**邮寄地址：广东省珠海市香洲区北京理工大学教师公寓二期3栋904房，陈灵 收，联系电话：15013355565**

**基于多模态融合的端到端中文唇读识别研究**

**陈焯辉1 林绰雅2 刘奕显2 王茗琛2 梁思敏2 陈灵\***

1. **澳门科技大学，澳门，999078**
2. **北京理工大学珠海学院，广东，珠海，519088**

**摘要：**为了更好地帮助健全人与听障或语言障碍人士进行交流，构建无障碍社会。本文构建了一个基于多模态融合的端到端音视频识别系统，实现中文唇语翻译功能。实验结果表明，将所提出的端到端视听语音识别结构体系应用于唇语识别模型，实现了8.0%的字符错误率。与之前的唇语识别模型相比，它在融合图像特征和音频特征方面表现出了良好的性能。

**关键词：**端到端音视觉语音识别结构体系；多模态融合；唇语识别

**Research on End-to-end Chinese Lip Reading Recognition Based on Multimodal Fusion**

**Chen Zhuohui1, Lin Chuoya2, Liu Yixian2, Wang Mingchen2, Liang Simin2, Chen Ling\*2**

1. **Macau University of Science and Technology, Macau 999078, China**
2. **Beijing Institute of Technology, Zhuhai 519088, China**

**Abstract:** To better help able-bodied people, communicate with hearing-impaired or speech-impaired people, and build a barrier-free society. This paper constructs an end-to-end audio and video recognition system based on multi-modal fusion to realize the translation function of Chinese lip language. Experimental results show that applying the proposed End-to-end Visual Speech Recognition Structure System to the lip recognition model achieves a character error rate of 8.0%. Compared with previous lip recognition models, it shows good performance in fusing image features and audio features.

**Keywords:** end-to-end visual speech recognition structure system; multi-modal fusion; lip recognition

**1.作者简介：**

**陈焯辉，男（1999.11-)，汉族，籍贯，广东广州，硕士，研究方向：人工智能**

**林绰雅，女（2000.12-)，汉族，籍贯，广东广州，本科，研究方向：数据科学与大数据技术**

**刘奕显，男（2000.04-)，汉族，籍贯，广东阳江，本科，研究方向：数据科学与大数据技术**

**王茗琛，男（2001.06-)，汉族，籍贯，广东湛江，本科，研究方向：数据科学与大数据技术**

**梁思敏，女（2001.02-)，汉族，籍贯，广东中山，本科，研究方向：数据科学与大数据技术**

**2.\*通讯作者：陈灵\*（1962.01-），汉族，教授，研究方向：人工智能**

**项目支持：2022年度广东省大学生创新创业训练项目——基于中文唇语翻译的听障人群无障碍交流系统（S202213675010）**

1. **引言**

音视觉语音识别是一种语音识别任务，它利用嘴唇运动的音频和视觉来进行识别翻译。 2018年，Afouras等人[1]正式提出第一个现代视听语音识别系统，他们使用了多种深度神经网络模型。他们[2]提出了Connectionist Temporal Classification Transformer (CTC Transformer)模型和Sequence-to-sequence Transformer (TM-seq2seq)模型，并对两种模型进行了比较。2019年，Makino等人[3]提出了一种基于递归神经网络转换器（RNN-T）结构的语音识别模型。Ma等人[4]在2021年提出了一种基于残差网络（Residual Network, ResNet）和卷积增强变换器（Conformer）的混合注意力机制模型，利用CTC和注意力机制的结合来学习识别字符。

本文提出的无障碍通信系统采用端到端视听语音识别架构（End-to-end Audio-visual Speech Recognition Architecture，EASRA），将视频拆分为音频波和图像序列输入 将视频内容转换为文本内容的模型。本文采用视听语音识别技术，利用视听语音识别解码器对音频特征进行解码。视觉语音识别解码器对视觉特征进行解码，然后在编码器的输出层利用自注意力寻找与当前听觉向量相关的视觉向量，然后与听觉向量进行拼接。利用视听语音识别技术对音视频内容进行处理后，可以在单模态或双模态的情况下完成文本翻译任务。

1. **中文普通话唇读数据集**
   1. **数据介绍**

本文使用的CMLR（Chinese Mandarin Lip Reading）数据集，它旨在促进视觉语音识别的研究。CMLR数据集来源于中央电视台2009年6月至2018年6月的新闻联播视频。该数据集共包含11位主机表达的102076个句子，每个句子最多包含29个汉字，不包括英文字母、阿拉伯数字和罕见的标点符号。此外，训练集、验证集和测试集按7:1:2的比例随机划分。详情如表1：

表1 数据集情况

|  |  |
| --- | --- |
| 数据集 | 句子个数 |
| 训练集 | 71448 |
| 验证集 | 10206 |
| 测试集 | 20418 |
| 总数 | 102072 |

* 1. **数据预处理**

数据集分为无声视频文件、有声音频文件和时间戳文件。为了使CMLR数据集方便后续的模型训练，需要对数据集进行预处理。具体处理流程如图1：

电视游戏的萤幕截图

描述已自动生成

图1 数据预处理流程

本文读取无声视频文件和有声音频文件的数据，利用FFMPEG软件将视频特征数据和音频特征数据结合生成有声视频文件。

1. **基于多模态融合的端到端中文唇读识别网络**

音视觉语音识别是一项多模态任务，从音频和视觉流中转录文本，同时利用人声的直观输入和嘴唇运动的视觉输入来完成唇读识别任务。在本文构建的听障人士无障碍交流系统中，识别模型采用的端到端视听语音识别结构体系如图2：

