arXiv:2603.01415v1 [eess.AS] 2 Mar 2026

中国科学技术大学 用于 CHiME-9 MCoRec 挑战赛的 C-NERCSLIP 系统

Ya Jiang[1], Ruoyu Wang[1], Jingxuan Zhang[2],Jun Du[1], Yi Han[1],Zihao Quan[1],Hang Chen[1],Yeran Yang[2], KongzhiZheng[2], Zhuo Chen[3], Yanhui Tu[4], Shutong Niu[5], Changfeng Xi[5], Mengzhi Wang[5], Zhongbin Wu[5], Jieru Chen[5], Henghui Zhi[5], Weiyi Shi[5],Shuhang Wu[5],Genshun Wan[5],Jia Pan[5],Jianqing Gao[5]*

[1]中国科学技术大学，中国合肥[2]陕西师范大学，中国西安 [3]莫斯科国立罗蒙诺索夫大学，俄罗斯莫斯科[4]安徽大学，中国合肥 [5]科大讯飞研究院，科大讯飞股份有限公司，中国合肥

摘要

本报告详细介绍了我们参加 CHiME-9 MCoRec 挑战赛的提交方案，任务是在室内社交场景中识别并聚类多个同时进行的自然对话。与围绕单一共同话题展开的传统会议不同，该场景包含多个并行对话，最多有八位说话人参与最多四场同时进行的对话，语音重叠率超过 90%。为应对这一挑战，我们提出了一种多模态级联系统，利用从同步的 360°视频中提取的各说话人视觉流，以及单通道音频。借助增强型音视频预训练模型，我们改进了处理流程中的三个组件：主动说话人检测（ASD）、音视频目标语音提取（AVTSE）和音视频语音识别（AVSR）。AVSR 模块还结合了 Whisper 和大语言模型技术，以提高转录准确率。我们的最佳单级联系统在开发集上的说话人词错误率（WER）为 32.44%。进一步使用 ROVER 融合不同前端和后端变体的输出后，我们将说话人 WER 降至 31.40%。值得注意的是，基于大语言模型的零样本对话聚类取得了 1.0 的说话人聚类 F1 分数，最终联合 ASR-聚类错误率（JACER）为 15.70%。

索引词— CHiME 挑战赛、活跃说话人检测、目标语音提取、语音识别、音视频预训练模型、大型语言模型

1. 系统描述

我们的系统独立处理CHiME-9 MCoRec 挑战赛[1,2]中的官方会话级裁剪中心轨道。首先，使用主动说话人检测（ASD）系统估计每个轨道视频的帧级说话人活动概率，并根据给定的起始和结束阈值生成分段。接下来，对每个分段的区域执行音视频目标语音提取（AVTSE），然后将提取的语音输入音视频语音识别（AVSR）系统。最后，会话聚类系统将说话人归入其对应的会话。

1.1. ASD

我们的 ASD 模块采用预训练音视频编码器（第 1.3 节中的基础版全脸编码器），以利用稳健的跨模态表征。该架构集成了一个用于目标检测的主要二分类器，以及一个用于强化模态特定判别的辅助视觉分类器，两者均采用二元交叉熵（BCE）损失进行优化。为在适应任务的同时保留预训练知识，我们采用两阶段迁移学习策略：首先冻结编码器并微调分类器，随后对整个网络进行端到端优化。

为应对 CHiME-9 复杂的声学场景，我们通过三种关键机制提升了模型的稳健性：**(1) 数据规范化**：我们引入了多种辅助数据集，包括 CHiME-9 官方数据集、AVA 数据集 [3], MSDWILD 数据集 [4],以及 M3SD 数据集 [5]。我们采用统一的预处理流程，裁剪无效片段，进行人脸几何标准化和重新跟踪以确保空间一致性，并利用标注提取说话人音频与人脸轨迹。随后将提取出的轨迹进一步切分为较短片段用于训练，最终获得约 200 小时的音视频 ASD 数据；**(2)在线增强**：通过强噪声注入模拟多说话人环境（2–4 名参与者）；**(3)时间平滑**：我们开发了后处理流程，通过应用 Softmax 归一化并合并零散的预测结果，确保说话人活动时间线符合物理规律。

