**Research on End-to-end Chinese Lip-Reading Recognition**

*Abstract*

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唇读识别架构(End-to-end Lip-Reading Recognition Architecture, ELRA)，实现中文唇语视频翻译功能。实验结果表明，将所提出的ELRA应用于CMLR数据集上，实现8.0%的字符错误率。与之前的唇语识别模型相比，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：**端到端唇读识别架构；多模态融合；唇语识别

1. Introduction

每一种信息源都可以称为一种模态。模态是一个人接收信息的方式，人有多种感知方式，如听、看、闻、触摸；当一个人仅通过嗅觉和触摸来准确地感知事物时，获取的信息存在一定程度损失。如今人们的日常生活中，视频是人们接触多媒体的主流，其包含文本、音频、视觉上的信息。因此，多模态学习已成为理解和分析多媒体内容中一个重要的工具。多模态学习由来自不同模块的信息组成，通常包含两种或多种模式，旨在共同表示来自不同模态的数据，捕捉内在的相关性，使每种模态的信息都能相互转换。尽管在缺乏某些获取信息的方式下，可以在交接中填补缺失的信息。多模态融合生物识别技术是一种结合了两种或多种生物特征的个人识别技术，也是一种数据融合技术，最大限度地利用每个生物提供的数据特征，使得最终的识别结果比单模态生物识别更加准确可靠，因此便展开了音视觉相关的多模态任务研究。

唇读识别，是利用说话者的嘴唇运动视觉特征和音频特征进行识别翻译。 2018年，Afouras等人[1]正式提出第一个现代视听语音识别系统，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)和Sequence-to-sequence Transformer (TM-seq2seq)，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）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）的混合注意力机制结构，利用CTC和注意力机制来学习识别字符。

本文提出的End-to-end Lip-Reading Recognition Architecture (ELRA)，其将有声视频拆分为音频波和图像序列输入，最终输入的视频内容转换为文本内容的架构。根据图像序列和音频波的数据形式，使用了视觉、音频前-后端对数据进行编码和解码从而提取数据特征，通过融合模块将两种不同形式的数据特征进行融合，最终通过解码和全连接层计算相关损失函数，从而实现模型反向传播以更新模型权重参数。训练完成后，利用该架构可以在无声视频或有声视频的情况下完成文本翻译任务。

1. Related work

本节首先概述了多模态融合研究，然后总结了唇语识别的相关工作。

1. *Multi-modal fusion*

多模态融合技术融合了听觉、视觉、嗅觉和触觉互动，可以更高效、更完整地呈现信息。由于多模态在描述物体特征方面的全面性，它在许多领域都有广泛的应用。

Truong等人[5]提出了视觉方面注意力网络，作为一种利用视觉数据进行情感分析的新技术。Le等人[6]设计了一种基于视频的对话系统，其中的对话取决于特定视频的视觉和听觉特征，因此比传统的基于图像或文本的对话系统更具挑战性。Cui等人[7]提出了一种用户注意力引导的多模态对话系统，利用结合不同模态信息的多模态对话格式，让用户更清楚地了解自己的表达。Zhang等人[8]提出了一种新的二维时空邻接网络，其核心思想是在二维时空图上检索时刻，将相邻的候选时刻视为时空上下文，这种模型可以扩展到其他时空定位任务中。浙江大学的Ya Zhao[9]提出了一种LIBS模型，将多模态视听识别纳入知识提炼结构，并计算帧、序列和文本层面的知识提取。

1. *Lip recognition*

为了进一步提高嘈杂环境下语音识别系统的准确性，研究人员开始尝试将不同模态的信息进行融合建模，以实现更高的识别率。视听语音识别（AVSR）系统就是在这种逐步探索的过程中诞生的。

