 Conformer-CTC 架构中引入了中间 CTC 损失、CTC 残差机制及补丁注意力，进一步提高了模型的性能[33]。Che 等人提出了自适应融合 Transformer，采用稀疏操作减少对非重要区域的过度关注，并利用自适应融合动态调整注意权重，从而提升多模态信息的捕获与融合效果，降低冗余信息对性能的影响[34]。为了增强音频信号与视频信号之间的交互，一些交叉融合模块被提出。例如，Sterpu 等人提出了 AV-Transformer，采用跨模态注意力权重融合机制直接对视听数据进行交互，有效增强了多模态信息的融合能力，进而提升了语音识别的精度与可靠性[35]。Wang 等人提出了 MLCA-AVSR 模型，该模型利用了多层交叉注意力模块和 Inter-CTC 损失，实现了跨模态特征在学习全流程的递进式交互，进一步提升了 AVSR 的识别性能[36]。

随着深度学习技术的广泛应用，AVSR 领域也得到了显著提高。但是现有的 AVSR 模型在不同语言与多口音环境、嘈杂或远场语音以及多说话者干扰等复杂环境下的的特征提取能力仍有限，容易产生大量冗余特征，从而难以捕捉到有效的特征。此外，在进行跨模态融合时，音频模态信息和视觉模态信息很难实现有效的交互。这在一定程度上限制了 AVSR 模型在复杂场景下的识别性能，因此亟需探索更有效的特征提取与跨模态融合方法，以增强 AVSR 模型的鲁棒性和泛化性能。

1.3 本文主要工作

针对上述挑战，本文提出了一种基于多注意力融合的多语言视听语音识别模型（MAF-AVSR），该模型可以在提取局部特征的同时显著提升全局上下文特征建模能力，并增强音频模态和视觉模态的交互，从而实现更强的特征表达能力。受自然界生物群体决策机制的启发，例如蜜蜂或蚂蚁在觅食时部分个体负责探索新资源，而其他个体维持已有资源，群体通过个体分工优化整体决策，本文提出了头部混合注意力（Mixture-of-Head Attention，MoH）作为编码器模块。MoH 将每个注意力头视为一个专家，其中共享头用来处理通用特征，始终处于激活状态，其余的专家头专注于处理输入特征的不同模式。音频特征和视觉特征可以动态选择最相关的专家头进行处理，各个专家头分工协作，整个编码器借助个体头的协作优化整体特征表示与识别性能，以有效增强模型的上下文建模能力。

在跨模态融合方面，本研究提出了一种多尺度稀疏交叉注意力（Multi-Scale Sparse Cross-Attention，MSSCA）模块，以实现音频与视觉信息的有效整合。MSSCA 借助稀疏策略过滤冗余特征，保留关键特征，并通过交叉注意力模块增强多模态特征的交互，从而进一步增强模型的表达能力。

本研究的主要贡献如下：

- 本文提出了一种新型 MAF-AVSR 模型，该模型不仅能够在捕捉局部细节的同时强化全局上下文建模能力，而且可以促进音频与视觉模态的深度交互，从而显著提升整体特征表示能力与语音识别性能。
- 首次提出了一种 MoH 模型，允许每个令牌自适应地选择适当的注意力头，通过单个注意力头的协通工作优化整个模型的特征表示，以增强全局上下文建模能力。
- 为了增强音频和视觉模态特征的融合，本文提出了 MSSCA 融合模块，该模块分别利用稀疏策略和交叉注意力融合策略来增强对关键特征的提取和交互，有效提升了模型的性能。

本文的其余部分组织如下：第二部分系统介绍了本文所涉及的相关算法与理论基础，包括多头自注意力和混合专家模型，在上述理论的基础上，本研究提出了一种基于多注意力融合的多语言视听语音识别模型。第三部分详细阐述了本文提出的视听语音识别模型，包括模型的整体架构、编码器设计、跨模态融合模块以及训练策略。第四部分首先介绍了语音识别领域数据集的数据来源、语言类型、视频时长及评价指标，随后介绍了实验所需的实验环境和训练配置。最后，本文对所提方法与多种当前最先进方法在数据集上的对比实验及消融实验结果进行了展示，并对结果进行了说明与分析。第五部分概述了本文的主要贡献，并对未来发展方向提出了参考意见。

2. 相关工作

2.1 多头自注意力

多头自注意力（Multi-Head Self-Attention，MHSA）是 Transformer 架构中的核心机制，它的思想是在同一层中并行使用多个独立的注意力头，每个注意力头都会在不同的子空间学习输入序列的关联信息[37]。具体而言，输入首先被映射为不同的查询（Q）、键（K）和值（V）向量，每个注意力头独立执行缩放点积注意力以生成加权输出，随后通过拼接各头输出并进行线性变换得到最终结果。多头注意力机制的示意如图 1 所示。

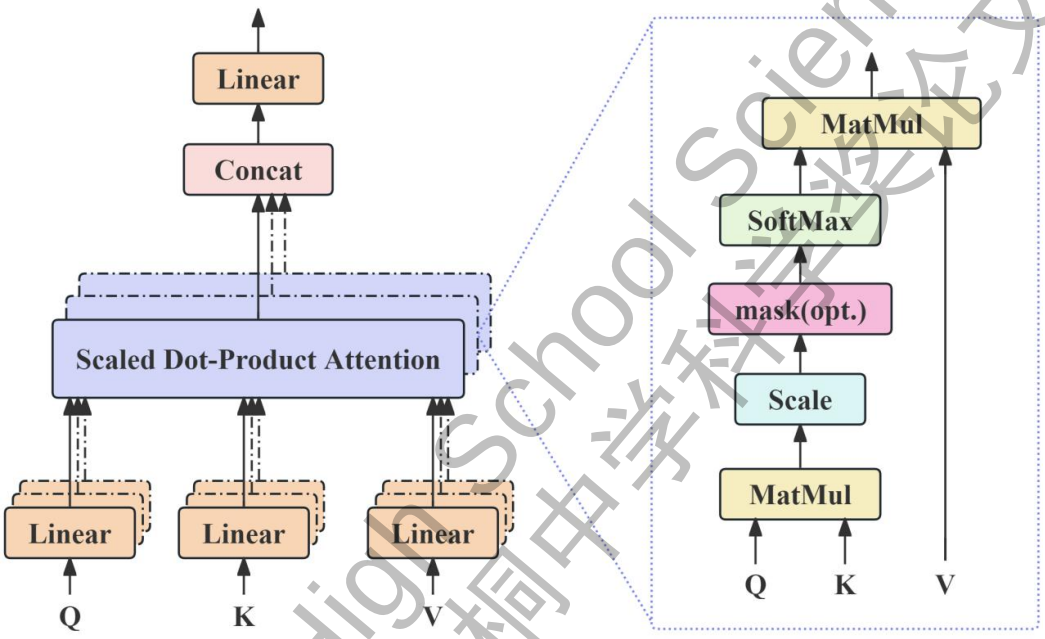

图 1 多头自注意力的结构图

该模型通过将输入标记映射到多个注意力头，实现了对数据中复杂依赖关系的有效建模。假设输入特征矩阵为 $\mathbf{A}$，对 $\mathbf{A}$ 进行线性映射后，可得到查询矩阵 $\mathbf{Q}$、键矩阵 $\mathbf{K}$ 与值矩阵 $\mathbf{V}$，如下所述：

$$\mathbf{Q}=\mathbf{A}\times\mathbf{W}_{\mathbf{Q}},\mathbf{K}=\mathbf{A}\times\mathbf{W}_{\mathbf{K}},\mathbf{V}=\mathbf{A}\times\mathbf{W}_{\mathbf{V}}, \tag{1}$$

式中，$\mathbf{W}_{\mathbf{Q}}$, $\mathbf{W}_{\mathbf{K}}$ 和 $\mathbf{W}_{\mathbf{V}}$ 是可学习权重矩阵。。经过线性变换后，利用缩放点积注意力对 $\mathbf{Q}$、$\mathbf{K}$ 和 $\mathbf{V}$ 进行自注意力运算，其表达式为：

$$\mathbf{H}=\text{Attention}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right)=\text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\mathrm{T}}}{\sqrt{D_k}}\right)\mathbf{V} \tag{2}$$

式中 D_k 表示注意力空间的维度。

多头自注意力机制中各个注意力头分别独立完成计算，随后将其结果在特征维度上进行拼接，形成整体的输出表示。其计算过程可以由以下公式表示：

$$\text{MHSA}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right)=\text{Concat}\left(\mathbf{H}_1,\mathbf{H}_2,\mathbf{H}_3,\cdots\cdots,\mathbf{H}_r\right)\mathbf{W} \tag{3}$$

式中 r 表示注意力头的数量。 $\mathbf{W}$ 表示可学习的参数矩阵。

通过这种方式，可获得更加丰富的上下文信息，从而显著提升模型的特征提取能力和泛化效果。相比单头注意力，多头机制使模型能够在同一层同时关注输入序列的不同方面，有效缓解信息瓶颈，并增强对语义信息的全面理解。

2.2 混合专家模型

混合专家模型（Mixture of Experts，MoE）是一类先进的神经网络架构，其核心在于通过整合多个子模型（即“专家”）的预测结果来提升整体性能[38]。该模型的核心机制包括两个部分：一是由多个并行的专家网络组成，每个专家子模型专注于特定的数据特征或任务；二是门控网络，根据输入特征动态分配权重，决定哪些专家被激活以及如何融合它们的输出。

