澳門科技大學
MACAU UNIVERSITY OF SCIENCE AND TECHNOLOGY

Research on Ene-to-end Lip-reading Recognition of Chinese

Student Name	:	Mingchen Wang
Student No	:	2230025907
Program	:	Fundamentals of Artificial Intelligence
Supervisor	:	Professor Kaizhou Gao

Research on End-to-end Lip-reading Recognition of Chinese

Abstract

In order to better help able-bodied people to communicate with hearing-impaired or speech-impaired people and build a barrier-free society. In this paper, an End-to-end Lip-Reading Recognition Architecture (ELRA) based on multimodal fusion is constructed to realize the Chinese lip-reading video translation function. Experimental results show that the proposed ELRA is applied to the CMLR dataset and 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 lip-reading recognition; Multi-modal fusion; Hearing-impaired

I. Introduction

Each source of information can be called a modality. Modality is the way a person receives information, and people have multiple ways of perception, such as hearing, seeing, smelling, and touching; there is a certain degree of loss in the information acquired when a person perceives things accurately only through smell and touch. Nowadays, in people's daily life, video is the mainstream of people's exposure to multimedia, which contains text, audio, and visual information. Therefore, multimodal learning has become an important tool in understanding and analyzing multimedia content. Multimodal learning consists of information from different modules, usually containing two or more modalities, aiming to jointly represent data from different modalities, capturing intrinsic correlations, and allowing the information from each modality to be transformed into each other. Although in the absence of some access to information, the missing information can be filled in the handoff. Multimodal fusion biometrics is a personal identification technique that combines two or more biometric features and a data fusion technique that maximizes the use of the data features provided by each organism, making the final identification result more accurate and reliable than unimodal biometrics, thus initiating the research on multimodal tasks related to audio-visual.

Lip-reading recognition, which uses visual features of the speaker's lip movements and audio features to recognize translations. In 2018, Afouras et al. [1] formally proposed the first modern audio-visual speech recognition system using a variety of deep neural network models. They [2] proposed

Connectionist Temporal Classification Transformer (CTC Transformer) and Sequence-to-sequence Transformer (TM-seq2seq) and compared the two models. In 2019, Makino et al. [3] proposed a speech recognition model based on Recurrent Neural Network Transformer (RNN-T). Ma et al. [4] in 2021 proposed a hybrid attentional mechanism structure based on Residual Network (ResNet) and Convolutional Enhancement Transformer (Conformer) using CTC and attention mechanism to learn to recognize characters.

This paper proposes End-to-end Lip-Reading Recognition Architecture (ELRA), which splits the audio video into audio wave and image sequence inputs, and finally the input video content is converted into text content architecture. According to the data forms of image sequences and audio waves, visual and audio front-back ends are used to encode and decode the data in order to extract the data features, and the two different forms of data features are fused through the fusion module, and finally the relevant loss function is computed through the decoding and full connectivity layer, thus realizing the model back propagation in order to update the model weighting parameters. After the training is completed, the text translation task can be accomplished in the case of silent video or audio video using this architecture.

II. Related work

This section first provides an overview of multimodal fusion research and then summarizes related work on lip recognition.

A. Multi-modal fusion

Multimodal fusion technology incorporates auditory, visual, olfactory, and tactile interactions to present information more efficiently and completely. Due to its comprehensiveness in characterizing objects, multimodality has a wide range of applications in many fields.

Truong et al. [5] proposed Visual Aspect Attention Networks as a new technique to utilize visual data for sentiment analysis. Le et al. [6] designed a video-based dialog system in which the dialog depends on the visual and auditory features of a particular video, and is therefore more challenging than traditional image- or text-based dialog systems. Cui et al. [7] proposed a user attention guided multimodal dialog system, which utilizes a multimodal dialog format that combines different modal information to give users a clearer understanding of their expressions. Zhang et al. [8] proposed a new 2D spatio-temporal adjacency network, the core idea of which is to retrieve moments on a 2D spatio-temporal graph,

and to consider neighboring candidate moments as spatio-temporal contexts, a model that can be extended to other spatio-temporal localization tasks. Ya Zhao [9] from Zhejiang University proposed a LIBS model that incorporates multimodal audiovisual recognition into the knowledge refinement structure and computes knowledge extraction at the frame, sequence and text levels.

