arXiv:2305.04542v1 [cs.CV] 8 May 2023

多时间维度-用于视觉语音识别的口唇—音频记忆

Jeong Hun Yeo, Minsu Kim, Yong Man Ro†

电气工程学院，KAIST，韩国{sedne246, ms.k, ymro}@kaist.ac.kr

摘要

视觉语音识别（VSR）是一项根据唇部运动预测句子或词的任务。近期已有一些研究使用音频信号来补充视觉信息。然而，现有方法仅利用了有限的信息，例如自动语音识别（ASR）网络生成的音素级特征和软标签。本文提出了一种多时序唇音记忆（MTLAM），充分利用音频信号来弥补唇部运动信息的不足。所提出的方法主要由两部分组成：1）MTLAM保存由短期和长期音频信号产生的多时序音频特征，并记忆视觉到音频映射，以便在推理阶段根据视觉特征加载存储的多时序音频特征。2）我们设计了一种音频时序模型，用于生成能够捕捉邻近词上下文的多时序音频特征。此外，为了构建有效的视觉到音频映射，音频时序模型能够生成与视觉特征时间对齐的音频特征。通过大量实验，我们验证了MTLAM的有效性，并在两个公开的VSR数据集上取得了最先进水平的性能。

索引词— 视觉语音识别，唇读，多时间尺度唇音记忆，记忆网络

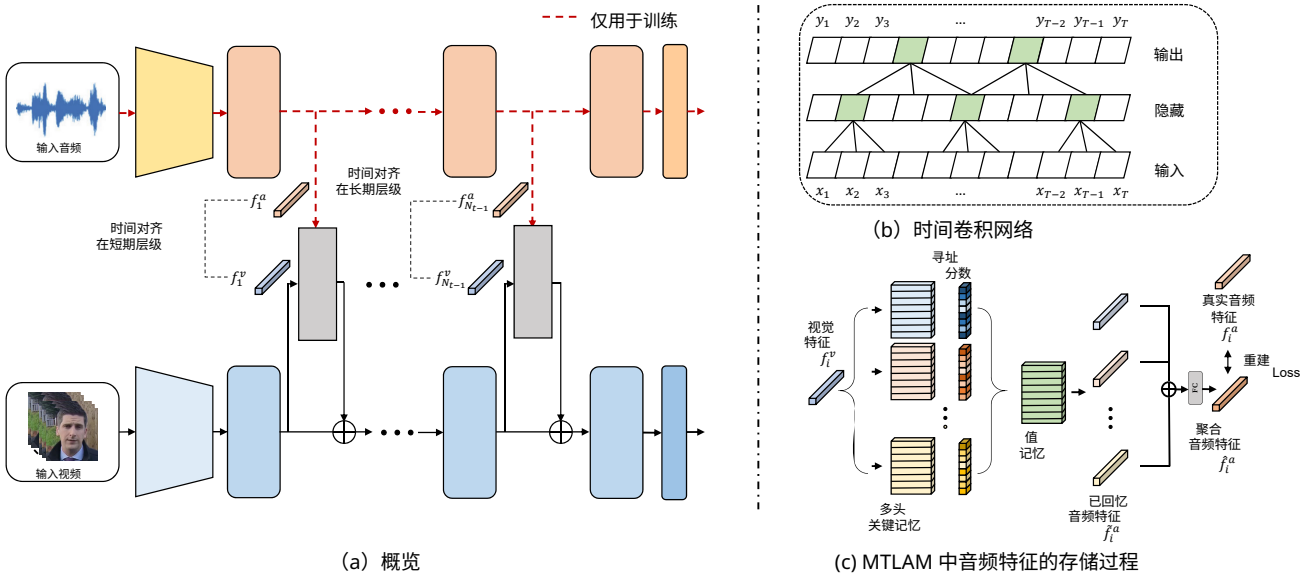
1. 引言

视觉语音识别（VSR） [1, 2, 3] 是一项通过解读无声视频中的唇部运动来识别词的任务。近年来，这项技术受到了广泛关注。视觉语音识别可用作无声电影的字幕工具、聋人的耳朵，甚至可用于在嘈杂环境中猜测说话者所说内容。然而，仅通过观察唇部运动来推断语音，可能不足以对语音内容进行建模，因为人类发声不仅依靠嘴唇，还依赖肌肉与多个器官的协调配合 [4]。

