微运动激励与时间感知的唇语识别方法

马金林[1],吕 鑫[1],马自萍[2],郭兆伟[1],吕 科[3]
(1. 北方民族大学计算机科学与工程学院,宁夏银川 750021;2. 北方民族大学数学与信息科学学院,宁夏银川 750021;
3. 中国科学院大学计算机与通信工程学院,北京 100049)

摘 要: 时序信息和唇部细微变化对唇语识别至关重要. 然而,现有唇语识别方法不能精准捕获时序信息和关注细微运动. 为此,提出一种关注微小唇部变化和增强时序信息的唇语识别方法 DMT-GhostNet. 首先,引入解藕时空增强块(Decoupled Spatio-Temporal Enhancement Block,DSTE),将单一 3D 卷积解藕为时间域和空间域;其次,基于运动激励(Motion Excitation,ME)与 Ghost 瓶颈块提出微运动瓶颈块(Micro-Motion Bottleneck,M-Ghost),捕捉唇部的微小运动;最后,提出时间感知模块(Transformer Multi-Scale Temporal Convolution Network,TransMS-TCN),聚焦重要时间序列,限制无关信息流入 MS-TCN. 实验结果表明,DMT-GhostNet 在 LRW 数据集上取得了 89.21% 的准确率,比基于 ResNet 的主流方法提升 3.91%,降低参数量近 6 M,能够更好地利用时序信息并聚焦唇部细节,显著提高唇语识别性能.

关键词: 唇语识别;GhostNetV2;时间维度;微运动激励
基金项目: 宁夏自然科学基金(No.2023AAC03264);北方民族大学中央高校基本科研业务费专项(No.2021KJCX09)
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2024)11-3657-12
电子学报 URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20230888

Micro-Motion Excitation and Time Perception for Lip Reading

MA Jin-lin[1], LÜ Xin[1], MA Zi-ping[2], GUO Zhao-wei[1], LÜ Ke[3]
(1. *School of Computer Science and Engineering*, *North Minzu University*, *Yinchuan*, *Ningxia* 750021, *China*;
2. *School of Mathematics and Information Science*, *North Minzu University*, *Yinchuan*, *Ningxia* 750021, *China*;
3. *College of Computing and Communication Engineering*, *University of Chinese Academy of Sciences*, *Beijing* 100049, *China*)

Abstract: Temporal information and subtle lip changes are crucial for lip reading. However, existing lip-reading methods have not accurately captured temporal information and focus on subtle movements. In response, we propose a lip-reading method named DMT-GhostNet that emphasizes minor lip variations and enhances temporal information. We introduce the decoupled spatio-temporal enhancement block (DSTE) to decouple the single 3D convolution into the time domain and the spatial domain. Based on motion excitation (ME) and the Ghost bottleneck block, we introduce the micro-motion bottleneck (M-Ghost) to detect subtle lip motions. The transformer multi-scale temporal convolution network (TransMS-TCN) is proposed to focus on important temporal sequences and restrict irrelevant information from flowing into MS-TCN. Experimental results show that DMT-GhostNet achieved an accuracy of 89.21% on the LRW dataset, which is an increase of 3.91% over mainstream methods based on ResNet and reduces the parameter count by nearly 6 M. This indicates that DMT-GhostNet effectively utilizes temporal information and focuses on lip details, significantly improving lip-reading performance.

Key words: lip-reading; GhostNetV2; time dimension; micro-motion excitation
Foundation Item(s): Natural Science Foundation of Ningxia (No.2023AAC03264); Basic Scientific Research in Central Universities of North Minzu University (No.2021KJCX09)

1 引言

唇语识别,是一项通过观察说话者的口型变化“读出”或“部分读出”其所说内容的任务[1],在嘈杂环境通信、安全应用、多模态语音识别[2]和听障辅助等诸多领

收稿日期:2023-09-22;修回日期:2024-06-03;责任编辑:孙瑶

域具有潜在应用价值. 然而,受较低准确率的制约,唇语识别技术仍然面临挑战,主要表现为现有模型难以准确捕捉视觉歧义词汇的细微唇形差异,且不能有效建模唇部运动的关键时序特征. 因此,唇部细微运动和关键时序特征的研究,对提升唇语识别性能具有重要意义.

近几年来,唇语识别模型形成了相对固定的三阶段端到端框架:第一阶段被称为前端,用于提取视觉时空特征;第二阶段被称为后端,用于整合时间上下文信息;最后一个阶段用于输出分类结果. 目前,前端主要使用3D CNN(Convolution Neural Network)和2D CNN联合的方法提取时空特征. 该方法通常先利用3D CNN获得连续视频帧的时间特征,再使用2D CNN提取细粒度空间特征. 例如,Feng等人[3]通过将ResNet的第一个2D卷积改为3D卷积的方式引入时间信息. Stafylakis等人[4]和Ma等人[5]通过在ResNet前额外添加一层5×7×7的3D卷积,提取浅层时空特征. 然而,一层3D卷积难以充分捕捉复杂的时序特征. 常用的后端主要有双向长短期记忆网络(Bidirectional Long Short-Term Memory network, Bi-LSTM)、双向门控循环单元(Bidirectional Gated Recurrent Unit, Bi-GRU)和多尺度时间卷积网络(Multi-Scale Temporal Convolutional Network, MS-TCN). 其中,Bi-LSTM[4]在捕捉上下文依赖关系方面表现出色,但其参数量庞大,增加了计算成本和训练时间. Bi-GRU[6]虽然一定程度上提高了时序特征提取能力,但在处理短序列数据时仍存在过拟合风险. MS-TCN[7]采用多尺度时间卷积混合长短期时序信息,但它未区分时序信息的重要程度,导致关键时序信息被忽略.

视频序列中的运动特征和时空特征蕴含了关键的互补信息[8]. 在动作识别领域,Jiang等人[9]融合时空特征和运动编码,打破了时空信息和运动信息分流学习的壁垒. 类似地,Li等人[8]和Wang等人[10]提出运动激励与时空特征联合学习的方法,提取时空特征的同时捕获时间序列的动态信息. 在唇语识别领域,难以识别的词语通常具有发音相似、唇部运动细微的特点,而这些细微的唇部运动蕴含着关键的辨别性信息. 然而,现有的唇语识别方法侧重于提取时空特征,未能充分关注唇部的细微运动.

针对上述问题,本文提出增强时序特征并关注唇部细微运动的识别方法DMT-GhostNet. 首先,利用解耦时空增强块增强时间维度的信息;然后,使用包含运动激励模块的M-GhostNet提取细微唇部特征;最后,使用TransMS-TCN建模时间序列的关联性. DMT-GhostNet的创新如下:

(1)引入解耦时空增强块,提高3D卷积的时间信息提取能力;

(2)优化GhostNetV2的堆叠方式,增强时空表征能力;

(3)设计微运动瓶颈块M-Ghost,捕捉唇部微小运动;

(4)提出时间感知模块TransMS-TCN,提高时序建模能力.

2 相关工作

2.1 GhostNetV2

在资源受限的设备上,将深度模型轻量化能够有效降低计算成本,实现高效运算[11,12]. GhostNet[13]便是为了减少模型参数量和计算复杂度而设计的一种神经网络架构. 马金林等人[14]将GhostNet作为前端骨干网络,加入空间和通道注意力机制强化特征提取能力,并使用同类自知识蒸馏训练,在LRW数据集上达到87.3%的准确率. Zhang等人[15]进一步在GhostNet的瓶颈块中引入ECA[16]模块,通过局部跨通道交互策略提升性能. GhostNetV2[17]在继承GhostNet高效性的基础上,增加了对长距离空间信息的处理能力,被广泛应用于人脸识别[18]、目标检测[19]等多个领域.

GhostNetV2的主要组成部分是Ghost模块和解耦全连接注意力. Ghost模块使用线性运算代替标准卷积,生成特征图. Ghost模块如图1所示.

