End-to-End Chinese Lip-Reading Recognition Based on Multi-modal Fusion

Yixian Liu [1,a]
[1] School of Applied Science and Civil Engineering, Beijing Institute of Technology, Zhuhai, China
[a] 1229773575@qq.com

Chuoya Lin[2,b]
[2] School of Applied Science and Civil Engineering, Beijing Institute of Technology, Zhuhai, China
[b] 1528762911@qq.com

Mingchen Wang[3,c]
[3] School of Applied Science and Civil Engineering, Beijing Institute of Technology, Zhuhai, China
[c] 2693970161@qq.com

Simin Liang[4,d]
[4] School of Applied Science and Civil Engineering, Beijing Institute of Technology, Zhuhai, China
[d] 1030729209@qq.com

Zhuohui Chen[5,e]
[5] Faculty of Innovation Engineering, Macau University of Science and Technology, China
[e] 2113950286@qq.com

Ling Chen[6,f*]
[6] School of Applied Science and Civil Engineering, Beijing Institute of Technology, Zhuhai, China
*[f] lingchensh@126.com
* Corresponding author: Ling Chen, lingchensh@126.com

2022 4th International Conference on Frontiers Technology of Information and Computer (ICFTIC) | 979-8-3503-2195-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICFTIC57696.2022.10075247

***Abstract*—With around 1.5 billion people worldwide suffering from hearing impairment, it is particularly important to communicate between non-disabled people and people with hearing or speech impairment and to build a barrier-free society. Multi-modal learning provides an excellent artificial intelligence channel for this purpose. In this article, we create an End-to-end Chinese Lip-Reading Recognition System based on multi-modal fusion to implement Chinese lip translation in order to facilitate communication between individuals with hearing impairment. Our system adopts the End-to-end Audio-visual feature fusion Lip-reading Recognition Architecture (EALRA), with feature extraction based on a MobileNet0.25 tuned CNN skeleton and the encoder back-end using the Conformer self-attentive convolution encoder for modelling. The largest Chinese Mandarin Lip-Reading (CMLR) was selected as the dataset for the empirical study, and the performance metric for Chinese lip recognition was the character error rate (CER). The results of our experiments show that the CER metric of EALRA in the lip-recognition model is 8.0, which is on average 23.74% lower than the CER metrics of other lip-recognition models, indicating that EALRA performs better in fusing image features and audio features.**

Keywords—Audio-visual speech recognition; EALRA; Hearing impairment; Multimodal fusion

I. INTRODUCTION

Each source of information can be referred to as a modality. A modality is a way in which a person receives information; people have a variety of perceptions, such as hearing, seeing, smelling, and touching to perceive things accurately when one modality of information is missing is a significant concern. As multimedia data is often a medium for the transmission of multiple information, for example, a video can contain textual, visual, and auditory information at the same time, multi-modal learning [1] has become an essential tool for the analysis and understanding of multimedia content. Multi-modal learning consists of information from different modalities, generally containing two or more modalities, and aims to jointly represent data from different modalities, capturing the intrinsic correlation between the different modalities, enabling information from each modality to be transformed into each other, although in the absence of certain modalities it may be possible to fill in the missing information in the handoff process. Multi-modal fusion biometric recognition is a technique that combines two or more biometric traits of the person to be recognized and is also a data fusion technique that makes maximum use of the data given by each biometric feature, making the final recognition result more accurate and reliable compared to single-modal biometric recognition.

Lip recognition [2], which is the process of understanding what a speaker says by observing the changes in their mouth pattern, is a multidisciplinary research field that involves a broad variety of theoretical knowledge in areas like pattern recognition, image processing, and machine learning. The lip recognition system uses machine vision technology to first recognize the face through the input video image, determine the speaker, locate the lip position and extract successive images of the speaker's lip changes. Then the continuously changing features are extracted by the neural network and input into the lip recognition model to identify the corresponding words or words of the speaker. Finally, calculate the most likely natural sentences according to the recognized sequence. Since manually extracted shape and texture features are challenging to characterize the complete information of lip-movement sequences, and since lip feature learning and classification learning should be learned as a whole, separate training will affect the machine learning effect, resulting in traditional lip recognition methods not being able to solve complex lip recognition problems well. The problem of recognizing the content of complicated video sequences is increasingly being

resolved due to the advent of deep neural network technology [3].