1.2. AVTSE

CHiME-9 MCoRec 挑战具有极其严重的声学干扰，语音持续受到竞争性对话的压制。我们开发了四种 AVTSE

*通讯作者。

系统（官方基线中未包含），利用不同的自监督预训练模型和多模态架构分离重叠的语音流，并对各系统的轮次级掩码平均结果进行集成。

数据模拟：我们使用公开的视听语料库——LRS3 [6], VoxCeleb2 [7],和 AVSpeech [8]——构建合成训练集，并加入 DNS-Noise [9]进行增强，预处理方法详见第 1.3 节。我们在数据集内部进行模拟，并留出 10% 的说话人作为不重叠的干扰源。干扰配置包括 45% 的单一干扰源、45% 的双重干扰源和 10% 的纯噪声源，混合时的信噪比从 $[-10, 10]$ dB 范围内均匀采样。原始的同步视听片段作为干净目标语音及其对应的视觉线索。尽管我们最初对 CHiME-9 中高密度重叠（最多 7 名说话人）进行了建模，但实证评估表明，使用此类高复杂度数据训练会降低泛化能力。因此，我们构建了包含 2–3 名说话人的混合数据集：LRS3 为 335 小时，VoxCeleb2 为 1150 小时，AVSpeech 为 1421 小时；其中，最后者仅用于**系统 IV**。

系统配置：系统I建立了 AVTSE 基线，采用 BRAVEn [10] 编码器（一种针对 ASR/视觉语音识别任务在 LRS3 上微调的自监督框架），从原始音频混合信号和目标唇部序列中提取语义音素表征。这些特征与经压缩时序卷积网络（S-TCN）和残差长短期记忆网络处理的对数功率谱（LPS）拼接后，输入 Conformer 网络进行时序建模。随后，S-TCN 解码器预测理想比率掩码（IRM），并通过均方误差损失进行优化。**系统 II**在基线架构上增加了一个大型全脸编码器，该编码器在第 1.3 节所述的大规模配对视听数据上进行预训练，从而构建三模态框架，提取超越 BRAVEn 语音内容的显式说话人身份线索。**系统 III** 引入了联合 TSE-ASR 优化：重建语音由**AVSR S3** 编码器编码，重建特征与干净特征之间的余弦相似度损失则作为正则化项，间接提升重建语音的可懂度。**系统 IV**采用另一种双塔[11] 架构，使用并行的 ResNet-18 音频流和视觉流骨干网络，通过多尺度 TCN 进行融合，并经过预训练以关联音素与视素。融合后的多模态嵌入与 LPS 特征

采用**系统I**范式的特征，为基于 BRAVEn 的方法提供了互补的特征空间。

1.3. 音视频语音识别

为确保在复杂声学环境中实现稳健转录，我们开发了一个多样化的集成系统，涵盖多种 AVSR 框架，并以一种新颖的自监督预训练策略为核心，包括使用随机初始化的解码器进行微调（**AVSR S1**）、与 Whisper [12] 融合（**AVSRS2**），以及集成基于大语言模型（LLM）的解码器（**AVSR S3** 和 **AVSRS3-UASR**）。最终转录结果通过分层组合以下三个不同框架的

