Multimodal Sparse Transformer Network for Audio-Visual Speech Recognition

Qiya Song, Bin Sun, *Member, IEEE*, and Shutao Li, *Fellow, IEEE*

***Abstract*— Automatic speech recognition (ASR) is the major human–machine interface in many intelligent systems, such as intelligent homes, autonomous driving, and servant robots. However, its performance usually significantly deteriorates in the presence of external noise, leading to limitations of its application scenes. The audio-visual speech recognition (AVSR) takes visual information as a complementary modality to enhance the performance of audio speech recognition effectively, particularly in noisy conditions. Recently, the transformer-based architectures have been used to model the audio and video sequences for the AVSR, which achieves a superior performance. However, its performance may be degraded in these architectures due to extracting irrelevant information while modeling long-term dependences. In addition, the motion feature is essential for capturing the spatio-temporal information within the lip region to best utilize visual sequences but has not been considered in the AVSR tasks. Therefore, we propose a multimodal sparse transformer network (MMST) in this article. The sparse self-attention mechanism can improve the concentration of attention on global information by selecting the most relevant parts wisely. Moreover, the motion features are seamlessly introduced into the MMST model. We subtly allow motion-modality information to flow into visual modality through the cross-modal attention module to enhance visual features, thereby further improving recognition performance. Extensive experiments conducted on different datasets validate that our proposed method outperforms several state-of-the-art methods in terms of the word error rate (WER).**

***Index Terms*— Audio-visual speech recognition (AVSR), cross-modal attention, motion information, multimodal, sparse transformer.**

I. Introduction

Automatic speech recognition (ASR), which aims to convert human speech information into readable text content and enables machines to “understand” human voices, is a key technology for human–machine interaction. Among the various expressions of human beings, the audio modality carries most of the useful information that is considered the most convenient and natural way of interaction. As deep learning improves by leaps and bounds, the latest ASR systems in a quiet environment can achieve a recognition accuracy of more than 95% [1], which has exceeded the recognition accuracy of humans. However, in the actual application of human–computer interaction, the speech information significantly deteriorates in noisy signals, which seriously affects speech recognition performance [2], [3]. Visual information is obtained through video and will not be affected in real-world environments. It also plays a crucial part in people’s communication and acoustic speech recognition [4]–[6].

Visual speech recognition (VSR), also called lip reading, is to identify utterances by analyzing the visual recording of the speaker’s mouth without any audio input. It has a variety of valuable applications, such as: 1) aids for people who cannot hear clearly (dysaudia) or generate the sounds (aphonia); 2) isolates the target speaker’s sound from the mixed voice; and 3) recovers high-quality voice in a noisy environment [7], [8]. Due to the multimodal way humans perceive the environment [9], the audio-visual speech recognition (AVSR) system uses visual information as a complementary modality to improve speech recognition performance. It has attracted considerable interest of many researchers for decades. The relative visual modality here mainly refers to the lip or face area recorded by visual sensor when the speaker pronounces. We can use the relative visual modality of speech content to disambiguate confusing sounds easily. For instance, the nasal sounds /m/ and /n/ that are auditorily similar in a noisy condition being visually distinct. Vice versa: the phonemes /p/ and /b/ look very similar in appearance, but they can be distinguished acoustically [10]. Moreover, how to effectively establish the correlation between visual and audio modalities and fuse the information of different modalities are the main challenges in AVSR. It should improve the recognition performance of the multimodal system compared with the unimodal system in the noisy or quiet scenario. The early research on audio-visual fusion methods is reviewed in [5] and [11]. We illustrate the outline of a traditional AVSR pipeline in Fig. 1. Among them, the mel-frequency cepstral coefficient (MFCC) operation is most successful to obtain the hand-crafted features in traditional audio signal processing. However, in the deep-learning-based research, the MFCC operation has been proven to be unnecessary, since it removes information and destroys spatial relations [12]. We select the short-time Fourier transform (STFT) to obtain audio features which have been effective in speech recognition tasks [10], [13].

In recent years, most of the existing approaches use recurrent neural networks (RNNs) as the sequence processing unit and the sequence-to-sequence network as the overall

Manuscript received 10 January 2021; revised 7 July 2021, 9 November 2021, and 18 January 2022; accepted 25 March 2022. Date of publication 12 April 2022; date of current version 1 December 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2018YFB1305200 and in part by the National Natural Science Fund of China under Grant 62171183. *(Corresponding author: Bin Sun.)*

The authors are with the College of Electrical and Information Engineering and the Key Laboratory of Visual Perception and Artificial Intelligence of Hunan Province, Hunan University, Changsha 410082, China (e-mail: sqyunb@hnu.edu.cn; sunbin611@hnu.edu.cn; shutao_li@hnu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TNNLS.2022.3163771.