在深度学习出现之前，大多数唇语识别都依赖于人工特征提取，而提取图像特征需要对许多帧进行预处理。Petajan等人[10]于1984年提出了第一个唇语识别系统，该系统使用传统的唇语识别方法获得唇语图片特征向量，然后计算数据库中单词的相似度，以相似度最高的单词作为输出。Goldschen等人[11]延续了Petajan等人的工作，受语音识别的启发，使用语音识别方法建立唇语识别模型，取得了良好的效果。Shaikh等人[12]提出了时空描述符和支持向量机（SVM）分类器，以促进唇语识别的发展。

21世纪之交，随着深度学习技术的出现，唇形识别算法获得了更大的发展空间。2011年，Ngiam等人[13]使用自动编码器和受限玻尔兹曼机（RBM）成为第一个基于深度学习的唇形识别算法。该方法通过将语音特征与图像特征相结合，融合不同模态的特征，提高了系统提取特征的能力。Wang等人[14]在2016年使用梯度直方图的特征作为长短期记忆（LSTM）网络的输入，但神经网络识别的准确率仅为79.6%。同年，Google的DeepMind团队与牛津大学合作开发了LipNet网络[15]，并取得了更加令人印象深刻的准确率。该算法采用时空图卷积网络（STGCN）、LSTM网络和联结时间分类（CTC）组成的结构来实现端到端的变长序列识别。 Chung和Zisserman [16]在一篇论文中解决了小数据集的问题，他们创建了500个单词的数据集LRW，并提出了一种结合卷积神经网络和循环神经网络的WLAS网络，将唇形识别技术与语音识别相结合，以提高识别能力在嘈杂环境中的速率，效果显着。2017年，Stafylakis等人[17]提出在时空产生式的基础上添加残差网络，并在序列建模部分使用双向LSTM来提高算法学习序列特征的能力。2018年，Afouras等人[18]在序列建模单元中使用了同样来自机器翻译领域的Transformer结构，但在特征提取部分仍然使用了时空卷积核残差网络的结构，算法取得了最高的识别精度 当时的任何唇部识别算法。从那时起，该领域一直处于快速发展阶段，大部分工作致力于架构改进。2019年，Zhang等人[19]正式提出时间焦点块和时空融合技术。该技术提出了用于描述短程关系的时间焦点块和用于保留局部空间信息并降低特征维度的时空融合模块（STFM）。同年，Shukla等人[20]首先探索了自监督学习在视听语音识别中的应用，其中使用跨模式设置从音频输入预测视频帧，即从音频输入预测嘴唇运动。2021年，Ma等人[21]首先使用Conformer声学模型和用于学习的混合CTC/注意力解码器实现了端到端LRS2。实验结果表明，新的前端在纯音频和纯视觉设置中均显着优于以前的前端，并在最终唇形识别方面取得了最新进展。

