Volume 11, Issue 11, November – 2026 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov599

Chinese Lip-Reading Based on Residual Network and Attention Mechanism

Chuan Qin[1]; ZiLong Liu[2*]

[1]School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai, China
[2]Department of Control Science and Engineering, University of Shanghai for Science and Technology, Shanghai, China

Publication Date: 2026/11/19

Abstract: To address the common issues in existing lip-reading recognition systems, such as insufficient feature extraction and visual ambiguities, a Chinese sentence-level lip-reading recognition system based on residual networks and attention mechanisms (Spatio-Temporal Cascade Network, STCNet) is proposed. This system can first translate continuous videos containing speakers into pinyin sequences sentence by sentence, and then convert them into continuous Chinese sentences. Firstly, the residual network in the front-end feature extraction module can achieve bidirectional transmission and fusion of multi-level features, enhancing the extraction capability of long-term temporal features. In the back-end pinyin classification module, a Temporal Convolutional Network (TCN) is used to extract long-term dependency information, and an attention mechanism is introduced to weight discriminative features at different time intervals to reduce visual ambiguities. Finally, the Transformer and language model in the Chinese character prediction module output the complete Chinese sentence. Experiments show that the proposed lip-reading recognition system reduces the prediction error rate on the Chinese sentence-level lip-reading dataset CMLR by 2.29% to 4.20% compared with existing mainstream algorithms.

Keywords: *Lip-Reading; Feature Extraction; Residual Network; Attention Mechanism; Deep Neural Network.*

How to Cite: Chuan Qin; ZiLong Liu (2026) Chinese Lip-Reading Based on Residual Network and Attention Mechanism. *International Journal of Innovative Science and Research Technology,* 11(11), 710-721. https://doi.org/10.38124/ijisrt/26nov599

I. INTRODUCTION

Lip-reading recognition technology is a method of understanding a speaker's meaning through visual information, and it has broad practical applications in fields such as public safety[1], assisting people with disabilities[2], and augmenting speech recognition[3].

Existing lip-reading recognition methods can be divided into two categories: word-level lip-reading recognition and sentence-level lip-reading recognition. The goal of word-level lip-reading recognition is to identify the word category corresponding to a video clip in a vocabulary, which falls under the category of video classification tasks[4],[5], whereas the goal of sentence-level lip-reading recognition is to translate a video clip into a complete sentence[6],[7]. This paper focuses on the sentence-level lip-reading recognition task and employs deep learning methods to achieve this goal.

Currently, the visual front-end of lip-reading technology mostly employs 2D or 3D convolutional architectures to extract features from sequences of lip video frames[8]. Xu et al.[9] proposed LCANet (Cascade Attention-CTC Deep Learning Model), which can process rich semantic features; they introduced a highway network into the LipNet front-end to enhance the model's feature selection capabilities. Stafylakis et al.[10] adopted a 3D CNN + ResNet34 architecture in the visual front-end, using grayscale images as model input to further improve the extraction of spatiotemporal features from lip-reading videos. In this approach, the 3D convolutional network is responsible for extracting short-term features, while ResNet34 captures spatiotemporal lip features and simultaneously reduces the spatial dimension of the feature vectors. To mitigate the vanishing gradient problem caused by RNNs during model training, Zhang et al.[11] employed a time-domain convolutional network in the front-end to model sequences. Ma et al.[12] proposed a network called DS-TCN, which utilizes a deep separable time-domain convolutional network to reduce the number of model parameters and employs it as a teacher model for

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov599

knowledge distillation, thereby improving prediction accuracy while achieving model lightweighting.

Among the numerous improvements to the ResNet architecture, ResNeXt[13] and iResNet[14] are two representative works. While maintaining the same network depth as the original ResNet, iResNet not only avoids introducing additional computational overhead but also significantly improves model accuracy, demonstrating its generalization performance across multiple computer vision tasks such as image classification, object detection, and video action recognition. This model successfully trains ultra-deep networks exceeding 3,000 layers—a depth at which most networks struggle to converge effectively. The core contributions of iResNet include: enhancing feature representation capabilities by restructuring information flow paths; strengthening the learning efficiency of residual modules with virtually no increase in parameters or computational cost; and consistently achieving superior network performance at equivalent depths[15].

The back-end of sentence-level lip-reading systems typically needs to address the mismatch between the length of the lip sequence and the text sequence. To this end, Assael et al.[16] proposed the LipNet model, which uses a 3D CNN for feature extraction and combines bidirectional gated recurrent units with CTC in the back-end decoder. This allows the model to process long lip-reading video sequences without the need for prior alignment of video and text sequences, achieving lip-reading accuracy that surpasses that of professional lip-readers. Chung et al.[17] proposed the WLAS network model, which utilizes a novel dual-attention mechanism and incorporates bimodal input. They demonstrated that, in the presence of audio, visual information helps improve speech recognition performance, resulting in a significant performance boost on the LRS dataset. However, this method to some extent overlooks the temporal alignment issue between the audio and video modalities. Furthermore, the model uses a cross-entropy loss function during training, with the optimization objective focused on character-level prediction accuracy, whereas model performance is typically evaluated based on the overall match of the entire output sequence, such as the word error rate. This optimization approach, which prioritizes local character accuracy, may introduce discrepancies with sentence-level evaluation criteria, thereby affecting the model's performance on the final task.

Currently, lip-reading technology still has many shortcomings in terms of linguistic complexity and environmental sensitivity. At the linguistic level, most consonants in human speech cannot be clearly distinguished based on lip shape alone and must be inferred from context[18], which fundamentally limits the upper bound of recognition accuracy. Furthermore, unlike some typical public datasets, such as MNIST and CIFAR-10, lip-reading datasets are relatively scarce and cannot support the use of an entire language corpus. Additionally, the multitude of regional dialects makes it difficult to establish a unified standard[19]; In terms of environmental sensitivity, lip-reading systems in practical applications struggle to withstand visual noise caused by factors such as varying light conditions, wrinkles, and facial hair. Furthermore, speech rate and redundant information at word boundaries also pose challenges for temporal modeling[20].

To address these issues, this paper proposes a Chinese sentence-level lip-reading system based on residual networks and attention mechanisms. The relevant work is as follows:

- To address the imbalance between spatial features and temporal dependencies in temporal-visual tasks, this paper proposes the Contextual Squeeze-and-Selection Inception (CSSI) module, a multi-path feature processing unit that balances the relationship between spatial and temporal dimensions through a parallel dual-branch architecture.
- Building on this, we further propose the Contextual Squeeze-and-Selection Inception ResNet (CSSI-ResNet). This model improves the structure of the input units and residual blocks in the traditional ResNet-18 and embeds CSSI between adjacent residual blocks to enhance the representation of long-range temporal dependencies, making it better suited for lip-reading applications than traditional residual networks.
- To improve responsiveness to key articulation moments, we propose the TMA mechanism, which can adaptively identify discriminative features across different time intervals within the lip-reading system, further enhancing recognition accuracy.
- Based on CSSI-ResNet and the TMA mechanism, we propose the Spatio-Temporal Cascade Network (STCNet), which achieves a Character Error Rate (CER) reduction of 2.29% to 4.20% compared to existing mainstream models on the largest Chinese lip-reading database, CMLR.