Digital Object Identifier 10.1109/TNNLS.2022.3163771

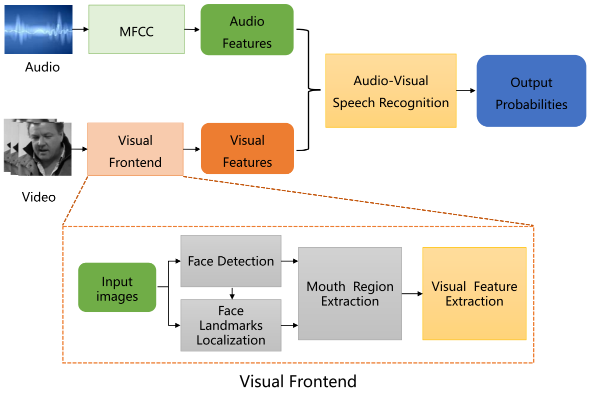


Fig. 1. Overview of a traditional AVSR pipeline based on hand-crafted features.

architecture to model the sequential information in the video. Petridis *et al.* [14] proposed a unified hybrid architecture, which utilized the neural network to extract visual and audio features and used a bidirectional long short-term memory (Bi-LSTM) network to model the temporal dynamics in each modality. Afouras *et al.* [15] compared two outstanding models on the basis of transformer self-attention architecture [16]. One is based on sequence-to-sequence loss (TM-seq2seq) and the other is a transformer with connectionist temporal classification loss (TM-CTC), which achieved a superior recognition performance. However, the transformer self-attention architecture can establish long-term context information, but it may extract the irrelevant information. In addition, the current AVSR methods do not make use of the lip motion features, which is important for converting the lip image to the fundamental element of visual speech cues.

Therefore, we introduce a multimodal sparse transformer network (MMST) in this article that can learn discriminative features of three different modalities (i.e., video, motion, and audio) to improve speech recognition performance. The motion features are used to enhance the visual features by the cross-modal attention mechanism, which models the correlation between sequential and motion features. MMST allows for complex reasoning over visual and audio features, which are jointly mapped into different subspaces and fused into the joint feature conditioned on the words previously decoded. Generally speaking, the contributions of the proposed method are as follows.

1) We introduce the multimodal sparse transformer network for AVSR, which improves the concentration of attention on context information by the explicit selection of the most relevant segments.
2) We merge motion features and visual features by the cross-modal attention mechanism to provide more diverse and discriminative visual representations. The optical flow is used to represent spatiotemporal motion-dependent information as complementary cues to the framewise features of the video sequence.
3) We conduct extensive experiments on the different datasets to verify the superiority of our method over several state-of-the-art methods in different noisy conditions.

The rest of this article is arranged as follows. The recent works about lip reading and AVSR are described in Section II. We describe the proposed multimodal sparse transformer method in detail in Section III. The experiments are introduced and the results are discussed in Section IV. Finally, the conclusive remarks are made and the future work is pointed out.

II. Related Work

In this section, we make a brief review of the previous studies on lip reading, AVSR, and attention models.

A. Lip Reading

Lip reading, also called VSR, aims to transcript the silent video to text sentences by capturing the motion of the speaker's lips. With the rapid development of deep learning in various aspects, it has been used in the lip-reading tasks and achieved considerable performance [17]–[19]. Chung and Zisserman [20] developed CNN architectures to map multiframe appearances to phonemes, which also proposed a pipeline to collect a large-scale dataset. Assael *et al.* [21] first proposed an end-to-end sentence-level model for lip reading, which employed a spatio-temporal CNN to extract powerful features and bidirectional gated recurrent unit sequence model to predict tokens.

Recently, a two-stream network that consists of the visual stream and optical flow modality is proposed in [22] and [23], which have achieved a remarkable performance for video tasks. They show that optical flow is an effective way to model spatio-temporal information within in adjacent images. Weng and Kitani [24] employed optical flow that contains motion information as a complementary modality to enhance recognition performance. They use the two-stream deep 3-D CNNs (C3D) to extract features (grayscale images and optical flow streams) and Bi-LSTM to model context dependences, which achieve superior results on word-level lip reading. However, the conclusion that the deep C3D are prone to overfitting during training is presented in [24]. Different from this work, we develop the two-stream C3D-pseudo-3D (P3D) as visual front end to alleviate this status and obtain more discriminative representations. Also, we use the cross-modal attention to enhance the visual feature, which can make the motion-modality information potentially flow into visual modality to represent visual features.

B. Audio-Visual Speech Recognition

The AVSR takes the visual modality as assisted information to improve speech recognition performance, especially in strong noise cases. Recently, it has made impressive progress and development. Chung *et al.* [25] proposed a WLAS (watch, listen, attend, and spell) model that has a two-stream attention mechanism including the visual (lip) stream and the audio (speech) stream. The LSTM is used to model the temporal information of the input sequences to obtain the context vectors, which are selected by the dual-attention mechanism to generate the result.