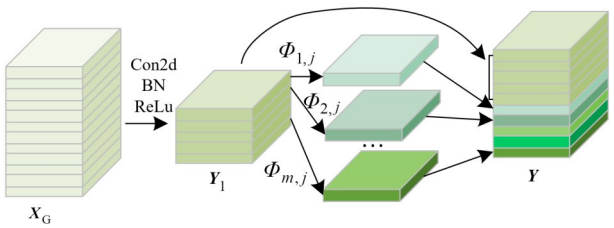


图1 Ghost模块

首先,特征图$\boldsymbol{X}_{\mathrm{G}}$($\boldsymbol{X}_{\mathrm{G}}\in\mathbb{R}^{H\times W\times C}$)卷积后生成特征图$\boldsymbol{Y}_1$($\boldsymbol{Y}_1\in\mathbb{R}^{H\times W\times m}$). 然后,对$\boldsymbol{Y}_1$的每个内在特征$y_i$进行线性操作$\boldsymbol{\Phi}_{i,j}$,生成扩展特征图$y_{ij}$. 最终,经$n=m\times k$个特征映射,输出特征$\boldsymbol{Y}=\left[y_{11},y_{12},\cdots,y_{mk}\right]$($\boldsymbol{Y}\in\mathbb{R}^{H\times W\times n}$). Ghost模块的操作如式(1)和式(2)所示:

$$\boldsymbol{Y}_1=\boldsymbol{X}_{\mathrm{G}}\times f \tag{1}$$

$$y_{ij}=\boldsymbol{\Phi}_{i,j}\left(y_i\right) \tag{2}$$

式中,$i\in[1,m]$;$j\in[1,k]$;y_i是$\boldsymbol{Y}_1$的第i个内在特征图;$\boldsymbol{\Phi}_{i,j}$(除$\boldsymbol{\Phi}_{i,k}$)表示第j个线性操作,$\boldsymbol{\Phi}_{i,k}$是y_i的恒等映射;f表示标准卷积操作.

GhostNetV2与GhostNet的最大区别在于,它引入了解藕全连接注意力(Decoupled Fully Connected attention, DFC)捕获远程像素之间的依赖关系. DFC将传统的全连接层(Fully Connected layer, FC)解耦为水平方向的FC和垂直方向的FC,这两个FC分别沿各自的方向

聚合特征,形成全局感受野.

图 2 展示了 DFC 的信息流. 首先,输入特征 $\boldsymbol{X}_{\mathrm{A}}$($\boldsymbol{X}_{\mathrm{A}} \in \mathbb{R}^{H\times W\times C}$)通过下采样来减小特征图的尺寸,再由一个 1×1 卷积将特征变换为 DFC 的输入. 然后,分别使用大小为 $1\times K_{\mathrm{w}}$ 和 $K_{\mathrm{h}}\times1$ 的深度卷积提取垂直方向和水平方向的语义信息. $\boldsymbol{X}_{\mathrm{A}}$ 除了进入 DFC 分支得到注意力矩阵 $\boldsymbol{A}$ 外,还并行进入如图 1 所示的 Ghost 分支,得到扩展特征 $\boldsymbol{Y}$. 最后,这两个分支的输出逐元素相乘得到最终特征 $\boldsymbol{O}_{\mathrm{G}}$($\boldsymbol{O}_{\mathrm{G}} \in \mathbb{R}^{H\times W\times C}$),如式(3)所示:

$$\boldsymbol{O}_{\mathrm{G}} = \mathrm{Sigmoid}(\boldsymbol{A}) \odot \boldsymbol{Y} \tag{3}$$

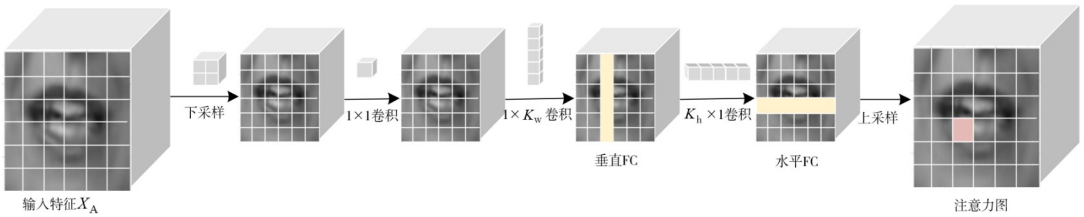


图 2　解藕全连接注意力(DFC)

2.2　多尺度时间卷积网络

与基于循环神经网络的 Bi-LSTM、Bi-GRU 不同,MS-TCN 和密集连接时间卷积网络(Densely Connected Temporal Convolutional Network,DC-TCN)是基于时间卷积的神经网络. Martinez 等人[7]提出了多尺度时间卷积网络 MS-TCN,更有效地建模时间序列特征. MS-TCN 在 LRW 数据集上表现出色,准确率达到 85.3%. 随后,Ma 等人[20]进一步扩展 MS-TCN,提出用于建模复杂的时间特征的 DC-TCN,并将准确率提高至 88.36%. 目前,更为常用的后端时序网络为 MS-TCN. MS-TCN 的多尺度特性体现在使用不同尺寸的一维卷积核提取时序特征,具体结构如图 3 所示.

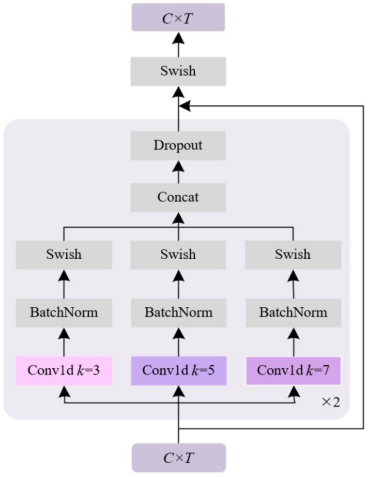


图 3　多尺度时间卷积块(MS-TCN 块)

MS-TCN 的每个分支都采用不同大小的卷积核,并在特征提取后使用了批归一化、Swish 激活和 Dropout. 本文遵循文献[7]的配置,分别采用大小为 3、5、7 的卷积核来混合长短期特征,并堆叠 2 次 MS-TCN 块.

虽然 MS-TCN 具有较强的时序建模能力,但无法有效捕获关键时序信息. 而 Transformer 具备联系上下文信息来捕获重要特征的突出优势[21]. 因此,本文使用 Transformer 对其进行优化,从而捕获关键时序特征、挖掘重要时序关系.

3　DMT-GhostNet

为解决时序信息建模能力不强和唇部细微变化关注度不足的问题,本文提出 DMT-GhostNet. DMT-GhostNet 由三部分组成:时空特征接收区、M-GhostNet 和 TransMS-TCN,其网络结构如图 4 所示. 时空特征接收区接收并增强时间信息,主要由 3D 卷积和解耦时空增强块组成;M-GhostNet 提取连续帧的视觉特征,主要由堆叠的 M-Ghost 瓶颈块组成;TransMS-TCN 聚合时间上下文信息,主要由 Transformer 编码块和 MS-TCN 组成.

3.1　解藕时空增强块

为进一步强化时间特征和空间特征,本文借鉴了伪 3D 残差卷积(Pseudo-3D,P3D)[22]解耦时间特征和空间特征的思想,提出并行和串行两种解耦时空增强块[23],其结构如图 5 所示. $T\times S\times S$ 的卷积核被拆分为两个卷积核,$1\times S\times S$ 卷积核增强空间特征,$T\times1\times1$ 卷积核增强时间特征,然后再次使用 $1\times S\times S$ 卷积核和 $T\times1\times1$ 卷积核强化空间和时间特征. 本文将两个 DSTE 增强块放置在 3D 卷积之后、最大池化之前,构成新的时空特征接收区,以弥补单一 3D 卷积提取时序特征能力不足的问题.