B. Lip recognition

In order to further improve the accuracy of speech recognition systems in noisy environments, researchers have begun to experiment with fusion modeling of information from different modalities to achieve higher recognition rates. The Audiovisual Speech Recognition (AVSR) system was born out of this gradual exploration process.

Before the advent of deep learning, most lip recognition relied on manual feature extraction, and extracting image features required preprocessing of many frames. Petajan et al. [10] proposed the first lip recognition system in 1984, which used a traditional lip recognition method to obtain a feature vector of lip pictures, and then computed the similarity of the words in the database, and used the word with the highest similarity as the output. Goldschen et al. [11] continued the work of Petajan et al. Inspired by speech recognition, Goldschen et al [11] used speech recognition methods to build a lip recognition model and achieved good results. Shaikh et al. [12] proposed spatio-temporal descriptors and Support Vector Machines (SVMs) classifiers to facilitate the development of lip recognition.

At the turn of the 21st century, lip recognition algorithms gained more ground with the emergence of deep learning techniques. In 2011, Ngiam et al. [13] used an autoencoder and a Restricted Boltzmann Machine (RBM) to become the first deep learning based lip recognition algorithm. The method improved the system's ability to extract features by combining speech features with image features and fusing features from different modalities. Wang et al. [14] used features from gradient histograms as inputs to a Long Short-Term Memory (LSTM) network in 2016, but the accuracy of neural network recognition was only 79.6%. In the same year, Google's DeepMind team developed the LipNet network in collaboration with the University of Oxford [15] and achieved even more impressive accuracy rates. The algorithm uses a structure consisting of a Spatio-Temporal Graph Convolutional Network (STGCN), an LSTM network, and Connected Time Classification (CTC) to achieve end-to-end variable-length sequence recognition. Chung and Zisserman [16] addressed the problem of small datasets in a paper where they

created a 500-word dataset LRW and proposed a WLAS network combining convolutional neural networks and recurrent neural networks to combine lip shape recognition techniques with speech recognition to improve the rate of recognition ability in noisy environments with significant results. 2017. Stafylakis et al. [17] proposed to add residual network on top of spatio-temporal generative and use bidirectional LSTM in the sequence modeling part to improve the algorithm's ability to learn sequence features. In 2018, Afouras et al. [18] used the Transformer structure, also from the field of machine translation, in the sequence modeling unit, but in the feature extraction part still using the structure of a spatio-temporal convolutional kernel residual network, the algorithm achieved the highest recognition accuracy of any lip recognition algorithm at the time. Since then, the field has been in a rapid development phase, with most of the work devoted to architectural improvements. In 2019, Zhang et al. [19] formally proposed temporal focus blocks and spatio-temporal fusion techniques. This technique proposes temporal focus blocks for describing short-range relationships and spatio-temporal fusion modules (STFM) for preserving local spatial information and reducing feature dimensionality. In the same year, Shukla et al. [20] first explored the application of self-supervised learning in audiovisual speech recognition, where a cross-modal setup was used to predict video frames from audio input, i.e., lip movements from audio input. In 2021, Ma et al. [21] first implemented an end-to-end LRS2 using the Conformer acoustic model and a hybrid CTC/attention decoder for learning. Experimental results show that the new front-end significantly outperforms the previous front-end in both audio-only and vision-only settings, with recent advances in final lip recognition.

III. Methodology

Audio-visual speech recognition is a multimodal task that transcribes text from audio and visual streams, and simultaneously utilizes intuitive inputs from the human voice and visual inputs from lip movements to accomplish the lip-reading recognition task. In the accessible communication system for hearing impaired people constructed in this paper, the end-to-end lip-reading recognition structure used in the recognition model is shown in Fig. 1.