Afouras *et al.* [15] first introduced the transformer, the well-known self-attention-based sequence-to-sequence architecture, for the AVSR. The self-attention represents each sequence's context information of variable length with a fixed dimension

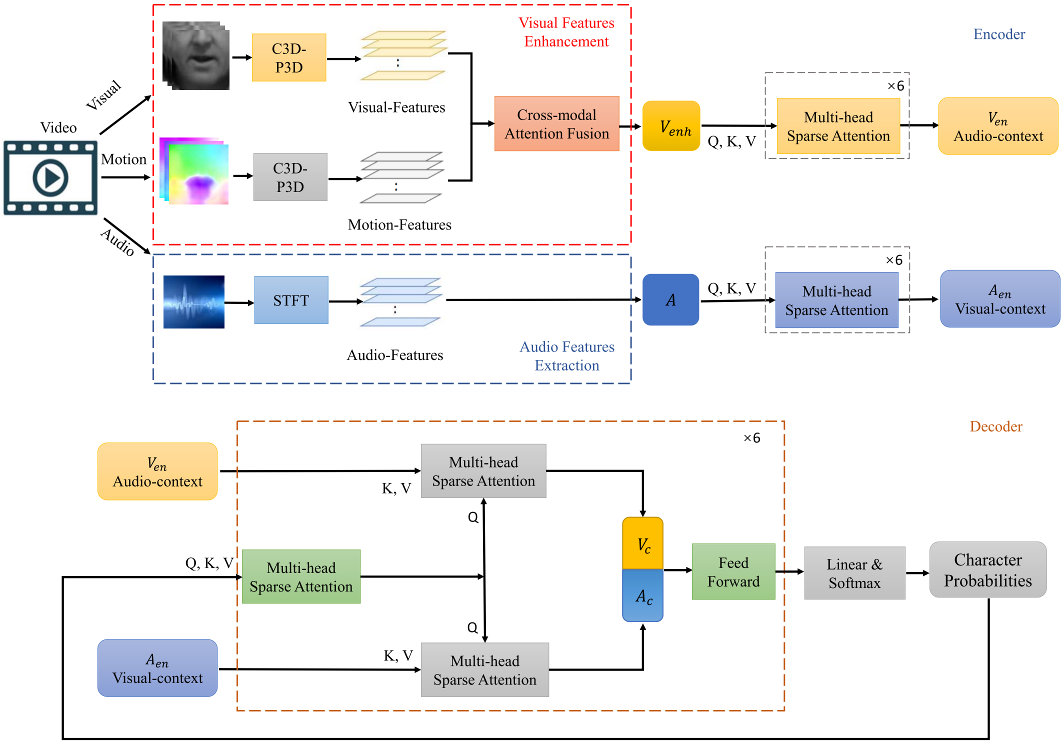


Fig. 2. Architecture of our proposed multimodal sparse transformer network for AVSR. Encoder: The block V_{enh} represents the enhanced visual feature, which is the output of the cross-modal attention fusion block aggregating the original visual feature and the optical flow. The block A is the audio features obtained from the STFT of the raw audio signal. Decoder: The Query (Q), Key (K), and Value (V) are tensors for each multimodal sparse attention block. In the encoder–decoder layers, the K and V tensors come from encoder layers in V_{en} or A_{en}. The Q tensors are the output of the previous decoding layer. In the self-attention layers, the Q, K, and V tensors are in the same layer.

vector on the encoder side. On the decoder side, the context-dependent vectors of different modalities are reweighted with the multihead attention layers and sent to the feed-forward layer to generate the probability of characters. However, several state-of-the-art methods use self-attention to build context dependence, but it will extract irrelevant information due to its global dependence. The sparse attention [26] is proposed to discard irrelevant segments in the context. The experimental results for three applications including neural machine translation, image captioning, and language modeling demonstrate the effectiveness of the sparse attention mechanism. Inspired by this, the sparse attention is introduced to improve the concentration of attention on context information by selecting the most relevant segments for AVSR.

C. Attention Models

The attention technique in deep learning models has achieved a great success in various tasks. The traditional attention is generally used in the encoder–decoder framework to capture information for context, which reweighs the importance of all input features for each output. Then, such an attention mechanism is expanded in [16] to capture the long-term dependences, which can attend a word to all other words for building relations within a sequence of features. Sterpu *et al.* [10] used LSTM to model the temporal information for the input sequence. They directly calculated the weight matrices between the visual and audio hidden features of the LSTM, which is like a standard unimodal attention decoder. Zhou *et al.* [27] proposed the modality attention mechanism, which aggregates information from various modalities through the attention weights, and merges the input multimodal information into a unified representation.

Unlike these previous works, we proposed multimodal sparse transformer network (MMST) in our work. The motion information is obtained from the movement of pixels in adjacent images, which have less repetition of information between them and the correlation is higher than the visual and audio modalities. Therefore, the cross-modal attention module is first used to enhance the representation of visual information due to the strong correlation between motion modality and visual modality. Then, the sparse transformer is used to attend the enhanced visual features and audio features to obtain the transcription result.

III. Proposed Approach

In this section, we illustrate the presented multimodal sparse transformer (MMST) as shown in Fig. 2 for AVSR, which mainly consists of multimodal features' encoder and characters' decoder. The main components of MMST are described in detail in the following.

A. Feature Representation

Three major modalities are considered in our method for AVSR: audio (A), visual (V), and motion (O) modalities. First, the audio modality is separated from the video of

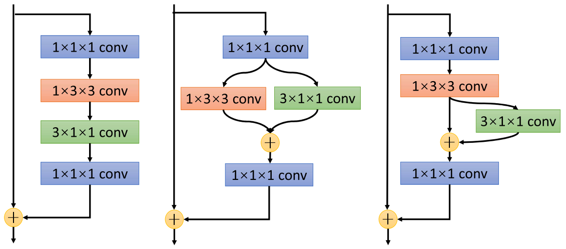


Fig. 3. Bottleneck building blocks of the P3D. From left to right, the three versions of P3D are P3D-A, P3D-B, and P3D-C, respectively. P3D ResNet is generated by interleaving the three versions sequentially, which can split the $3 \times 3 \times 3$ convolution into $1 \times 3 \times 3$ spatial convolution and $3 \times 1 \times 1$ temporal convolution.