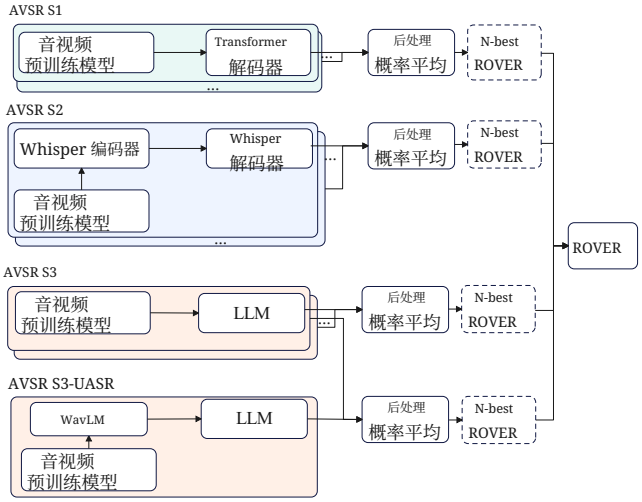


图 1。音视频语音识别方法概览。N-best ROVER仅用于我们的提交3。

输出，并采用后验概率平均和 ROVER [13] 方法生成。

数据准备： 除了 CHiME 数据集外，我们还使用了五个公开可用的音视频数据集进行音视频语音识别训练：LRS2 [14] 的预训练子集（195 小时）和训练子集（28 小时）；LRS3 [6] 的预训练子集（ 408 小时）以及训练 – 验证子集（ 30 小时）；遵循 Shi 等人的方法使用 VoxCeleb2 [7] 的英语部分，共计 1,326 小时（文本标签由 Whisper-large-v2 转录）；AVSpeech [8], ，按照 LRS2 和 LRS3 的流程进行分段和预处理，然后使用 Whisper-large-v2 进行语言识别和转录，仅保留英语部分，得到 1,436 小时的训练数据；以及 AVYT [15] 的 talking-face 子集（1,449 小时）。预处理后，在排除 AVYT 的情况下，我们共获得 3,530 小时的音视频数据，其中包含全脸和唇部感兴趣区域视频。加入 AVYT 后，进一步增加了 1,436 小时的仅含唇部视频数据，使音视频训练数据总量达到 4,966 小时。

音视频预训练：我们的骨干网络采用受 DistillAV [16], 启发的掩码预测目标，并进行了三项关键架构改进。首先，我们将 ConvNeXt [17]前端与增强型Conformer 模块 [18] 结合。其次，我们不采用梅尔频谱图作为目标，而是使用 BEST-RQ [19] 式损失：模型预测离散编码（词表大小为 8,192），这些编码由量化冻结的 Whisper 模型的隐藏表示得到。第三，根据我们此前的研究 [20, 21], ，为捕捉唇部区域之外更广泛的面部线索，我们采用双路视觉输入——全脸和唇部 ROI——分别由独立的前端处理，并沿特征维度拼接。

系统配置：AVSR S1将随机初始化的 Transformer 解码器连接到预训练编码器上。整个网络使用混合CTC/注意力损失进行微调。为缓解过拟合，我们对视频应用自适应时间掩蔽，对音频应用 SpecAugment。**音视频语音识别**

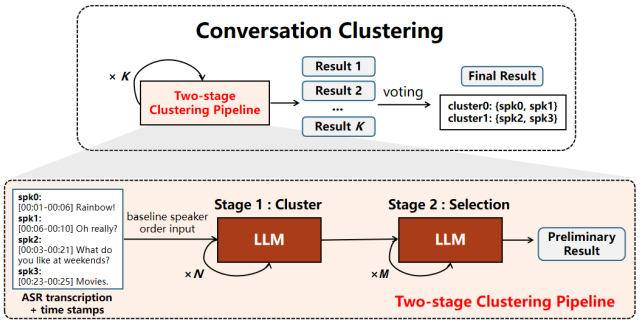


图 2. 我们的对话聚类方法概览。

S2 以 Whisper [12] 骨干网络为模型。我们使用预训练模型（仅视频分支）提取视觉表征，并通过 Flamingo 风格的 [22, 23] 交叉注意力层将其注入 Whisper 编码器。这些交叉注意力层插入每个奇数编号的 Transformer 块，使冻结的 Whisper 骨干网络能够显式关注视觉线索。**AVSRS1** 和**AVSR S2** 先在内部数据集上进行微调，然后再针对 CHiME-9 进行领域自适应。**AVSR S3** 将微调后的 S1 编码器的输出映射到通义千问-2.5 [24]的词元嵌入空间。我们采用分阶段训练方案：先训练适配器，再将注入大语言模型的低秩适配（LoRA） [25]参数纳入其中进行联合优化。此外，**AVSRS3-UASR** 采用我们先前的 UASR-LLM 框架 [26], ，通过 Flamingo 层将视觉特征注入 WavLM 编码器，再映射到大语言模型。模型采用两阶段策略进行训练，包括视觉注入预训练和 AVSR 微调。