3.2　微运动瓶颈块

微运动瓶颈块是 M-GhostNet 的基本构建单元. 通过将运动激励模块(Motion Excitation,ME)[8,10]加入 Ghost 瓶颈块来融合运动特征和时空特征,以便更敏锐地感知唇部微小运动. M-Ghost 瓶颈块的网络结构如图 6 所示. 当第二个 Ghost 模块的 stride=1 时,M-Ghost 的输入和输出特征图大小不变. 当 stride=2 时,M-Ghost 通过深度可分离卷积实现空间下采样,特征图尺寸减半. 因此,残差连接时需要借助 stride=2 和 stride=1 的卷积来匹配特征图的维度.

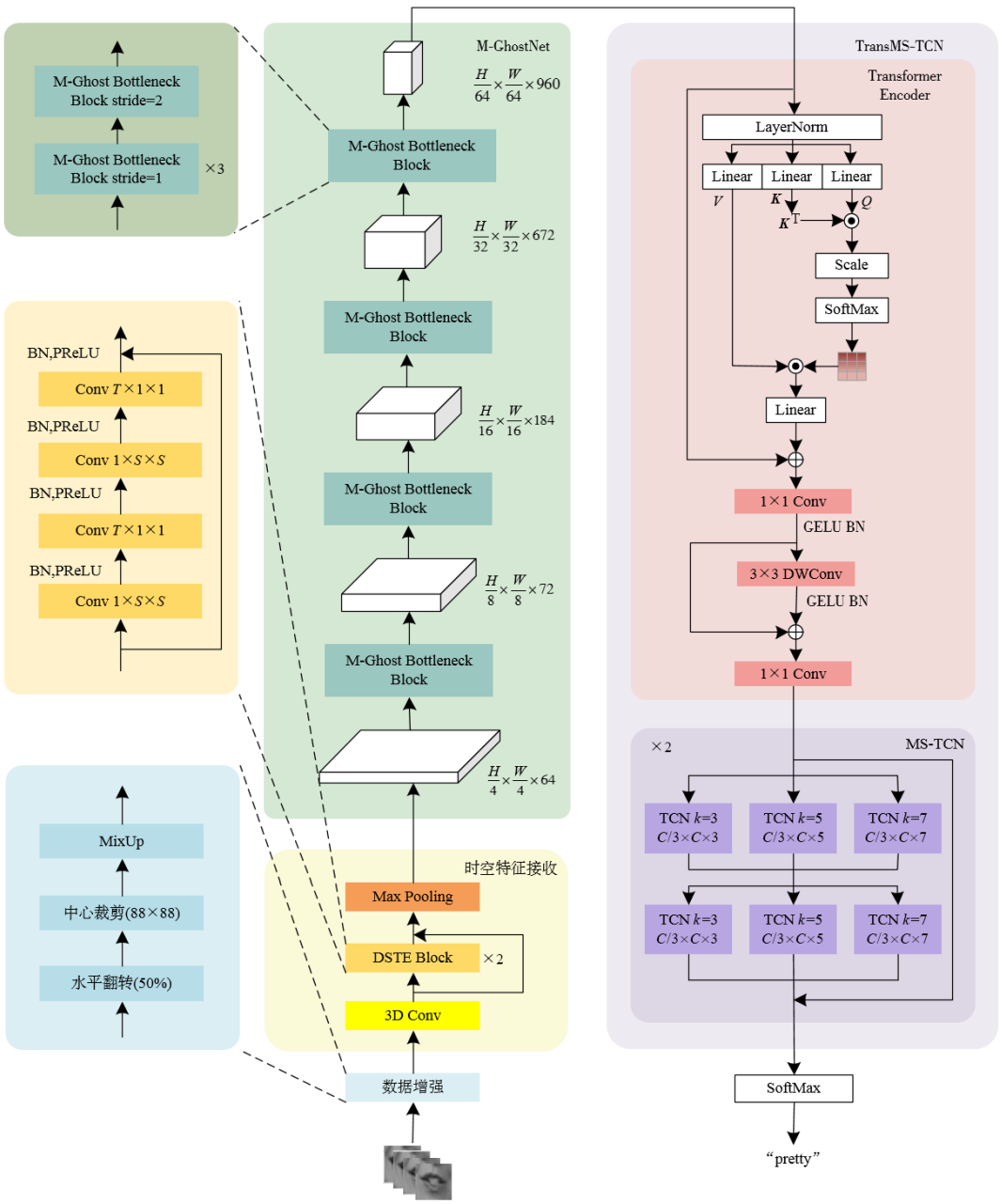
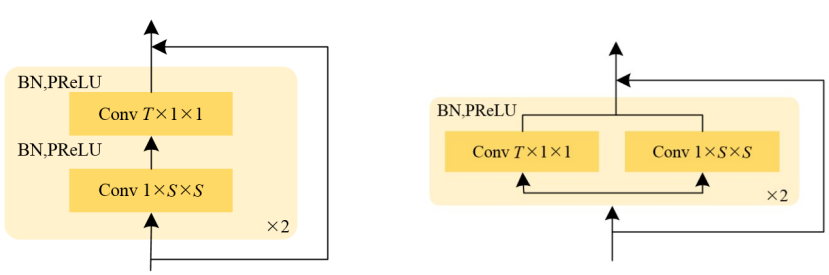


图4 DMT-GhostNet结构图



(a) 顺序DSTE块 (b) 并行DSTE块

图5 顺序和并行解耦时空增强块的结构图

ME模块计算特征的时间差异,并利用时间差异激发运动敏感通道,进而捕捉唇部运动的细节信息. ME的结构如图6(*a*)的左侧部分所示,其计算过程如下.

(1)使用1×1卷积减少输入特征图$\boldsymbol{X}_{\mathrm{M}}$($\boldsymbol{X}_{\mathrm{M}}\in\mathbb{R}^{\mathrm{BT}\times C\times H\times W}$)的通道数量,得到特征图$\boldsymbol{X}_r$($\boldsymbol{X}_r\in\mathbb{R}^{B\times C/r\times H\times W}$):

$$\boldsymbol{X}_r=\mathrm{conv}_{1\times1}\left(\boldsymbol{X}_{\mathrm{M}}\right) \tag{4}$$

(2)在每个时间步t,运动特征$\boldsymbol{M}(t)$($\boldsymbol{M}(t)\in\mathbb{R}^{B\times C/r\times H\times W}$)表示相邻两帧$\boldsymbol{X}_r(t+1)$和$\boldsymbol{X}_r(t)$的内容位移.

$$\boldsymbol{M}(t)=\mathrm{conv}_{3\times3}\left(\boldsymbol{X}_r(t+1)\right)-\boldsymbol{X}_r(t) \tag{5}$$

式中,$1\leqslant t\leqslant T-1$. 所有时间步的运动矩阵拼接在一起,构成运动矩阵$\boldsymbol{M}$:

$$\boldsymbol{M}=\left[\boldsymbol{M}(1),\boldsymbol{M}(2),\cdots,\boldsymbol{M}(T)\right] \tag{6}$$

(3)使用平均池化忽略空间布局,得到汇总信息$\boldsymbol{M}_{\mathrm{S}}$($\boldsymbol{M}_{\mathrm{S}}\in\mathbb{R}^{\mathrm{BT}\times C/r\times1\times1}$):

$$\boldsymbol{M}_{\mathrm{S}}=\mathrm{Pool}(\boldsymbol{M}) \tag{7}$$

(4)使用1×1卷积扩充运动特征的通道维度,得到特征$\boldsymbol{M}_{\mathrm{e}}$($\boldsymbol{M}_{\mathrm{e}}\in\mathbb{R}^{\mathrm{BT}\times C\times1\times1}$):