因此，研究人员尝试利用语音音频信号来弥补视觉信息的不足。一种常见的方法是通过知识蒸馏来利用音频知识 [5,6], ，其中包含一个视觉学生网络和一个音频教师网络。视觉学生网络经过训练，以跟随由音频语料数据预训练的音频教师网络所提供的软标签或中间特征。然而，由于视觉模态与音频模态之间存在异质性鸿沟，音频网络可能只能蒸馏出有限的信息。为绕过异质性鸿沟问题， [7, 8]提出了由记忆网络增强的VSR模型。记忆网络可利用构建的视觉与音频之间的映射绕过异质性鸿沟，从而在推理阶段无需输入真实音频即可加载音频特征。然而，现有的记忆增强VSR模型仅利用了音频的部分信息，例如从短时音频信号中编码得到的音素级表示。由于短时音频几乎不包含上下文信息，包含邻近词上下文信息的长时音频信号可能更适合用于构建准确的视觉到音频映射。

为解决这一问题，我们提出多时间尺度唇部-音频记忆网络（MTLAM），利用从短期到长期音频信号中编码的多时序音频特征，弥补视听语音识别模型视觉信息不足的问题。为了利用在多个时间尺度上编码的音频特征，我们考虑以下两个关键点：1）需要额外的音频时序模型，以从长期音频信号中捕获上下文信息。2）音频时序模型各层编码的音频特征应与每个视觉特征在时间上对齐，从而在记忆网络中构建有效的视觉到音频映射。为此，我们设计了一种音频时序模型，能够从一系列音频特征中捕获上下文信息，这些特征与视觉特征具有相同的时序信息，因此可以获得与视觉特征时间对齐的音频特征。时间对齐的音频特征能够提升记忆网络中视听特征映射的准确性。最后，MTLAM 存储由音频时序模型编码的多时序音频特征，从而能够利用从短期到长期的多时间尺度上下文。我们通过两个视听语音识别基准数据集 LRW 和 LRW-1000 验证了 MTLAM 的有效性，并证明了利用音频信号上下文信息的重要性，

本研究获得韩国政府（韩国科学技术信息通信部）资助的 IITP 项目支持（编号：2020-0-00004，基于长期视觉记忆网络的预测智能开发）。 †通讯作者。



（a）概览　　（b）时间卷积网络　　(c) MTLAM 中音频特征的存储过程

图 1。(a) 所提出的采用多时间层次注意力模块的视听语音识别框架的整体架构；(b) 随着层数加深，时间模型逐渐扩展时间感受野；(c) 多时间层次注意力模块使用多头键和一个值记忆的存储过程。

这对于构建准确的视觉到音频映射至关重要。此外，所提出的 MTLAM 在两个公开视听语音识别基准数据库上取得了最先进的性能。

2. 方法

2.1. 概览

所提出方法的整体流程如图 1（a）所示。首先，为了从唇部运动视频和原始音频信号中分别生成时空视觉特征和音频特征，我们分别采用视觉编码器和音频编码器。接下来，我们使用视觉时序模型和音频时序模型，生成包含短期至长期上下文信息的多时间尺度视觉特征和多时序音频特征，并通过多时间层次注意力模块对生成的多时间尺度视觉特征进行多时序音频特征补充。最后，将补充后的视觉特征输入线性层，以预测与输入唇部运动对应的真实文本。此外，为了引导音频特征形成有意义的语音表征，我们还在训练期间训练音频流来预测文本。