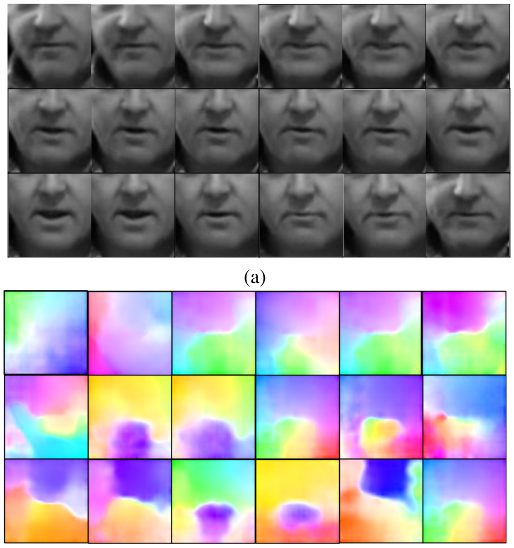
an utterance. Then, we use the OpenCV-toolkit to extract the visual frames. Finally, the obtained visual modality is employed to generate motion information. We obtain the input features' sequences from these three modalities as follows.

1) Audio Features: For the input audio representation, the STFT is utilized to extract spectral magnitudes with a 25-ms window length and 10-ms hop length from a waveform signal at 16-kHz sampling rate. Due to the video frame rate at 25 frames/s, we concatenate multiple audio frame features in a sequence of 4 to obtain the same temporal scale for visual and audio modalities.

2) Visual Features: The input video frame rate is 25 frames/s. We first use OpenCV-toolkit, which contains dlib face [28] and landmark detector to obtain the 68 facial key points and choose the mouth as the region of interest (ROI). Then, lip-centered patches of 112 × 112 pixels are cropped from the raw video frames. To extract the visual features from the obtained lip-centered image sequence, a (C3D) [29]-P3D [30] network is employed as the front end. Among them, P3D is generated by interlacing its three versions in turn as shown in Fig. 3. The network can capture more powerful visual spatio-temporal feature than C3D + 2-D ResNet which is widely used in many lip-reading works [15], [31], [32].

C3D–P3D is based on [29], which alleviates computational cost by replacing part of the C3D layer with P3D. By decomposing the convolution factor, the P3D ResNet can split the $3 \times 3 \times 3$ convolution into $1 \times 3 \times 3$ spatial convolution and $3 \times 1 \times 1$ temporal convolution. The network applies the 3-D convolution layer with 64 kernels of size $5 \times 7 \times 7$ and executes a 50-layer P3D ResNet that is a mixture of three blocks to gradually reduce the depth of spatial dimensions and preserving temporal dimensions. For the input image frames of $T \times H \times W$, the output of the visual front end is average-pooled in spatial dimension and uses the 512-D vector to represent per video frame.

3) Motion Features: The optical flow can clearly capture the movement of pixels in adjacent images, expressed by the direction and size of each pixel's magnitude. Some examples are shown in Fig. 4. We first employ the PWC-Net [33] pretrained on MPI Sintel dataset [34] to calculate the dense optical flow corresponding to the input lip-centered images. Then, we also use the other P3D front end (using the same frame as the visual stream) to extract the motion features from the optical-flow stream. Finally, the obtained visual features and motion features are fed to the cross-modal attention fusion module (CMAF) to enhance visual representation. However, the introduction of optical flow increases the overall parameter budget that will cost more storage and computational load. In our work, we focus on the improvement of AVSR performance brought by this method.


(a)

(b)

Fig. 4. Examples of the inputs of the visual stream and the optical flow stream. (a) Lip-centered image sequence in the visual stream. (b) Corresponding optical flow sequence obtained by PWC-Net.

B. Cross-Modal Attention Fusion

The optical-flow information between adjacent frames was not considered in previous works, which results in suboptimal performance. In this article, the CMAF is proposed to obtain the enhanced visual features. As shown in Fig. 5, given a query of visual modality, the cross-modal attention measures its correlation with all elements in the optical-flow modality, which makes the motion-modality information potentially flow into visual modality to update the visual features. Then, the linear transformation is used to obtain discriminative visual features. We consider the visual-modality features $V_\alpha \in \mathbb{R}^{l_\alpha \times d_\alpha}$ and motion-modality features $O_\beta \in \mathbb{R}^{l_\beta \times d_\beta}$, where l_α and l_β are the sequence lengths of different modalities and d_α and d_β represent the feature dimensions of different modalities. Some early works in [24] and [35] have shown that the motion information can improve the performance of VSR. The effective way to fuse cross-modal information through a potential cross-modal adaptation has been verified in [36] and [37].