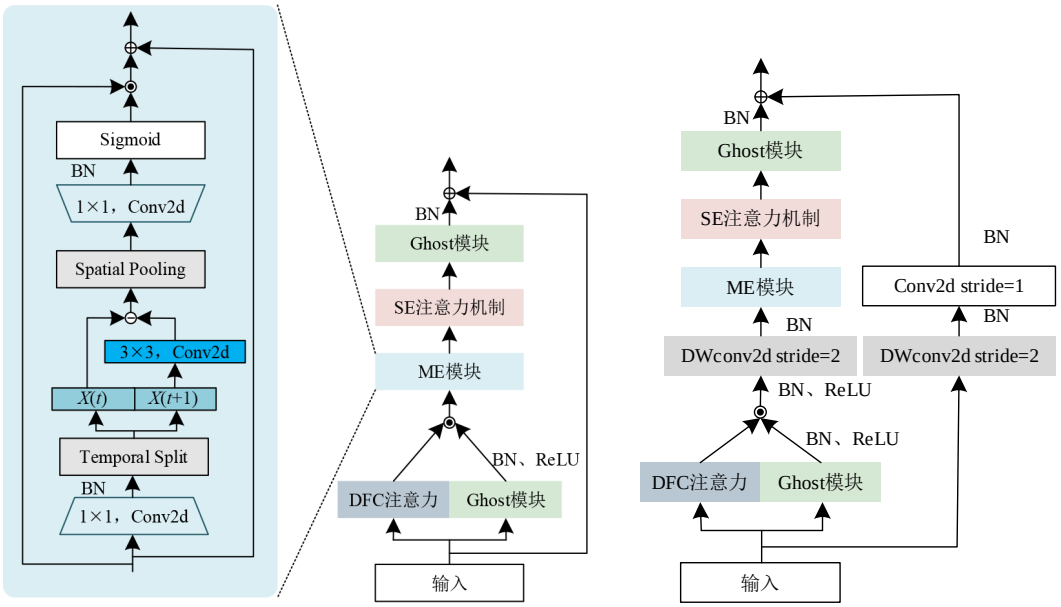
$$\boldsymbol{M}_{\mathrm{e}}=\mathrm{conv}_{1\times1}\left(\boldsymbol{M}_{\mathrm{S}}\right) \tag{8}$$

(5)通过Sigmoid得到运动注意力权重$\boldsymbol{A}$($\boldsymbol{A}\in\mathbb{R}^{\mathrm{BT}\times C\times1\times1}$),量化元素对应位置上运动差异的重要程度. 然后将$\boldsymbol{A}$与输入特征$\boldsymbol{X}_{\mathrm{M}}$逐通道相乘,以激发运动敏感通道. 最终将激发后的特征与原始特征$\boldsymbol{X}_{\mathrm{M}}$相加,得到微运动特征$\boldsymbol{O}_{\mathrm{M}}$($\boldsymbol{O}_{\mathrm{M}}\in\mathbb{R}^{\mathrm{BT}\times C\times H\times W}$):

$$\boldsymbol{A}=\mathrm{Sigmoid}\left(\boldsymbol{M}_{\mathrm{e}}\right) \tag{9}$$

$$\boldsymbol{O}_{\mathrm{M}}=\boldsymbol{X}_{\mathrm{M}}\odot\boldsymbol{A}+\boldsymbol{X}_{\mathrm{M}} \tag{10}$$

式中,Sigmoid激活函数的取值范围为(0,1). $\boldsymbol{A}$中的激发权重越接近1,表示对应位置上的运动差异越关键,应该被保留或增强. 相反,如果激发权重接近0,这意味着对应位置上的运动差异是不重要的,应该被削弱或忽略.



(*a*) 步长=1　　(*b*) 步长=2

图6　微运动瓶颈块(M-Ghost)

3.3　时间感知模块

时间感知模块TransMS-TCN旨在更有效地进行时序建模,由Transformer编码器和MS-TCN构成. TransMS-TCN借助Transformer的自注意力机制,有效聚焦重要时间序列信息,限制无关时序信息进入MS-TCN,迫使模型依据更重要的特征进行决策. TransMS-TCN的主要过程如下:

(1)使用多头自注意力关注关键时序特征. 首先,$\boldsymbol{Q}$与$\boldsymbol{K}^{\mathrm{T}}$相乘并缩放,得到注意力分数,以衡量当前时间步与其他时间步之间的关联性. 然后,应用Softmax将注意力分数归一化为注意力权重,并与$\boldsymbol{V}$加权求和,所得的值(Att)反映了模型对不同时序特征的关注程度.

$$\mathrm{Att}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_K}}\right)\boldsymbol{V} \tag{11}$$

式中,d_K为$\boldsymbol{K}$的维度.

(2)改进传统MLP层并将其用于特征转化. 首先,使用1×1卷积扩充特征的通道数. 之后,采用3×3深度可分离卷积提取局部时序特征. 最后,通过1×1卷积完成特征的细化和重塑. 为获得最佳性能,TransMS-TCN

将第一个1×1卷积的输出和3×3卷积的输出相加.

(3)使用MS-TCN增强时序特征,得到增强的时间序列特征$\boldsymbol{O}_{\mathrm{T}}$($\boldsymbol{O}_{\mathrm{T}}\in\mathbb{R}^{B\times C\times T}$).

4 实验设置

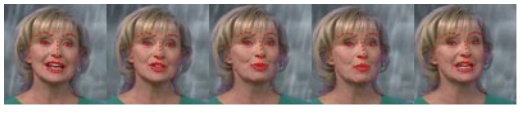
4.1 数据集

本文采用LRW数据集[24]训练和评估DMT-GhostNet. LRW数据集的视频数据提取自BBC电视节目,包含500类单词和173小时视频. 该数据集被分为488 766个训练视频、25 000个验证视频和25 000个测试视频. LRW数据集的挑战性主要体现在视觉歧义、光照变化和背景噪声等方面.

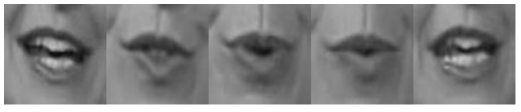
4.2 数据预处理

数据预处理过程如图4所示,包括水平翻转、中心裁剪和MixUp. 首先,使用Dlib[25]检测68个面部关键点,精确定位唇部感兴趣区域,并裁剪和灰度化为96×96的图像. 其次,应用水平随机翻转(翻转率为50%)、中心裁剪(裁剪尺寸为88×88)和MixUp增加数据的多样性. 图7展示了LRW数据集中"Pretty"的预处理过程.

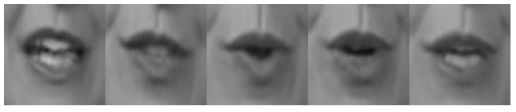
(a) 原始视频帧序列



(b) 人脸关键点检测



(c) 唇部中心裁剪



(d) MixUp数据增强

图7 "Pretty"的预处理过程

4.3 评价指标

单词级的唇语识别本质上是一个多分类问题. 本文采用分类准确率(Accuracy)评估DMT-GhostNet的预测性能. 准确率反映了正确预测单词与总单词之间的比例关系,如式(12)所示:

$$\mathrm{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{12}$$

式中,TP、TN、FP和FN分别代表真正例、真负例、假正例和假负例.

此外,本文使用参数量评估模型的规模.

4.4 实验设置

本文实验基于PyTorch 1.8.1+cu111和NVIDIA GeForce RTX3080运行,使用余弦调度器和AdamW优化器进行训练,设置初始学习率为0.000 3,权重衰减率为0.02,批大小为32,Epoch为120.