1. Methodology

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唇读识别结构如图1：

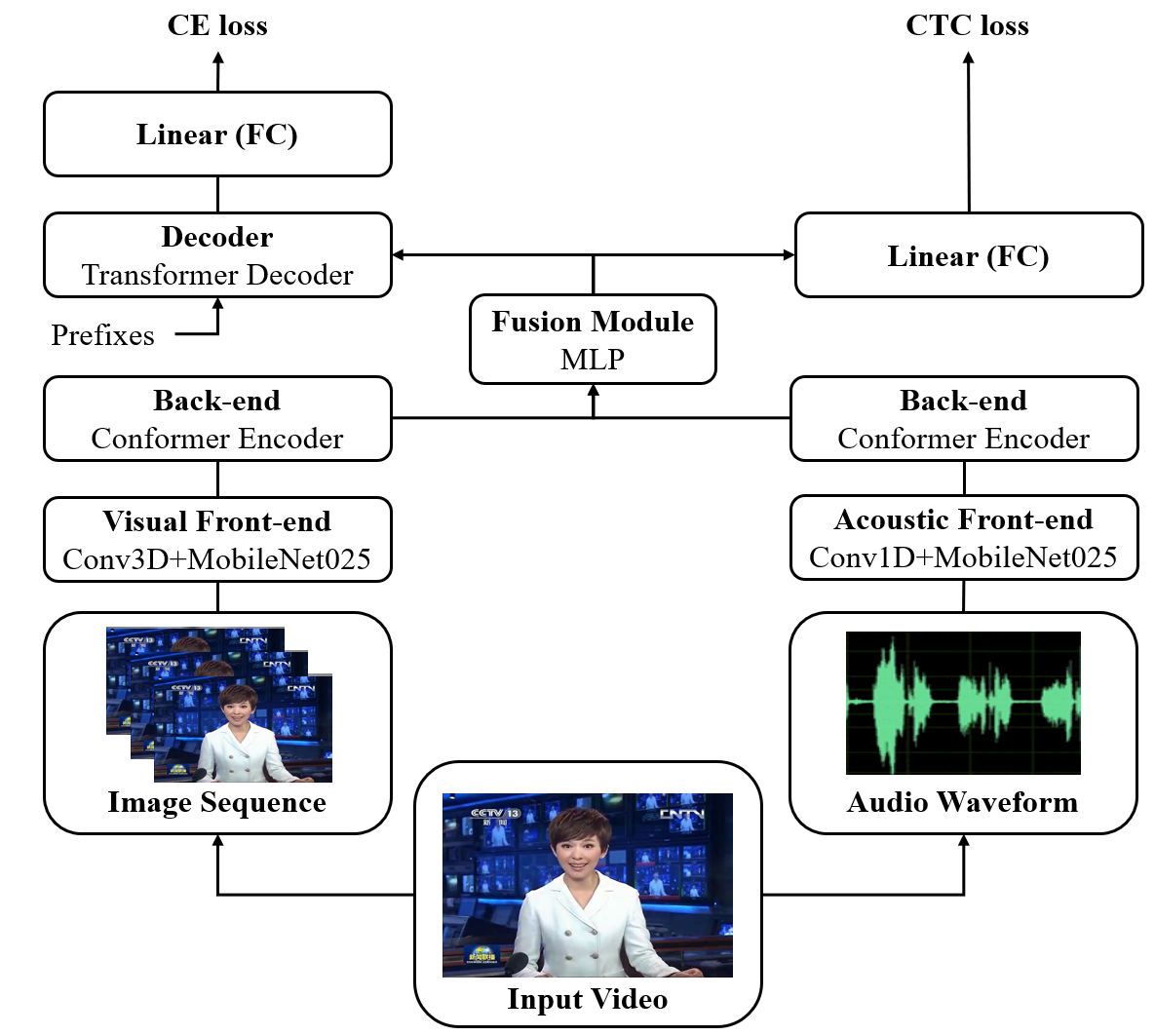


Figure 1. End-to-end Lip-reading Recognition Architecture.

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. *Front-end module*

视觉和音频提取模块的时间建模前端均采用卷积神经网络骨架的形式，在输出部分去除最后用于输出各概率的全连接层，使用剩下结构前向计算得到的多通道特征图完成任务。在前端模块中，将基于不同数据结构调整简化的MobileNet架构，该模型最初用于移动和嵌入式视觉中的图像分类应用。假设输入224×224的三通道图像，图2展示了模型的结构。