2.2. 音频时序模型

不同于先前采用固定短期音频特征的工作 [8]，我们构建了一个额外的音频时序模型，以考虑具有短期和长期上下文信息的多时序音频特征。该音频时序模型由时序卷积网络（TCN） [9], 组成，如图 1(b)所示。TCN 沿时间维度执行一维卷积，并从特征序列中捕获上下文信息。步幅为 1 时，每个 TCN 层的感受野大小 R 可计算为:

$$R = (k-1)(d-1)+k, \tag{1}$$

其中，k 表示滤波器尺寸，d 表示膨胀尺寸。例如，滤波器尺寸为 3、膨胀尺寸为 1、步幅为 1 时，感受野为 3，如图 1（b）中隐藏层所示。通过堆叠 TCN 层，我们可以利用逐渐增大的感受野，从音频输入中提取短期和长期上下文。此外，为了在 MTLAM 中构建有效的视觉到音频映射，音频时序模型被设计为生成与视觉特征时间对齐的音频特征。具体而言，我们设计 TCN 层，使音频时序模型中每一层的感受野与视觉时序模型中对应层的感受野相同，从而使这两种模态的表征在时间上保持同步。

2.3. 多时序唇语-音频记忆网络

我们提出 MTLAM，通过多时序音频特征补充视觉信息不足的问题。这些特征包含由音频时序模型编码的语音音频的短期和长期上下文。我们将音频时序模型除最后一层外的 N_{t-1} 层所编码的多时序音频特征记为 $\{f_i^a\}_{i=1}^{N_{t-1}}$，其中 $f_i^a \in \mathbb{R}^{T\times D}$，$T$ 是音频特征序列的长度，D 是嵌入维度。同理，将视觉时序模型生成的多时间尺度视觉特征记为 $\{f_i^v\}_{i=1}^{N_{t-1}}$，其中 $f_i^v \in \mathbb{R}^{T\times D}$。

为存储 N_{t-1} 多时序音频特征，MTLAM 采用基于 [8] 的架构，其中包含多头键记忆和值记忆，以将视觉到音频映射视为一对多映射。因此，MTLAM 共采用 N_{t-1} 个多头键记忆$M_{k,\ i} \in \mathbb{R}^{h\times N\times D}$ 以及 N_{t-1}个值记忆$M_{v,\ i} \in \mathbb{R}^{N\times D}$，以构建多时间层级的视觉到音频映射，其中 h 是头数， N 是记忆槽数。

首先，我们通过 MT-LAM 中的多头键记忆构建 N_{t-1} 多时间尺度的视觉到音频映射 N_{t-1} 。具体而言，多头键记忆的每个 j槽位都与值记忆的每个 j槽位相连。因此，可以映射来自同一词或短语的视觉和音频表征。为计算键记忆与值记忆之间的所有连接，我们通过以下公式得到寻址分数$A_{k,\ i} \in R^{h\times N}$：

$$A_{k,i} = \sigma(\alpha \cdot d(M_{k,i}, W^T_{q,i} f^v_i)), \tag{2}$$

其中，$d\ (\cdot)$ 表示余弦相似度，σ 表示 Softmax 函数，α表示缩放因子，$W_{q,\ i} \in R^{h\times D}$ 表示嵌入矩阵。

然后，通过使用寻址分数$A_{k,\ i}$ 访问值记忆，可以取回存储的音频特征 $\tilde{f}^a_i \in R^{h\times D}$，如下所示：

$$\tilde{f}^a_i = A_{k,i} \cdot M_{v,i}. \tag{3}$$

取回的音频特征 h通过聚合层 $W_a \in \mathbb{R}^{Dh\times D}$聚合为单个音频特征$\hat{}\ f^a_i \in \mathbb{R}^{T\times D}$，并与输入视觉特征时间对齐。该过程如图 1 (c) 所示。随后，每一层中的聚合多时间尺度音频特征$\hat{}\ f^a_i$通过求和与多时间尺度视觉特征 f^v_i融合，以补充视觉语音信息。