5 实验结果和分析

5.1 时空特征提取网络的性能

由图8可知,GhostNetV2较ResNet-18的性能提升了1.7%,这归因于GhostNetV2具有更深的网络结构,从而增强了层级表达能力和特征提取能力;DenseNet-121取得了最高的准确率,但参数量比GhostNetV2多2.796 M;随着版本增加,MobileNet系列和ShuffleNet系列的准确率也稳步提升,但是它们的表现仍落后于GhostNetV2,原因是其对特征之间的相关性和冗余性的处理不够理想,而特征映射中的冗余性是深度神经网络成功的重要因素[15];GhostNetV2比GhostNetV1的识别精度提高了1.08%,这得益于解耦注意力有效捕捉了长程依赖关系;FasterNet[26]的核心算子是部分卷积(Partial Convolution,PConv),虽然PConv本身是轻量级卷积,但在实际应用中,需要与其他卷积一同嵌入模型,导致以FasterNet为前端的唇语识别模型的参数量高达50.709 M,约比GhostNetV2高20 M.

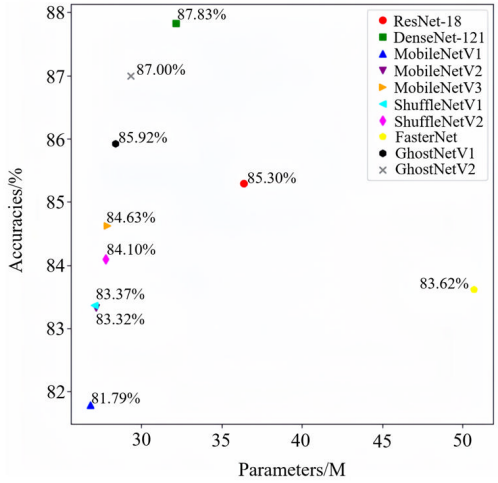


图8 时空特征提取网络的性能对比

综上所述,GhostNetV2能获取全局时空信息,精度高且参数量适中,是最佳的时空特征提取网络.

5.2 GhostNetV2的瓶颈块配置

本实验主要探索GhostNetV2的瓶颈块的堆叠数量和排列方式对模型性能的影响,实验结果如表1所示.表1中,配置[1,1,1,2]×4表示将瓶颈块按照3个stride=1和1个stride=2的顺序堆叠4次.Baseline为未改变堆叠数量和排列方式的GhostNetV2网络.本实验在后端为MS-TCN下进行.

表1　堆叠数量和排列方式对模型性能的影响

排列方式×堆叠数量	参数量/M	准确率/%
Baseline	29.335	87.00
[2,1,1,1]×4	29.384	85.61
[1,2,1,1]×4	29.376	86.97
[1,1,2,1]×4	29.400	87.54
[1,1,1,2]×2	**25.462**	86.76
[1,1,1,2]×4	29.397	88.02
[1,1,1,2]×5	31.686	**88.23**

(1)排列方式实验.比较分别使用排列方式为[2,1,1,1]、[1,2,1,1]、[1,1,2,1]和[1,1,1,2]的模型性能.实验结果显示,stride=2的瓶颈块位置越靠后,识别效果越好,这是因为过早下采样会丢失关键空间信息.

(2)堆叠数量实验.在瓶颈块的排列方式为[1,1,1,2]的基础上,分别比较堆叠2、4、5次的模型性能.结果显示,2次堆叠的模型具有较少的参数量,但准确率比4次堆叠的模型低1.26%,这表明骨干网络的总层数不宜过少,网络深度的减少导致特征的表达能力不足.5次堆叠的模型具有最大的参数量,而准确率仅比4次堆叠的模型高0.21%.

综合参数量和准确率的关系,本文使用[1,1,1,2]×4作为GhostNetV2的瓶颈块配置.

5.3 时序网络的性能

本实验考察不同时序网络下的模型识别性能.以GhostNetV2为前端特征提取网络,[1,1,1,2]×4为瓶颈块配置方式,通过更换不同后端网络,比较不同时序网络的参数量、准确率和收敛速度,实验结果如表2和图9所示.

表2　不同时序网络的性能对比

后端时序网络	参数量/M	准确率/%
Bi-LSTM	68.207	83.52
Bi-GRU	52.466	85.91
TCN	**6.254**	81.37
MS-TCN	29.397	88.02
TransMS-TCN	29.546	**88.59**

由表2可以看出,采用多尺度卷积策略的TransMS-TCN和MS-TCN的识别精度高于采用循环神经网络的Bi-LSTM和Bi-GRU.具体地,MS-TCN比Bi-LSTM高4.50%,比Bi-GRU高2.11%,TransMS-TCN比Bi-LSTM高5.07%,比Bi-GRU高2.68%.此外,采用双向结构的Bi-LSTM和Bi-GRU还增加了模型的复杂度和参数量.相比之下,TCN的参数量最小,但精度最低,主要原因是TCN使用单一卷积核提取时序信息,既缺乏MS-TCN的多尺度时间特征提取能力,又不具备TransMS-TCN过滤无效信息的能力.TransMS-TCN的识别精度最高,参数量仅比MS-TCN多0.149 M,说明Transformer使MS-TCN具备了关注重要信息的能力.

由图9可知,循环神经网络Bi-LSTM和Bi-GRU在前50轮的性能优于时间卷积网络TCN、MS-TCN和TransMS-TCN,并在70轮时迅速收敛.这是因为Bi-LSTM和Bi-GRU的双向结构同时利用过去和未来的时序信息捕捉序列依赖关系,这能够迅速学习数据的模式和规律.然而,随着训练轮次数的增加,MS-TCN和TransMS-TCN的优势逐渐显现,并远远超过Bi-LSTM和Bi-GRU.

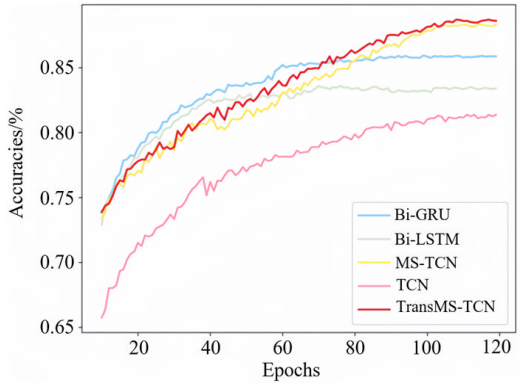


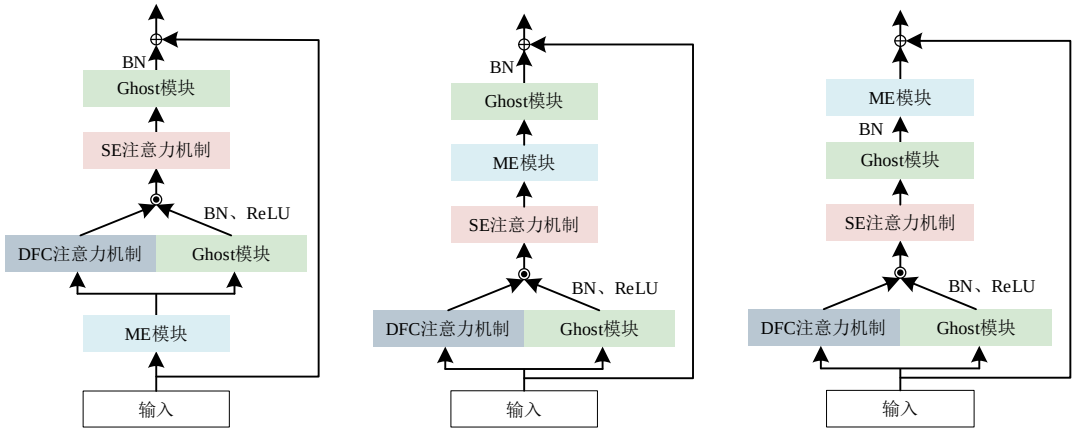
图9　不同时序网络的epochs-accuracies曲线

综上所述,TransMS-TCN可以有效增强模型的时序建模能力.