模型集成： 我们采用两级融合策略。首先，通过后验概率平均 组合使用相同分词器的模型（系统内），以稳定预测结果。其次，使用 ROVER 汇总不同架构 **(AVSRS1–S3)** 及其 N-best 束搜索假设的文本级输出，生成最终提交方案。

1.4. 对话聚类

语义线索——例如主题连贯性和问答依赖关系——对于将说话人分组成不同的对话聚类至关重要。为利用这些语义特征，我们采用了大语言模型（LLM），包括 Qwen 2.5 (70B) [1] 和 DeepSeek R1 (671B) [2]；这些模型擅长上下文理解，能够直接根据自动语音识别转录文本推断说话人之间的关系。我们提出了一种稳健的两阶段集成流程，如图 2 所示。在最初的聚类阶段，我们分别提示大语言模型 N 次，以生成多种候选说话人—对话分配方案，输入包括自动语音识别转录文本、时间戳和基线说话人顺序。

[1]https://huggingface.co/通义千问 [2] https://huggingface.co/deepseek-ai

表 1. $\mathcal{D}_{\text{dev}}$ 和 $\mathcal{D}'_{\text{dev}}$上各提交方案的词错误率（%）结果

Set	提交方案 1	提交 2	提交 3
$\mathcal{D}_{\text{dev}}$	32.03	31.43	**31.39**
$\mathcal{D}'_{\text{dev}}$	29.98	29.36	**29.35**

在选择阶段，模型从候选结果中找出最优聚类结果。为过滤随机错误，我们重复这一选择 M 次，并选取出现频率最高的结果。最后，整个两阶段流程执行 K 次，随后进行投票融合。最终提交汇总了基线、通义千问和 DeepSeek 系统的预测。由于 6710亿参数模型的计算和时间成本高昂，DeepSeek-R1 的推理仅限于有限次数的迭代（$M = 5$, $K = 5$）。具体提示设计如下：（1）**阶段 1 提示 -** 请根据转录结果进行对话聚类。您可以根据对话内容、共同话题、对话结构以及话语之间的时间重叠等因素作出判断；（2）**阶段 2 提示 -** 请从候选聚类结果中选出最合理的一项。您的判断可以依据对话内容、共同话题、对话结构以及话语之间的时间重叠等因素。请从这些候选结果中选出最有可能的一项。

2. 结果

我们将经过精炼的开发集记为 $\mathcal{D}'_{\text{dev}}$，它是从原始音频集 $\mathcal{D}_{\text{dev}}$ 中剔除三个样本后得到的：第53场会话（说话人2）、第54场会话（说话人4）和第55场会话（说话人3）。由于发现这些样本的标注与对应视频之间存在错位，因此将其剔除。

2.1. 总体结果

所有提及的视听语音识别系统均已按照第1.3节（模型集成）所述，通过后验概率平均在内部完成融合。详细的后验概率平均结果见表4。我们三项提交方案的构建细节如下。**提交1**：基于原始音频输入，将来自**AVSR S1、S2、S3** 和**S3-UASR**的输出假设使用ROVER进行融合。**提交2**：针对原始音频以及四个语音分离系统，我们使用**AVSR S1、S2、S3** 和**S3-UASR**生成20个解码结果。采用基于ROVER的启发式搜索策略来确定最优系统组合。所选的AVTSE–AVSR组合包括以下九个系统：**系统 III - S3-UASR** 、**系统 IV - S3-UASR** 、**系统 II - S3** 、**系统 IV - S3**、**原始音频 - S3**、**系统 III - S2**、**系统 IV - S2**、**系统 II - S1**和**原始音频 - S2**。**提交3**：对于提交2中所述的九个系统，我们首先使用N-best=30 对每个系统进行单系统ROVER融合，