II. STCNET

- *Overall Structure*

The sentence-level Chinese lip-reading recognition system STCNet proposed in this paper consists of a front-end preprocessing module, a feature extraction module, and a back-end pinyin classification module and Chinese character prediction module. It can process continuously input lip-reading videos, and its structure is shown in Fig .

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov599

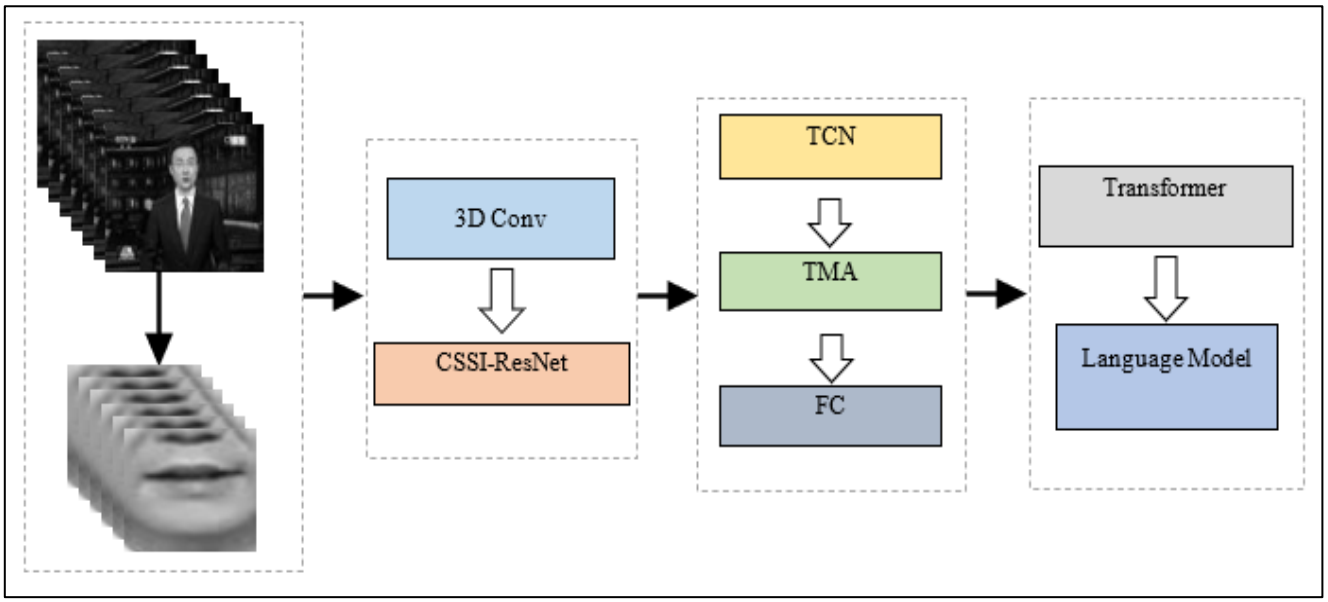


Fig 1 Overall Structure of STCNet.

On the front-end, the preprocessing module first extracts the frame sequence from the video and uses the Dlib library to locate and crop the lip images, then feeds the cropped lip image sequence into the subsequent spatiotemporal feature extraction module. The spatiotemporal feature extraction module uses a 3D convolutional neural network to perform local spatiotemporal modeling on consecutive lip frames, capturing short-term lip motion and spatial texture information. The results are then fed into the spatiotemporal interaction residual network (CSSI-ResNet), which employs the cross-stage spatiotemporal interaction unit (CSSI) to enable bidirectional propagation and fusion of multi-level features. This enhances the model's ability to represent long-range temporal dependencies while maintaining the optimization stability provided by residual connections.

The back-end pinyin classification module is based on a time-domain convolutional network (TCN) to construct a context-aware model with long-range dependencies[21], and incorporates a TMA mechanism to adapt to discriminative features across different time intervals, thereby improving responsiveness to key articulation moments. This module is responsible for mapping temporal features to pinyin-level probability distributions, providing sequence priors for subsequent text generation. The Chinese character prediction module uses a Transformer to perform contextual semantic modeling of the pinyin sequence and combines it with the CTC loss function to learn sequence alignment between pinyin and Chinese characters; during the inference stage, a language model is introduced to reorder candidate sequences, thereby improving the semantic coherence and linguistic fluency of the final Chinese character sentence.

- *CSSI*

The CSSI (Contextual Squeeze-and-Selection Inception) module is a multi-path feature processing unit designed specifically for temporal visual tasks, as shown in Fig . Its core lies in the simultaneous processing of spatial features and temporal dependencies through a parallel dual-path architecture. The input features to this module first undergo channel-dimension reduction via a 1×1×1 convolutional layer, followed by batch normalization. After normalization, the feature stream is split into two independent and complementary processing paths.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov599