5.4 ME模块的位置

本实验旨在探究运动激励模块ME在微运动瓶颈块M-Ghost中的最佳插入位置,以优化运动特征和时空特征的提取效果.ME模块有三种不同的插入位置,分别位于两个Ghost模块之前、之间和之后,如图10所示.

表3显示了后端为TransMS-TCN时,不同ME位置下的模型识别性能.由表3可知,ME模块位于Ghost模块之间比位于其他位置的准确率高,但参数量稍大.ME模块位于Ghost模块之前的准确率仅为88.61%.这是由于第一个Ghost模块的主要作用是扩充通道的维度.第一个Ghost块之后插入ME模块会激励更多的运动敏感通道,增强模型对运动信息的感知能力.综上,本文将ME模块插入两个Ghost模块之间,充分利用第一个Ghost模块的通道维度扩充特性,增强模型对唇部



(*a*) 位于两个Ghost模块之前 (*b*) 位于两个Ghost模块之间 (*c*) 位于两个Ghost模块之后

图10 ME模块的位置

细微运动的感知能力.

表3 ME模块的位置对识别性能的影响

ME模块的位置	参数量/M	准确率/%
两个Ghost模块之前	**29.567**	88.61
两个Ghost模块之间	30.176	**88.83**
两个Ghost模块之后	29.570	88.75

5.5 DSTE的形式和数量

本实验考查不同形式和数量的DSTE对模型性能的影响.

(1)不同形式DSTE块的实验. 基础模型使用未经ME增强的GhostNetV2作为前端,TransMS-TCN作为后端. 由表4的实验结果可知,经顺序或并行DSTE块增强的模型比基础模型的识别精度高,说明DSTE可提高模型初期的时空特征提取能力. 顺序DSTE块与并行DSTE块的参数量均为29.743 M,但顺序DSTE块增强的模型比并行DSTE块增强的模型的精度高0.18%,这是因为顺序DSTE块的有序时空特征提取操作增强了块内不同层的信息交互能力,而并行DSTE块同层级的并列操作在一定程度上限制了时空特征的交互和整合. 本文选择顺序DSTE块构建浅层时空特征接收区,以有序提取时空特征.

(2)堆叠数量实验. 基础模型使用M-GhostNet作为前端,TransMS-TCN为后端. 由表5可知,当顺序DSTE块的堆叠数量为1时,模型性能未明显提升;当堆叠数量为2时,模型的精度明显提升;当堆叠数为3时,模型的参数量和准确率几乎与堆叠数量为2时持平. 这说明,一味增加堆叠数量不能明显提升模型的识别效果,反而增加了模型复杂度. 综上,本文使用2个顺序DSTE块增强模型的时空特征表达能力.

表4 DSTE的形式对识别性能的影响

DSTE形式	堆叠数量	参数量/M	准确率/%
无DSTE	0	**29.546**	88.59
并行DSTE	2	29.743	88.72
顺序DSTE	2	29.743	**88.90**

表5 DSTE的堆叠数量对识别性能的影响

DSTE形式	堆叠数量	参数量/M	准确率/%
无DSTE	0	**30.176**	88.83
顺序DSTE	1	30.275	88.89
顺序DSTE	2	30.374	89.21
顺序DSTE	3	30.472	**89.22**

5.6 消融实验

为探究DMT-GhostNet各模块对模型识别性能的影响,在前端网络为GhostNetV2,后端网络为MS-TCN的Baseline上进行消融实验. 在LRW数据集上的实验结果如表6所示. 由表可知,集成瓶颈块排列、TransMS-TCN、M-Ghost、DSTE这4个改进的DMT-GhostNet,比Baseline的准确率提高2.21%,参数量仅增加1.039 M,在相对较少的参数增长下提高了识别准确率.

“Baseline+瓶颈块排列”模型在Baseline上改变Ghost瓶颈块的排列方式,与原始GhostNet网络相比,参数量没有明显变化,但识别准确率提高了1.02%,表明Ghost瓶颈块排列方式的改变能够缓解过早下采样导致关键特征丢失的问题,有助于网络捕获更丰富的特征. “Baseline+瓶颈块排列+TransMS-TCN”将后端网络由MS-TCN改为TransMS-TCN,获得了0.57%的准确率提升,表明Transformer编码器与MS-TCN的结合在唇语识别中具备优势. “Baseline+瓶颈块排列+TransMS-TCN+M-Ghost”进一步将动作激励模块ME引入Ghost模块,比未引入ME模块的“Baseline+瓶颈块排列+TransMS-TCN”模型和“基础模型+瓶颈块排列”模型

表6　消融实验

模型	瓶颈块排列	TransMS-TCN	M-Ghost	DSTE	参数量/M	准确率/%
Baseline	—	—	—	—	**29.335**	87.00
Baseline+瓶颈块排列	√	—	—	—	29.397	88.02
Baseline+瓶颈块排列+TransMS-TCN	√	√	—	—	29.546	88.59
Baseline+瓶颈块排列+TransMS-TCN+M-Ghost	√	√	√	—	30.176	88.83
DMT-GhostNet	√	√	√	√	30.374	**89.21**

的准确率分别提高了0.24%和0.81%,说明M-Ghost模块能够更准确地捕获和分析唇部的微小运动. DMT-GhostNet在前述模型的基础上进一步加入DSTE块,准确率提高了0.38%,而参数量仅增加了0.198 M,表明DSTE块在增加少量参数的情况下,有效提高了时空信息提取能力. 综上所述,DMT-GhostNet能够更充分地提取语义特征、更高效地建模动态时间序列和更准确地捕获唇部的细微运动.

5.7　与主流唇语识别方法的对比

为验证DMT-GhostNet的性能,本实验对比近5年的主流唇语识别方法,结果如表7所示.

由表7可知,DMT-GhostNet的准确率明显高于其他对比方法,而参数量适中. 由表7还可以看出:

(1)大多数方法仅使用简单3D卷积处理时间维度的信息,这严重限制了模型的时序信息提取能力;

(2)ResNet为主流的前端时空特征提取网络,一些方法尝试使用GhostNet作为视觉前端,然而,多数方法因未充分关注视频帧的运动信息而难以捕获唇部的细微运动;

(3)后端时序建模模块存在多种变体,如Conv-BLSTM、Bi-GRU、MS-TCN、Transformer等,表7所列的对比方法均独立使用这些变体作为后端网络,它们未能深入挖掘不同变体在时序建模上的互补性.

表7　主流唇语识别方法对比

方法	年份	特征提取网络	时序建模网络	参数量/M	准确率/%
Multi-grained[27]	2019	ResNet-34+DenseNet3D	Conv-BLSTM	331.603	83.34
SE+MixUp[3] + Cosine LR+LS	2020	3D Conv+SE-ResNet-18	Bi-GRU	—	85.00
Temporal Convolution [7]	2020	3D Conv+ResNet-18	MS-TCN	36.361	85.30
Multi-Stage Distillation [5]	2020	3D Conv+ShuffleNetV2	MS-TCN	28.800	85.50
Dense Temporal Convolution [20]	2021	3D Conv+ResNet-18	DC-TCN	—	88.36
同类自知识蒸馏[14]	2022	3D Conv+iGhostneck	Bi-GRU	38.723	87.30
解耦同类自知识蒸馏[28]	2022	3D Conv+TSM+Ghostneck	—	**20.310**	85.20
Variational Temporal Mask[29]	2022	3D Conv+SE-ResNet-18	VTM+Transformer	—	85.18
self-attention+self-distillation[30]	2023	3D Conv+ResNet-18+Resformer	Transformer	—	85.25
DMT-GhostNet	2023	3D conv+DSTE+GhostNet+ME	Transformer+MS-TCN	30.374	**89.21**

5.8　仿真实验

为深入验证DMT-GhostNet在真实场景下的性能,录制了5名不同说话者的视频片段. 说话者分别录制单词“there”“pretty”“about”“phone”和“costumers”,其中,“there”为视觉歧义词汇,并选择轻微嘈杂环境中发音的“pretty”,略带口音发音的“costumers”,以更接近真实环境. 经数据预处理后输入DMT-GhostNet,识别结果如表8所示. 结果显示,DMT-GhostNet成功识别了这5个单词. 这表明,DMT-GhostNet在真实场景中具有较强的识别能力,即使在面对口音差异、环境噪声等挑战时也能准确识别.