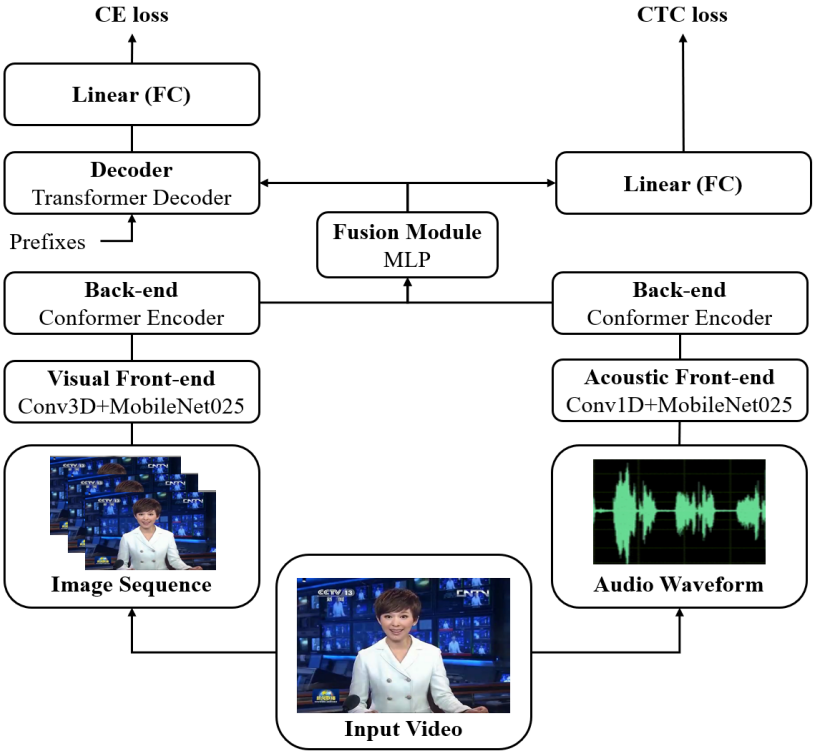

Figure 1. End-to-end Lip-reading Recognition Architecture.

Among them, the structure of both visual and audio feature extraction modules includes a convolutional front-end and an encoder back-end. The role of the convolution front-end is to extract features from image sequences or audio waveforms, and then encode the features through the encoder back-end to generate two kinds of feature information, and then fuse the visual and audio features through the fusion module to get the fused features for calculating the cross-entropy loss and CTC loss during the training process, so as to realize the back-propagation algorithm of the model.

A. Front-end module

The temporal modeling front-end of both the vision and audio extraction modules takes the form of a convolutional neural network skeleton, removing the last fully-connected layer used for outputting each probability in the output section, and completing the task using the multichannel feature maps obtained from the forward computation of the remaining structures. In the front-end module, a simplified MobileNet architecture will be adapted based on different data structures, a model originally used for image classification applications in mobile and embedded vision. Assuming a 224×224 three-channel image as input, Fig 2. illustrates the structure of the model.

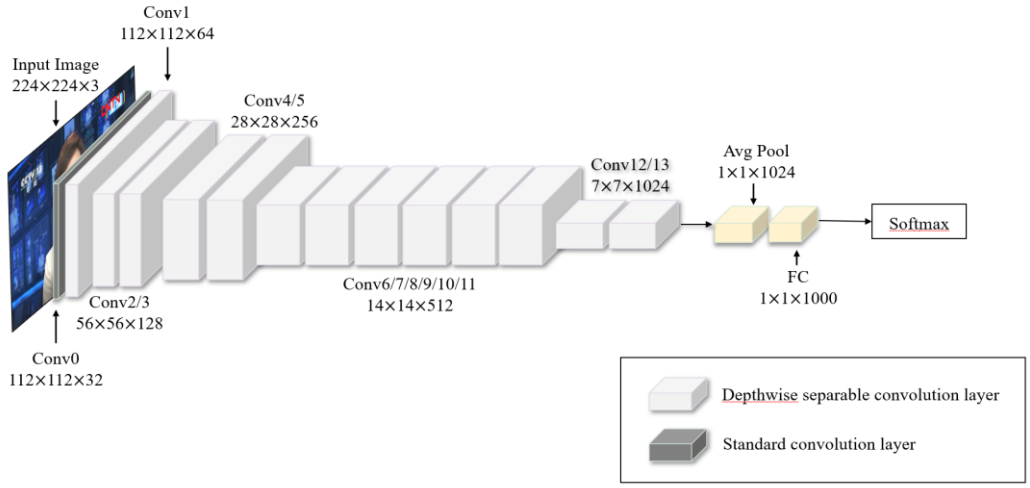

Figure 2. MobileNet original model structure.