According to the survey, people and their environment are multi-modal, and separate lip-reading and speech recognition have some defects. Separate speech recognition mainly has the following two defects: first, the speech is ambiguous, that is to say, different words sound similar; secondly, the interference of environmental noise, noise has a serious impact on speech recognition, so the recognition rate is shallow. Lip recognition is an efficient auxiliary technology that can compensate for current speech recognition deficiency. Still, lip recognition has some problems: lip shape is also ambiguous if there is no tongue position information, and it is challenging to distinguish simply by using visual features, which will cause information recognition errors, so multi-modal audio-visual recognition is an inevitable result. One of the first applications of multi-modal research was audio-visual speech recognition, as first demonstrated by the McGurk effect, in which many of the speakers to be recognized perceive a speech "ba" and a visual "ga" as "da", as a result of the interaction the relationship between of visual and auditory senses in speech perception. In response, researchers began to merge visual and sound modalities in the sound identification process, resulting in an enormous leap forward over the original system with only a single sound modal input, and multi-modal learning began to demonstrate its excellent learning capabilities.

In summary, we can see that lip features as an aid to speech recognition have become an indispensable part of the process. In recent years, the advancement of deep learning has substantially assisted the development of multi-modal audio-visual recognition. Under this current situation and background, we propose implementing an accessible communication system for the hearing impaired based on Chinese lip translation, the system adopts the End-to-end Audio-visual feature fusion Lip-reading Recognition Architecture (EALRA). The EALRA splits video into the audio wave and image sequence input models, translating video content into textual content. Traditional methods usually up-sample or down-sample the video content as well as the audio content into the same frame rate for direct stitching. At the same time, we chose to use the Audio-Video Dual-Modal Speech Recognition technique, which uses an ASR (Audio Speech Recognition) decoder to decode the audio features and a VSR (Visual Speech Recognition) decoder to decode the visual features, followed by self-attentiveness at the output layer of the encoder to find the visual vector associated with the current auditory vector and then splice it with the acoustic vector after finding the visual vector associated with the current auditory vector at the output layer of the encoder. The processing of audio and video content using audio-video bimodal speech recognition technology allows text translation to be performed in either unimodal or bimodal situations.

In order to determine the efficacy of our presented EALRA method, we chose the largest Chinese Mandarin Lip-reading (CMLR) dataset as the experimental subject. We compared EALRA with five Chinese lip-reading models and then assessed lip-reading performance using a single performance metric, Character Error Rate (CER). The experimental results demonstrate that EALRA can outperform these models.

II. Related Work

This section begins with a summary of multi-modal fusion research. Then the related work on lip recognition is summarised.

A. Multi-modal fusion

Multi-modal fusion technology, which incorporates auditory, visual, olfactory, and tactile interaction, allows for a more efficient and complete representation of information. Due to multimodality comprehensiveness in characterizing objects, it offers a vast array of applications in numerous domains.

The visual aspect attention network was proposed by Truong et al. [4] as a new technique for sentiment analysis utilizing visual data. Le et al. [5] designed a video-based dialogue system in which the dialogue is dependent on the visual and aural characteristics of a specific video, making it more challenging than the traditional image- or text-based dialogue systems. Cui et al. [6] proposed a user-attention-guided multi-modal dialogue system to c provide a clearer understanding of the user's expression using a multi-modal dialogue format that combines information from different modalities. Zhang et al. [7] proposed a new 2D temporal adjacency network with the core idea of retrieving a moment on a 2D temporal graph that treats neighboring candidate moments as temporal contexts, a model that can be extended to other temporal localization tasks. Ya Zhao [8] of Zhejiang University proposed a LIBS model that incorporates multi-modal audio-visual recognition into the knowledge distillation structure and computes knowledge extraction at the frame, sequence, and text levels.

B. Lip recognition

With the aim of further improving the accuracy of speech recognition systems in noisy environments, researchers have started to experiment with the fusion modelling of information from different modalities to achieve higher recognition rates. The Audio-visual speech recognition (AVSR) system was created in the process of this gradual exploration.