表8　模拟场景验证实验结果

单词	识别结果
there	there
pretty	pretty
about	about
phone	phone
costumers	costumers

6　困难样本识别

本实验将基线方法[7]在LRW数据集上的测试结果划分为4个难度,统计各难度下被正确识别的单词数. 识别难度的划分依据为:每种方法的识别准确率在90%至100%的单词定义为简单单词,80%至90%为中等难度单词,60%至80%为困难单词,低于60%的单词视为极难单词. 然后,考查DMT-GhostNet对极难单词的识别能力.

6.1　不同识别难度的单词分布

为进一步验证DMT-GhostNet的识别能力,对比

ResNet+MS-TCN[7]、改变堆叠方式的 GhostNet+MS-TCN 和 DMT-GhostNet 三种方法在 4 种难度单词上的识别能力，如图 11 所示.

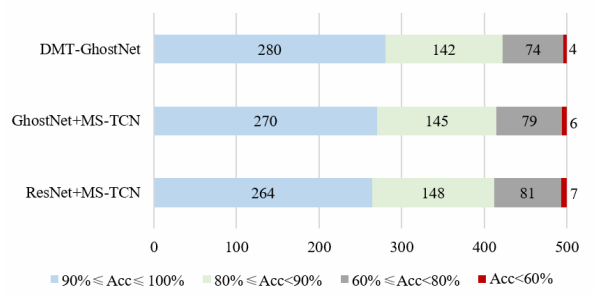


图 11 不同正确率区间的单词数量分布

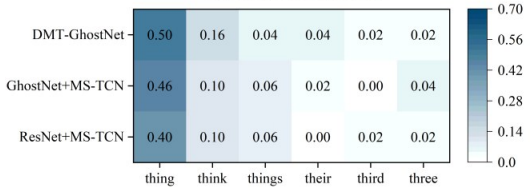
由图 11 可知，DMT-GhostNet 将超过一半单词的识别准确率提升至 90% 以上，对比 ResNet+MS-TCN 和 GhostNet+MS-TCN 方法，被定义为简单单词的数量分别增加了 16 种和 10 种. 此外，DMT-GhostNet 使 422 种单词的识别准确率达到了 80% 以上，超越了其他两种方法. DMT-GhostNet 在困难单词和极难单词的辨识上具备优势，500 种单词中有 74 种单词被定义为困难单词，仅 4 种单词被定义为极难单词. 相比另外两种方法，定义为困难单词和极难单词的数量明显降低.

6.2 极难单词识别能力对比

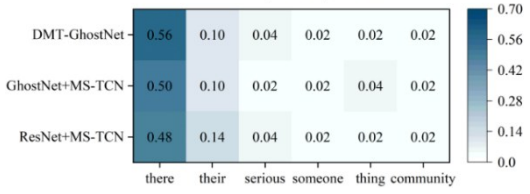
为验证 DMT-GhostNet 识别极难单词的能力，选择了 ResNet+MS-TCN 认为极难的 7 种单词，与 GhostNet+MS-TCN 和 DMT-GhostNet 方法进行比较. 这三种方法针对每种极难单词识别出的前 6 个单词几乎一致，图 12(a)~(g)分别展示 7 种单词识别率最高的前 6 个结果，并按照 DMT-GhostNet 的识别结果进行排序. 图 12(k)给出了 7 种单词在三种方法下的最高识别率.

由图 12 可以观察到：

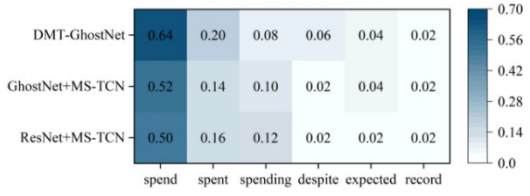
(1) DMT-GhostNet 的识别准确率明显高于 GhostNet+MS-TCN 和基线方法(ResNet+MS-TCN)，这证明 DMT-GhostNet 能有效区分视觉歧义词汇，具有较强的



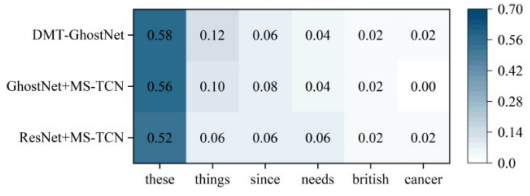
(a) “thing”的识别概率



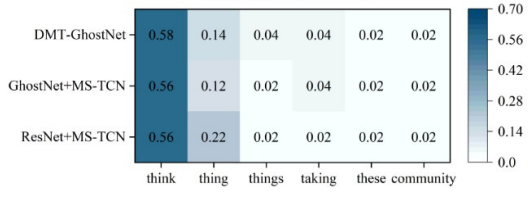
(b) “there”的识别概率



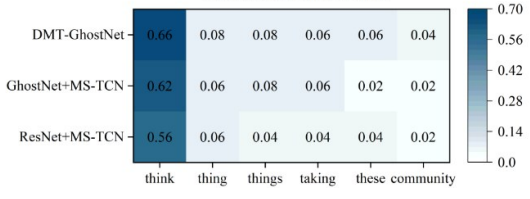
(c) “spend”的识别概率



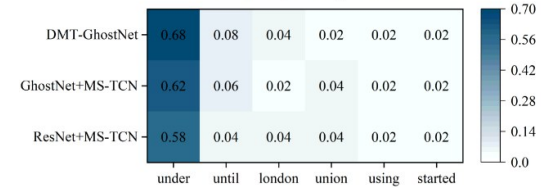
(d) “these”的识别概率



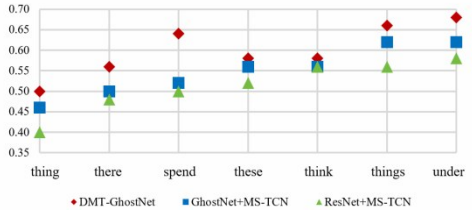
(e) “think”的识别概率



(f) “things”的识别概率



(g) “under”的识别概率



(h) 7 种单词在 3 种方法下的最高识别率

图 12 极难样本识别能力对比

困难样本识别能力;除“think”外,GhostNet+MS-TCN在其他单词上的识别率也均高于基线方法,说明调整排列方式后的GhostNetV2比ResNet更能有效地捕获时空特征.

(2)ResNet+MS-TCN与DMT-GhostNet极难识别的单词不一致.基线方法极难识别的单词“spend”被DMT-GhostNet识别为困难单词.可见基线方法极难识别的单词,DMT-GhostNet也能以较高的准确率识别.

(3)各子图的第2列的误识别率最高,第3~6列的误识别率明显降低.此外,这3种方法的误识率最高的单词基本相同.例如在3种方法下,“under”均最多被误识为“until”,“thing”均被误识为“think”,“there”被误识为“their”.这说明,极难单词之所以难以识别,是因为其与某一单词的发音和唇部动作高度相似,增加了识别难度.

因此,与其他方法相比,DMT-GhostNet对极难单词的识别能力更强.

7　结论

针对现有唇语识别方法对微小唇部运动的关注能力和时序特征捕获能力不足的问题,提出DMT-GhostNet,通过引入解耦时空增强块(DSTE)、微运动瓶颈块(M-Ghost)和时间感知模块(TransMS-TCN),有效强化模型对时序信息的感知能力和对微小运动的捕获能力.模型在LRW数据集上的测试与评估结果表明,DMT-GhostNet可以准确捕捉唇部微小运动和优化关键时序信息建模能力,显著提高唇语识别性能.