2.4. 联合损失函数

所提出的利用多时序音频特征的方法以端到端方式进行联合训练。首先，为了存储和重构多时序音频特征，我们采用重构损失和对比损失。重构损失旨在缩小重构的多时序音频特征与原始多时序音频特征之间的距离，可表示为：

$$\mathcal{L}_{recon} = \frac{1}{N_{t-1}} \sum_{i=1}^{N_{t-1}} ||1 - d(\hat{f}^a_i, f^a_i)||_1. \tag{4}$$

接下来，采用对比损失将不同的音频特征存储到值记忆的不同槽位中，其形式可表示为：

$$\mathcal{L}_{cont} = \frac{1}{N_{t-1}} \sum_{i=1}^{N_{t-1}} (\sum_{p\neq q} d(M^p_{v,i}, M^q_{v,i}) + \sum_{p=q} d(M^p_{v,i}, M^q_{v,i})). \tag{5}$$

表 1。音频时序模型各层生成的音频特征的有效性

基线[10]	MTLAM			Top-1 准确率 (%)
	#1	#2	#3	
✓	-	-	-	89.6
✓	✓	-	-	90.9
✓		✓	-	91.1
✓	-	-	✓	90.8
✓	✓	✓	-	91.3
✓	-	✓	✓	91.4
✓	✓	✓	✓	**91.7**

最后，为了训练视觉模型和音频模型，分别为视听语音识别（VSR）和自动语音识别（ASR）使用两个分类损失。$\mathcal{L}^v_{cls} = \mathcal{L}_{CE}(\hat{}y_v,\ y)$，$\mathcal{L}^a_{cls} = \mathcal{L}_{CE}(\hat{}y_a,\ y)$，其中 $\hat{}y_v$ 和 $\hat{}y_a$ 分别是由 VSR 网络和 ASR 网络预测的标签。 y 是真值。此外，设计了一个分类损失 $\hat{}\mathcal{L}^{v,\ a}_{cls} = \mathcal{L}_{CE}(\hat{}y_{v,\ \hat{a}},\ y)$，用于训练所提出的 MT-LAM，其中 $\hat{}y_{v,\ \hat{a}}$ 是使用视觉特征和重建的多时序音频特征预测的标签。因此，总损失可表示为：

$$\mathcal{L}_{total} = \mathcal{L}^v_{cls} + \mathcal{L}^a_{cls} + \mathcal{L}^{v,\hat{a}}_{cls} + \mathcal{L}_{recon} + \mathcal{L}_{cont}. \tag{6}$$

3. 实验结果

3.1. 实验设置

数据集。LRW [11] 是一个英语词级视觉语音识别数据集，包含 500 个词，每类有 800∼1,000 个视频片段。每个视频片段包含 29 帧，时长为 1.16 秒。我们对视频进行预处理，使其聚焦于唇部运动。视频数据裁剪为以唇部为中心的 136×136 像素区域，不对齐人脸，并调整大小至 96×96 像素。音频信号的采样率为 16,000 Hz。

LRW-1000 [12] 是一个普通话词级视听语音识别数据集，包含1,000个词，共计718,808个视频片段。LRW-1000的视频已经裁剪完成，我们只将视频片段调整为 96×96 像素。为匹配视频与音频之间的时间对齐关系，我们将视频片段的时长调整为长0.2秒。

实现细节。 视觉编码器由 ResNet-18 和三维卷积网络组成，视觉时序模型由 4 个 DC-TCN [13]层组成。类似地，我们将 ResNet-18 与一维卷积网络结合用作音频编码器，音频时序模型与视觉时序模型具有相同的结构。为了存储多时序音频特征，我们在视觉时序模型与音频时序模型之间设置了 3 个 MTLAM，但最后一个 DC-TCN 层除外。我们使用 8 张 NVIDIA GeForce RTX 3090 GPU 训练模型。采用端到端方式训练 80 个训练轮次，单个模型大约需要 7 天。批量大小设为 32，并采用余弦退火策略。