Before the advent of deep learning, most lip recognition relied on manual feature extraction and extracting image features required pre-processing many frames. Petajan et al. [9] proposed the first lip recognition system in 1984, in which a vector of lip picture features was obtained using traditional lip recognition methods, and then the similarity measures of the words in the database were calculated, with the word with the highest similarity as the output. Goldschen et al. [10] continued the work of Petajan et al., inspired by speech recognition using a speech recognition approach to build lip recognition models with good results. Shaikh et al. [11] proposed Spatio-temporal descriptors and Support Vector Machine (SVM) classifiers to facilitate the development of lip recognition.

With the emergence of deep learning techniques at the turn of the 21st century, lip recognition algorithm has gained more room for development. In 2011, Ngiam et al. [12] used an auto encoder and Restricted Boltzmann Machine (RBM) to become the first deep learning-based algorithm for lip recognition. By combining voice features with image features and fusing features from different modalities, the method improves the system's capacity to extract features. Wand et al. [13] used

features from gradient histograms as input 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 and Oxford University collaborated to develop the LipNet network [14] and achieved a more impressive accuracy rate. The algorithm uses a structure consisting of a Spatial-Temporal Graph Convolutional Networks (STGCN), a LSTM network, and Connectionist Temporal Classification (CTC) to enable end-to-end variable-length sequence recognition. The problem of small datasets was addressed in a paper by Chung and Zisserman [15], who created a dataset LRW of 500 words and proposed a WLAS network combining convolutional neural networks and recurrent neural networks to combine lip recognition techniques with speech recognition to improve recognition rates in noisy environments with significant results. In 2017, Stafylakis et al. [16] proposed to add residual networks on top of Spatio-temporal productions and to use bi-directional LSTM in the sequence modelling part to improve the ability of the algorithm to learn sequence features. In 2018, Afouras et al. [17] used the Transformer structure, also from the field of machine translation, for the sequence modelling unit, but still used the structure of a Spatio-temporal convolutional kernel residual network for the feature extraction part, and the algorithm achieved the highest recognition accuracy of any lip recognition algorithm at the time. Since then, the field has been in a phase of rapid development, with most of the work devoted to architectural improvements. In 2019, Zhang et al. [18] formally proposed temporal focus blocks and spatial-temporal fusion techniques. This technique presents a temporal focus block for describing short-range relationships and a space-time fusion module (STFM) for preserving local spatial information and reducing feature dimensionality. In the same year, Shukla et al. [19] first explored the application of self-supervised learning to audio-visual speech recognition, where video frames were predicted from audio input using a cross-modal setting, i.e., predicting lip motion from the audio input. In 2021, Ma et al. [20] first implemented end-to-end LRS2 using the Conformer acoustic model and a hybrid CTC/attention decoder for learning. Experimental results showed that the new front-end significantly outperformed previous front-ends in both audio-only and visual-only settings and achieved recent advances in final lip recognition.

III. DATASET

A. CMLR dataset

The Visual Intelligence and Pattern Analysis (VIPA) group at Zhejiang University collected CMLR, the biggest public Chinese mandarin lip-reading dataset, in order to promote research on visual speech recognition, often known as automatic lip-reading. The CMLR dataset is created from video of China Central Television (CCTV) news broadcasts between June 2009 and June 2018. The dataset contains a total of 102,076 sentences expressed by 11 presenters, each of which has a maximum of 29 Chinese characters, eliminating English letters, Arabic numbers, and uncommon punctuation marks. Additionally, in order to facilitate the capture of video sequence matching to the various Chinese words, the dataset also provides the corresponding timestamps for the words in each sentence to enable the classification of Chinese words. The training set, validation set, and test set were randomly divided in the ratio of 7:1:2. Table 1 shows the details of the CMLR dataset.

TABLE I. 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

B. Data generation

As the characteristics of Chinese Mandarin were considered, the VIPA group optimized some of the steps to produce a better quality CMLR Chinese lip-reading dataset. Fig. 1 shows the specific data generation process.