通过这种设计，MoE 能够确保不同输入被分配给最适合的专家进行处理，从而在保持计算效率的同时显著提升预测的准确性和模型的泛化能力。近年来，MoE 已在大规模预训练模型和多模态任务中展现出巨大潜力，例如在视 AVSR 领域，通过为音频和视觉模态分别设计专家网络，实现灵活分工与融合[39]。这种策略不仅可以拓展模型容量，同时也为多模态信息的联合建模提供了创新思路。

3. 基于多注意力融合的多语言视听语音识别模型

3.1 整体框架

所提 MAF-AVSR 模型的整体框架如图 2 所示。该框架主要包括四个阶段，前端特征提取阶段，特征编码阶段，融合阶段，和解码阶段。该模型可以充分提取视觉和音频模态的局部细节和全局上下文特征，并增强模态间的交互以提升模型的特征表达能力。

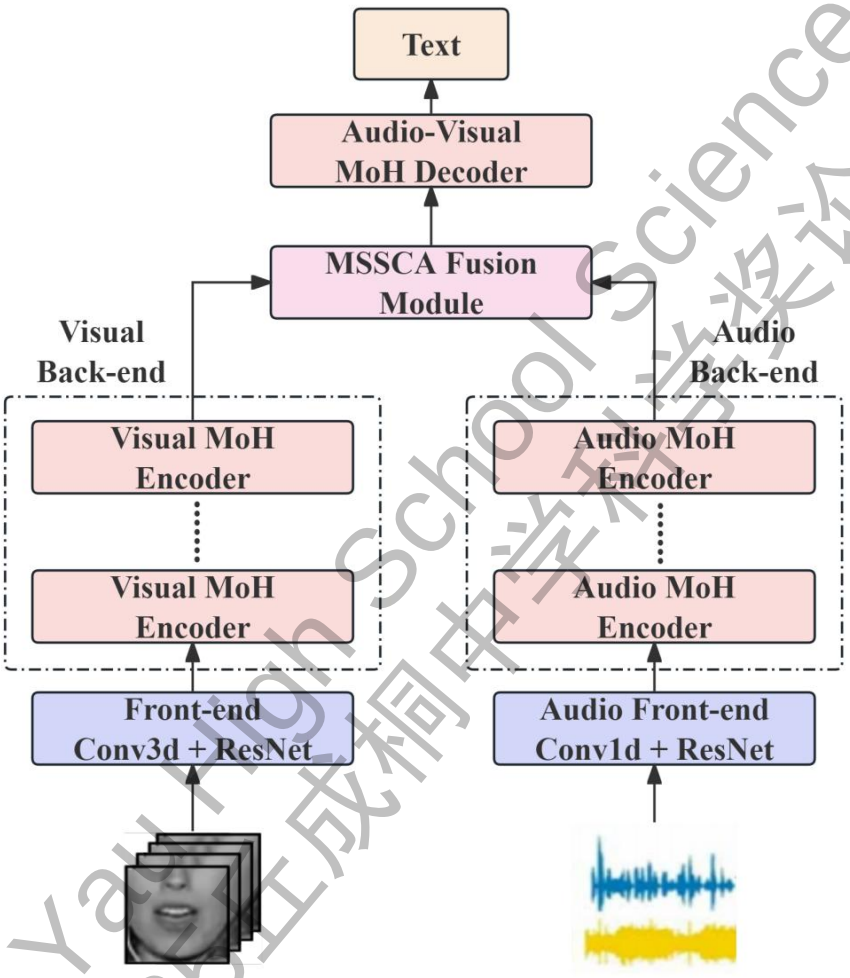

图 2 MAF-AVSR 模型的整体框架

在前端特征提取环节，视觉和音频模态各自采用独立网络进行处理。视觉端以裁剪后的唇部区域为输入，首先通过三维卷积网络提取时空特征，以捕捉唇部运动的动态变化，随后结合残差网络以提升视觉特征的层次化表示能力。音频端以带噪声的语音信号作为输入，利用一维卷积网络提取局部时序特征，并结合残差网络对复杂声学模式进行深层建模。该前端设计确保两类模态在进入编码器阶段时具备充分且鲁棒的特征表示。

在特征编码阶段，视觉与音频特征分别引入 MoH 编码器。MoH 作为一种基于注意力的编码器，其核心目标在于提升模型对输入序列上下文特征的捕捉能力。在视觉端，它有助于建模唇部运动序列的时间一致性；在音频端，则能够提取音频信号的长程上下文信息，同时提取语音学特征。两个编码器独立运行，保证各模态内部的特征依赖得到充分挖掘。

为实现跨模态信息交互与融合，本文构建了 MSSCA 融合模块。该模块通过多尺度建模在不同时间尺度上建立音频与视觉特征的对应关系，从而实现精细的跨模态对齐与信息互补。交叉注意力机制进一步增强两模态间的特征交互，同时结合稀疏策略抑制冗余和噪声特征，最大限度降低干扰。MSSCA 模块有效提升了音频与视觉模态融合的表达能力，进而提升了了整个模型的性能。

在解码阶段，多模态融合信息被输入音视频联合 MoH 解码器。该解码器利用多头自注意力对跨模态特征进行综合建模，并结合上下文信息逐步生成最终文本输出。通过这一解码过程，模型能够将复杂的多模态特征精准映射为可读文本，从而提升语音识别的准确性与鲁棒性。

3.2 多头自注意力求和形式

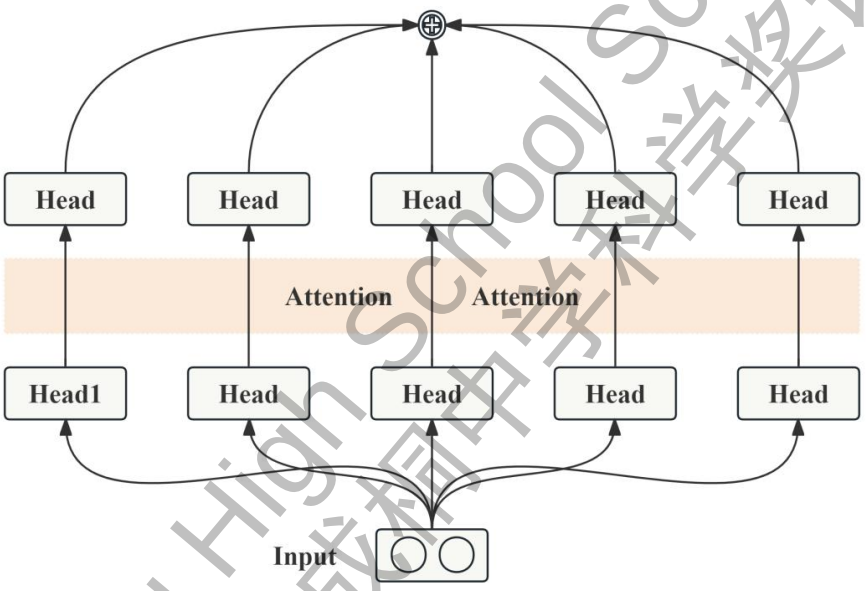

图 3 多头自注意力求和形式

MHSA 一般是拼接形式，即将各个注意力头的输出拼接后，通过一个线性映射生成最终的表示。然而，从另一种视角出发，MHSA 也可以用求和的形式来理解[40]，标准 MHSA 求和形式的结构图如图 3 所示。详细来说，我们将投影矩阵 $\mathbf{W}$ 按行进行分解，就可以用求和的形式来表达多头注意力。具体来说，将 $\mathbf{W}$ 按行划分为 r 个子矩阵，即：

$$\mathbf{W}=\begin{bmatrix}\mathbf{W}_1\\ \mathbf{W}_2\\ \vdots\\ \mathbf{W}_r\end{bmatrix} \tag{4}$$

式中，每一个 $\mathbf{W}_i$ 对应一个注意力头的输出映射矩阵，在这种划分方式下，MHSA 的输出可以看作是各个注意力头经过各自的线性变换后，在向量空间中进行逐元素相加而形成的结果。这种求和形式可由下式表示：

$$\mathrm{MHSA}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right)=\sum_{i=1}^{r}\mathbf{H}_i\mathbf{W}_i \tag{5}$$

依据上述公式，标准 MHSA 在各个头上并行计算，每个头独立产生子表示，最终通过聚合得到整体的上下文表示。从表示角度看，求和形式可视为拼接的一种等价变体，各注意力头工作在互补的子空间，其维度之和恰好等于隐藏维度的大小。因此，只要后接合适的线性映射，按头求和与拼接操作在表达上可以视作同一类聚合方式的不同实现。

基于“各头独立工作”的特性，本文进一步提出了 MoH 机制。该机制在处理输入序列时，会根据令牌的上下文语义特征动态计算各注意力头的相关性得分，并据此选择性地激活最具贡献的注意力头，同时抑制或跳过冗余度高或贡献较低的头，最终输出由被选中注意力头的加权组合构成。通过这一方式，模型在信息表达上能够保留关键注意力头的贡献，同时降低冗余干扰，从而释放注意机制在全局特征建模上的潜力。