表 2。与现有视听语音识别方法的比较

方法	Top-1 准确率（%）	
	LRW	LRW-1000
多粒度 + C-BiLSTM [14]	83.3	36.9
R34 + BGRU [12]	-	38.2
R18 + BGRU + GRU [15]	83.5	38.7
T I3D + BiLSTM [16]	84.1	-
T R18 + 双向门控循环单元 [17]	84.1	41.2
R18 + 双向门控循环单元 + LSTM [18]	84.4	38.8
P3D R50 + 双向长短期记忆网络 [19]	84.8	-
R18 + 双向门控循环单元 / 人脸抠图 [20]	85.0	45.2
R18 + 多阶段时序卷积网络 [21]	85.3	41.4
R18 + 多阶段时序卷积网络 / Born-Again [22]	87.9	46.6
R18 + 多阶段时序卷积网络 / LiRA [23]	88.1	-
R18 +多阶段时序卷积网络 / WPCL + APFF [24]	88.3	-
R18 + DC-TCN [13]	88.4	43.7
R18 + 多阶段时序卷积网络 / MVM [8]	88.5	53.8
R18 + 多阶段时序卷积网络 / NetVLAD [25]	89.4	-
R18 + DC-TCN / 训练策略 [10]	91.6*	-
所提方法（MTLAM）	**91.7**	**54.3**

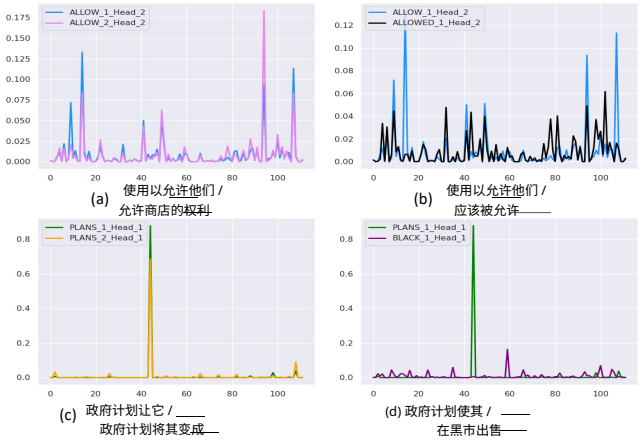


图 2。（a）和（c）为音频对具有相似上下文信息时的寻址分数；（b）和（d）为音频对具有不同上下文信息时的寻址分数。

3.2. 结果

消融实验。所提出的 MTLAM 旨在利用包含短期至长期音频信号上下文信息的多时序音频特征。MTLAM 的层次越深，存储的音频知识越长。为了验证存储在多时间层次注意力模块中的多时序音频特征的有效性，我们在 LRW 数据集上使用 6 个变体模型进行实验，结果如表 1 所示。基线 VSR 模型 [10]取得了 89.6% 的词准确率。当我们仅在不同的时序层次使用一个 MTLAM 时，MTLAM#1、MTLAM#2 和 MTLAM#3 的词准确率分别达到 90.9%、91.1% 和 90.8%。结果表明，当 VSR 模型将一个时序层次的音频特征存储到 MTLAM 中时，MTLAM#2 中保存的音频知识最为有效。当使用两个 MTLAM 时，MTLAM#1 与 MTLAM#2 结合可达到 91.3% 的准确率，MTLAM#2 与 MTLAM#3 结合可达到 91.4% 的准确率。这些结果表明，使用多时序音频特征优于使用单个 MTLAM。最后，通过考虑从短期到长期的上下文信息，使用全部多时序音频特征取得了 91.7% 的最佳准确率，这就是最终提出的方法。