In this case, the grey squares represent the standard convolutional layer, a combination of the ordinary convolutional layer and the normalized layer. The purple squares represent the Depthwise Separable Convolution of MobileNet basic unit, which consists of Depthwise Convolution and Pointwise Convolution. The former uses different convolution kernels for the three channels of the RGB image in the convolution operation, while the latter employs ordinary convolution with 1×1 convolution kernels. The design achieves the effect of standard convolution on the basis of reduced computation and model parameters. In practice, Batch Normalization (BN) and ReLU activation functions will be implemented to accelerate the convergence of training.

As mentioned above, image sequences and audio waves belong to different modalities of data, so the input dimensions in the visual and audio front-end modules are different, so the 2D convolutional layer needs to be adjusted, and the adjusted MobileNet Backbone structure is shown in Table 1.

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

Unit	**Layers**	**Input audio waveform**	**Stride**	**Layers**	**Input image Sequence**	**Stride**
C0	Conv1D	80×32	4	Conv3D MaxPool3D	$5 \times 7^2 \times 32$ 1×3^2	1×2^2 1×3^2
C1	Conv1D dw Conv1D	3×32 dw $1 \times 32 \times 64$	1 1	Conv2D dw Conv2D	$3^2 \times 32$ dw $1^2 \times 32 \times 64$	1^2 2^2
C2	Conv1D dw Conv1D	3×64 dw $1 \times 64 \times 128$	2 1	Conv2D dw Conv2D	$3^2 \times 64$ dw $1^2 \times 64 \times 128$	1^2 1^2
C3	Conv1D dw Conv1D	3×128 dw $1 \times 128 \times 256$	1 1	Conv2D dw Conv2D	$3^2 \times 128$ dw $1^2 \times 128 \times 128$	1^2 1
C4	Conv1D dw	3×128 dw	2	Conv2D dw	$3^2 \times 128$ dw	2^2

	Conv1D	$1 \times 128 \times 256$	1	Conv2D	$1^2 \times 128 \times 256$	1^2
C5	Conv1D dw Conv1D	3×256 dw $1 \times 256 \times 256$	1 1	Conv2D dw Conv2D	$3^2 \times 256$ dw $1^2 \times 256 \times 256$	1^2 1^2
C6	Conv1D dw Conv1D	3×256 dw $1 \times 256 \times 512$	2 1	Conv2D dw Conv2D	$3^2 \times 256$ dw $1^2 \times 256 \times 512$	2^2 1^2
C7-C11	Conv1D dw Conv1D	3×512 dw $1 \times 512 \times 1024$	1 1	Conv2D dw Conv2D	$3^2 \times 512$ dw $1^2 \times 512 \times 512$	1^2 1^2
C12	Conv1D dw Conv1D	3×512 dw $1 \times 512 \times 1024$	1 1	Conv2D dw Conv2D	$3^2 \times 512$ dw $1^2 \times 512 \times 1024$	1^2 1^2
C13	Conv1D dw Conv1D	3×1024 dw $1 \times 1024 \times 1024$	2 1	Conv2D dw Conv2D	$3^2 \times 1024$ dw $1^2 \times 1024 \times 1024$	2^2 1^2

Where Conv1D dw represents 1D depth convolution. Conv2D dw represents 2D deep convolution. For the vision front-end, the first convolutional layer (C0) of the MobileNet backbone is replaced with a combination of a 3D convolutional layer of $5 \times 5 \times 7$ kernel size and a 3D maximal pooling layer of 1×32 kernel size, converted from $B \times Tv \times W \times H$ (B is the batch size) to $(B \times Tv) \times W \times H$, and the temporal dimension of the image sequences is integrated into the batch number dimension. For the acoustic front-end, since the audio waveforms are 1D data, the whole network needs to be modified from 2D convolution to 1D convolution, and C0 is adjusted to a 1D convolution with a core size of 80 (5ms), and the step size is set to 4 so that the final acoustic features are sampled at 25 frames per second to match the frame rate of the visual features. Since visual feature extraction is not a major part of audiovisual recognition, a width multiplier α is introduced to reduce the number of model parameters for the front-end module, with values in the range $(0,1]$. The effect is to minimize the number of channels of the feature map that we let $\alpha = 0.25$, from 1024 to 256. as mentioned above, MobileNet 0.25 in ELRA is derived from this.