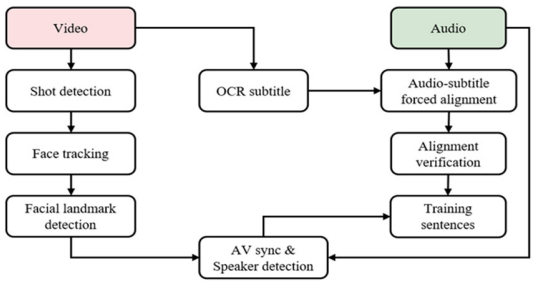

Figure 1. Data generation process.

The video footage was taken from the national news program " CCTV News" recorded from June 2009 to June 2018. Firstly, a Histogram of Oriented Gradient (Hog) -based face detection method was used for face recognition and alignment using an open-source platform to intercept video clips of the 11 presenters broadcasting the news. As there were no subtitles and text annotations in the original program, FFMPEG was used to extract the corresponding audio tracks from the clip set of the video, and the text annotations corresponding to the video clip set were obtained by iFLYTEK3 ASR, removing noise labels such as script letters, Arabic numerals, and rare punctuation marks to obtain a purer lip-reading dataset in Mandarin Chinese. Finally, the video clip sets were intercepted according to the text annotation information containing the time nodes, and the video start and end frames were intercepted with specific adjustments to obtain the corresponding text-annotated word, phrase, or sentence video clips.

C. Data preprocessing

The dataset is divided into silent video files, audio files with sound, and text annotation files containing time nodes. The dataset needs to be pre-processed to enable the CMLR dataset to facilitate subsequent model training. Fig. 2 shows a specific dataset pre-processing process.

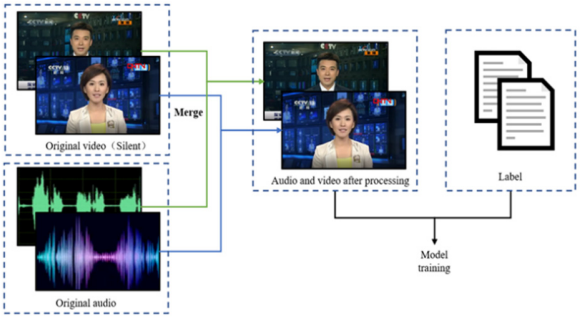

Figure 2. Data preprocessing process.

As shown above, the silent video file and the audio file data with sound are the first to read. We use FFMPEG to combine the visual feature data with the audio feature data to generate the video file with sound.

IV. Methodology

Transcribing text from audio and visual streams is the objective of Audio-visual speech recognition (AVSR). At the same time, it uses the intuitive input of human voice and the visual input of lip movements to complete the lip recognition task. We intend to implement an accessible communication system for the hearing impaired, where the recognition model takes the structure of EALRA, as shown in Fig. 3.

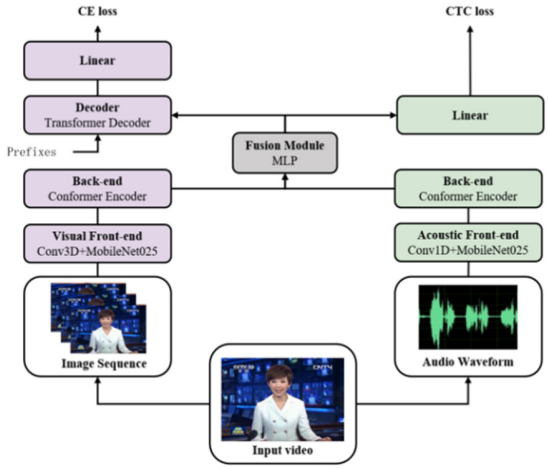

Figure 3. End-to-end Audio-visual feature fusion Lip-reading Recognition Architecture.

The structure of the visual and audio feature extraction modules both contain a convolutional front-end and an encoder back-end. The convolutional front-end serves to extract features from image sequences or audio waveforms and then encodes the features through the encoder back-end, thus generating two features of information, which are fused by the fusion module. The Cross-Entropy (CE) loss and the Connectionist Temporal Classification (CTC) loss are calculated using the fused features during the training process, thus realizing the back-propagation algorithm of the model.