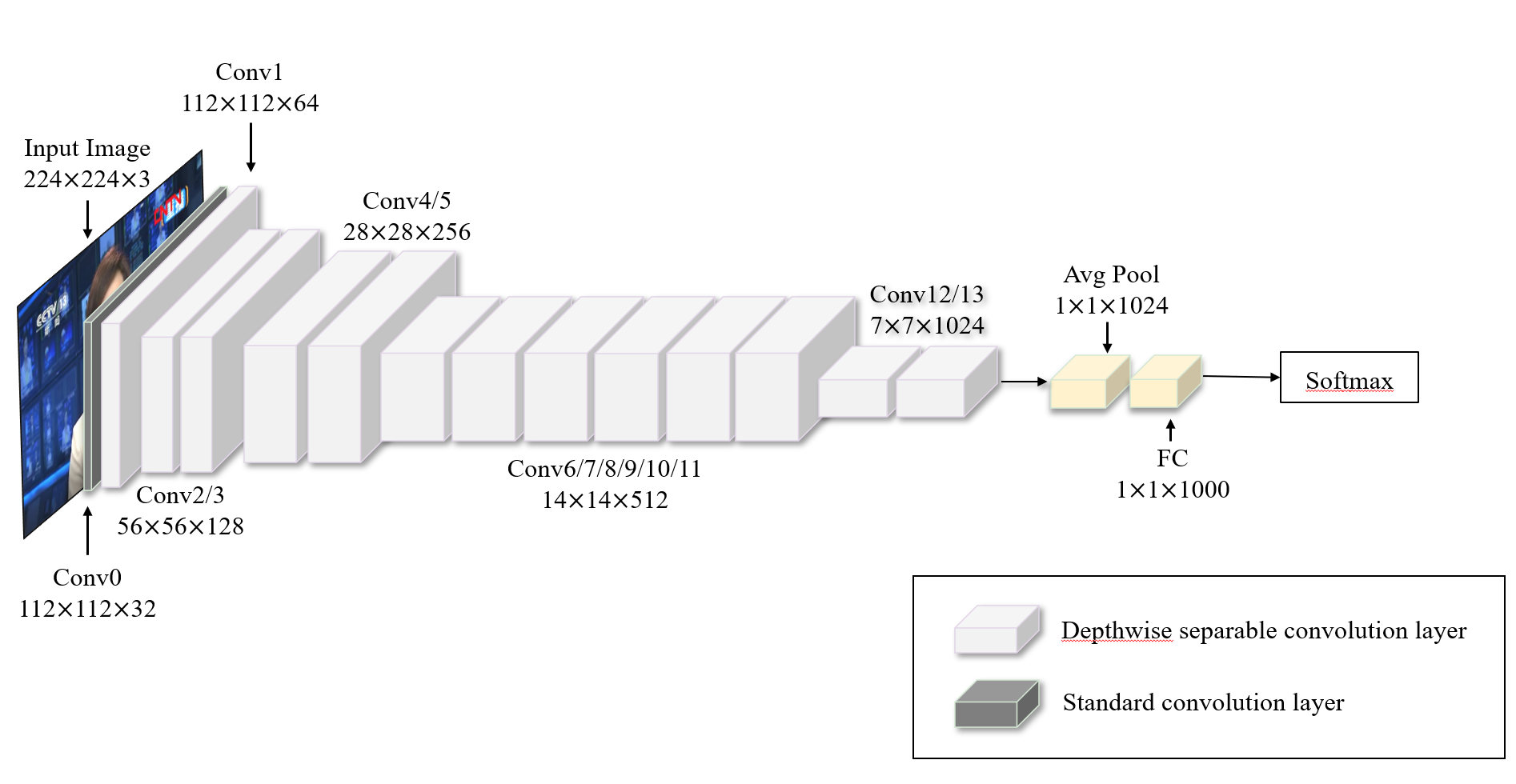


Figure 2. MobileNet original model structure.

其中，灰色方块代表标准卷积层，普通卷积层和标准化层的组合。紫色方块代表MobileNet基本单元的Depthwise Separable Convolution，由Depthwise Convolution和Pointwise Convolution组成。前者在卷积运算中对RGB图像的三个通道使用不同的卷积核，而后者则采用1×1卷积核的普通卷积。 该设计在减少计算量和模型参数的基础上达到了标准卷积的效果。 在实际应用中，将实现Batch Normalization（BN）和ReLU激活函数来加速训练的收敛。

上文提到，图像序列和音频波属于不同模态的数据，因此在视觉、音频前端模块的输入维度不同，因此需要对2D卷积层进行调整，调整后的MobileNet Backbone结构如表1。