B. Back-end module

Gulatiet et al. [22] proposed a new architecture that integrates a self-attention mechanism and convolution into an ASR model, called the Conformer encoder. We use the Conformer encoder to extract image sequence features and audio waveforms at the back-end of visual and acoustic temporal modeling, the architecture of which is shown in Figure 3.

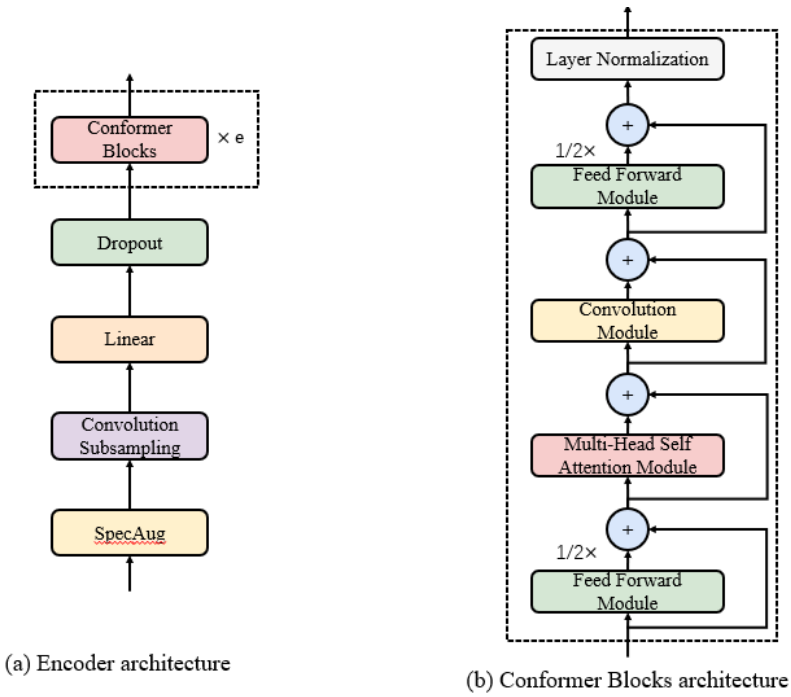

Figure 3. Encoder and Conformer architecture.

The Conformer encoder first performs a simple data enhancement of the input features and then downsamples the convolutional layer to enhance the timing of the data. The internal structure of a single Conformer module is a "sandwich" structure, with the convolutional and self-attention modules sandwiched between two feedforward modules.

C. Fusion module

The fusion module employs a Multi-Layer Perceptron (MLP) to fuse image sequences and audio waveforms by means of features extracted from the front-end module and projected into a d_k dimensional space. The MLP output size is composed of a fully-connected layer with a size of $4 \times d_k$, a BN layer, a ReLU activation function, and a fully-connected layer with a size of d_k output.

D. Decoder

The decoder used after the fusion module is a Transformer-tuned based decoder consisting of an embedding layer, a multi-head self-attention module and a feed-forward neural network. The output sequence generated in this paper refers to the word vector at the current prediction moment, embedded with the generated output sequence and relative position encoding. In the multi-head self-attention module, it mainly consists of a masked multi-head attention mechanism, a multi-head attention mechanism and a feed-forward neural network, the structure of which is shown in Fig. 4.