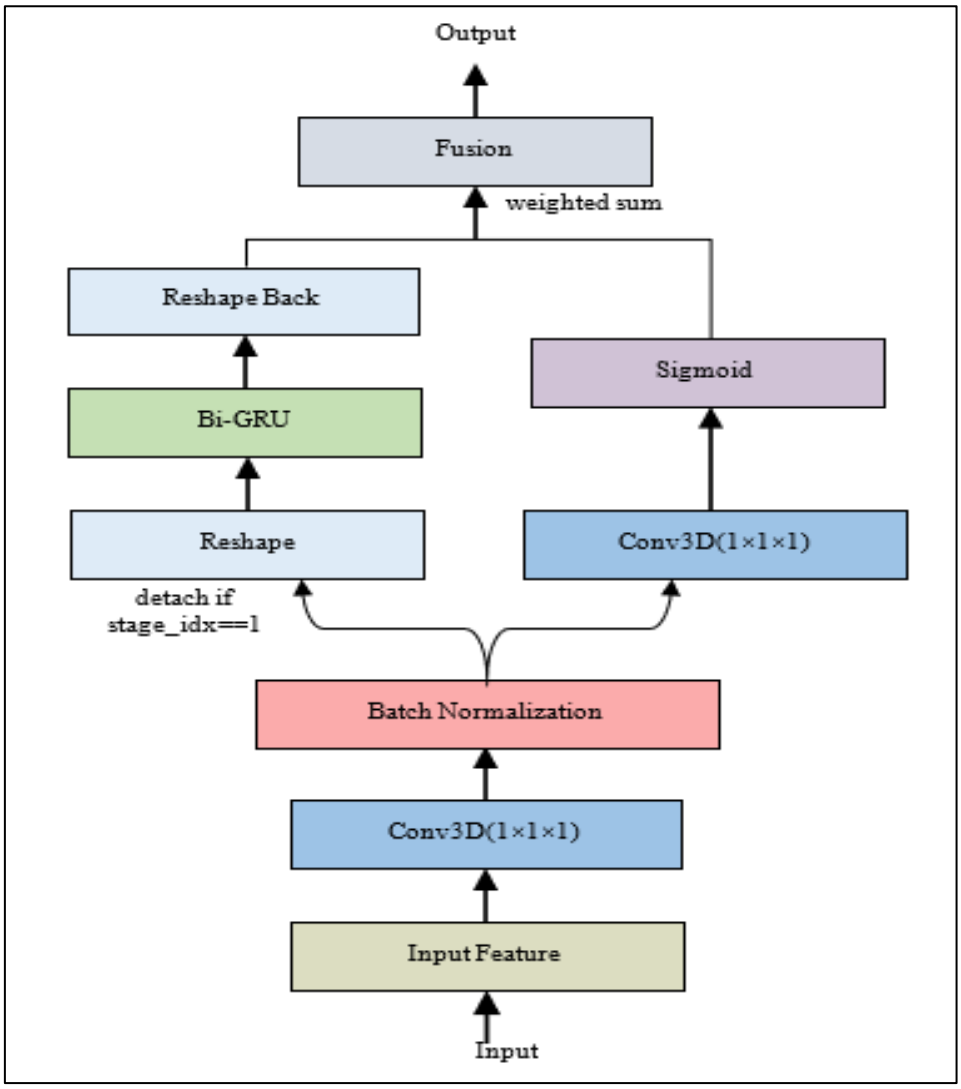


Fig 2 CSSI Network Architecture.

The left branch focuses on temporal modeling, using an optional `detach` operation to control gradient backpropagation. It then uses a `reshape` operation to convert the features into a sequence format, which is fed into a bidirectional gated recurrent unit (Bi-GRU) to capture long-range temporal dependencies. Finally, a `reshape back` operation restores the features to their original shape. The right branch uses a 1×1×1 convolution followed by a Sigmoid activation function to generate a weight value between 0 and 1 for each spatial location and channel; this value represents the importance of temporal information at each location, as assessed by the network. The outputs from the left and right branches are weighted and aggregated in the fusion layer to form output features that simultaneously contain rich spatiotemporal contextual information, as shown by the formula:

$$Output = gate \times GRU_Output \quad (1)$$

Here, `Output` represents the aggregated output of the left and right branches; `gate` represents the weight values generated by the right branch; and `GRU_Output` represents the output of the left branch.

This design enables the CSSI module to effectively coordinate the spatial correlations and temporal dynamics of visual features, making it suitable for tasks such as lip-reading that require precise spatiotemporal modeling.

- *CSSI-ResNet*

CSSI-ResNet adopts a hierarchical cascaded architecture consisting of four feature stages and three interleaved CSSI modules, progressively extracting high-level semantic features from low-level visual information, as shown in Fig .

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov599

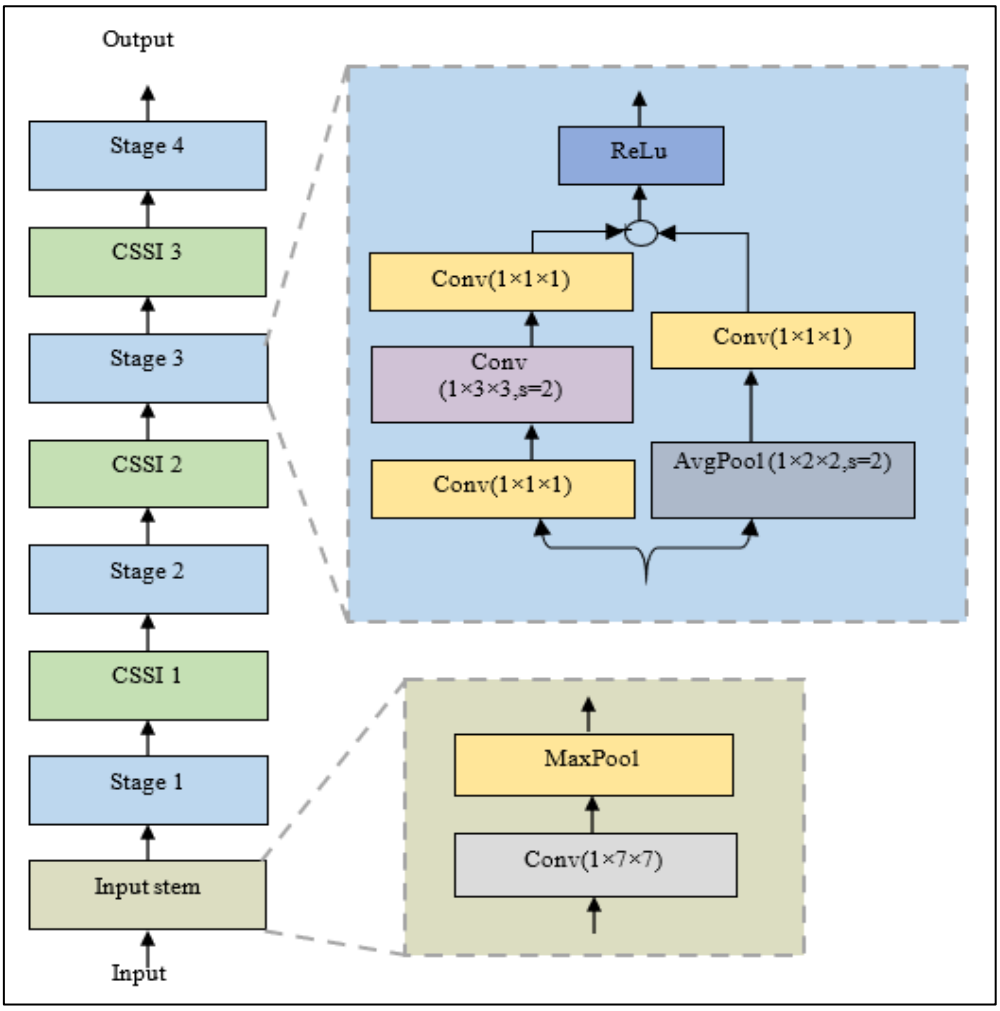


Fig 3 CSSI-ResNet Network Architecture.

The input lip-frame sequence first passes through the Input Stem, which performs preliminary feature mapping and dimensionality reduction using a 1×3×3 convolution and max-pooling layer. In Stage 1, the left branch employs a 1×3×3 convolution for downsampling, separating dimensionality reduction from spatial downsampling to reduce information loss. Meanwhile, the right branch introduces a 1×2×2 average-pooling layer to perform downsampling in the identity mapping path, preventing the subsequent 1×1×1 convolution from discarding critical lip information. The outputs of both branches are added and activated by ReLU. The subsequent three stages follow the same procedure.

The output of Stage 1 is then fed into CSSI 1. During early training, the GRU module may produce unstable gradients due to random initialization. Therefore, the left branch temporarily blocks gradient propagation to the preceding network, stabilizing training and reducing interference from unreliable temporal features. Subsequent CSSI modules adopt the same structure without gradient blocking.