3.3 头部混合注意力

近年来，MoE 方法因其在处理复杂任务时能够灵活调度计算资源、提高模型精度而受到广泛关注。MoE 通过将输入按特征动态分配给不同的专家子模型，每个专家专注处理特定任务或数据类型，随后通过加权整合各专家输出生成最终预测。这种机制能够充分利用专家的优势，实现更精确的建模和推理。考虑到 MHSA 在本质上可以视为求和操作，并结合 MoE 的优势，本节提出了 MoH 模型，该模型把每个注意力头视为一个专家，并且设计了共享头，两阶段路由，负载均衡损失来增强对视听特征的上下文建模能力。MoH 模型的算法结构图如图 4 所示。

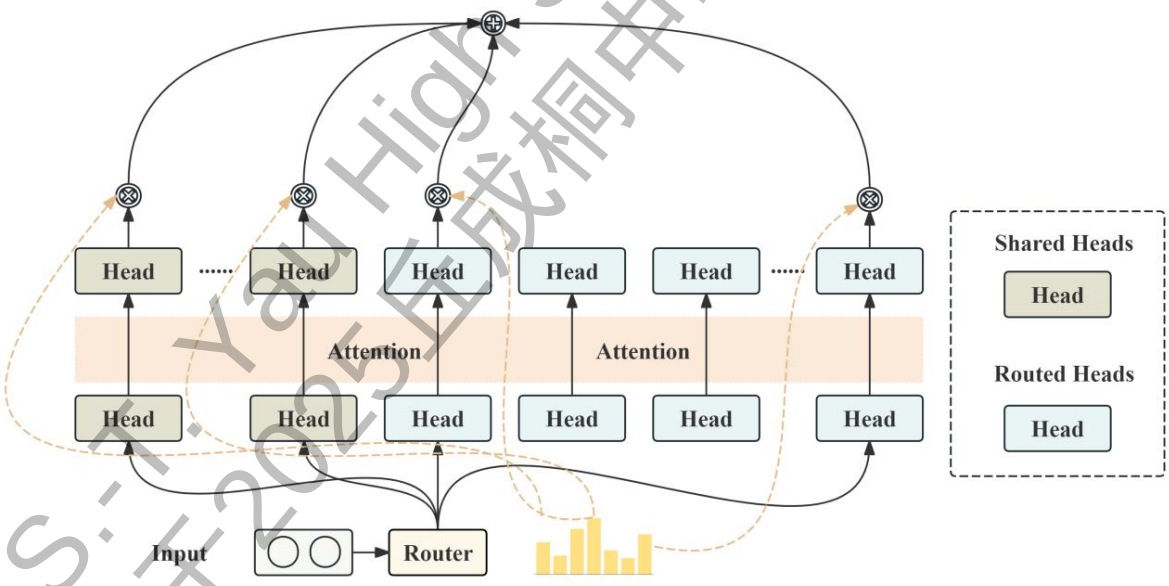

图 4 头部混合注意力的算法结构图

（1）注意力头作为专家：受自然界群体决策的启发，本文将每个注意力头视为一个专家，使其在特征提取过程中能够分工协作，从而实现最优性能。在具体实现中，我们结合 MoE 和 MHSA，提出了一种基于 MoH 的方法，它将 MHSA 中的每个注意头视为一个独立的专家。MoH 包含一组注意头 $\mathbf{H}=\{\mathbf{H}_1,\mathbf{H}_2,\mathbf{H}_3,\cdots\cdots,\mathbf{H}_r\}$ 以及一个用于激活 TopK 注意力头的路由器模块。具体来说，给定输入令牌 $\mathbf{X}$，路由器会为每个注意头计算路由分数，选择

得分最高的 K 个注意头参与计算。随后，这些注意力头的输出通过加权求和融合，得到最终的注意力表示。该过程可以由以下公式表示：

$$\mathrm{MoH}(\mathbf{X})=\sum_{i=1}^{r} s_i \mathbf{H}_i \mathbf{W}_i \tag{6}$$

式中，只有在第 i 个注意力头被激活时输出才会取非零值，这样的设计可以大大增强模型的表达能力。在标准多头注意力中，所有注意力头的输出被简单拼接或平均，这种处理方式难以区分不同头的重要性。而 MoH 的加权融合机制能够根据路由得分动态调整每个激活头在最终表示中的贡献大小，从而释放注意机制更强的潜在能力。这种加权策略不仅能突出关键头的作用，还能有效抑制噪声头对整体结果的干扰。

（2）共享头：在 MHSA 机制中，不同的注意力头通常承担着不同的任务。然而，部分注意力头在面对不同类型的输入时，可能学习到高度相似甚至完全相同的特征模式，从而导致计算资源浪费或信息冗余。这类特征往往具有跨任务、跨模态甚至跨语言的一致性，例如音频特征中的句法规则，视觉特征中的形状轮廓与边缘信息。由于这些模式在不同上下文中均可重复利用，它们可以被视为通用特征的载体。为此，我们将这类信息集中于一个固定的子集—共享头，并在训练过程中保持它们始终处于激活状态，从而减少其他路由头在学习过程中的特征冗余问题。

这种设计的优势在于，它能够将跨上下文的稳定模式“锁定”在共享头中，减少其他注意力头在学习过程中的信息重叠与功能冗余，从而释放更多的计算资源给动态路由头。动态路由头则可以根据具体输入自适应选择激活的子集，更专注于提取与当前任务或上下文强相关的特征。共享头的引入可以提高模型的表达多样性与泛化能力，使其在 AVSR 任务中表现更为优异。

（3）两阶段路由：在前期工作中，本文提出了共享头用来学习音频和视觉特征中的通用特征，路由头通过路由机制来选择哪些注意力头参与计算。为了在共享头与路由头之间实现更合理的权重分配，本文设计了一种两阶段路由策略。在该策略中，最终的路由得分并非仅由单个注意力头对输入令牌的响应强度决定，而是由每个注意力头的个体分数与该头部所属类型（共享或路由）的分数共同决定。这种双重得分的共同作用，使得模型在特征选择上既能保持共享头的稳定贡献，又能充分利用路由头的动态适应能力。具体而言，对于输入序列 $\mathbf{X}\in\mathrm{R}^{T\times d}$ 中的第 m 个令牌 $x_m\in\mathrm{R}^d$，路由得分 s_i 可以由以下公式定义:

$$s_i=\begin{cases}\beta_1\mathrm{Softmax}\left(\mathbf{W}_e x_m\right)_i & if\ 1\le i\le h_e\\ \beta_2\mathrm{Softmax}\left(\mathbf{W}_r x_m\right)_{i-h_e} & if\ Head\ i\ is\ activated\\ 0 & if\ \text{not}\end{cases} \tag{7}$$

式中，h_e 代表共享头的个数，$\mathbf{W}_e\in\mathrm{R}^{h_e\times d}$ 表示共享头的投影矩阵，$\mathbf{W}_r\in\mathrm{R}^{(r-h_e)\times d}$ 表示路由头的投影矩阵。仅当满足 $\left(\mathbf{W}_r x_m\right)_{i-h_e}\in\mathrm{Topk}\left(\left\{\left(\mathbf{W}_r x_m\right)_{i-h_e}\middle| h_e+1\le i\le r\right\}\right)$ 时，路由头 i 才能被激活。系数 β_1 与 β_2 用于平衡共享头与路由头的相对贡献，其定义可以由以下公式表示：

$$[\beta_1,\beta_2]=\text{Softmax}\left(\mathbf{W}_f x_m\right) \tag{8}$$

式中，$\beta_1+\beta_2=1$，$\mathbf{W}_f\in\mathrm{R}^{2\times d}$ 为可训练的投影矩阵，d 表示的隐藏维度的大小。这一机制能够根据不同输入自适应调整共享头与动态路由头的参与比例，从而有效提升信息利用效率和模型的上下文特征建模能力。

（4）**负载均衡损失：**在 MoE 中，直接进行训练往往会出现路由机制偏置问题，即大多令牌会集中被分配给少数几个专家。这种“专家负载不均衡”的情况会造成两个直接后果：一方面，被频繁选中的专家参数更新过快，容易导致过拟合或训练不稳定；另一方面，其余很少被选中的专家得不到足够的梯度信号和数据样本，几乎处于“闲置”状态，难以发挥 MoE 的并行建模优势。这种不均衡不仅降低了模型整体的参数利用率，也削弱了 MoE 设计所追求的稀疏高效特性。为了减轻所提出的 MoH 中可能出现的负载分配不均问题，本文借鉴了以往 MoE 方法的做法，在训练过程中引入负载均衡损失。具体而言，对于输入序列 $\mathbf{X}\in\mathrm{R}^{T\times d}$ 中的第 m 个令牌 $x_m\in\mathrm{R}^d$，这里定义了负载均衡损失 L_b，以确保不同专家能在训练过程中均衡接收令牌，从而避免部分专家过度饱和而其他专家利用不足的情况，定义如下：

$$\begin{aligned}
L_b &= \sum_{m=h_e+1}^{r}\mathbf{Q}_m\mathbf{F}_m \\
\mathbf{Q}_m &= \frac{1}{T}\sum_{m=1}^{T}\text{Softmax}\left(\mathbf{W}_r x_m\right)_{i-h_e} \\
\mathbf{F}_m &= \frac{1}{T}\sum_{m=1}^{T}1\left(x_m\cdot Head_i\right)
\end{aligned} \tag{9}$$