表1. Model architecture for audio and visual front ends.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Unit** | **Layers** | **Input audio waveform** | **Stride** | **Layers** | **Input image Sequence** | **Stride** |
| C0 | Conv1D |  | 4 | Conv3D  MaxPool3D |  |  |
| C1 | Conv1D dw  Conv1D | dw | 1  1 | Conv2D dw  Conv2D | dw |  |
| C2 | Conv1D dw  Conv1D | dw | 2  1 | Conv2D dw  Conv2D | dw |  |
| C3 | Conv1D dw  Conv1D | dw | 1  1 | Conv2D dw  Conv2D | dw |  |
| C4 | Conv1D dw  Conv1D | dw | 2  1 | Conv2D dw  Conv2D | dw |  |
| C5 | Conv1D dw  Conv1D | dw | 1  1 | Conv2D dw  Conv2D | dw |  |
| C6 | Conv1D dw  Conv1D | dw | 2  1 | Conv2D dw  Conv2D | dw |  |
| C7 -C11 | Conv1D dw  Conv1D | dw | 1  1 | Conv2D dw  Conv2D | dw |  |
| C12 | Conv1D dw  Conv1D | dw | 1  1 | Conv2D dw  Conv2D | dw |  |
| C13 | Conv1D dw  Conv1D | dw | 2  1 | Conv2D dw  Conv2D | dw |  |

其中Conv1D dw表示1D深度卷积。 Conv2D dw 代表2D深度卷积。 对于视觉前端，MobileNet骨架的第一个卷积层（C0）被替换为核心尺寸的3D卷积层和核心尺寸的3D最大池层的组合，转换为（为batch size）变为，并将图像序列的时间维度整合到batch数量维度中。对于声学前端，由于音频波形是1D数据，所以整个网络需要将2D卷积修改为1D卷积，并将C0调整为80（5ms）核心大小的1D卷积，步长设置为4，以便最终的声学特征采样为每秒 25 帧，以匹配视觉特征的帧速率。 由于视觉特征提取不是视听识别的主要部分，因此引入宽度乘数来减少前端模块模型参数的数量，其取值范围为(0,1]。作用是最小化我们让，特征图的通道数从1024减少到256。如上所述，ELRA中的MobileNet0.25就是由此推导出来的。

1. *Back-end module*

Gulatiet等人[21]提出了一种新的架构，将自注意力机制和卷积集成到 ASR 模型中，称为 Conformer 编码器。 我们使用Conformer编码器在视觉和声学时间建模的后端提取图像序列特征和音频波形，其架构如图3所示。

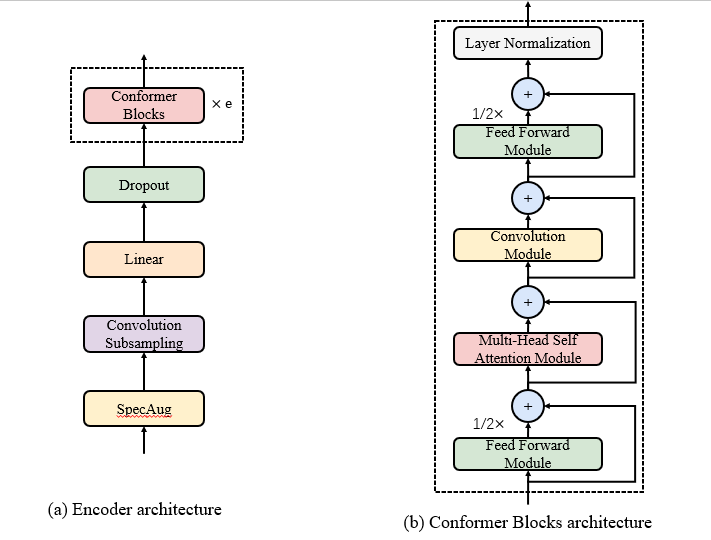


Figure 3. Encoder and Conformer architecture.