Compared with standard ResNet, CSSI-ResNet offers three main advantages. First, it decouples dimensionality reduction from downsampling and uses average pooling in the identity branch, reducing information loss. Second, the CSSI module employs 3D convolutions and sigmoid-based gating to generate dynamic weights for adaptive spatiotemporal feature fusion. Third, cascading CSSI modules with residual stages enables progressive multi-level feature integration. These designs enhance the modeling of subtle lip movements and improve robustness to rapid mouth-shape changes and complex lighting conditions.

➢ *TMA*

To address the computational efficiency and feature representation requirements for processing long-sequence video data in lip-reading tasks, this paper proposes a lightweight Temporal Multiscale Attention (TMA) mechanism. This module is deployed in the back-end of a temporal convolutional network and is designed to enhance the model's ability to focus on key frames of lip movements; its structure is shown in Fig .

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov599

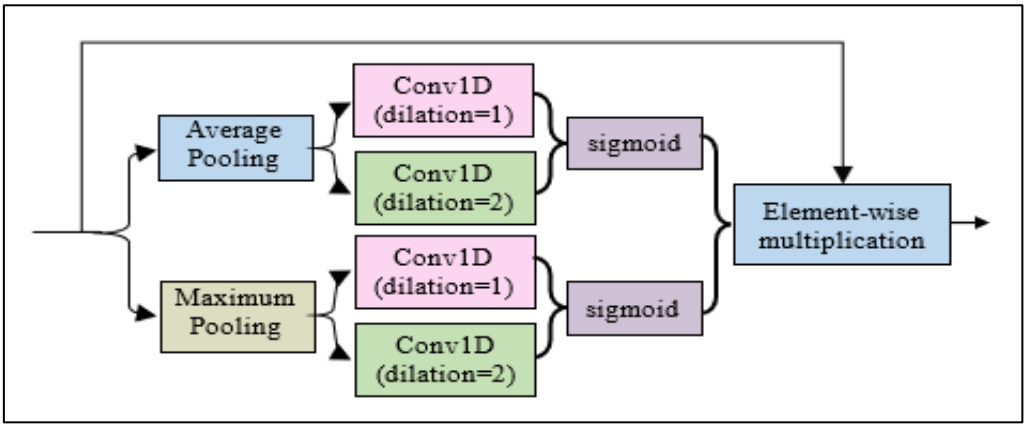


Fig 4 Schematic Diagram of the TMA Mechanism.

Given the input features $X \in R^{B\times C\times T}$, where B is the batch size, C is the number of channels, and T is the number of time steps. X first undergoes max-pooling and average-pooling operations in parallel to extract, respectively, salient local features and global statistical features from the sequence, denoted by

$$F_{\max} = MaxPool(X), \tag{2}$$

$$F_{avg} = AvgPool(X). \tag{3}$$

Subsequently, the two sets of pooled features are fed into two parallel branches for processing. Each branch contains two one-dimensional convolutional layers with different dilation rates (1 and 2), designed to capture time-dependent relationships at different scales. The specific calculation formula is

$$C_{d1}^{\max} = Conv1D(F_{\max}, K=3, d=1, p=1), \tag{4}$$

$$C_{d2}^{\max} = Conv1D(F_{\max}, K=3, d=2, p=2), \tag{5}$$

$$C_{d1}^{avg} = Conv1D(F_{avg}, K=3, d=1, p=1), \tag{6}$$

$$C_{d2}^{avg} = Conv1D(F_{avg}, K=3, d=2, p=2). \tag{7}$$

Here, d and p represent the dilation factor and padding, respectively. Dilated convolutions can increase the receptive field while keeping the number of parameters constant, effectively capturing short- and medium-range temporal contextual information[23]. Subsequently, the convolutional outputs with different dilation factors within each branch are added element-wise to fuse multiscale information, as shown in Equations:

$$M_{\max} = C_{d1}^{\max} + C_{d2}^{\max}, \tag{8}$$

$$M_{avg} = C_{d1}^{avg} + C_{d2}^{avg}. \tag{9}$$

The fused multiscale features are fed separately into the Sigmoid activation function to generate attention weight maps based on different statistical characteristics and contextual scales, as shown in Equations:

$$A_{\max} = \sigma(M_{\max}) = \frac{1}{1+e^{-M_{\max}}}, \tag{10}$$

$$A_{avg} = \sigma(M_{avg}) = \frac{1}{1+e^{-M_{avg}}}. \tag{11}$$

In particular, the sigmoid function maps values to the interval (0,1) and serves as a key metric for each time step.

Finally, by fusing the two sets of attention weights through element-wise multiplication, we obtain a comprehensive attention mask A with multiscale perception capabilities, as shown in Equation:

$$A = A_{\max} \otimes A_{avg}. \tag{12}$$

This operation ensures that the final mask reflects regions of common interest that encompass both local salient features and global average features, thereby enhancing its discriminative power. After broadcasting A along the channel dimension, it is element-wise weighted with the original input feature X to obtain the refined output feature Y, as shown in Equation:

$$Y = X \otimes A. \tag{13}$$

Compared to traditional multi-head self-attention, TMA offers significant advantages in lip-reading. Traditional methods struggle to efficiently process long video sequences due to their quadratic computational complexity, whereas TMA achieves linear complexity based on one-dimensional convolution and pooling, significantly reducing computational and memory overhead. In terms of feature modeling, TMA uses max-pooling to highlight locally salient features and employs convolutions with varying dilation rates to simultaneously capture both local details and long-range dependencies,

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov599

thereby more accurately modeling subtle changes in mouth shape and temporal dynamics. Furthermore, its inherent pooling operations enhance the model's robustness against static frames and background noise, allowing the attention to focus more on speech-relevant segments and improving generalization performance in real-world scenarios.

➢ *Chinese Character Prediction Module*

The goal of the Chinese character prediction module is to generate corresponding Chinese sentences based on an existing pinyin sequence, incorporating contextual semantic information. Its structure is shown in Fig . Let $X_p = \left\{x_{p_1}, x_{p_2}, \cdots, x_{p_{L_1}}\right\}$ be the pinyin sequence output by the previous module, and $Y_c = \left\{y_{c_1}, y_{c_2}, \cdots, y_{c_{L_2}}\right\}$ be the target Chinese character sequence, where L_1 is the length of the pinyin sequence and L_2 is the length of the Chinese character sequence.