式中，1(*) 表示指示函数。

一般来说，整体训练的目标函数包含两个主要部分：一是面向具体下游任务的损失项，用于直接优化模型在该任务上的性能；二是负载均衡损失，用于约束专家的使用分布，防止调用集中或分配不均。最终的训练损失为这两类损失的加权组合，通过合理设定权重系数，在保证模型任务性能的同时，也可以提升注意力头之间的利用率与训练的稳定性。

3.4 多尺度稀疏交叉注意力

在利用 MoH 编码器分别提取音频特征和视觉特征后，需要对上述多模态特征进行融合。为此，本文提出了一种新型的 MSSCA 融合模块，该模块不仅可以通过稀疏方案减少无关信息的影响，从而最小化干扰项的影响，并且可以有效建模单一数据源特征图的空间位置信息，并在二维平面上与音频特征及视觉特征的特征图实现交互。该模型的基本框架如图 5 所示。

在具体实践中，我们首先通过对输入特征图应用不同窗口大小的平均池化操作，生成多个池化特征图 。这些池化操作能提取输入数据的多尺度信息，较大的窗口关注整体结构特征，而较小的窗口则保留细节信息。随后，将这些多尺度池化得到的特征图进行融合，以形成一个统一且信息丰富的表示，为后续的特征提取和注意力计算提供输入。为提取视听特征的自注意力表示，先计算 $\mathbf{Q}$ 与 $\mathbf{K}$ 的转置矩阵乘积，再经 softmax 进行激活，该过程

可以由以下公式表示：

$$\begin{aligned}\mathbf{S}_a &= \mathrm{Softmax}\left(\mathbf{Q}_a \otimes \mathbf{K}_a{}^{\mathrm{T}}\right),\\ \mathbf{S}_v &= \mathrm{Softmax}\left(\mathbf{Q}_v \otimes \mathbf{K}_v{}^{\mathrm{T}}\right),\end{aligned} \tag{10}$$

式中，⊗表示矩阵乘法。通过上述矩阵乘法，可以有效捕获特征之间的远距离依赖关系和内部相关性，从而获得全局特征。

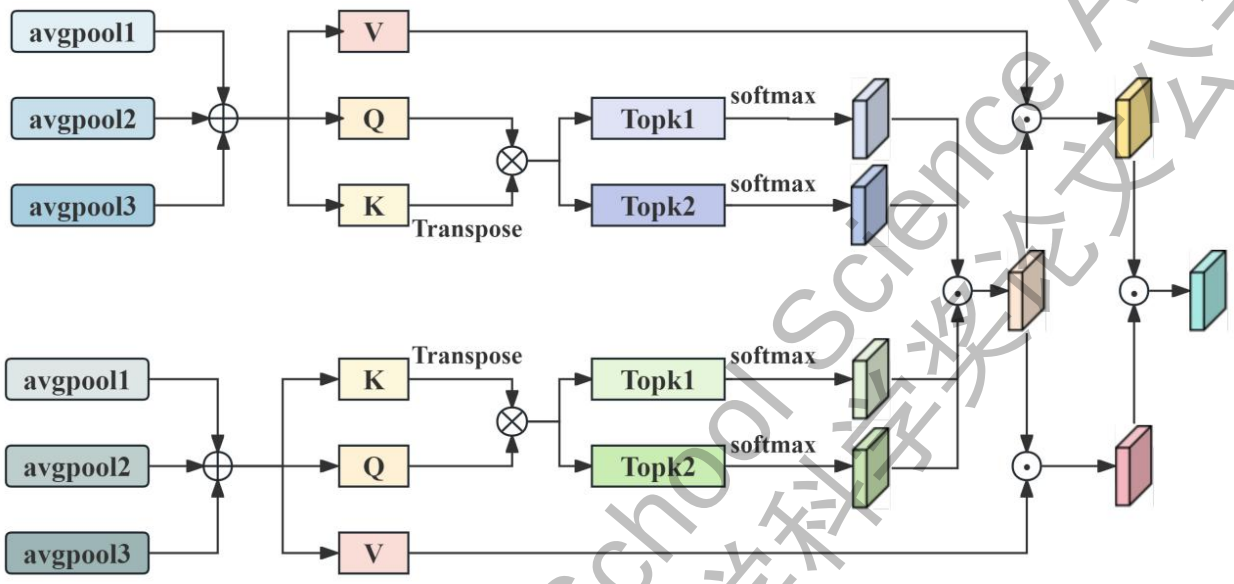

图 5 多尺度稀疏交叉注意力融合模块的结构图

为了减少无关信息的干扰，该模型在每个模态引入了两次 Topk 稀疏操作，稀疏操作的原理如图 6 所示。该操作通过保留每行中相关性最高的 k 比例元素，将低相关性的元素置为负无穷，使得在后续的 softmax 计算中，这些无关元素会被转化为 0，从而有效抑制干扰信息的影响。该过程可以定义为：

$$\begin{aligned}\mathbf{S}_a &= \mathrm{Softmax}\left(\mathrm{Topk}_{\mathrm{n}}\left(\mathbf{S}_a\right)\right),\\ \mathbf{S}_v &= \mathrm{Softmax}\left(\mathrm{Topk}_{\mathrm{n}}\left(\mathbf{S}_v\right)\right),\end{aligned} \tag{11}$$

式中 n 表示稀疏操作的次数，本文中采用了 2 次稀疏操作。这里对稀疏操作后的注意力图应用可学习参数和，以自适应调整两种稀疏程度的比例：

$$\begin{aligned}\mathbf{S}_a &= \lambda \cdot \mathrm{Softmax}\left(\mathrm{Topk}_1\left(\mathbf{S}_a\right)\right) + \eta \cdot \mathrm{Softmax}\left(\mathrm{Topk}_2\left(\mathbf{S}_a\right)\right),\\ \mathbf{S}_v &= \lambda \cdot \mathrm{Softmax}\left(\mathrm{Topk}_1\left(\mathbf{S}_v\right)\right) + \eta \cdot \mathrm{Softmax}\left(\mathrm{Topk}_2\left(\mathbf{S}_v\right)\right),\end{aligned} \tag{12}$$

式中，$\mathbf{S}_a$ 和 $\mathbf{S}_v$ 分别表示经过两次稀疏操作的音频特征和视觉特征，λ 和 η 分别设置为 0.5 和 0.7。

由于音频信息和视觉信息包含的特征并不相同，且具有互补性，将二者进行融合有助于提高语音识别的准确性。通过高层特征交叉融合得到的注意图可以表示如下:

$$\mathbf{S}_{av} = \mathbf{S}_a \odot \mathbf{S}_v. \tag{13}$$

最后，将该注意力图分别和音频和视觉特征图进行加权融合，随后对加权以后的特征图再次进行交叉融合，得到最终的联合注意力图，公式如下：

$$\begin{aligned}\mathbf{Att}_a &= \mathbf{S}_{av} \odot \mathbf{V}_a, \\ \mathbf{Att}_v &= \mathbf{S}_{av} \odot \mathbf{V}_v, \\ \mathbf{Att} &= \mathbf{Att}_a \odot \mathbf{Att}_v,\end{aligned} \tag{14}$$

式中,$\odot$ 表示 Hadamard 乘法,$\mathbf{Att}_a$ 和 $\mathbf{Att}_v$ 分别表示音频特征和视觉特征的加权特征图。**Att** 表示联合注意力图。综上所述，该模块可以减少无关信息的影响，更有效的融合音频模态信息和视觉模态信息。

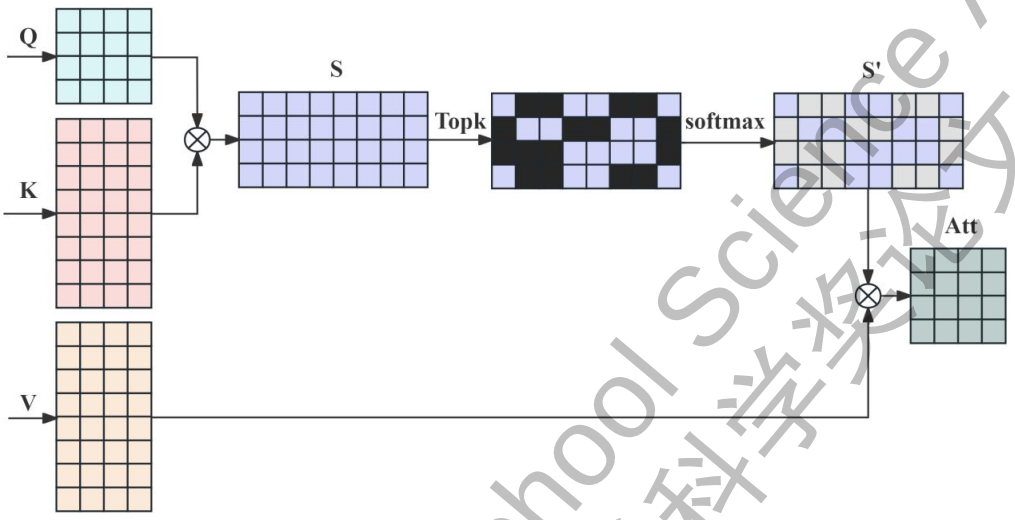

图 6 稀疏操作原理

3.5 损失函数

在 CTC 中，输入序列与输出序列之间无需依赖人工对齐，而是通过模型自动完成对齐过程。CTC 的核心思想是在给定输入序列的条件下，计算所有可能输出序列的概率，以得到特定输出的最终概率。为了解决连续重复字符的识别问题，CTC 在解码过程中引入了一个特殊的空白符号 ε。借助该机制，模型在预测阶段能够先生成包含重复字符与空白符的中间序列，再通过合并重复字符并去除空白符的操作，得到期望的输出结果。该方法确保了输入序列与输出序列在时间映射上的单调性，同时允许输入与输出之间存在多对一对应关系，因此在大多数情况下，输入序列的长度通常大于输出序列。