Conformer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. *Fusion module*

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，投射到维空间。MLP输出大小为的全连接层、BN层、ReLU激活函数和输出大小为的全连接层组成。

1. *Decoder*

融合模块之后使用的解码器是基于Transformer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，其结构如图4所示。

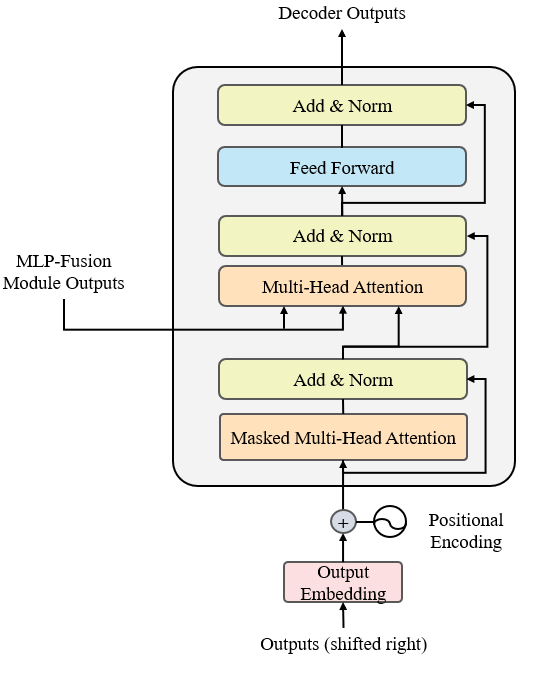


Figure 4. Transformer decoder architecture.

1. Experiments
2. *Loss function*

采用混合 CTC/attention 损失函数。假设 是融合模块中Transformer解码器的输入帧序列，是输出目标，其中和表示输入长度和目标长度。CTC 损失中，假设每个输出的预测值都是相互独立的，如公式 (1) 所示：

相比之下，基于注意力机制的模型消除了这个假设并估计后验基于链式法则的概率，如公式（2）所示：

由此，可以得到计算总损失的公式如式（3）：

其中是CTC损失函数的权重因子与混合CTC/attention中的注意力机制对比机制。重量不仅仅是两个损失的综合函数转化为单个训练损失，但也将两个解码过程中的预测和需求。

1. *Evaluation metric*

为了验证端到端唇读识别结构的识别效果，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式(4)：

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。

1. *CMLR Dataset*

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2。

Table 2. statistical information about the dataset

|  |  |  |  |
| --- | --- | --- | --- |
| **Dataset** | **Sentence** | **Phrase** | **Symbol** |
| Training | 71.448 | 22,959 | 3,360 |
| Validation | 10,206 | 10,898 | 2,540 |
| Test | 20,418 | 14,478 | 2,834 |
| **Total** | **102,072** | **25,633** | **3,517** |

1. *Results*

为了评估 ELRA 认可的有效性，我们选择了五个不同的模型来进行比较，它们是WAS、LipCH-Net、CSSMCM、LIBS 和 CTCH。 他们都是唇形识别方法。 WAS是该领域的经典方法句子级唇形识别，将用于直接识别汉字； LipCH-Net 和CSSMCM是中国人句子级唇识别模型； LIBS是一种实现唇读的方法提取多粒度信息从演讲唇语识别器，可用于识别普通话数据集。 以上所有模型均在CMLR 数据集测试的结果如表3。

Table 3. Performance comparison of different lip recognition models on the CMLR data set

|  |  |  |
| --- | --- | --- |
| **Methods** | **Training Set** | **CER** |
| WAS | CMLR | 38.93 |
| LipCH-Net | CMLR | 34.07 |
| CSSMCM | CMLR | 32.48 |
| LIBS | CMLR | 31.27 |
| CTCH | CMLR | 9.1 |
| ELRA (ours) | CMLR | **8.0** |

可以看出，本文使用端到端唇读识别结构的字符错误率为8.0，是六种模型里效果最好的。本文使用的端到端唇读识别结构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. Conclusions