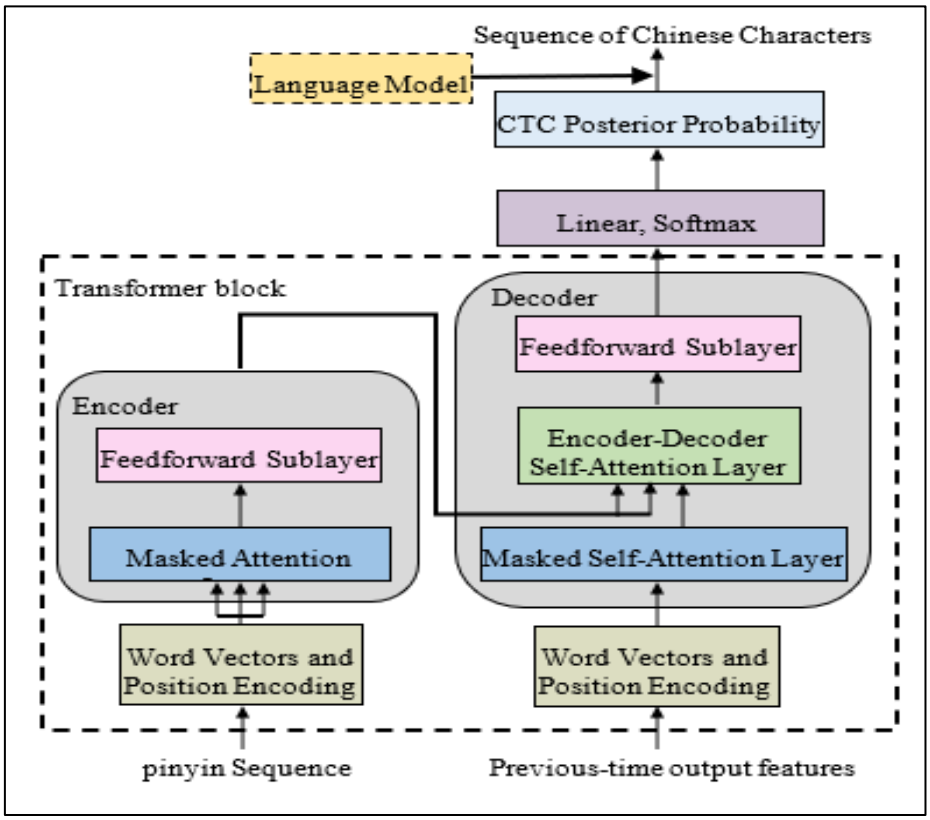


Fig 5 Structure Diagram of the Chinese Character Prediction Module.

First, the pinyin sequence is split into syllables, and then each pinyin syllable is mapped to a 512-dimensional continuous vector

$$x_i = embedding\left(x_{p_i}\right) \tag{14}$$

through an embedding layer.

After adding the position code, the final input is

$$e_i = x_i + p_i . \tag{15}$$

Here, x_{p_i} represents the i-th pinyin syllable, and p_i represents the positional embedding of the i-th word vector. With this approach, the model can obtain a semantic representation of each pinyin syllable while also explicitly capturing its relative position within the sequence.

During the encoding phase, the input sequence passes through a Transformer encoder to generate a contextual semantic representation

$$Z = Encoder\left(E\right). \tag{16}$$

The encoder consists of several stacked layers, each of which includes multi-head self-attention sublayers and feedforward network sublayers, combined with residual connections and normalization to enhance sequence modeling capabilities and stabilize the training process. Subsequently, the decoder generates the probability distribution of Chinese character categories at each time step based on the contextual features output by the encoder and the features from the previous time step, as described by equations

$$H_D = Decoder\left(y_{c<i}, Z\right), \tag{17}$$

$$h_i^D = DecoderLayer\left(y_{c<i}, Z\right). \tag{18}$$

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov599

The decoder output is then passed through a linear mapping and a Softmax layer to obtain the probability distribution

$$P\left(y_{c_i} \mid y_{c<i}, X_p\right) = \text{soft}\max\left(W_0 h_i^D + b_0\right) \tag{19}$$

for each Chinese character category.

To further constrain sequence alignment relationships, this paper introduces CTC posterior probabilities to optimize the Chinese character prediction stage. Let ω_c be an alignment path of length L, and let $F_c(\cdot)$ denote the mapping that produces the actual sequence of Chinese characters after removing duplicates and blank characters.; then, the conditional probability of the target Chinese character sequence can be written as

$$P_{char}\left(Y_c \mid X_p\right) = \sum_{\omega_c \in F_c^{-1}(Y_c)} \prod_{t=1}^{T} q_{\omega_{c_t}}^t \ . \tag{20}$$

The corresponding loss function is defined as

$$L_{char} = -\ln P_{char}\left(Y_c \mid X_p\right). \tag{21}$$

Here, T represents the time step length and denotes the predicted Chinese character at the t-th time step; $q_{\omega_{c_t}}^t$ represents the corresponding class probabilities output by the softmax layer; and $F_c^{-1}\left(y_c\right)$ is the set of all CTC paths that can be mapped to the true Chinese character sequence y_c.

During the prediction phase, the pinyin sequence is input into the pre-trained Chinese character prediction subnetwork, and beam search is used to find the optimal Chinese character sequence among the candidate paths. Specifically, beam search retains several high-probability candidate results at each time step and continuously filters them based on cumulative probability, ultimately yielding the complete output.

To further incorporate linguistic constraints, this paper performs shallow fusion[6] between the external language model and the decoding results; the combined score can be expressed as

$$Score\left(Y_c\right) = \log P_{ctc}\left(Y_c \mid X_p\right) + \lambda \log P_{LM}\left(Y_c\right). \tag{22}$$

Here, λ represents the language model weights, and $P_{LM}\left(Y_c\right)$ denotes the prior probability assigned by the language model to the candidate Chinese character sentences. This strategy leverages the patterns of Chinese word order to filter out candidate results that do not conform to grammatical or semantic rules, thereby improving the accuracy of Chinese character recognition in long-sentence scenarios. The language model used in this paper is a Chinese pre-trained model open-sourced by the Harbin Institute of Technology–iFlytek Joint Laboratory (HFL).

III. EXPERIMENTAL RESULTS AND ANALYSIS

➢ *Dataset*

The dataset used in this paper is CMLR, currently the largest open-source Chinese sentence-level dataset. It contains 102,072 sentences spoken by 11 speakers from recordings of the "Xinwen Lianbo" news program between 2009 and 2018[24]. Each sentence contains up to 29 Chinese characters and excludes English words, Arabic numerals, and rare punctuation marks. The dataset was randomly split into training, validation, and test sets in a 7:1:2 ratio. This dataset provides the model with a wealth of data on dynamic changes in lip movements[25].

➢ *Data Preprocessing*

This paper converts the video into a continuous sequence of frames at a rate of 25 fps and uses it as input data; each video sample is transformed into a grayscale feature matrix containing information on lip motion. The preprocessing workflow is shown in Fig . First, dlib is used to detect and track 68 facial landmarks[26] (facial landmarks). Based on these landmarks, a fixed 64×128-pixel bounding box is used to locate and crop the lip region. Finally, OpenCV is used to convert the cropped RGB image into a grayscale image, which is then normalized based on the mean and variance of the entire training set to eliminate lighting variations and accelerate model convergence.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov599