CTC 中的一个核心前提是，在任意时刻，输入与输出条件独立。基于这一假设，可以对给定的输入序列 $U=\left[u_1,u_2,u_3,\cdots\cdots,u_Y\right]$ 和目标输出序列 $Z=\left[z_1,z_2,z_3,\cdots\cdots,z_L\right]$ 来计算其条件概率。这里的 Y 表示输入长度，L 表示目标长度，CTC 损失可以定义为：

$$L_{CTC}=P_{CTC}\left(z|u\right)\approx\sum_{t=o}^{Y}P\left(z_t|u\right) \tag{15}$$

由于本文还使用了负载均衡损失，所以本文的总训练损失是负载均衡损失和 CTC 损失的加权和：

$$loss=L_{CTC}+\rho L_b \tag{16}$$

式中，ρ 被用作控制路由崩溃风险的平衡因子。在本研究中，负载均衡损失权重 ρ 取值为 0.01。

4. 实验设置与结果分析

4.1 数据集介绍

当前可用的 AVSR 数据集（如 VoxCeleb 和 LRS3）主要集中于英语语料。实际上，现有的非英语 AVSR 数据集不仅时长有限，而且大多在受控实验环境下采集，这在一定程度上限制了模型在跨语言及多口音场景中的泛化能力。

鉴于上述限制，本研究采用了由 Narayan 等人构建的最新 ViSpeR 多语言数据集[41]。该数据集利用 YouTube 视频平台，通过搜索 API 获取视频内容，并结合关键词（如“采访”“讨论”）以及目标语言（如中文、西班牙语、法语和阿拉伯语）进行筛选，以获得最相关的视频。同时，通过比对 YouTube 视频的 ID，有效避免了与现有多语言 AVSR 数据集（如 AV-Speech 和 VoxCeleb2）内容的重叠。

ViSpeR 数据集主要由 TED 演讲视频和真实生活场景（如街头采访、日常讨论、非官方会议）的视频构成。数据集涵盖四种语言，其中训练数据中中文约 787 小时，西班牙语约 794 小时，法语约 872 小时，阿拉伯语约 1200 小时，并且真实生活场景视频的数量远超 TED 演讲视频，具体数据见表 1。测试数据同样包括 TED 演讲视频和真实生活场景视频，详见表 2。

表 1 ViSpeR 数据集训练数据的比较。ViSpeR 数据集比现有的非英语数据集数量更庞大，涵盖了之前没有的中文数据集。TED 表示演讲视频，WILD 表示真实生活场景视频，ALL 表示所有视频。其中视频时长以 h 为单位，括号里的表示视频剪辑的数量。

Dataset	Chinese	Spanish	French	Arabic
AVSpeech	\	270	122	\
VoxCeleb2	\	42	124	\
MuAVIC	\	178	176	16
ViSpeR(TED)	129 (143k)	207 (151k)	192 (160k)	49 (48k)
ViSpeR(WILD)	658 (593k)	587 (383k)	680 (481k)	1152 (1.01M)
ViSpeR(ALL)	787 (736k)	794 (534k)	872 (641k)	1200 (1.06M)

表 2 ViSpeR 数据集测试集比较。其中视频时长以 h 为单位，括号里的内容表示视频剪辑的数量。

Dataset	Chinese	Spanish	French	Arabic
ViSpeR(TED)	0.37 (387)	0.65 (429)	0.31 (221)	0.26 (208)
ViSpeR(WILD)	3.30 (2989)	1.21 (828)	2.01 (1442)	1.19 (745)
ViSpeR(ALL)	3.67 (3376)	1.86 (1257)	2.32 (1663)	1.45(953)

4.2 评估指标

在 AVSR 中，错误率是衡量系统性能的核心指标之一，其基本思路是通过比较模型输出的识别结果与人工标注的参考文本，计算出识别过程中的替换、删除和插入情况。错误率能够清晰地表征系统在语音内容还原方面的精度，同时也是评价语音识别技术成熟度和

实用价值的重要依据。在本研究中，主要选用词错误率（Word Error Rate，*WER*）和字符错误率（Character Error Rate，*CER*）作为评价指标[42]。

WER 适用于衡量词层面的识别精度，是衡量 AVSR 模型的性能的较为常用的指标。该指标通过统计识别结果与参考文本之间的不一致部分来计算，其中包括插入、删除和替换三类错误。*WER* 的计算公式如下式所示：

$$WER = \frac{S + D + I}{N_{words}} \tag{17}$$

式中，S 表示替换次数，D 表示删除次数，I 表示插入次数，N_{words} 表示参考文本中的总词数。*WER* 数值越低，表示模型在词层面上的识别效果越佳。

CER 是与 *WER* 类似的一种评估指标，但能更好地反映在字符级别上的识别效果，*CER* 的计算公式如下式所示：

$$CER = \frac{S + D + I}{N_{chars}} \tag{18}$$

式中，N_{chars} 为参考文本中的总字符数。*CER* 更关注识别结果在细粒度上的偏差，它直接统计参考文本与识别输出之间的字符差异。

4.3 实验设置

4.3.1 实验环境

本研究实验在配备 64 位 Windows10 操作系统的服务器上进行。实验环境由 Anaconda 管理，使用 Python3.8.20 开发，以保证环境的独立性与可复现性。深度学习框架采用 PyTorch 2.1.2，并利用八块 16GB NVIDIA GeForce RTX 4060 Ti GPU 加速模型训练，同时结合 CUDA 11.8 对计算进行优化以提升效率。具体的软硬件环境配置如表 3 所示。

表 3 实验的硬件和软件配置

Category	Name	Version
Hardware	CPU	3.40 GHz Intel(R) Core(TM) i7-14700KF
	GPU	16 GB NVIDIA GeForce RTX 4060 Ti × 8
Operating System	Windows system	Windows 10 Professional
Software	Python	3.8.20
	Pytorch	2.1.2
	Numpy	1.23.5
	Sentencepiece	0.1.97
	hydra-core	1.1.1
	omegaconf	2.1.1

4.3.2 训练配置

模型在多语言环境下进行训练，其中编码器由 6 层堆叠而成，解码器包含 4 层。隐藏层维度、MLP 分别设置为 768 和 3072。自注意力头的数量设置为 12。其中汇表规模为 21,000，

Unigram 分词器针对所有语言联合训练。训练时，采用 AdamW 优化器，epoch 设置为 140，学习率设为 0.001，权重衰减系数为 0.1。

4.3.3 对比方法

为了评估所提模型的有效性，本研究将一些当前最先进的方法和我们的方法在 ViSpeR 数据集上进行了对比实验。这些方法主要包括 TM-seq2seq[43]，AV Transformer[35]，AVEC[33]，MLCA-AVSR[36]和 DCIM-AVSR[44]：

- **TM-seq2seq：**该方法以自注意力模块为核心，属于一种端到端训练框架。它借助预训练模型完成特征的提取与保存，并在此基础上利用自注意力机制对各模态特征进行处理，从而生成相应的上下文表示。
- **AV Transformer：**该 AVSR 方法采用跨模态注意力权重融合机制，不依赖预训练策略，而是直接利用视听数据进行建模。该模型同时引入了基于视觉表征的回归动作单元作为 Transformer 框架中的辅助损失，从而提升训练效果。
- **AVEC：**该研究通过结合音频与视觉模态，在 Conformer-CTC 架构中引入了中间 CTC 损失、CTC 残差机制及补丁注意力，以提升日常交流环境中的的语音识别性能。
- **MLCA-AVSR：**该方法将交叉注意力模块嵌入音频/视觉编码器的中间层，将每个中间层融合模块的输出直接作为下一层编码器的输入，同时引入 Inter-CTC 损失对中间层融合特征进行多阶段监督，实现了跨模态特征在学习全流程的递进式交互。
- **DCIM-AVSR：**该方法将将音频设为主要模态、视觉设为辅助模态，提出双 Conformer 交互模块（DCIM），将跨模态融合分布式拆解到多个层级，实现音视频特征从低到高逐步互补，并且引入“信息补全 + 净化”双功能适配器，将单模态有用特征注入另一模态互补信息，以解决音视频时序差异问题，有效提升了语音识别效果。