表 2。ASD 模型在$\mathcal{D}_{\text{dev}}$ 上的召回率和精确率结果，以及在$\mathcal{D}'_{\text{dev}}$上的相应词错误率（WER，%）。

系统	召回率 (%)	精确率（%）	词错误率（WER，%）
基线	75.51	94.95	33.62
Ours	**82.74**	**95.92**	**31.23**

表 3。在 $\mathcal{D}'_{\text{dev}}$ 上，基于我们的 ASD 和**AVSRS3**系统实现的不同 AVTSE 系统的词错误率（%）结果。

系统	I	II	III	IV	原始音频
WER	30.90	30.82	30.89	31.04	31.23

并分别完成。随后将得到的九个融合假设进一步使用ROVER组合，以获得最终输出。详细评估结果见表1。

2.2. 详细结果

ASD 的实验结果见表 2。与基线模型 light-ASD [27]，相比，我们的模型在召回率和精确率方面均有显著提升，绝对增幅分别为 7.23% 和 0.97%。

对于 AVTSE，我们在自有 ASD 和 **AVSR S3**模型上比较了四种 AVTSE 系统，结果如表 3 所示。实验结果表明，与原始混合输入相比，经不同 AVTSE 系统提取的音频均带来了不同程度的改善；而 AV-Speformer、USEV 和 Seanet 等常用提取方案在这种具有挑战性的条件下均导致性能下降。

对于音视频语音识别，我们评估了两种预训练规模——**基础版**（12层，1.1亿参数）和**大型版**（18层，3.3亿参数）——以及三种视觉配置：**人脸**（全脸）、**唇部**（唇部 ROI）和 **人脸与唇部**。依赖人脸的模型使用了一个 3530 小时的子集（不含 AVYT），而基于唇部的模型则使用了完整的 4966 小时语料。我们使用官方 ASD 分段（拼接为 8–16 秒）和原始远场音频，在 $\mathcal{D}'_{\text{dev}}$ 上进行评估，且未使用 N-best ROVER。如表4所示，基于 Whisper 的架构（**S2**）从显式唇部特征中获益显著，而基于大语言模型的模型（**S3**）在扩展单个组件时则表现出性能饱和。关键的是，将 **S3-UASR** 框架与基于大语言模型的模型融合，取得了整体最佳性能。这一提升凸显了**S3** 与 **S3-UASR** 框架之间的互补性，优于对**S3**模型进行同质集成。此外，后验概率平均始终能够提升所有框架（**S1–S3**）的鲁棒性。

对于对话聚类，基于**AVSR S1** 结果的基于大语言模型的对话聚类方法取得了较高的准确率；通过融合不同的大语言模型与基线方法，开发集上的准确率达到 100%，如表 5 所示。

表4。 $\mathcal{D}'_{\text{dev}}$上的词错误率（%）结果。E1/E2 和 E3/E4 使用不同的超参数设置进行训练。

Exp. ID	框架	编码器	解码器	词错误率（%）
E1	AVSR S1	大型面部	Transformer	36.71
E2				36.50
E3		大型面部与唇部		36.27
E4				36.51
–	后验概率平均 (E1+E2+E3+E4)			**34.92**
E5	AVSR S2	大型面部 + Whisper 编码器	Whisper 解码器	36.48
E6		大型面部与唇部 + Whisper 编码器		35.30
–	后验概率平均 (E5+E6)			**34.21**
E7	AVSR S3	大型面部	Qwen 2.5-7B	35.64
E8		大型面部	Qwen 2.5-14B	35.72
E9		大型面部与唇部	Qwen 2.5-7B	35.86
–	后验概率平均 (E7+E8+E9)			**34.20**
E10	AVSR S3 -UASR	面部与唇部基础型 + WavLM	Qwen 2.5 7B	36.57
–	后验概率平均 (E8+E10)			**33.62**