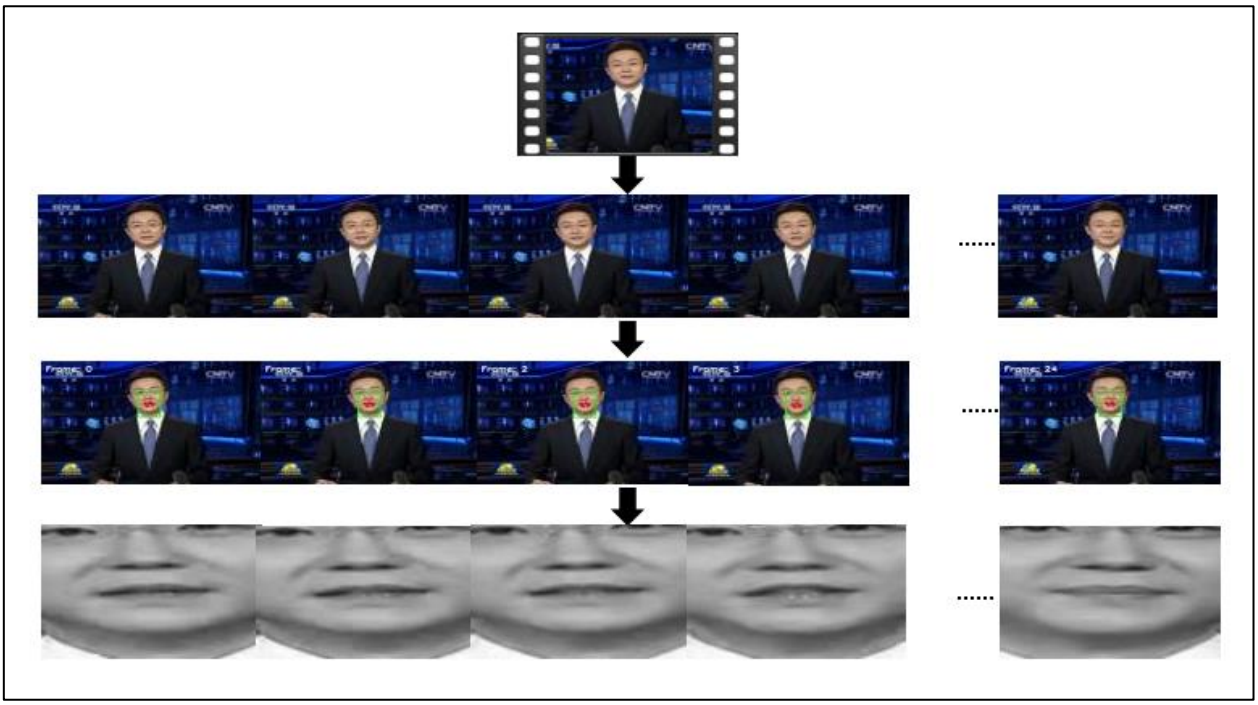


Fig 6 Video Data Preprocessing.

To address the lack of pinyin annotations in the CMLR dataset, this paper uses the pinyin library during the data preprocessing stage to automatically convert Chinese character text into corresponding pinyin sequences. Since the semantic unit division in pinyin labels can effectively enhance a lip-reading model's ability to model Chinese word structure, the model retains word boundary spaces[27] during data preprocessing to assist the model in learning pronunciation patterns.

➢ *Training Setup*

This study implemented the STCNet model using the PyTorch deep learning framework. The experimental environment consisted of an NVIDIA GeForce GTX 4060 GPU and the Python 3.10.16 programming environment. During model training, the pinyin classification module was optimized using the CTC loss function and the stochastic gradient descent (SGD) algorithm. The initial learning rate was set to 10^{-3}, and a dynamic adjustment mechanism was incorporated: if the training loss did not decrease in any of four consecutive training epochs, the learning rate was reduced by 50%. During the Chinese character prediction training phase, the CTC loss function was used in conjunction with the Adam optimizer for parameter optimization. The initial learning rate was set to 10^{-4}, and the dynamic adjustment mechanism was as follows: if the loss value did not decrease over six consecutive training epochs, the learning rate was reduced by 50%, eventually reaching 10^{-5}.

➢ *Comparison with Existing Methods*

To evaluate the performance of STCNet, we selected the following classic methods for comparison:

- LipCH-Net[7]. An end-to-end Chinese sentence-level lip-reading method that predicts pinyin from video and then predicts Chinese characters from the pinyin.
- CSSMCM[30]. A triple Seq2Seq method based on encoder-decoder attention, which proceeds from video to pinyin, then to tone, and finally to Chinese characters.
- CHSLR-VP[31]. An end-to-end Chinese sentence-level lip-reading method that proposes visual pinyin as an intermediate representation, predicting visual pinyin from video and then predicting Chinese characters from visual pinyin.
- 3DT-CHLipNet[28]. A Chinese sentence-level lip-reading method that replaces LSTM with TCN, processing video to pinyin to Chinese characters.
- APAT-LR[8]. An end-to-end lip-reading method based on an adaptive pooling attention Transformer that directly predicts Chinese characters from video.

Table . A "–" in the table indicates that the metric was not used in the original literature. The evaluation metrics are the pinyin error rate (PER) and the character error rate (CER). As shown in the table, STCNet achieved the best PER and CER on the CMLR dataset, at 22.39% and 26.11%, respectively. Compared to APAT-LR, the current state-of-the-art model, the CER was reduced by 2.29%. This is primarily due to CSSI-ResNet enhancing

The prediction accuracy of each method on the CMLR dataset is shown in

spatiotemporal feature interaction capabilities, the TMA mechanism strengthening the modeling of key phonetic segments, and the Transformer and language model further utilizing contextual semantic constraints on candidate sequences, thereby effectively reducing recognition errors caused by visual ambiguity.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov599

Table 1 Comparison of Pinyin and Character Error Rates Between the Algorithm Described in This Paper and the Comparison Algorithm.

Model	PER/%	CER/%
LipCH-Net	41.49	54.30
CSSMCM	36.22	32.48
CHSLR-VP	24.52	32.62
3DT-CHLipNet	24.47	30.31
APAT-LR	-	28.40
STCNet	**22.39**	**26.11**

➢ *Ablation Experiments*

- *Residual Networks and Attention Mechanisms*

To verify the effectiveness of the CSSI-ResNet and TMA mechanisms in the field of lip-reading, this study conducted ablation experiments focusing on these two modules while keeping the Chinese character prediction module unchanged, using a 3D CNN+ResNet18+TCN architecture as the baseline model. The results are shown in

Table . The baseline model achieved a PER of 30.67% and a CER of 33.97% on the CMLR dataset. Upon introducing the CSSI-ResNet module to replace the original ResNet18 backbone network, both the model's PER and CER decreased, confirming that the CSSI module can effectively enhance the model's low-level representation capabilities and recognition accuracy by strengthening spatiotemporal feature interaction. Furthermore, when only the TMA mechanism was integrated into the baseline model, the PER and CER decreased to 24.83% and 27.26%, respectively, highlighting the significant role of TMA in focusing on key phonetic segments and suppressing irrelevant information. Finally, the best performance was achieved when the CSSI-ResNet and TMA mechanisms were applied jointly to the model, with the PER and CER reduced to 22.39% and 26.11%, respectively. This set of results not only quantitatively validates the independent effectiveness of each module but also demonstrates that the two are functionally complementary; working in concert, they can maximize the overall performance of the system.

Table 1 Results of Ablation Experiments on Residual Networks and Attention Mechanisms.