4.4 对比实验

在这一部分，我们使用 ViSpeR 数据集分别在视觉模态，音频模态和视听模态进行了对比实验。对于中文（Chinese）数据，我们使用 *CER* 进行评估。对于西班牙语（Spanish），法语（French），和阿拉伯语（Arabic）数据，我们使用 *WER* 和 *CER* 进行评估。表 4-5 分别给出了所有对比方法在 ViSpeR 数据集上的 *WER* 和 *CER*。

从表 4 中可以看出，本文提出的的 MAF-AVSR 模型在所有模态下的 *WER* 均为最低，取得了最优的语音识别效果。在视觉模态中，本文的方法在三种语言上的 *WER* 分别达到了 27.35%，29.64%，34.74%，显著低于对比方法。同样，在音频模态中，我们的方法在三种语言上的 *WER* 分别为了 8.91%，9.43%，10.83%，同样实现了最佳性能。这主要得益于模型首先通过卷积提取到了局部特征，随后利用 MoH 模块将每个注意力头视为一个专家，各个注意力头协同配合，可以更加充分的提取上下文特征，增强了对全局和局部视觉特征和音频特征的提取。相比之下，TM-seq2seq 采用的普通注意力机制无法实现注意力头之间的协作，因此其 *WER* 明显偏高。AVEC 虽然引入补丁注意力降低了复杂度，但存在局部时序

信息丢失的问题，导致性能变差。在视听模态下，本文对视觉特征和音频特征进行了融合，在三种语言上分别实现了 2.72%，3.15%，4.47%的*WER*，远远优于其余的方法。这是因为 MSSCA 模块可以抑制干扰信息，增强对重要特征的关注度，从而增强视听特征之间的交互。AV-Transformer 虽然使用了跨模态注意力权重融合机制增强融合能力，但仍受干扰信息影响，性能受到限制。MLCA-AVSR 采用交叉注意力将每个中间层融合模块的输出直接作为下一层编码器的输入，但是这样的设计在层数较多时会产生过多冗余特征，进而造成语音识别性能衰减。DCIM-AVSR 将视觉模态仅作为辅助模态，忽视了部分视觉特征，因此识别效果不及本文模型。从表 5 中可以看出，所提 MAF-AVSR 模型的*CER*在所有模态中都是最低的。在视听模态中，所提模型在四种语言上分别达到了 3.45%，1.59%，2.14%，2.95%的*CER*，这进一步证明了 MAF-AVSR 模型卓越的特征提取能力和跨模态交互能力。

此外，在单模态视觉输入条件下，所有对比模型的错误率都相对较高，说明仅依靠唇部运动信息进行语音识别仍存在局限性。在单模态音频输入条件下，错误率明显下降，表明音频信号提供了更为丰富、更具判别性的语音特征，有助于降低识别的错误率。相比之下，当采用视觉特征和音频特征同时作为输入时，模型错误率进一步降低，这表明视觉与音频特征之间存在显著的互补效应，从而有助于提升语音识别的鲁棒性。这充分证明了视听融合在提升语音识别精度方面的显著优势，也验证了本文所提出的 MAF-AVSR 模型的优越性。

表 4 ViSpeR 数据集在所有对比方法上的 WER 对比

Modality	**Method**	**WER across Different Languages(%)**		
		Spanish	**French**	**Arabic**
VSR	TM-seq2seq	51.21	53.92	63.53
	AV-Transformer	52.72	51.81	55.84
	AVEC	43.43	42.69	45.47
	MLCA-AVSR	41.95	44.39	41.38
	DCIM-AVSR	40.51	38.73	43.91
	MAF-AVSR	**27.35**	**29.64**	**34.74**
ASR	TM-seq2seq	14.87	15.12	17.75
	AV-Transformer	14.94	16.70	19.67
	AVEC	12.60	14.25	15.81
	MLCA-AVSR	13.77	12.80	16.97
	DCIM-AVSR	12.27	11.73	14.35
	MAF-AVSR	**8.91**	**9.43**	**10.83**
AVSR	TM-seq2seq	9.52	8.46	10.73
	AV-Transformer	8.55	8.81	9.44
	AVEC	8.17	7.39	8.74
	MLCA-AVSR	7.74	7.10	7.68
	DCIM-AVSR	7.90	8.29	8.51
	MAF-AVSR	**2.72**	**3.15**	**4.47**

表 5 ViSpeR 数据集在所有对比方法上的 CER 对比

Modality	Method	CER across Different Languages(%)			
		Chinese	Spanish	French	Arabic
VSR	TM-seq2seq	55.87	42.47	45.59	56.65
	AV-Transformer	51.26	41.93	40.16	47.71
	AVEC	41.70	38.76	35.92	40.47
	MLCA-AVSR	43.18	34.4	37.93	35.39
	DCIM-AVSR	39.84	35.62	34.45	36.18
	MAF-AVSR	**31.52**	**25.54**	**28.73**	**30.17**
ASR	TM-seq2seq	17.55	12.32	13.67	13.81
	AV-Transformer	16.78	12.29	14.42	15.63
	AVEC	13.89	10.49	11.18	12.74
	MLCA-AVSR	12.12	10.65	10.97	11.39
	DCIM-AVSR	14.93	9.61	10.05	11.78
	MAF-AVSR	**10.96**	**7.62**	**7.46**	**8.17**
AVSR	TM-seq2seq	8.11	6.54	5.15	6.19
	AV-Transformer	7.28	6.19	5.45	5.24
	AVEC	6.48	5.28	6.73	4.76
	MLCA-AVSR	7.36	4.37	4.16	5.19
	DCIM-AVSR	6.55	5.92	5.65	4.87
	MAF-AVSR	**3.45**	**1.59**	**2.14**	**2.95**

4.5 噪声环境下的有效性评估

为了验证所提 MAF-AVSR 模型在噪声环境下的有效性，本文在音频模态和视听模态下引入 Babble 噪声进行了对比实验。这里采用了不同的信噪比（Signal-to-Noise Ratio，SNR）进行该实验，并采用 *CER* 作为评估指标。测试结果如表 6 所示。

Babble 噪声是一类典型的非平稳背景噪声，由多人同时讲话的声音叠加而成，常见于餐厅、课堂、车站等嘈杂环境中[45]。由于其在时域和频域上均具有高度随机性和复杂性，Babble 噪声会对语音信号造成显著干扰，给语音识别任务带来较大挑战。同时由于 Babble 噪声能够较为真实地模拟日常生活中嘈杂人声环境下的听觉条件，因而更贴近实际应用需求，常被作为典型的测试场景加以使用。

表 6 ViSpeR 数据集在 Babble 噪声下的 CER 对比

Modality	ViSpeR	-10dB	-5dB	0dB	5dB	Clean
ASR	Chinese	97.29	89.93	58.85	18.74	**10.96**
	Spanish	91.17	88.71	40.75	8.59	**7.62**
	French	92.89	86.68	36.29	10.61	**7.46**
	Arabic	94.48	87.74	38.17	13.82	**8.17**
AVSR	Chinese	89.73	78.81	24.51	9.83	**3.45**
	Spanish	82.62	71.83	15.26	5.74	**1.59**
	French	83.45	77.19	13.87	6.37	**2.14**
	Arabic	82.16	75.96	19.93	6.54	**2.95**

从表 6 可以看出，在 SNR 较小的情况下，模型的*CER*较高，语音识别能力较差。随着 SNR 的增大，模型的*CER*越来越小，性能提升比较明显。此外，无论是在多大的 SNR 下，视听模态都比音频模态的*CER*更小，这证明了在含噪声情况下将视觉和音频特征进行融合的有效性。为了更清晰的显示不同 SNR 下性能的变化，我们将上述实验结果进行了可视化，如图 7 所示。