图片包含 图形用户界面

描述已自动生成

图2 端到端的音视觉语音识别结构体系

其中，视觉和音频特征提取模块结构均包括卷积前端和编码器后端。卷积前端的作用是从图像序列或音频波形中提取特征，然后通过编码器后端对特征进行编码，生成两种特征信息，通过融合模块融合视觉和音频特征，得到融合特征用于计算训练过程中的交叉熵损失和CTC损失，从而实现模型的反向传播算法。

* 1. **前端模块**

本文使用的骨架基于流线型的MobileNet[5]架构。该架构是一种轻量级的卷积神经网络架构，相比于传统的卷积神经网络，MobileNet具有更少的参数和更小的模型体积，因此可以在移动设备等资源受限的环境下进行实时图像识别和分类任务。MobileNet已经被广泛应用于移动设备、嵌入式设备等场景，可以用于图像分类、目标检测、人脸识别等任务。

* 1. **编码器后端模块**

本文中Conformer编码器作为后端模块，Conformer编码器用于在视觉和听觉时间建模的后端提取图像序列和音频波形的特征。Converter编码器首先对输入的特征进行简单的数据增强，然后对卷积层进行下采样，以增强数据的时序。单个Conformer编码模块的内部结构采用“夹心式”结构，将卷积模块和自注意力模块夹在两个前馈模块之间。

* 1. **多层感知机融合模块**

融合模块采用多层感知器（Multi-Layer Perceptron, MLP），通过前端模块提取的特征将图像序列和音频波形进行融合，投射到高维空间中。多层感知器由两个全连接层、批标准化层和高斯误差线性单元激活函数组成。

* 1. **解码器模块**

融合模块之后使用的解码器是基于Transformer[7]调整的解码器，由一个嵌入层、多头自注意力模块和前馈神经网络组成。本文生成的输出序列是指当前预测时刻的词向量，嵌入了生成的输出序列和相对位置编码。在多头自注意力模块中，主要由一个带掩码多头注意力机制、多头注意力机制和前馈神经网络组成。

1. **实验与结果分析**

为了验证端到端的音视觉语音识别结构体系的识别效果，本文选取了5种不同的模型在CMLR数据集上的字符错误率（CER）进行比较，字符错误率的计算公式如下：

 （1）

其中，*S*表示替换数，*D*表示删除数，*I*表示从预测序列到标准序列的插入数，*N*表示预测序列中的单词数。字符错误率的比较结果如表2：

表2 不同唇语识别模型在CMLR数据集上的性能比较

|  |  |  |
| --- | --- | --- |
| 模型 | 数据集 | 字符错误率/% |
| Watch Listen Attend and Spell[8] | CMLR | 38.93 |
| Lipreading Chinese Network[9] | CMLR | 34.07 |
| CSSMCM[10] | CMLR | 32.48 |
| Lip by Speech[11] | CMLR | 31.27 |
| CTC & Attention[12] | CMLR | 9.1 |
| EASRA | CMLR | **8.0** |

可以看出，本文使用端到端的音视觉语音识别结构体系的字符错误率为8.0，是六种模型里效果最好的。本文使用的端到端的音视觉语音识别结构体系比之前以往的唇读模型更好，在融合图像特征和音频特征表现的性能更好，并且能够很好的完成中文的唇读任务。

1. **结论**

为了更好地帮助健全人与听障或语言障碍人士进行交流，构建无障碍社会，本文实现了基于多模态融合端到端音频的汉语唇语翻译功能和视频识别系统。 通过实验得出结论，将所提出的端到端视听语音识别结构体系应用于唇语识别模型时，可以取得更好的效果。

**参考文献**

1. Afouras T, Chung J S, Zisserman A, Deep Lip Reading: A Comparison of Models and an Online Application[J]. 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018.
2. Afouras T, Chung J S, Senior A, et al. Deep Audio-visual Speech Recognition[J]. 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018.
3. Makino T, Liao H, Assael Y, et al. Recurrent Neural Network Transducer for Audio-Visual Speech Recognition[J]. 2019 IEEE Automatic Speech Recognition and Understanding Workshop, 2019, 905-912.
4. Ma P, Petridis S, Pantic M, et al. End-to-end Audio-visual Speech Recognition with Conformers[J]. 2021 IEEE Conference on Computer Vision and Pattern Recognition, 2021, 7613-7617.
5. Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted Residuals and Linear Bottlenecks[J]. 2018 IEEE Conference on Computer Vision and Pattern Recognition, 2018, 4510-4520.
6. Gulati A, Qin J, Chiu C, et al. Conformer: Convolution-augmented Transformer for Speech Recognition[J]. 2020 Interspeech, 2020, 5036-5040.
7. Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need[J]. 2017 Neural Information Processing Systems, 2017, 6000-6010.
8. Chung J S, Senior A, Vinyals O, et al. Lip Reading Sentences in the Wild[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, 3444-3453.
9. Zhang X, Gong H, Dai X, et al. Understanding Pictograph with Facial Features: End-to-end Sentence-level Lip Reading of Chinese[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2019, 33(01): 9211-9218.
10. Zhao Y, Xu R, Song M. A Cascade Sequence-tosequence Model for Chinese Mandarin Lip Reading[J]. Proceedings of the ACM International Conference on Multimedia, 2019, 1-6.
11. Zhao Y, Xu R, Wang X C, et al. Hearing Lips: Improving Lip Reading by Distilling Speech Recognizers[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 6917-6924.
12. Ma P, Petridis S, Pantic M. Visual Speech Recognition for Multiple Languages in the Wild[J]. 2022 IEEE Conference on Computer Vision and Pattern Recognition, 2022.