Network Architecture	PER/%	CER/%
3D CNN+ResNet18+TCN	30.67	33.97
3D CNN+CSSI-ResNet+TCN	25.94	29.07
3D CNN+ResNet18+TCN+TMA	24.83	27.26
3DCNN+CSSI-ResNet+TCN+TMA	**22.39**	**26.11**

- *Language Models*

In both the pinyin classification and Chinese character prediction modules, the CTC loss function is used for parameter optimization. The CTC algorithm is based on the assumption that each time step in the input sequence is independent of the others; however, in actual speech or video sequences, adjacent time steps often exhibit a certain degree of semantic continuity. To make better use of contextual information during decoding, an external language model is typically introduced during the CTC beam search phase to improve the overall performance of sequence recognition. However, because the features extracted by the pinyin classification module lack explicit linguistic constraints, the performance improvement achieved by introducing a language model is relatively limited for this module. In contrast, in the subsequent Chinese character prediction module, the language model can effectively leverage the correspondence between pinyin sequences and Chinese characters, thereby significantly improving the system's language generation quality and recognition accuracy[28].

From a system architecture perspective, the integration of a language model provides the Chinese character prediction module with additional linguistic prior information. Without a language model, the system relies solely on the mapping from visual features to pinyin, making it prone to misclassifying homophones due to similar mouth shapes or environmental noise. By analyzing statistical language patterns, the language model can reorder candidate pinyin sequences based on confidence scores, suppressing low-probability and syntactically anomalous outputs, thereby reducing the sentence-level recognition error rate. As shown in Fig , after integrating the language model, the Chinese character prediction module's error rate for sentence prediction decreased significantly, with the reduction being particularly pronounced in medium- to long-sentence scenarios. This is because long sequences contain richer contextual information, allowing the language model to perform more accurate semantic disambiguation and infer sentence completeness; consequently, it enhances the overall robustness of recognition more effectively than in short sentences.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov599

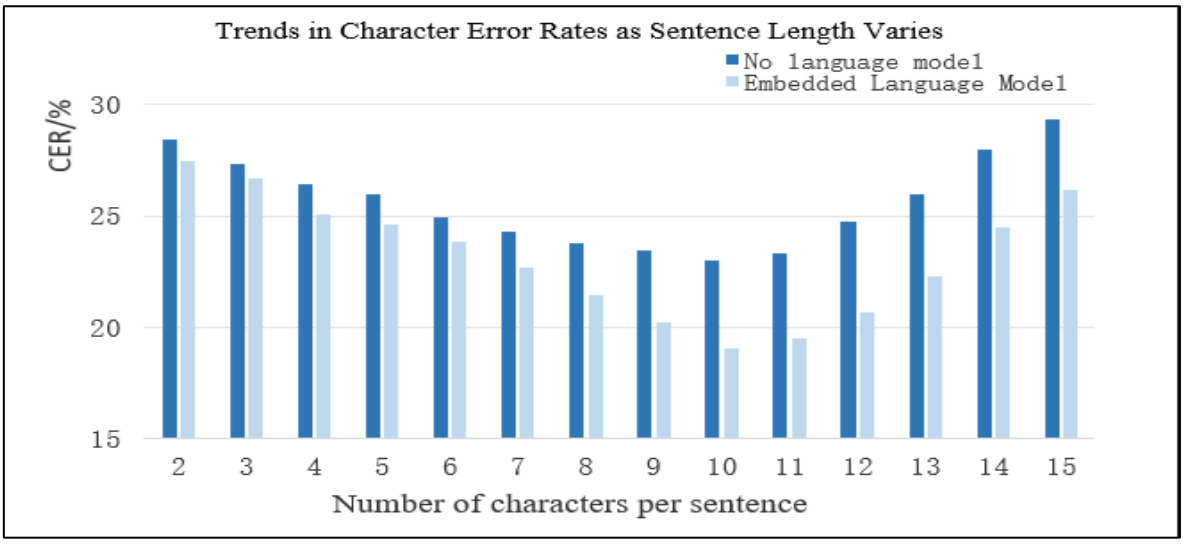


Fig 7 The Impact of Language Models on Predicting Chinese Sentences of Different Lengths.

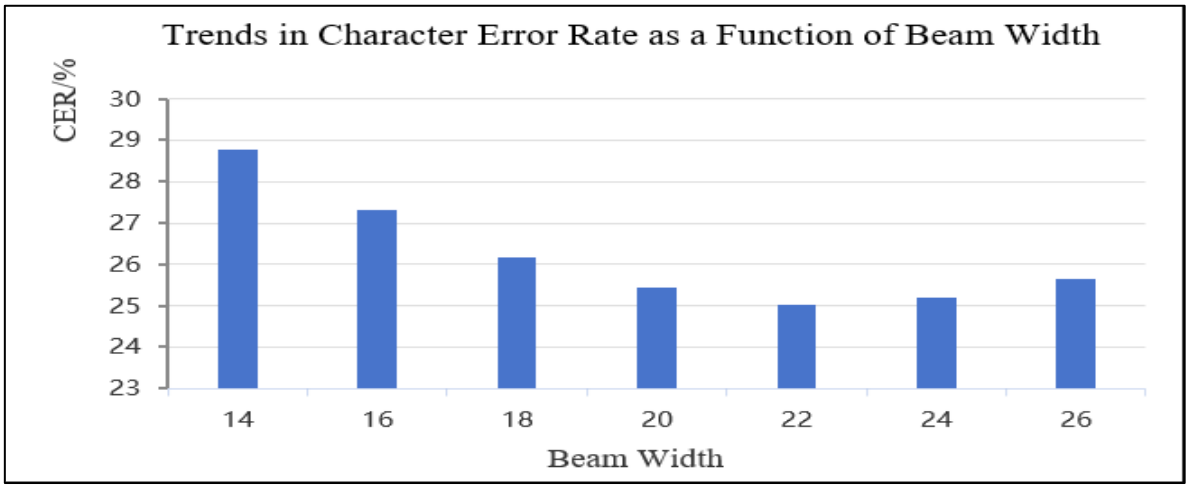


Fig 8 The Effect of Different Beam Widths on Character Error Rates.

Furthermore, the beam width of the beam search algorithm is a key factor affecting decoding efficiency and performance. Generally, increasing the beam width preserves more candidate paths, thereby improving the global optimality of the generated sequence; however, it also leads to a linear increase in computational complexity and memory usage[29]. Excessive computational complexity introduces more uncertainty, which reduces the accuracy of Chinese character prediction. Therefore, in resource-constrained scenarios, a wider beam is not necessarily better. Through experimentation, this paper found that expanding the beam width to 22 yields the optimal character error rate. The experimental results are shown in Fig , and their quantitative relationship further validates the positive effect of expanding the search space on long-sequence decoding after incorporating language priors.