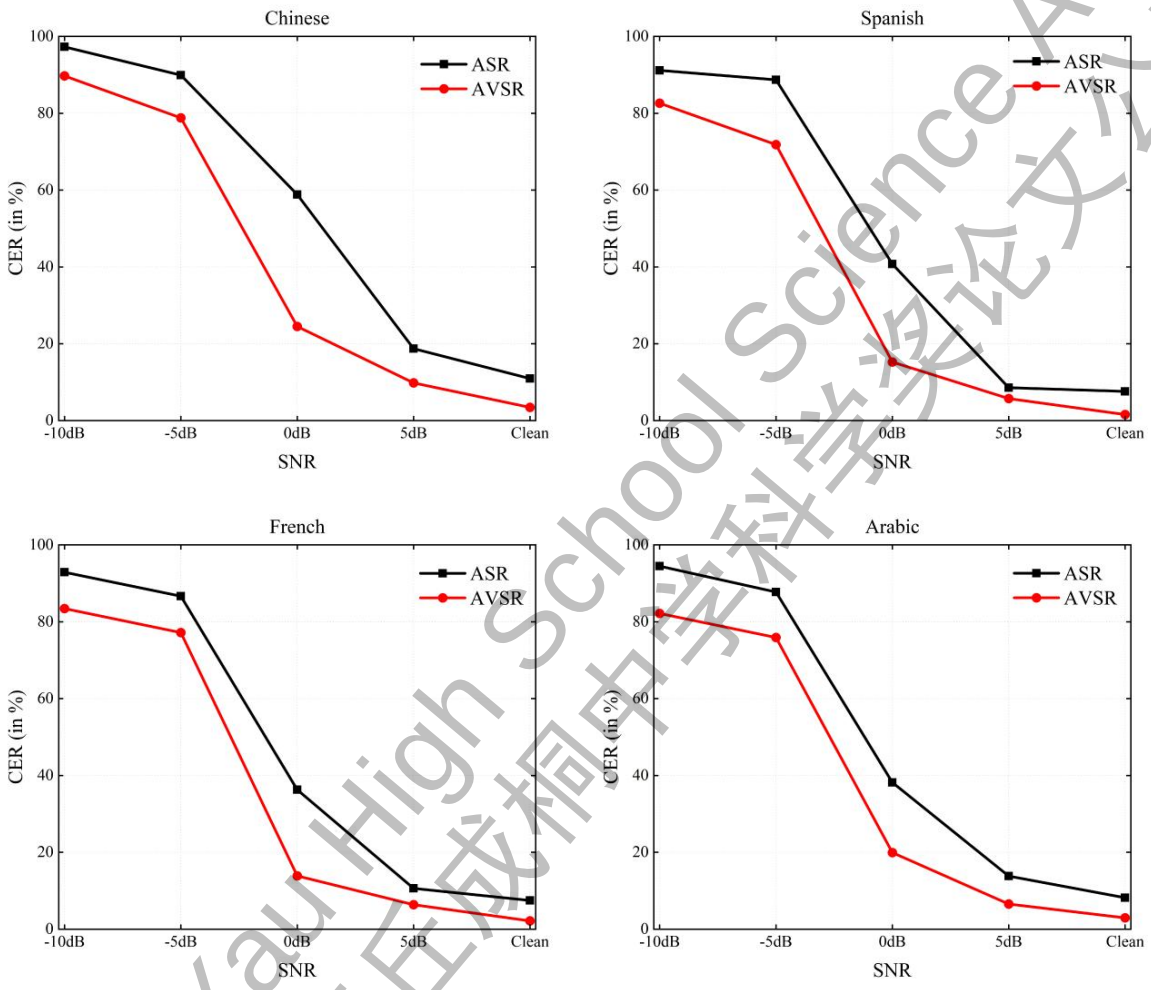

图 7 不同 SNR 下的 CER 对比实验

4.6 消融实验

4.6.1 编码器模块的消融实验

本文使用 MoH 模块作为编码器进行特征提取，该模块可以在保留最重要信息的同时实现不同注意力头之间的分工协作，从而增强模型的上下文特征提取能力。为了验证 MoH 模块的有效性，本文在保证参数不变的情况下，将编码器分别替换为 Conformer，Branchformer，和 E-Branchformer，在视听模态下评估模型的*CER*。实验结果如表 6 所示。从表 6 中可以明显看出，E-Branchformer 编码器的性能明显优于另外两个编码器，与 Branchformer 相比，E-Branchformer 在四种语言上的的*CER*分别低了 2.48%，1.35%，6.11%，2.77%，这得益于 E-Branchformer 更大的感受野。相比之下，我们的 MoH 编码器实现了最低的*CER*，这是因

为 MoH 模块实现了所有注意力头的协通工作，有效增强模型的上下文建模能力，这也证明了 MoH 模块的有效性。

表 7 不同 Encoder 的 CER 对比实验

Modality	Encoder	CER across Different Languages (%)			
		Chinese	Spanish	French	Arabic
AVSR	Conformer	12.76	9.17	9.69	10.52
	Branchformer	10.35	8.25	11.73	9.51
	E-Branchformer	7.87	6.97	5.62	6.74
	MoH	3.45	1.59	2.14	2.95

4.6.2 融合模块的消融实验

为了更好地展示 MSSCA 融合模块的优势，我们对两种常见的融合方法进行对比，其一是将音频与视觉编码器的输出进行直接相加；其二是将两者的输出在特征维度上拼接，并利用 MLP 完成融合。不同融合策略下的实验结果如表 8 所示。显然可以看出，直接相加的方法效果最差，MLP 融合方法由于基本较难实现视觉特征和音频特征的交互，所以效果也不如所提方法。MSSCA 融合模块由于不仅使用稀疏策略增强了对关键特征的提取，而且利用交叉注意力增强了视觉特征和音频特征的交互，实现了最好的语音识别效果。结果进一步验证了 MSSCA 融合模块在特征融合中的有效性。

表 8 不同融合模块的 CER 对比实验

Modality	Fusion	CER across Different Languages (%)			
		Chinese	Spanish	French	Arabic
AVSR	ADD	7.18	3.54	4.97	5.92
	MLP	5.71	2.16	2.97	3.88
	MSSCA	3.45	1.59	2.14	2.95

5. 总结与展望

本文提出了一种新型的 MAF-AVSR 模型，旨在解决多语言 AVSR 模型中在不同语言与多口音环境、嘈杂或远场语音以及多说话者干扰等复杂环境下特征提取能力不足及跨模态交互不充分的问题。受自然界群体决策机制的启发，模型在编码端引入 MoH 模型，将每个注意力头视为专家，并通过共享头与路由头的协同机制，提升了对视听特征的全局上下文建模能力。同时，MSSCA 融合模块结合稀疏策略与交叉注意力机制，有效抑制冗余信息、突出关键信息，并促进多模态特征的深层交互，从而显著增强了模型的表达能力与鲁棒性。在 ViSpeR 数据集上与当前最先进方法的对比实验证明了 MAF-AVSR 模型在复杂场景下的卓越性能。噪声环境下的测试进一步验证了多模态融合的优势，而消融实验则证明了 MoH 和 MSSCA 模块在整体性能提升中的关键作用。

总体而言，本研究不仅丰富了多语言 AVSR 数据集的应用，还为深度学习在多模态语音识别中的创新提供了新范式。未来工作可进一步探索模型在低资源语言或实时部署场景下的优化，以推动 AVSR 技术在听力障碍辅助领域的实际应用。

参考文献

[1] 刘恒, 林玮. 联合训练方法在带噪语音识别中的优化研究[J]. 电子器件, 2025, 48(04): 814-820.

[2] 王华朋, 冯嘉琪. 基于深度学习的语音增强方法综述[J]. 科学技术与工程, 2025, 25(20): 8331-8346.

[3] 张晋宁. 基于神经网络的视觉语音识别系统[J]. 电声技术, 2023, 47(11): 101-104.

[4] 赵小芬, 彭朋. 基于多模态视听融合的 Transformer 语音识别算法研究[J]. 传感器与微系统, 2025, 44(02): 48-52.

[5] 吴兰, 杨攀, 李斌全, 王涵. 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(01): 52-60.

[6] 马晗, 唐柔冰, 张义, 张巧灵. 语音识别研究综述[J]. 计算机系统应用, 2022, 31(01): 1-10.

[7] Davis KH, Biddulph R, Balashek S. Automatic recognition of spoken digits[J]. The Journal of the Acoustical Society of America. 1952, 24(6): 637-642.

[8] Forgie JW, Forgie CD. Results obtained from a vowel recognition computer program[J]. The Journal of the Acoustical Society of America. 1959, 31(11): 1480-1489.

[9] Olson H, Belar H. Time compensation for speed of talking in speech recognition machines[J]. IRE Transactions on Audio. 2003, 29(3): 87-90.

[10] Baum LE, Eagon JA. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology[J], Bulletin of the American Mathematical Society. 1967, 73(3): 360-363.

[11] Lee KF. On large-vocabulary speaker-independent continuous speech recognition[J]. Speech Communication. 1988, 7(4): 375-379.

[12] Han W, Zhang Z, Zhang Y, Yu J, Chiu CC, Qin J, Gulati A, Pang R, Wu Y. Contextnet: Improving convolutional neural networks for automatic speech recognition with global context[C]. In Proceedings of the International Speech Conference. 2020: 3610-3614.

[13] Dahl GE, Yu D, Deng L, Acero A. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition[J]. IEEE Transactions on Audio, Speech, and Language Processing. 2011, 20(1): 30-42.

[14] Graves A, Fernández S, Gomez F, Schmidhuber J. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks[C]. In Proceedings of the 23rd International Conference on Machine Learning, 2006: 369-376.

[15] Graves A. Sequence transduction with recurrent neural networks[J]. arXiv preprint arXiv:1211.3711, 2012.

[16] Eyben F, Wöllmer M, Schuller B, Graves A. From speech to letters-using a novel neural network architecture for grapheme based asr[C]. In IEEE Workshop on Automatic Speech Recognition. 2009: 376-380.

[17] Yao Z, Kang W, Yang X, Kuang F, Guo L, Zhu H, Jin Z, Li Z, Lin L, Povey D. CR-CTC: Consistency regularization on CTC for improved speech recognition[J]. arXiv preprint arXiv:2410.05101, 2024.

[18] 冯小丹, 云利军, 高海峰, 孟凤菊. 机器视觉注意力机制研究综述[J]. 云南民族大学学报(自然科学版), 2025, 34(04): 453-463.

[19] Zhang Q, Lu H, Sak H, Tripathi A, McDermott E, Koo S, Kumar S. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss[J]. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2020: 7829-7833.