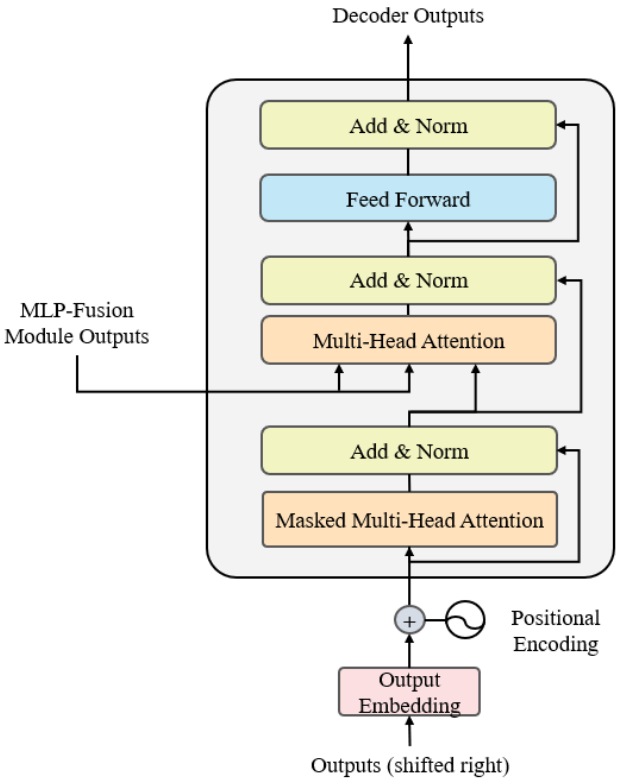

Figure 4. Transformer decoder architecture.

IV. Experiments

A. *Loss function*

A hybrid CTC/attention loss function is used. Suppose $X = [x_1, \cdots x_T]$ is the input frame sequence of Transformer decoder in fusion module and $Y = [y_1, \cdots y_L]$ is the output target, where T and L denote the input length and the target length.In the CTC loss, it is assumed that the predicted values of each output are independent from each other as shown in equation (1).

$$P_{\mathrm{CTC}}(Y|X) \approx \prod_{t=1}^{L} P(y_t|X) \tag{1}$$

In contrast, the model based on the attention mechanism eliminates this assumption and estimates the probability that the posterior is based on the chain rule, as shown in Equation (2).

$$P_{\mathrm{CE}}(Y|X) = \prod_{l=1}^{L} P(y_l|y < 1, X) \tag{2}$$

From this, the formula for calculating the total loss can be obtained as in equation (3).

$$L = \lambda \log P_{\mathrm{CTC}}(Y|X) + (1-\lambda) \log P_{\mathrm{CE}}(Y|X) \tag{3}$$

Where λ is the weight factor of the CTC loss function versus the attention mechanism in hybrid CTC/attention. The weight is not only the combined function of the two losses transformed into a single training loss, but also the predictions and demands of the two decoding processes

B. Evaluation metric

In order to verify the recognition effect of the end-to-end lip-reading recognition structure, the character error rate (CER) of five different models on the CMLR dataset is selected for comparison in this paper, and the character error rate is calculated as in equation (4).

$$\mathrm{CER} = \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{N}} \tag{4}$$

where S denotes the number of substitutions, D denotes the number of deletions, I denotes the number of insertions from the predicted sequence to the standard sequence, and N denotes the number of words in the predicted sequence.

C. CMLR dataset

This paper uses the Chinese Mandarin Lip Reading (CMLR) dataset, which is designed to facilitate research in visual speech recognition. The CMLR dataset is derived from the CCTV news broadcast videos from June 2009 to June 2018. The CMLR dataset is a collection of the CCTV news broadcast videos. The dataset contains a total of 102,076 sentences expressed by 11-bit hosts, and each sentence contains at most 29 Chinese characters, excluding English letters, Arabic numerals and rare punctuation marks. In addition, the training, validation and test sets are randomly divided in the ratio of 7:1:2. The details are shown in Table 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**

D. Results

In order to evaluate the validity of ELRA recognition, we have chosen five different models for

comparison, they are WAS, LipCH-Net, CSSMCM, LIBS and CTCH. They are all lip recognition methods. WAS is the classical method sentence-level lip recognition in this field, which will be used to recognize Chinese characters directly; LipCH-Net and CSSMCM are the Chinese sentence-level lip recognition models; LIBS is an implementation of lip-reading method to extract multi-granularity information from speech lip recognizer, which can be used to recognize Mandarin datasets. The results of all the above models tested on the CMLR dataset are shown in Table 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**

It can be seen that the character error rate of the end-to-end lip reading recognition structure used in this paper is 8.0, which is the best among the six models. The end-to-end lip-reading recognition structure used in this paper is better than previous lip-reading models, performs better in fusing image features and audio features, and is able to accomplish the task of lip-reading in Chinese well.

V. Conclusions

In order to better help able-bodied people communicate with hearing-impaired or speech-impaired people and build a barrier-free society, this paper realizes an end-to-end Chinese lip-reading translation function and video recognition system based on multimodal fusion. It is concluded through experiments that better results can be achieved when the proposed end-to-end visual lip-reading recognition structure is applied to the lip recognition model.

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 Artificial 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. (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.