参考文献

[1] 姚鸿勋, 高文, 王瑞, 等. 视觉语言: 唇读综述[J]. 电子学报, 2001, 29(2): 239-246.
YAO H X, GAO W, WANG R, et al. A survey of lipreading—One of visual languages[J]. Acta Electronica Sinica, 2001, 29(2): 239-246. (in Chinese)

[2] 陈雁翔, 刘鸣. 基于发音特征的音视频说话人识别鲁棒性的研究[J]. 电子学报, 2010, 38(12): 2920-2924.
CHEN Y X, LIU M. Research on robustness of audio-visual speaker recognition based on articulatory features[J]. Acta Electronica Sinica, 2010, 38(12): 2920-2924. (in Chinese)

[3] FENG D L, YANG S, SHAN S G, et al. Learn an effective lip reading model without pains[EB/OL]. (2020) [2023]. http://arxiv.org/abs/2011.07557.

[4] STAFYLAKIS T, TZIMIROPOULOS G. Combining residual networks with LSTMs for lipreading[EB/OL]. (2017)[2023]. http://arxiv.org/abs/1703.04105.

[5] MA P C, MARTINEZ B, PETRIDIS S, et al. Towards practical lipreading with distilled and efficient models[C]// ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Piscataway: IEEE, 2021: 7608-7612.

[6] 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: IEEE, 2018: 6548-6552.

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

[8] LI Y, JI B, SHI X T, et al. TEA: Temporal excitation and aggregation for action recognition[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 906-915.

[9] JIANG B Y, WANG M M, GAN W H, et al. STM: Spatiotemporal and motion encoding for action recognition[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2019: 2000-2009.

[10] WANG Z W, SHE Q, SMOLIC A. ACTION-net: Multipath excitation for action recognition[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2021: 13209-13218.

[11] 申屠敏健, 朱强, 朱树元, 等. 基于视频先验信息的轻量化去噪卷积神经网络[J]. 电子学报, 2023, 51(6): 1510-1517.
SHENTU M J, ZHU Q, ZHU S Y, et al. A priori information-based lightweight convolutional neural network for video denoising[J]. Acta Electronica Sinica, 2023, 51(6): 1510-1517. (in Chinese)

[12] 张淑军, 彭中, 李辉. SAU-Net: 基于U-Net和自注意力机制的医学图像分割方法[J]. 电子学报, 2022, 50(10): 2433-2442.
ZHANG S J, PENG Z, LI H. SAU-net: Medical image segmentation method based on U-net and self-attention[J]. Acta Electronica Sinica, 2022, 50(10): 2433-2442. (in Chinese)

[13] HAN K, WANG Y H, TIAN Q, et al. GhostNet: More features from cheap operations[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2020: 1577-1586.

[14] 马金林, 刘宇灏, 马自萍, 等. HSKDLR: 同类自知识蒸馏的轻量化唇语识别方法[J]. 计算机科学与探索, 2023, 17(11): 2689-2702.
MA J L, LIU Y H, MA Z P, et al. HSKDLR: Lightweight lip reading method based on homogeneous self-knowledge distillation[J]. Journal of Frontiers of Computer Science and Technology, 2023, 17(11): 2689-2702. (in Chinese)

[15] ZHANG G Y, LU Y Y. Research on a lip reading algorithm based on efficient-GhostNet[J]. Electronics, 2023, 12(5): 1151.

[16] WANG Q L, WU B G, ZHU P F, et al. ECA-net: Efficient channel attention for deep convolutional neural networks[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE,

2020: 11531-11539.
[17] TANG Y H, HAN K, GUO J Y, et al. GhostNetv2: Enhance cheap operation with long-range attention[C]//Advances in Neural Information Processing Systems 35. New Orleans: NeurIPS, 2022: 9969-9982.
[18] ALANSARI M, HAY O A, JAVED S, et al. GhostFaceNets: Lightweight face recognition model from cheap operations[J]. IEEE Access, 2023, 11: 35429-35446.
[19] WANG H B, HAN J. Research on military target detection method based on YOLO method[C]//2023 IEEE 3rd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA). Piscataway: IEEE, 2023: 1089-1093.
[20] MA P C, WANG Y J, SHEN J, et al. Lip-reading with densely connected temporal convolutional networks[C]//2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE, 2021: 2856-2865.
[21] 周登文, 李文斌, 李金新, 等. 一种轻量级的多尺度通道注意图像超分辨率重建网络[J]. 电子学报, 2022, 50(10): 2336-2346.
ZHOU D W, LI W B, LI J X, et al. Image super-resolution reconstruction based on lightweight multi-scale channel attention network[J]. Acta Electronica Sinica, 2022, 50(10): 2336-2346. (in Chinese)
[22] QIU Z F, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 5534-5542.
[23] FENG J F, LONG R H. Cross-language lipreading by reconstructing Spatio-Temporal relations in 3D convolution[J]. Displays, 2023, 76: 102357.
[24] CHUNG J S, ZISSERMAN A. Lip reading in the wild[C]//Computer Vision - ACCV 2016: 13th Asian Conference on Computer Vision. Cham: Springer International Publishing, 2017: 87-103.
[25] KING D E. Dlib-ml: A machine learning toolkit[J]. Journal of Machine Learning Research, 2009, 10: 1755-1758.
[26] CHEN J R, KAO S H, HE H, et al. Run, don´t walk: Chasing higher FLOPS for faster neural networks[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2023: 12021-12031.
[27] WANG C H. Multi-grained spatio-temporal modeling for lip-reading[EB/OL]. (2019)[2023]. http://arxiv.org/abs/1908.11618.
[28] 马金林, 刘宇灏, 马自萍, 等. 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. DOI: 10.13700/j.bh.1001-5965.2022.0931.
MA J L, LIU Y H, MA Z P,et al. Lightweight lip recognition method based on decoupling homogeneous self-knowledge distillation[J]. Journal of Beijing University of Aeronautics and Astronautics. DOI: 10.13700/j.bh.1001-5965.2022.0931.(in Chinese)
[29] SHENG C C, LIU L, DENG W X, et al. Importance-aware information bottleneck learning paradigm for lip reading[J]. IEEE Transactions on Multimedia, 2023, 25: 6563-6574.
[30] XUE J X, HUANG S B, SONG H W, et al. Fine-grained sequence-to-sequence lip reading based on self-attention and self-distillation[J]. Frontiers of Computer Science, 2023, 17(6): 176344.

作者简介

马金林 男, 1976 年 3 月出生, 宁夏青铜峡人. 现为北方民族大学计算机科学与工程学院副院长、副教授、硕士生导师. 主要研究方向为唇语识别、人工智能、计算机视觉与图像处理.
E-mail: majinlin@nmu.edu.cn

吕 鑫 女, 1998 年 8 月出生, 山西汾阳人. 2024 年毕业于北方民族大学计算机科学与工程学院. 主要研究方向为唇语识别、计算机视觉.
E-mail: 20217433@stu.nmu.edu.cn

马自萍 女, 1977 年 1 月出生, 宁夏吴忠人. 现为北方民族大学数学与信息科学学院副教授、硕士生导师. 主要研究方向为计算机视觉、图像处理.
E-mail: 2006041@nmu.edu.cn

郭兆伟 男, 1997 年 10 月出生, 甘肃庆阳人. 2024 年毕业于北方民族大学计算机科学与工程学院. 主要研究方向为唇语识别、计算机视觉.
E-mail: 20217417@stu.nmu.edu.cn

吕 科 男, 1971 年出生, 宁夏人. 现为中国科学院大学计算机与通信工程学院教授、博士生导师. 主要研究方向为计算机视觉、3D 图像建模、计算机图形学.
E-mail: luk@ucas.ac.cn