A. Front-end module

The temporal modeling front-ends of the visual and audio extraction modules both take the form of a convolutional neural network skeleton, removing the completely connected layer from the output portion and using the feature maps obtained by reasoning in the upper part of the output layer to implement the proposed task. The skeleton of our architecture is based on the streamlined MobileNet architecture, a model originally used for image classification applications in mobile and embedded vision. Assuming an input of 224 × 224 three-channel images, Fig. 4 demonstrates the structure of the model.

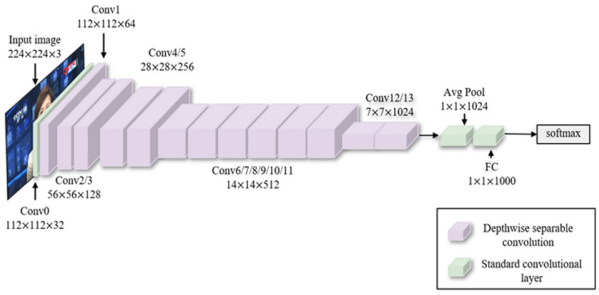

Figure 4. MobileNet original model structure.

Among them, the green square represents the standard convolution layer, the combination of the ordinary convolution layer, and the standardization layer. The purple square represents the Depthwise Separable Convolution of the basic unit of MobileNet, composed 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 adopts the ordinary convolution of 1 × 1 convolution kernel. This design achieves the effect of standard convolution based on reducing the amount of calculation and model parameters. In practical application, Batch Normalization (BN) and ReLU activation functions will be implemented to accelerate the convergence of training.

For the End-to-end Audio-visual feature fusion Lip-reading Recognition Architecture used in this paper, the input dimensions of the visual and acoustic front-end modules are different, and only the extracted feature map is needed, so the adjusted MobileNet skeleton, such as Table 2.

TABLE II. 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 × 32 × 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 × 64 × 128	2 1	Conv2D dw Conv2D	3^2 × 64 dw 1^2 × 64 × 128	1^2 1^2
C3	Conv1D dw Conv1D	3 × 128 dw 1 × 128 × 128	1 1	Conv2D dw Conv2D	3^2 × 128 dw 1^2 × 128 × 128	1^2 1^2
C4	Conv1D dw Conv1D	3 × 128 dw 1 × 128 × 256	2 1	Conv2D dw Conv2D	3^2 × 128 dw 1^2 × 128 × 256	2^2 1^2
C5	Conv1D dw Conv1D	3 × 256 dw 1 × 256 × 256	1 1	Conv2D dw Conv2D	3^2 × 256 dw 1^2 × 256 × 256	1^2 1^2
C6	Conv1D dw Conv1D	3 × 256 dw 1 × 256 × 512	2 1	Conv2D dw Conv2D	3^2 × 256 dw 1^2 × 256 × 512	2^2 1^2
C7-11	Conv1D dw Conv1D	3 × 512 dw 1 × 512 × 512	1 1	Conv2D dw Conv2D	3^2 × 512 dw 1^2 × 512 × 512	1^2 1^2
C12	Conv1D dw Conv1D	3 × 512 dw 1 × 512 × 1024	1 1	Conv2D dw Conv2D	3^2 × 512 dw 1^2 × 512 × 1024	1^2 1^2
C13	Conv1D dw Conv1D	3 × 1024 dw 1 × 1024 × 1024	2 1	Conv2D dw Conv2D	3^2 × 1024 dw 1^2 × 1024 × 1024	2^2 1^2

Where Conv1D dw represents 1D depth convolution. Conv2D dw stands for 2D depth convolution. For the visual front end, the first convolution layer (C0) of the MobileNet skeleton is replaced by a combination of a 3D convolution layer of 5 × 5 × 7 core size and a 3D maximum pool layer of 1 × 32 core size, converting B × Tv × W × H (B is the batch size) to (B × Tv) × W × H, and integrating the time dimension of the image sequence into the batch quantity dimension. For the acoustic front end, because the audio waveform is 1D data, the whole network needs to modify 2D convolution into 1D convolution, and C0 is adjusted to the 1D convolution of 80 (5ms) core size, and the step is set to 4, so that the final acoustic feature is sampled to 25 frames per second to match the frame rate of the visual feature.