与最先进方法的比较。为验证所提方法的有效性，我们在 LRW 和 LRW-1000 数据集上，将其性能与现有最先进模型进行比较。表 2 展示了比较结果。采用 DC-TCN 和 MS-TCN 时序模型时，我们提出的方法在 LRW 上分别达到了 91.7% 和 90.1% 的准确率，并取得了优于基线方法的最佳性能。尤其是，所提方法的表现甚至优于一种最先进方法[10]，该方法采用了 4 个不同训练模型的集成。此外，我们在 LRW-1000 上达到了 54.3 % 的词准确率。这表明，在记忆网络中利用同时包含短期和长期上下文信息的多时序音频特征来构建视觉到音频映射，可以提升视听语音识别的性能。

MTLAM 中的长期音频知识。为验证具有长期上下文信息的音频特征是否能很好地存储在 MTLAM 中，我们将四组不同音频对的 MTLAM#2 寻址分数可视化。图 2（a）和（c）展示了音频对在句子中包含相同短语和上下文时的寻址分数。相比之下，图 2（b）和（d）展示了具有不同短语和上下文信息的音频对的寻址分数。两个句子越相似，寻址分数的分布就越一致。这表明，除短期上下文外，所提出的 MTLAM 还能够在长期上下文中构建视觉到音频映射。

4. 结论

本文提出了多时序唇音记忆（MTLAM），利用多时序音频知识来弥补视听语音识别（VSR）中不足的视觉信息。此外，MTLAM的音频时序模型生成了与视觉特征时间对齐的多时序音频特征，从而有效地在视觉特征与音频特征之间建立映射。我们证明，通过考虑音频特征和视觉特征中的多时间尺度上下文，可以增强视觉特征与音频特征之间的映射。我们在两个视听语音识别基准数据库上取得了最先进水平的性能，验证了所提出方法的有效性。