The cross-modal attention fusion module first produces a set of query, key, and value by linear transformations. We define the query from visual modality as $Q_\alpha = V_\alpha W_{Q_\alpha}$, the pairs of key–value from the motion modality denoted $K_\beta = O_\beta W_{K_\beta}$, and $V_\beta = O_\beta W_{V_\beta}$, respectively [16], [36]. The attention weights from O_β to V_α are defined as follows:

$$\begin{aligned} \mathrm{CMA}_{O_\beta \to V_\alpha} &= \mathrm{SoftMax}\left(\frac{Q_\alpha K_\beta^\top}{\sqrt{d_k}}\right) \\ &= \mathrm{SoftMax}\left(\frac{V_\alpha W_{Q_\alpha} W_{K_\beta}^\top O_\beta^\top}{\sqrt{d_k}}\right) \end{aligned} \tag{1}$$

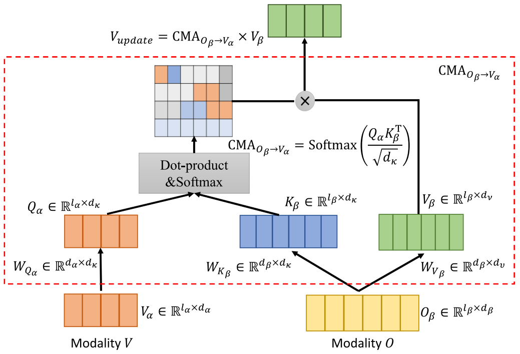

$V_{update} = \mathrm{CMA}_{O_\beta \to V_\alpha} \times V_\beta$
$\mathrm{CMA}_{O_\beta \to V_\alpha}$
$\mathrm{CMA}_{O_\beta \to V_\alpha} = \mathrm{Softmax}\left(\frac{Q_\alpha K_\beta^{\mathrm{T}}}{\sqrt{d_\kappa}}\right)$
Dot-product &Softmax
$Q_\alpha \in \mathbb{R}^{l_\alpha \times d_\kappa}$
$K_\beta \in \mathbb{R}^{l_\beta \times d_\kappa}$
$V_\beta \in \mathbb{R}^{l_\beta \times d_v}$
$W_{Q_\alpha} \in \mathbb{R}^{d_\alpha \times d_\kappa}$
$W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_\kappa}$
$W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$
$V_\alpha \in \mathbb{R}^{l_\alpha \times d_\alpha}$
$O_\beta \in \mathbb{R}^{l_\beta \times d_\beta}$
Modality V
Modality O

Fig. 5. Crossmodal attention $\mathrm{CMA}_{O_\beta \to V_\alpha}$ between the feature sequences of V_α and O_β, where $\mathrm{CMA}_{O_\beta \to V_\alpha}$ is weighting for aggregating information from motion to visual; V_α and O_β represent the visual-modality features and motion-modality features, respectively; V_{update} is the updated visual features by using CMA.

where $\mathrm{CMA}_{O_\beta \to V_\alpha} \in \mathbb{R}^{l_\alpha \times l_\beta}$ is weighting for aggregating information from motion to visual. The weights $W_{Q_\alpha} \in \mathbb{R}^{d_\alpha \times d_k}$ and $W_{K_\beta} \in \mathbb{R}^{d_\beta \times d_k}$ are parameters to be learned. To be specific, the dot-product attention executes the scale $1/\sqrt{d_k}$ and the SoftMax function to normalize the inner-product values. The CMA module can capture the relationship between each visual feature and all motion-modality information by weighting the sum of the motion value feature V_β. We illustrate the motion information flow to update visual features by using CMA as V_{update}

$$\begin{aligned} V_{\mathrm{update}} &= \mathrm{CMA}_{O_\beta \to V_\alpha} \times V_\beta \\ &= \mathrm{Softmax}\left(\frac{V_\alpha W_{Q_\alpha} W_{K_\beta}^\top O_\beta^\top}{\sqrt{d_k}}\right) O_\beta W_{V_\beta} \end{aligned} \tag{2}$$

where the weight parameters $W_{V_\beta} \in \mathbb{R}^{d_\beta \times d_v}$, and the length of $V_{\mathrm{update}} \in \mathbb{R}^{l_\alpha \times d_v}$ is equal to Q_α. Then, we merge the raw visual features V_α and the represented visual features V_{update}. A linear transformation is employed to project the concatenated features to output enhanced features V_{enh} with

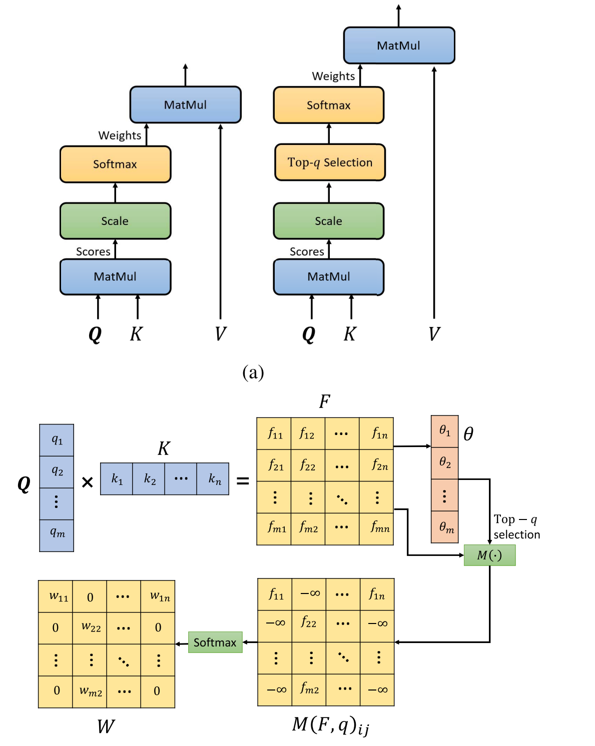
$$V_{\mathrm{enh}} = \mathrm{Linear}\left(\left[V, V_{\mathrm{update}}\right]\right) \tag{3}$$