Because visual feature extraction is not the main part of audio-visual recognition, width multiplier α is introduced to decrease the amount of front-end module model parameters, and its value range is (0,1]. The function is to minimize the number of channels proportionally. We let α = 0.25, and the number of channels of the characteristic graph is decreased from 1024 to 256. As mentioned above, the MobileNet0.25 in the EALRA is derived from this.

B. Back-end module

Gulatiet et al. [21] presented a new architecture that integrates the self-attention mechanism and convolution in the ASR model called Conformer encoder. We use the Conformer encoder is used to extract image sequence characteristics and audio waveform in the back-end of visual and acoustic temporal modelling, and its architecture is shown in Fig. 5.

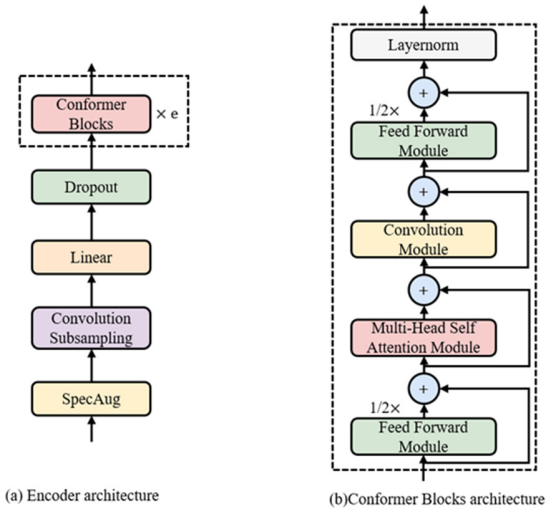

Figure 5. Conformer architecture.

Among them, the left side is the overall architecture of the Conformer; first, a simple data enhancement of the input features through the convolution layer sampling to enhance the timing of the data; then Dropout layer prevents over-fitting and entering the Conformer block. The dotted frame on the right represents the internal structure of a single Conformer coding block, which adopts a "sandwich" structure, in which the convolution module and the self-attention module are sandwiched between two feedforward modules, the internal construction of these sub-modules is shown in Fig. 6.

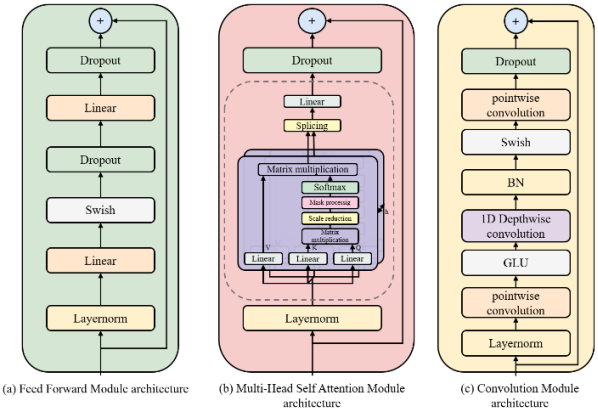

Figure 6. Conformer sub-module architecture.

(a) the subgraph represents the internal structure of a single feedforward sub-module. The input features are first Layer Normalization (LN), the first linear conversion is completed through the first full connection layer, and the second linear conversion is completed after the Swish activation function and Dropout layer to complete Feed Forward Network (FFN).

(b) the subgraph represents the multi-head self-attention module with the characteristics of Conformer, which follows the multi-head self-attention mechanism of Transformer. The attention mechanism of reducing dot product is adopted, which is easy to implement, and the shape of the obtained vector is consistent with that of the input vector.

(c) the subgraph represents the internal architecture of the convolution module in Conformer, which is like MobileNet but also uses point-by-point convolution and depth convolution. In practice, BN and LN are added to accelerate the convergence of the model.

C. Fusion module

The fusion module uses Multi-Layer Perceptron (MLP) to fuse the features extracted from the image sequence and audio waveform through the front-back module and project them to the d_k dimensional space. MLP consists of a full connection layer with an output size of $4 \times d_k$, a BN layer, a ReLU activation function, and a full connection layer with an output size of d_k.

D. Transformer decoder

The decoder used after the fusion module is based on the Transformer adjusted decoder, which consists of an embedded layer, a set of multi-head self-attention modules, and a feedforward network, which contains two sets of residual connections and LN layers, as shown in Fig. 7.