表 5。对话聚类系统比较（迭代次数 K，选择 M，聚类数 N）。

模型	Para	K	M	N	F1分数 (%)
基线	-	-	-	-	81.53
通义千问 2.5	72B	10	10	5	98.89
深度求索 R1	671B	10	10	5	100.00
融合					**100.00**

3. 结论

CHiME-9 MCoRec 挑战赛涉及大量语音重叠和并发的多方对话，这给准确的自动语音识别带来了重大限制。我们通过一个融合音视频自监督学习（AV-SSL）与大语言模型的集成框架应对这些挑战。我们的处理流程依次采用数据正则化音视频说话人检测（AV-ASD），以提供精确的时间边界；在大量合成数据集上训练的语义感知音视频语音分离（AVTSE）；以及稳健的基于 Whisper-LLM 的音视频语音识别后端。我们的主要发现如下：大规模音视频预训练能够减轻低信噪比和语音重叠造成的性能下降；融入全面的面部特征比仅使用唇部表征更有利于身份识别；预训练多模态表征显著提升了检测、提取和自动语音识别的稳健性，JACER 达到 15.7%。

4. 致谢

本工作得到国家自然科学基金委员会U25A20409号项目、国家自然科学基金委员会62401348号项目、中央高校基本科研业务费GK202406005号项目以及中国陕西省科学技术协会青年人才托举计划（项目编号：20250126）的资助。

5. 参考文献

[1]Thai-Binh Nguyen、Katerina Zmolikova、Pingchuan Ma、Ngoc Quan Pham、Christian Fuegen和Alexander Waibel，《鸡尾酒会基准：多模态数据集与比较评估结果》，*arXiv*预印本 *arXiv:2510 23276*，2025。。[2] Thai-Binh Nguyen、Ngoc-Quan Pham和Alexander Waibel，《鸡尾酒会音视频语音识别》，*arXiv*预印本 *arXiv:2506.02178*，2025。[3] Sourish Chaudhuri、Joseph Roth、Daniel P. W. Ellis等，《Ava-speech：电影中语音活动的密集标注数据集》，*CoRR*，第abs/1808.00606卷，2018年。[4] Tao Liu、Shuai Fan、Xu Xiang等，《MSDWild：野外多模态说话人日志数据集》，载于*Interspeech 2022*，2022年，第1476–1480页。[5]Shilong Wu，《M3sd：多模态、多场景和多语言说话人日志数据集》，*arXiv*预印本 *arXiv:2506.14427*，2025。[6] Triantafyllos Afouras、Joon Son Chung和Andrew Zisserman，《Lrs3-ted：用于视觉语音识别的大规模数据集》，*arXiv*预印本 *arXiv:1809.00496*，2018。[7] J. S. Chung、A. Nagrani和A. Zisserman，《Voxceleb2：深度说话人识别》，载于*INTERSPEECH*，2018年。[8] Ariel Ephrat、Inbar Mosseri、Oran Lang、Tali Dekel、Kevin Wilson、Avinatan Hassidim、William T. Freeman和Michael Rubinstein，《鸡尾酒会现场的聆听：一种独立于说话人的音视频语音分离模型》，*ACM Trans.Graph.*，第37卷第4期，2018年7月。[9] Harishchandra Dubey、Ashkan Aazami、Vishak Gopal、Babak Naderi、Sebastian Braun、Ross Cutler、Hannes Gamper、Mehrsa Golestaneh和Robert Aichner，《Icassp 2023深度噪声抑制挑战赛》，载于*ICASSP*，2023年。[10]Alexandros Haliassos、Andreas Zinonos、Rodrigo Mira、Stavros Petridis和Maja Pantic，《Braven：改进视觉与听觉语音识别的自监督预训练》，载于*ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*。IEEE，2024年，第11431–11435页。[11] Hang Chen、Jun Du、Yu Hu、Li-Rong Dai、Bao-Cai Yin和Chin-Hui Lee，《将子词发音与唇形关联起来，以实现嵌入感知的音视频语音增强》，*NeuralNetworks*，第143卷，第171–182页，2021年。[12] Alec Radford、Jong Wook Kim、Tao Xu、Greg Brockman、Christine McLeavey和Ilya Sutskever，《通过大规模弱监督实现鲁棒语音识别》，载于国际机器学习会议。PMLR，2023年，第28492–28518页。[13] J.G. Fiscus，《一种用于降低词错误率的后处理系统：识别器输出投票错误降低（ROVER）》，载于*1997 IEEE*自动语音识别与理解研讨会论文集，1997年，第347–354页。[14] Triantafyllos Afouras、Joon Son Chung、Andrew Senior、Oriol Vinyals和Andrew Zisserman，《深度音视频语音识别》，*IEEE transactions on pattern analysis and machine intelligence*，第44卷第12期，第8717–8727页，2018年。