where $V_{\mathrm{enh}} \in \mathbb{R}^{l_\alpha \times d_{\mathrm{model}}}$, and d_{model} represents the common dimension of each modality features. Then, the output features through the CMA model are fed into the sparse transformer network (STN). It can establish the intramodality information relations and capture the context-dependent relations.

C. Sparse Transformer Network

The transformer architecture is an end-to-end sequence model, which utilizes the global attention to model the long-term dependences. In this article, the STN is introduced to overcome the shortcoming that the transformer [15] may extract the irrelevant information for AVSR. As shown in Fig. 6(a), we conduct a sparse scaled dot-product attention based on top-q selection to reserve the most important segments.

1) Sparse Scaled Dot-Product Attention: The scaled dot-product attention is a universal attention model, which has mainly three components, the queries, keys, and values, to describe the source context. The execution of sparse attention is provided in Fig. 6(b). In the implementation, the sparse scaled dot-product also receives query $Q \in \mathbb{R}^{m \times d_q}$, key $K \in \mathbb{R}^{n \times d_k}$, and value $V \in \mathbb{R}^{n \times d_v}$ as inputs, where n is the length of key–value pairs, m is the length of query, and d is the dimension of corresponding features. Normally, $d_q = d_k$. To be specific, the query Q, key K, and value V are linear mapping of input features x, so that $Q = W_Q x$, $K = W_K x$, and $V = W_V x$. The attention mechanism computes the dot-products of Q and K, and divides each by $\sqrt{d_k}$ to generate the attention scores F as follows:



(a)

(b)

Fig. 6. Difference between the attentions of the transformer and sparse transformer, and the description of the sparse attention module. (a) From left to right: original scaled dot-product attention and sparse scaled dot-product attention. (b) Core step of sparse attention is to select the top-q most contributing elements.

$$F = \frac{QK^T}{\sqrt{d_k}}. \tag{4}$$

The scores F reflect the degree of relation between the features. We assess it based on assumption that the large values of scores have higher relevance. Then, the sparse attention is executed, which uses a masking function $M(\cdot)$ on the scores F to widely select the top-q contributive values, where q is a hyperparameter. Especially, the model defines a position matrix $\mathrm{p}(i, j)$ to save the top-q largest values of each row in F. The masking function is described as follows:

$$M(F, q)_{ij} = \begin{cases} F_{ij}, & \text{if } F_{ij} \geq \theta_i \\ -\infty, & \text{if otherwise} \end{cases} \tag{5}$$

where θ_i corresponds to qth largest element of row i, which is considered to be the threshold of each row. When the value of θ_i is smaller than the jth component, record the position (i, j). Afterward, all the thresholds are concatenated to make a vector $\theta = [\theta_1, \theta_2, \ldots, \theta_m]$. Unlike the dropout which randomly deletes the elements, we obtain the high attention values through the top-q selection. Finally, we use a SoftMax function to normalize the scores

$$W = \mathrm{SoftMax}(M(F, q)) \tag{6}$$

where W is a normalized value. With the masking function $M(\cdot)$, the scores that are smaller than threshold value will be assigned with negative infinity (approximate 0). The attended features A can be calculated as follows:

$$A = WV. \tag{7}$$

2) Multihead Sparse Attention: In order to improve the representation ability of a single attention model, multihead sparse attention (MSA) performs multiple sparse scaled-dot-product operations in parallel. The MSA projects the diverse information into h subspaces and then uses a linear transformation to combine the results. The attended features of MSA are provided as follows:

$$\mathrm{MSA}(Q, K, V) = [\mathrm{h}_1; \ldots; \ \mathrm{h}_h]W^O \tag{8}$$

$$\mathrm{h}_i = A\left(QW_i^Q, K \ W_i^K, V \ W_i^V\right) \tag{9}$$

where the matrices $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d_{\mathrm{model}} \times d_h}$ are learnable weight parameters, $W^O \in \mathbb{R}^{hd_h \times d_{\mathrm{model}}}$ merges the information from the h heads, i represents the ith head, and $d_h = d_{\mathrm{model}}/h$ is the dimension of the output feature of each head.

The STN uses the MSA to establish the context dependence between input features, and then further increases the nonlinearity of the model through the feed-forward layers to get attended features. In addition, when the MSA is on the encoding or decoding stage, all of queries, keys, and values are in the same place and have equal dimensions. In encoder–decoder attention layers, the MSA produces one or multiple query vectors in the decoder stage, to control attention values from sequence of the encoder state.