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唇读识别结构应用于唇语识别模型时，可以取得更好的效果。

Reference

1. Afouras T, Chung J S, Zisserman A. (2018) Deep Lip Reading: A Comparison of Models and an Online Application. IEEE Conference on Computer Vision and Pattern Recognition.
2. Afouras T, Chung J S, Senior A, et al. (2018) Deep Audio-visual Speech Recognition. IEEE Conference on Computer Vision and Pattern Recognition.
3. Makino T, Liao H, Assael Y, et al. (2019) Recurrent Neural Network Transducer for Audio-Visual Speech Recognition. IEEE Automatic Speech Recognition and Understanding Workshop, 905-912.
4. Ma P, Petridis S, Pantic M, et al. (2021) End-to-end Audio-visual Speech Recognition with Conformers[J]. 2021 IEEE Conference on Computer Vision and Pattern Recognition, 7613-7617.
5. Truong Q T, Lauw H W. (2019) Vistanet: visual aspect attention network for multi-modal sentiment analysis. AAAI Conference on Artificial Intelligence, 33(1): 305-312.
6. LE H，SAHOO D，CHEN N F，et al. (2019) Multi-modal transformer networks for end-to-end video-grounded dialogue systems. arXiv:1907.01166.
7. CUI C, WANG W, SONG X, et al. (2019) User attention-guided multi-modal dialog systems. ACM SIGIR Conference on Research and Development in Information Retrieval, 445-454.
8. ZHANG S, PENG H, FU J, et al. (2020) Learning 2d temporal adjacent networks for moment localization with natural language. AAAI Conference on Arti-ficial Intelligence, 12870-12877.
9. Zhao Y, Xu R, Wang X, et al. (2020) Hearing Lips: Improving Lip-reading by Distilling Speech Recognizers, 6917-6924.
10. Petajan E, Bischoff B, Bodoff D, et al. (1988) An improved automatic lipreading system to enhance speech recognition. ACM, 19-25.
11. Goldschen A J, Garcia O N, and Petajan E D. (1997) Continuous automatic speech recognition by lipreading. Computational Imaging and Vision, 321-343.
12. Shaikh A A, Kumar D K, Yau W C, et al. (2010) Lip-reading using optical flow and support vector machines. IEEE International Congress on Image and Signal Processing, 1: 327-330.
13. Ngiam J, Khosla A, Kim M, Nam J, Lee H, and Ng A Y. (2011) Multi-modal deep learning. International Conference on Machine Learning (ICML).
14. Wand M, Koutník J, and Schmidhuber J. (2016) Lipreading with long short-term memory. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6115-6119.
15. Assael Y M, Shillingford B, Whiteson S, and De Fre-itas N. (2016) LipNet: End-to-end sentence-level lip-reading, arXiv preprint arXiv:1611.01599.
16. Chung J S, Zisserman A. (2016) Lip-reading in the wild. Asian Conference on Computer Vision, 87-103.
17. Stafylakis T and Tzimiropoulos G. (2017) Combining residual networks with lstms for lipreading, Interspeech.
18. Afouras T, Chung J, Senior A, et al. (2018) Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1-1.
19. Zhang X X, Cheng F, Wang S L. (2019) Spatio-temporal fusion based convolutional sequence learning for lip-reading. IEEE/CVF International Conference on Computer Vision, 713-722.
20. Shukla A, Vougioukas K, Ma P, et al. (2020) Visually guided self supervised learning of speech representations. IEEE International Conference on Acoustics, 6299-6303.
21. Ma P, Petridis S, Pantic M. (2021) End-to-end audio-visual speech recognition with conformers. IEEE International Conference onAcoustics, 7613-7617.
22. Gulati A, Qin J, Chiu C, Parmar N, Zhang Y, et al. (2020) Conformer: Convolution-augmented transformer for speech recognition. Interspeech, 5036-5040.