IV. CONCLUSION

To address insufficient accuracy in lip-reading, this study proposes STCNet, which integrates residual networks and attention mechanisms. CSSI-ResNet enables multi-level spatiotemporal feature fusion to capture long-range dependencies, while TMA enhances key articulation segments within the TCN framework. A Transformer and an external language model are further employed for contextual modeling and semantic refinement. Experiments on the CMLR dataset show that STCNet achieves a recognition accuracy of 73.89%, demonstrating its effectiveness in Chinese sentence-level lip-reading.

Although developed for Chinese lip-reading, STCNet is not inherently language-specific. Its spatiotemporal feature extraction, multiscale temporal modeling, and attention-based fusion mechanisms rely primarily on visual motion patterns, offering potential applicability to other languages. Moreover, its modular architecture allows language-specific components to be adapted or replaced for different linguistic systems. However, cross-lingual generalizability requires further validation. Meanwhile, the relatively large model size presents a challenge for lightweight deployment. Future work will focus on model compression, computational efficiency, and multilingual evaluation.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/26nov599

REFERENCES

[1]. Y. Q. Ren, G. D. Tian, X. D. Zhou, et al., "Lip-reading algorithm in face recognition systems for high security," *Application Research of Computers* 34(4), 1221-1225+1230 (2017).

[2]. J. H. Lei, "Experimental study on lip-reading Chinese character pronunciation recognition in hearing-impaired students," Dissertation, East China Normal University, Shanghai (2006).

[3]. Y. M. Wang, K. Chen, Aihaitilamu Saabdusalam, "End-to-end audio-visual speech recognition based on SDBN and BLSTM attention fusion," *Telecommunications Science* 35(12), 79-89 (2019).

[4]. J. S. Chung, A. Zisserman, "Lip reading in profile," in *Proc. 28th British Machine Vision Conference*, London, UK, ChungZ17 (2017).

[5]. S. Yang, Y. Zhang, D. Feng, et al., "LRW1000: A naturally-distributed large-scale benchmark for lip reading in the wild," in *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, 1-8, IEEE (2019).

[6]. T. Afouras, J. S. Chung, A. Senior, et al., "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell*. 44(12), 8717-8727 (2018).

[7]. X. Zhang, H. Gong, X. Dai, et al., "Understanding pictograph with facial features: End-to-end sentence-level lip reading of Chinese," in *Proc. AAAI Conf. Artif. Intell*. 33(01), 9211-9218 (2019).

[8]. Y. Yao, Z. X. Hu, T. Deng, et al., "A lip reading method based on adaptive pooling attention transformer," *Chinese Journal of Intelligent Science and Technology* 7(2), 211-220 (2025).

[9]. K. Xu, D. W. Li, N. Cassimatis, et al., "LCANet: End-to-end lipreading with cascaded attention-CTC," in *Proc. 2018 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG 2018)*, 548-555, IEEE Press, Piscataway (2018).

[10]. T. Stafylakis, G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," *arXiv preprint* (2017) [doi:arXiv:1703.04105].

[11]. T. Zhang, L. He, X. D. Li, et al., "Efficient end-to-end sentence-level lip reading with temporal convolutional networks," *Appl. Sci*. 11(15), 6975-6982 (2021).

[12]. P. C. Ma, B. Martinez, S. Petridis, et al., "Towards practical lipreading with distilled and efficient models," in *Proc. ICASSP 2021 - 2021 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, 7608-7612, IEEE Press, Piscataway (2021).

[13]. S. N. Xie, R. Girshick, P. Dollár, et al., "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 5987-5995, IEEE Computer Society Press, Los Alamitos (2017).

[14]. I. C. Duta, L. Liu, F. Zhu, et al., "Improved residual networks for image and video recognition," in *Proc. 25th Int. Conf. Pattern Recognit*., 9415-9422, IEEE Computer Society Press, Los Alamitos (2021).

[15]. J. L. Ma, Z. W. Guo, Z. P. Ma, et al., "Multi-scale gated spatio-temporal enhancement for lip recognition," *Journal of Computer-Aided Design & Computer Graphics* 37(7), 1229-1238 (2025).

[16]. Y. M. Assael, B. Shillingford, S. Whiteson, et al., "LipNet: End-to-end sentence-level lipreading," *arXiv preprint* (2016) [doi:arXiv:1611.01599].

[17]. J. S. Chung, A. Senior, O. Vinyals, et al., "Lip reading sentences in the wild," in *Proc. 2017 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 3444-3453, IEEE Press, Piscataway (2017).

[18]. B. X. Wang, "Design and development of a lip-reading recognition system based on multi-feature fusion inference," Dissertation, Beijing University of Posts and Telecommunications, Beijing (2025) [doi:10.26969/d.cnki.gbydu.2025.002166].

[19]. W. L. Wu, R. K. Li, S. Y. Xu, et al., "End-to-end lip-reading algorithm based on adaptive sequence frame length," *Life Science Instruments* 21(4), 35-39 (2023).

[20]. Y. Y. Wang, P. Wang, K. C. Wu, "Research on lip reading based on self-attention sequence model," *Chinese Journal of Electron Devices* 44(3), 624-627 (2021).

[21]. C. Wang, "Research on lip reading methods based on deep learning," Dissertation, Shenyang University of Technology, Shenyang (2025) [doi:10.27322/d.cnki.gsgyu.2025.000966].

[22]. M. Miao, "Research on multimodal Chinese lip reading methods," Dissertation, Inner Mongolia University, Hohhot (2025) [doi:10.27224/d.cnki.gnmdu.2025.001040].

[23]. Y. W. Zhang, "Design and implementation of a lip-reading algorithm based on multi-scale feature fusion," Dissertation, Beijing University of Posts and Telecommunications, Beijing (2025) [doi:10.26969/d.cnki.gbydu.2025.001578].

[24]. W. J. Xu, K. Li, "Sentence-level lip reading based on optical flow and attention mechanism," Chinese Journal of Electron Devices 46(5), 1339-1348 (2023).

[25]. J. He, X. Wang, "Research progress on Chinese lip-reading corpus," Journal of Information Engineering University 26(6), 690-696 (2025).

[26]. C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, et al., "300 faces in-the-wild challenge: The first facial landmark localization challenge," in Proc. IEEE Int. Conf. Comput. Vis. Workshops (2013).

[27]. A. Cutler, Prosody and the word boundary problem, Psychology Press, Germany (2014).

[28]. P. P. Liu, "Research on Chinese sentence-level lip reading based on Transformer," Dissertation, North China Electric Power University (Beijing), Beijing (2023) [doi:10.27140/d.cnki.ghbbu.2023.001292].

[29]. P. P. Liu, J. P. Jia, "Chinese sentence-level lip reading algorithm based on temporal convolutional network," Application Research of Computers 40(9), 2596-2602 (2023) [doi:10.19734/j.issn.1001-3695.2023.02.0051].

[30]. Y. Zhao, R. Xu, M. Song, "A cascade sequence-to-sequence model for Chinese Mandarin lip reading," in Proc. ACM Multimedia Asia, 1-6, ACM, New York (2019).

[31]. S. He, J. B. Yuan, Y. Y. Lu, "Research on lip-reading method based on visual characteristics of Chinese pronunciation," Computer Engineering and Applications 58(4), 157-162 (2022).