D. Overall Architecture of MMST

The overall architecture of the proposed MMST is described in Fig. 2. We get three main modalities: audio modality (A), visual modality (V), and motion modality (O) from the input videos. Among them, the motion modality is generated by computing the optical flow of the visual modality. The three input modalities that correspond to the same utterance are first encoded as the feature of different dimensions. We standardize all input features into the same dimensions to facilitate subsequent processing and denote the input feature sequences with $X_{\{V,O,A\}} \in \mathbb{R}^{l_{\{V,O,A\}} \times d_{\mathrm{model}}}$. To ensure that the input feature sequences of each modality have temporal information to be aware of neighboring elements, the positional encodings (PEs) are added to $X_{\{V,O,A\}}$, as follows:

$$\{V_\alpha, O_\beta, A_\gamma\} = X_{\{V,O,A\}} + \mathbf{PE} \tag{10}$$

$$\mathbf{PE}_{(\mathrm{pos},i)} = \begin{cases} \sin\left(\mathrm{pos}/10\,000^{\frac{2i}{d}}\right), & 0 \le i < d/2 \\ \cos\left(\mathrm{pos}/10\,000^{\frac{2i}{d}}\right), & d/2 \le i < d \end{cases} \tag{11}$$

where $\{V_\alpha, O_\beta, A_\gamma\} \in \mathbb{R}^{l_{\{\alpha,\beta,\gamma\}} \times d_{\mathrm{model}}}$ contain elements' position information at every timestep, pos represents the position in sequence, $\mathrm{pos} = 0, \ldots, l-1$, and $d = d_{\mathrm{model}}$.

TABLE I

STATISTICS INFORMATION OF THREE LARGE-SCALE LIP-READING DATASETS. THESE DATASETS ARE PUBLICLY AVAILABLE. UTTER: UTTERANCES. VOCAL: VOCABULARY SIZE OF THE DATASET. SENT.: SENTENCE-LEVEL DATASET

Dataset	Type	Split	Vocal	# Utter	# Hours
LRW [20]	Words	Train-val	500	514k	165
		Test		25k	8
LRS2-BBC [15]	Sent.	Pre-train	41k	96k	195
		Train-val	18k	47k	29
		Test	1,693	1,243	0.5
LRS3-TED [38]	Sent.	Pre-train	52k	132k	444
		Train-val	17k	32k	30
		Test	2,136	1,452	1

Then, the motion and visual features are fed into the CMA module, which can generate attention flow to pass information. Each visual feature will select all motion features by attention weight to capture the discriminative visual features V_{enh}. Finally, the STN is implemented by combining enhanced visual and audio representations. In the STN encoder stage, sparse attention can select most concerned attention weights for allowing information within each input modality to build the context-dependent relations. The query-aware attention is added to the context vector of input modalities in the STN decoder stage. Then, the attended multifeatures are concatenated to the output character probabilities.

IV. EXPERIMENTAL SETUP

To evaluate the effectiveness of the proposed method, extensive experiments are conducted on three public benchmark datasets.[1] The details of the dataset are shown in Table I. In addition, the proposed method is compared with several state-of-the-art methods through widely used quantitative metrics.

A. Datasets and Evaluation Metric

1) LRW Dataset: The Lip Reading in the Wild (LRW) [20] is a large-scale word-level dataset, which contains 514 000 utterances with a vocabulary size of 500 and spoken by over 1000 speakers from BBC broadcasts. Each sample in the dataset consists of a short video clip with the duration of 1.16 s. It is a challenging dataset due to the variability of appearance and pose of speakers. This dataset is employed to train the P3D-based visual front end.

2) LRS2-BBC Dataset: The Lip Reading Sentences2 (LRS2) is a large-scale publicly available database for research from BBC program, mainly news and TV shows. The LRS2 [15] dataset contains more than 14 000 utterances and two million words with a vocabulary size of 41 000. This is a very challenging dataset due to the variations of the head posture, lighting conditions, the large number of speakers, and low image resolution.

3) LRS3-TED Dataset: LRS3-TED [38] is one of the largest available audio-visual datasets for lip-reading (VSR) task. It contains 5594 YouTube videos and more than 400 video clips extracted from TED English talks and TEDx talks. The dataset is divided into three sets: pretrain, train-val, and test.

[1] https://www.robots.ox.ac.uk/vgg/data/

TABLE II

TRAINING STRATEGIES FOR DIFFERENT DATASETS

Modules Datasets	Visual Front-end	CMAF	Sparse-transformer
LRW	pre-trained	-	-
LRS2-BBC	frozen	from scratch	from scratch
LRS3-TED	frozen	fine-tuned	fine-tuned

TABLE III

QUANTITATIVE METRICS OF DIFFERENT LIP-READING APPROACHES ON THE LRW DATASET. THE INPUT OF THE TWO-CHANNEL NETWORK IS COMPOSED OF THE MOTION AND VISUAL MODALITIES

Method	WER (%)
WAS [25]	23.8
Res+LSTM [31]	17.0
Multi Graned [40]	16.7
C3D-P3D [41]	15.2
Two-stream (Concat fusion)	**13.1**

It contains more than 4.2 million words and the vocabulary size is 51 000.

4) Evaluation Metric: We evaluate the effectiveness of our method with the word error rate (WER). It is a common evaluation criteria for AVSR, which compares the reference to the hypothesis with: $\text{WER} = (S + D + I)/\text{Num}$, where Num is the number of words in the reference, and S, D, and I are the numbers of substitution, deletion, and insertion operations to reedit the hypothesis sentence as exactly the same with the reference one.