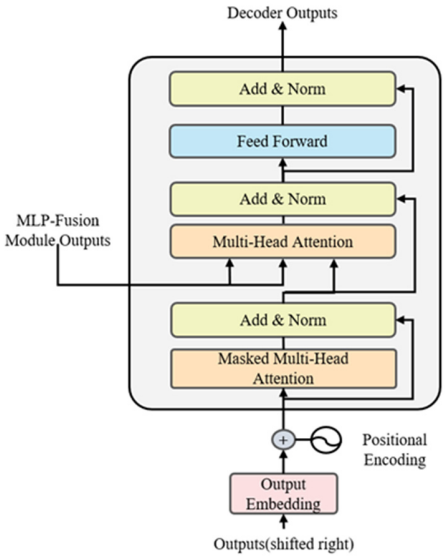

Figure 7. Transformer decoder architecture.

We use the overall architecture of the Transformer decoder, which consists mainly of an embedding layer and N blocks of multi-headed self-attentive modules. The generated output sequence refers to the word vector at the current prediction moment. Moreover, the resulting output sequence contains the relative position encoding. In the multi-headed self-attention module, the main components are a unique masked multi-headed attention mechanism, a general multi-headed attention mechanism, and a feed-forward module.

V. Experiments

A. Loss function and Evaluation metric

1) Loss function

EALRA uses a hybrid CTC/attention loss function. Let $X = [x_1, \ldots, x_T]$ is the input frame sequence at the input of the Transformer decoder in the fusion module, and $Y = [y_1, \ldots, y_L]$ is the target of the output, where T and L represent the input's length and the target's length, respectively.

The CTC loss assumes that the predicted values of each output are independent of each other in the form shown in equation (1).

$$P_{CTC}(Y \mid X) \approx \prod_{t=1}^{T} P(y_t \mid X) \tag{1}$$

In contrast, the attention mechanism-based model eliminates this assumption and estimates a posteriori probability based on the chain rule, shown in equation (2).

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

Therefore, the formula for calculating the total loss is shown in equation (3).

$$L = \lambda \log P_{CTC}(Y \mid X) + (1 - \lambda) \log P_{CE}(Y \mid X) \quad (3)$$

Where λ is the weight factor of the CTC loss function versus the attention mechanism in the hybrid CTC/attention mechanism. The weight integrates not just the two loss functions into a single training loss, but also the two forecasts and needs during the decoding process.

2) Evaluation metric

We use the CMLR dataset to evaluate the proposed method. Because the Chinese statement is expressed as a contiguous string, which is not separated by spaces, and there is no division of word boundaries, we use CER to evaluate the performance of the method, and it assesses how near the anticipated character sequence and the target character sequence are to one another, as shown in equation (4).

$$CER = \frac{(S + D + I)}{N} \quad (4)$$

Where S denotes the amount of replacements, D denotes the number of deletions, I reflects the amount of insertions into the standard sequence from the prediction sequence, and N denotes the amount of words inside the prediction sequence.

B. Results

In order to assess the effectiveness of EALRA's recognition, we select five different models to compare them, which are WAS, LipCH-Net, CSSMCM, LIBS, and CTCH. They are all lip-recognition methods. WAS is a classic method in the field of sentence-level lip recognition, which will be used to recognize Chinese characters directly; LipCH-Net and CSSMCM are end-to-end Chinese sentence-level lip recognition models; LIBS is a method to realize lip-reading by extracting multi-granularity information from speech recognizer to lip reader, which can be used to recognize Mandarin data sets. All the above models are compared on the CMLR dataset.

TABLE III. COMPARISON OF THE PERFORMANCE OF SEVERAL LIP RECOGNITION ALGORITHMS ON THE CMLR DATASET

Methods	Training Set	CER
WAS	CMLR	38.9
LipCH-Net	CMLR	34.0
CSSMCM	CMLR	32.5
LIBS	CMLR	31.3
CTCH	CMLR	22.0
EALRA(Ours)	CMLR	8.0

Figure 8. Experimental results.