[15] Thai-Binh Nguyen、Ngoc-Quan Pham和Alexander Waibel，“鸡尾酒会音视频语音识别”，载于*Interspeech 2025*，2025年，第1828–1832页。[16] Jing-Xuan Zhang、Genshun Wan、Jianqing Gao和Zhen-Hua Ling，“通过语音基础模型知识蒸馏进行音视频表征学习”，模式识别，第162卷，第111432页，2025年。[17] Zhuang Liu、Hanzi Mao、Chao-Yuan Wu、Christoph Feichtenhofer、Trevor Darrell和Saining Xie，“面向2020年代的卷积网络”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第11976–11986页。[18] Anmol Gulati、James Qin、Chung-Cheng Chiu、Niki Parmar、Yu Zhang、Jiahui Yu、Wei Han、Shibo Wang、Zhengdong Zhang、Yonghui Wu和Ruoming Pang，“Conformer：用于语音识别的卷积增强型Transformer”，载于*Interspeech 2020*，2020年，第5036–5040页。[19]Chung-Cheng Chiu、James Qin、Yu Zhang、Jiahui Yu和Yonghui Wu，“面向语音识别的随机投影量化器自监督学习”，载于国际机器学习会议。PMLR，2022年，第3915–3924页。[20] Jing-Xuan Zhang、Genshun Wan和Jia Pan，“唇部感兴趣区域是否足以进行唇读？”，载于*2022*年国际多模态交互会议论文集，2022年，第368–372页。[21] Jing-Xuan Zhang、Tingzhi Mao、Longjiang Guo、Jin Li和Lichen Zhang，“通过音视频自蒸馏预训练与说话人适配进行目标说话人唇读”，专家系统及其应用，第272卷，第126741页，2025年。[22] Jean-Baptiste Alayrac、Jeff Donahue、Pauline Luc等，“Flamingo：用于少样本学习的视觉语言模型”，载于神经信息处理系统进展，S. Koyejo、S. Mohamed、A. Agarwal、D. Belgrave、K. Cho和A. Oh编，2022年，第35卷，第23716–23736页，Curran Associates, Inc.[23] Andrew Rouditchenko、Yuan Gong、Samuel Thomas、Leonid Karlinsky、Hilde Kuehne、Rogerio Feris和James Glass，“Whisper-Flamingo：将视觉特征融入Whisper以进行音视频语音识别与翻译”，载于国际语音通信协会年会论文集，2024年，第2420–2424页。[24] Qwen团队等，“Qwen2.5技术报告”，2025年。[25] Edward J Hu、Yelong Shen、Phillip Wallis、Zeyuan Allen-Zhu、Yuanzhi Li、Shean Wang、Lu Wang、Weizhu Chen等，“LoRA：大语言模型的低秩适配”，*ICLR*，第1卷，第2期，第3页，2022年。[26] Jing-Xuan Zhang、Genshun Wan、Jin Li和Jianqing Gao，“利用大语言模型适配语音基础模型以实现统一语音识别”，*arXiv*预印本 *arXiv:2510.22961*，2025年。[27] Junhua Liao、Haihan Duan、Kanghui Feng、Wanbing Zhao、Yanbing Yang和Liangyin Chen，“一种用于主动说话人检测的轻量级模型”，*arXiv*电子预印本，第arXiv:2303.04439页，2023年3月。