* 4 个训练模型的集成性能。

5. 参考文献

[1] Pingchuan Ma、Stavros Petridis和Maja Pantic，“使用Conformer的端到端视听语音识别”，载于*ICASSP 2021-2021 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2021年，第7613–7617页。[2] KR Prajwal、Triantafyllos Afouras和Andrew Zisserman，“利用视觉注意力进行子词级唇读”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2022年，第5162–5172页。[3] Minsu Kim、Hyunjun Kim和Yong Man Ro，“采用用户相关填充的说话人自适应唇读”，载于计算机视觉–*ECCV 2022*：第*17*届欧洲会议，以色列特拉维夫，*2022*年*10*月*23–27*日，论文集，第*XXXVI*部分。Springer，2022年，第576–593页。[4] Robert T Sataloff，“人类声音”，《科学美国人》，第267卷，第6期，第108–115页，1992年。[5] Ya Zhao、Rui Xu、Xinchao Wang、Peng Hou、Haihong Tang和Mingli Song，“聆听唇语：通过蒸馏语音识别器改进唇读”，载于*AAAI*人工智能会议论文集，2020年，第34卷，第6917–6924页。[6] Sucheng Ren、Yong Du、Jianming Lv、Guoqiang Han和Shengfeng He，“向大师学习：为唇读蒸馏跨模态高级知识”，载于*IEEE/CVF*计算机视觉与模式识别会议论文集，2021年，第13325–13333页。[7] Minsu Kim、Joanna Hong、Se Jin Park和Yong Man Ro，“Cromm-vsr：跨模态记忆增强的视觉语音识别”，*IEEE*多媒体汇刊，2021年。[8] Minsu Kim、Junghun Yeo和Yong Man Ro，“使用多头视觉—音频记忆区分同音异义词”，载于*AAAI*人工智能会议论文集，2022年。[9] Colin Lea、Michael D Flynn、Rene Vidal、Austin Reiter和Gregory D Hager，“用于动作分割与检测的时间卷积网络”，载于*IEEE*计算机视觉会议论文集，2017年，第156–165页。[10] Pingchuan Ma、Yujiang Wang、Stavros Petridis、Jie Shen和Maja Pantic，“改进唇读的训练策略”，载于*ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022年，第8472–8476页。[11] Joon Son Chung和Andrew Zisserman，“野外唇读”，载于亚洲计算机视觉会议。Springer，2016年，第87–103页。[12] Shuang Yang、Yuanhang Zhang、Dalu Feng、Mingmin Yang、Chenhao Wang、Jingyun Xiao、Keyu Long、Shiguang Shan和Xilin Chen，“LRW-1000：用于野外唇读的自然分布式大规模基准”，载于*2019*年第*14*届*IEEE*自动人脸与手势识别国际会议（*FG 2019*）。IEEE，2019年，第1–8页。[13] Pingchuan Ma、Yujiang Wang、Jie Shen、Stavros Petridis和Maja Pantic，“使用密集连接的时序卷积网络，”载于*IEEE/CVF*冬季计算机视觉应用会议，2021年，第2857–2866页。[14] Chenhao Wang，“用于唇读的多粒度时空建模”，*arXiv*预印本 *arXiv:1908.11618*，2019年。[15] Mingshuang Luo、Shuang Yang、Shiguang Shan 和 Xilin Chen，“用于序列到序列唇读的伪卷积策略梯度”，载于*2020*年第*15*届*IEEE*国际自动人脸与手势识别会议（*FG 2020*）。IEEE，2020年，第273–280页。[16] Xinshuo Weng 和 Kris Kitani，“利用双流深度3D卷积神经网络学习唇读的时空特征”，*arXiv*预印本 *arXiv:1905.02540*，2019年。[17] Jingyun Xiao、Shuang Yang、Yuanhang Zhang、Shiguang Shan 和 Xilin Chen，“基于形变流的唇读双流网络”，载于*2020*年第*15*届*IEEE*国际自动人脸与手势识别会议（*FG 2020*）。IEEE，2020年，第364–370页。[18] Xing Zhao、Shuang Yang、Shiguang Shan 和 Xilin Chen，“用于有效唇读的互信息最大化”，载于*2020*年第*15*届*IEEE*国际自动人脸与手势识别会议（*FG 2020*）。IEEE，2020年，第420–427页。[19] Bo Xu、Cheng Lu、Yandong Guo 和 Jacob Wang，“判别式多模态语音识别”，载于*IEEE/CVF*计算机视觉与模式识别会议，2020年，第14433–14442页。[20] Yuanhang Zhang、Shuang Yang、Jingyun Xiao、Shiguang Shan 和 Xilin Chen，“我们能否超越嘴唇读取语音？重新思考深度视觉语音识别中的感兴趣区域选择”，载于*2020*年第*15*届*IEEE*国际自动人脸与手势识别会议（*FG 2020*）。IEEE，2020年，第356–363页。[21] Brais Martinez、Pingchuan Ma、Stavros Petridis 和 Maja Pantic，“利用时间卷积网络进行唇读”，载于*ICASSP 2020-2020 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2020年，第6319–6323页。[22] Pingchuan Ma、Brais Martinez、Stavros Petridis 和 Maja Pantic，“通过蒸馏和高效模型实现实用唇读”，载于*ICASSP 2021-2021 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2021年，第7608–7612页。[23] Pingchuan Ma、Rodrigo Mira、Stavros Petridis、Bj¨orn W Schuller 和 Maja Pantic，“LiRA：通过自监督从音频中学习视觉语音表征”，*arXiv*预印本 *arXiv:2106.09171*，2021年。[24] Weidong Tian、Housen Zhang、Chen Peng 和 Zhong-Qiu Zhao，“基于整体-部分协同学习的唇读模型”，载于*ICASSP 2022-2022 IEEE*国际声学、语音与信号处理会议（*ICASSP*）。IEEE，2022年，第2425–2429页。[25] Heng Yang、Tao Luo、Yakun Zhang、Mingwu Song、Liang Xie、Ye Yan 和 Erwei Yin，“利用时间收缩网络和NetVLAD改进词级唇读”，载于*2022*年国际多模态交互会议论文集，2022年，第504–508页。