Table 3's comparative results allow us to derive the following inferences: by comparing the CERs of the six models, the EALRA model that we used has a CER of 8.0, which is the best result among the six models. Compared to the CTCH model, the CER was decreased by 13%; compared to the WAS model, it was lowered by 30.9%; and compared to the prior five models, the CER was reduced by an average of 23.74%. This indicates that the EALRA model we used is better than previous lip-reading models for Chinese lip recognition, performs better in fusing image features and audio features, and is able to perform the Chinese lip-reading task well.

From the comparison between the actual results of Fig. 8 and the predicted results, the conclusion may be drawn from the experimental findings of incorrect recognition, and the actual results are particularly similar to the predicted pronunciation of Chinese characters and the lip shape of the pronunciation of the Chinese characters. For example, the Chinese characters "Main business" and "Best wishes", " He has" and "And" are very similar in their standard mandarin pronunciation and the lip shape of the Chinese characters. In the future, We will do more training in this region to increase the model's recognition accuracy.

VI. CONCLUSION

In order to better help non-disabled people communicate with deaf and hard of hearing people or language-impaired people and build a barrier-free society, we build an End-to-end Chinese Lip-Reading Recognition System based on multi-modal fusion to achieve the function of Chinese lip translation. Through experiments, it is concluded that the proposed End-to-end Audio-visual feature fusion Lip-reading Recognition Architecture EALRA can be applied to the lip recognition model to achieve better results.

There are still shortcomings in this study. Fewer experimental subjects were selected, and only one data set was used as the experimental subject, which led to the low robustness of the model; fewer experimental strategies were adopted, and only audio and video strategies were chosen for the experiments, which led to the contingency of the experiment. In the future, it is hoped that some extensions to this work will be made:

- Collecting a bigger corpus of Chinese Mandarin lip-reading data in order to increase the model's robustness with additional data;
- Optimize the model structure and explore more suitable methods for lip feature extraction or utterance parsing to solve the existing problems of the model;
- More experimental strategies are chosen for comparison. For example, three different modalities - audio-only, visual-only, and audio-visual can be added for comparison experiments to improve the model's performance.

ACKNOWLEDGMENT

This work is partly supported by the "Barrier-free communication system for hearing impaired people based on Chinese lip translation" (Grant NO. 2022-06) under the Innovation and Entrepreneurship Training Programme for University Students.

REFERENCES

[1] Ramachandram D, Taylor G W. (2017) Deep multi-modal learning: a survey on recent advances and trends. IEEE Signal Processing Magazine, 34(6): 96-108.

[2] Matthews I, Cootes T F, Bangham J A, et al. (2002) Extraction of visual features for lipreading. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(2): 198-213.

[3] Deng L, Yu D. (2014) Deep learning: methods and applications. Foundations & Trends in Signal Processing, 7(3): 197-387.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] Zhao Y, Xu R, Wang X, et al. (2020) Hearing Lips: Improving Lip-reading by Distilling Speech Recognizers, 6917-6924.

[9] Petajan E, Bischoff B, Bodoff D, et al. (1988) An improved automatic lipreading system to enhance speech recognition. ACM, 19-25.

[10] Goldschen A J, Garcia O N, and Petajan E D. (1997) Continuous automatic speech recognition by lipreading. Computational Imaging and Vision, 321-343.

[11] 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.

[12] 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).

[13] 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.

[14] 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.

[15] Chung J S, Zisserman A. (2016) Lip-reading in the wild. Asian Conference on Computer Vision, 87-103.

[16] Stafylakis T and Tzimiropoulos G. (2017) Combining residual networks with lstms for lipreading, Interspeech.

[17] Afouras T, Chung J, Senior A, et al. (2018) Deep audio-visual speech recognition. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1-1.

[18] 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.

[19] Shukla A, Vougioukas K, Ma P, et al. (2020) Visually guided self supervised learning of speech representations. IEEE International Conference on Acoustics, 6299-6303.

[20] Ma P, Petridis S, Pantic M. (2021) End-to-end audio-visual speech recognition with conformers. IEEE International Conference on Acoustics, 7613-7617.

[21] Gulati A, Qin J, Chiu C, Parmar N, Zhang Y, et al. (2020) Conformer: Convolution-augmented transformer for speech recognition. Interspeech, 5036-5040.