[20] Gulati A, Qin J, Chiu CC, Parmar N, Zhang Y, Yu J, Han W, Wang S, Zhang Z, Wu Y, Pang R. Conformer: Convolution-augmented transformer for speech recognition[J]. arXiv preprint arXiv:2005.08100, 2020.
[21] Peng Y, Dalmia S, Lane I, Watanabe S. Branchformer: Parallel mlp-attention architectures to capture local and global context for speech recognition and understanding[C]. In Proceedings of the International Conference on Machine Learning, 2022: 17627-17643.
[22] Kim K, Wu F, Peng Y, Pan J, Sridhar P, Han KJ, Watanabe S. E-branchformer: Branchformer with enhanced merging for speech recognition[C]. In Proceedings of the IEEE Spoken Language Technology Workshop, 2023: 84-91.
[23] Chiou GI, Hwang JN. Lipreading from color video[J]. IEEE Transactions on Image Processing, 1997, 6(8): 1192-1195.
[24] Xue H, Yang Q, Chen S. SVM: Support vector machines[D]. In the Top Ten Algorithms in Data Mining. 2009: 51-74.
[25] Assael YM, Shillingford B, Whiteson S, De Freitas N. Lipnet: End-to-end sentence-level lipreading[J]. arXiv preprint arXiv:1611.01599, 2016.
[26] Stafylakis T, Tzimiropoulos G. Combining residual networks with LSTMs for lipreading. arXiv preprint arXiv:1703.04105, 2017.
[27] Martinez B, Ma P, Petridis S, Pantic M. Lipreading using temporal convolutional networks[C]. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2020: 6319-6323.
[28] Prajwal KR, Afouras T, Zisserman A. Sub-word level lip reading with visual attention[C]. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, 2022: 5162-5172.
[29] 马金林, 刘宇灏, 马自萍, 郭兆伟, 吕鑫. 解耦同类自知识蒸馏的轻量化唇语识别方法[J]. 北京航空航天大学学报, 2024, 50(12): 3709-3719.
[30] Baaloul A, Benblidia N, Reguieg FZ, Bouakkaz M, Felouat H. An arabic visual speech recognition framework with CNN and vision transformers for lipreading[J]. Multimedia Tools and Applications, 2024, 83(27): 69989-70023.
[31] Dupont S, Luettin J. Audio-visual speech modeling for continuous speech recognition[J]. IEEE Transactions on Multimedia. 2000, 2(3): 141-151.
[32] Ma P, Petridis S, Pantic M. End-to-end audio-visual speech recognition with conformers[C]. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2021: 7613-7617.
[33] Burchi M, Timofte R. Audio-visual efficient conformer for robust speech recognition[C]. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023: 2258-2267.
[34] 车娜, 朱奕明, 赵剑, 孙磊, 史丽娟, 曾现伟. 基于联结主义的视听语音识别方法[J]. 吉林大学学报(工学版), 2024, 54(10): 2984-2993.
[35] 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[J]. arXiv preprint arXiv:2005.09297, 2020.
[36] Wang H, Guo P, Zhou P, Xie L. Mlca-avsr: Multi-layer cross attention fusion based audio-visual speech recognition[C]. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2024: 8150-8154.

[37] Ni R, Jiang H, Zhou L, Lu Y. Lip recognition based on Bi-GRU with multi-head self-attention[C]. In Proceedings of the IEEE International Conference on Artificial Intelligence Applications and Innovations, 2024: 99-110.
[38] Masoudnia S, Ebrahimpour R. Mixture of experts: a literature survey[J]. Artificial Intelligence Review. 2014, 42(2): 275-293.
[39] Zhang D, Song J, Bi Z, Yuan Y, Wang T, Yeong J, Hao J. Mixture of experts in large language models[J]. arXiv preprint arXiv:2507.11181, 2025.
[40] Jin P, Zhu B, Yuan L, Yan S. Moh: Multi-head attention as mixture-of-head attention[J]. arXiv preprint arXiv:2410.11842, 2024.
[41] Narayan S, Djilali YA, Singh A, Bihan EL, Hacid H. ViSpeR: Multilingual audio-visual speech recognition[J]. arXiv preprint arXiv:2406.00038, 2024.
[42] Song Q, Sun B, Li S. Multimodal sparse transformer network for audio-visual speech recognition[J]. IEEE Transactions on Neural Networks and Learning Systems. 2022, 34(12): 10028-10038.
[43] Afouras T, Chung JS, Senior A, Vinyals O, Zisserman A. Deep audio-visual speech recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2018, 44(12): 8717-8727.
[44] Wang X, Jiang H, Huang H, Fang Y, Xu M, Wang Q. DCIM-AVSR: Efficient audio-visual speech recognition via dual conformer interaction module[C]. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 2025: 1-5.
[45] Krishnamurthy N, Hansen JH. Babble noise: modeling, analysis, and applications[J]. IEEE Transactions on Audio, Speech, and Language Processing. 2009, 17(7): 1394-1407.

致 谢

1. 论文选题来源，研究背景，指导老师与参赛学生的关系，在论文写作过程中所起的作用，指导是否有偿；以及他人协助完成的研究成果

本研究最初的选题是因为我本人从小有听力障碍而产生的想法。根据在网上查找有关的资料，了解到截至目前全球已有超过4.3亿人受到严重听力障碍的影响，其中大多数人都像我一样需要专业助听设备辅助。在我的平时生活中，戴助听器在嘈杂的环境中，甚至是刮稍微大点的风的时候，助听器处的嗖嗖音很大时，很难听清别人的说话，给我日常生活中也带来了很大的不便。那时候我就想，要是我有能读懂别人的唇语的“特异功能”就好了。此外，我发现谷歌翻译眼镜是一款外观类似普通眼镜的智能穿戴设备，通过智能语音交互系统将语音转换为对应文本显示，一定程度上可以协助听障人士识别语音交互内容。但是在嘈杂噪声的复杂环境中，也会导致其性能会大幅度下降。这让我萌生了一个想法：是不是可以把唇语识别的功能和谷歌翻译眼镜中的语音识别的功能融合在一起，做一个类似于谷歌眼镜的基于唇语视觉与听觉融合的助听眼镜呢？这样一来，即使在嘈杂的环境中，无法有效识别别人语音信息时，也可以通过识别说话人的唇语，也能准确地“听清”别人的语音信息（这一想法，我也在指导老师—梁联晖老师的指导和帮助下，申请了欧盟发达国家发明专利：Hearing Aid System Based on The Fusion of Vision and Hearing，PMT330CN25124，已授权。同时也正在准备申请国内发明专利）。因此，探索将视觉与音频特征作为互补信息进行融合的视听语音识别技术，对提升听障人群的日常交流能力具有重要意义。我希望设计出一种基于视觉与听觉融合的助听系统，通过自己的研究，结合语音识别与唇语识别提高整体识别的准确性，为听障人士提供更好的助听服务。

本研究的主要指导老师是广西大学的梁联晖老师，他是我在一次课外活动中认识的指导老师。在本研究的整个过程中，梁老师在了解到我的想法后，在有关资料的调研、深度学习网络框架的搭建、实验设计、论文撰写、欧盟发明专利的书写和申请等各个环节，梁老师都给予了我全面的指导。当我在科研受挫时（在刚入门深度学习，进行深度学习环境的配置总是报错。搭建模型总是达不到理想结果等时，让我感到比较失落），梁老师总是能耐心地帮助我梳理问题、分析问题，并及时地给出宝贵的建议和指导。在此，我衷心的感谢梁老师的无偿指导。正是因为有了梁老师的悉心指导，不仅帮助我顺利完成了本研究，让我离我最终的目标又近了一步，也让我在学术能力和思维方式上得到了极大的提升。同时，也让我懂得了遇到科研难题时，如何去思考问题、分析问题、一步一步地解决问题，极大地锻炼了解决问题的能力。

本研究成果均为本人在导师指导下独立完成。在此，也要特别感谢在Github上无私分享和开源自己源代码的学者们，这极大地帮助了我的算法模型的代码实现，谢谢您们。

2. 本文局限与展望

本文局限与展望：由于受限于本人高中阶段所学的知识、专业技能水平及课外学习的时间，本人当前的工作主要借鉴了一些计算机领域的顶级会议论文中的idea和Github上的一些开源代码，进行了本项目的一个开发和定制。虽然本文所提方法与其他方法相比，具良好的语音识别性能，但在计算复杂度方面，还暂时无法满足我最终想在智能穿戴设备上应用的目标（最终想做一个类似于谷歌翻译眼镜的智能穿戴助听设备）。后续本人将继续利用课余时间研究一些轻量级的视听语音识别模型，也希望在大学期间能够读人工智能、计算机等方面的一些专业，争取早日把基于视觉（唇语）与听觉融合的智能穿戴助听设备研发出来，为听障人士提供更好的助听服务。

3. 指导老师简介

梁联晖，助理教授，广西大学电气工程学院硕士研究生导师，广西大学光谱智能信息处理实验室负责人，广西人工智能协会高级专家，American Journal of Remote Sensing期刊副主编，IEEE JSTARS客座编辑。一直从事人工智能、计算机视觉、图像分析和处理、嵌入式设计等方面的研究工作，担任IEEE TIP、IEEE TCSVT、IEEE TII、IEEE TGRS、IEEE JSTARS等20余个SCI期刊审稿人。近年来，先后主持国家自然科学基金青年基金项目、广西高校中青年教师基础能力提升项目等国家级/省部级项目5项，相关工作在国际权威期刊和会议上发表SCI/EI论文20余篇，申请/授权发明专利10余项。