B. Implementation Details

Our network is implemented with PyTorch library and trained on two GeForce GTX 1080 Ti GPU with 11-GB memory. The model uses the Adam [39] optimizer and has an initial learning rate of 0.0001, which is decayed by 0.5 when the validation error does not show improvement in three epochs. To prevent over-fitting, we use the label smoothing and set the dropout to 0.1. The model is trained with the teacher-forced strategy, where the label of the previous decoding step is used as the input of the next step. $d_{\text{model}} = 512$ and the number of heads is 8. We set output size of the network to 40, including the 26 characters in the alphabet, the ten digits, and four special tokens [pad], [space], [bos], and [eos]. More details about our method and some demos can be found in https://github.com/UNBSQY/MMST.

C. Training Strategy

Table II provides the training strategies of the proposed method on different datasets. The visual feature extractors are first pretrained on the LRW dataset. After that, the visual and motion features are generated using the frozen feature extractors' front end for the LRS2-BBC [4] and LRS3-TED [8] datasets. Specifically, the CMAF and sparse transformer modules are trained from scratch for the LRS2-BBC dataset, which are further fine-tuned for the LRS3-TED dataset.

1) Front-End Feature Extraction: The feature extractors with C3D-P3D, which are pretrained on the LRW dataset with word-level recognition of 500 classes, are used to obtain initial feature representation ability. The network consists of a C3D-P3D front end to extract spatial-temporal feature and a two-layer Bi-LSTM for modeling the temporal dependence similar to [31] and [41]. First, the sequences of lip-centered and optical flow are, respectively, fed into feature extractors to form a two-stream network, which is a 3-D convolutional network followed by a 50-layer P3D to capture the short-term dynamics of visual and motion. Second, the motion and visual features are concatenated and fed into the two-layer of Bi-LSTM, which is followed by the linear and SoftMax layer. The output character probabilities directly match the ground truth labels and are trained through the cross-entropy loss function.

2) Data Simulation: In real life, the performance of speech recognition deteriorates due to the background noise and multispeakers. To reduce the influence of noise and improve the performance of speech recognition, we follow the mixed-noise audio method of [15] to train our network. The Babble noise with different signal-to-noise ratios (SNR) from −5 to 10 dB is additive to the audio stream with probability $P_n = 0.25$, and the Babble noise samples are synthesized by mixing 20 different audio samples in LRS2 and 30 different audio samples in LRS3. We also extract and save the features of raw audio and the mix-noise speech by using STFT.

3) MMST: We extract and save the visual and motion features by employing the feature front end frozen. The CMAF is employed to compute the relationships between each visual feature and all motion features and generate enhanced visual features. The model employs the attention mechanism to integrate the motion information. After getting the enhanced visual features, we train MMST with the enhanced visual features and audio features as input. A sequence-to-sequence loss [43] is used in this model. However, the seq2seq processing networks have been mentioned that the converge is very difficult due to the long timesteps. Thereby, a curriculum learning [44] strategy is used. The training starts with the input samples of single word and then increases the length of samples gradually. The strategy accelerates the training procedure and alleviates the over-fitting problem. The pretraining sample sets of the LRS2-BBC and LRS3-TED datasets are employed to train the network first. Then, we fine-tune the model on the train-val parts from LRS2-BBC and LRS3-TED datasets separately. In order to handle the different lengths of the input sequence, we make them to a maximum sequence length by zero-padding.

D. Compared Methods

To verify the effectiveness of our method, several state-of-the-art methods are compared with experiments on two large datasets with different noise levels. The baselines for comparison are listed as follows.

1) Lips Only: We have some baselines for VSR. The results of all baselines are reported in the original literature.

a) WAS: The method in [25] is an attention-based encoder–decoder category. In the encoder stage, this method uses LSTM to model the input features at every timestep and to generate a context-dependent state vector. In the decoder stage, the attention mechanism is added to wisely select the encoder state vector. Then, this model fuses the hidden state of LSTM to the output characters.

b) Multigrained: The multigrained method is provided by Wang's [40] spatio-temporal convolution module to obtain the fine-grained and medium-grained visual features. The extracted different-level features are fed into bidirectional ConvLSTM to build the context-information visual features of the entire input sequence. Finally, the output results are obtained through the fully connected layer.

c) Res-lstm: The system in [31] consists of C3D convolutions, 34-layer ResNet, and two-layer Bi-LSTM. This model applies the C3D followed by the ResNet front end to the image sequence and Bi-LSTM to model the context dependence of visual features. The SoftMax function is utilized for every timestep.

2) Audio-Visual: We compare several state-of-the-art methods for AVSR. The comparison methods are reimplemented with the same parameters from original literature by using the Pytorch framework.

a) TM-seq2seq: The core of this method in [15] is self-attention blocks. This method uses the pretrained vision model to extract and save visual features and the self-attention mechanism directly trained on features to generate the context-representation vectors of each modality. The whole network is trained end-to-end with sequence-to-sequence loss. The model is trained by using the ADAM optimizer and the width of beam search is set to 35. We reimplement the method in [15] under identical conditions.

b) AV transformer: Sterpu *et al.* [42] proposed the fusion strategy with the cross-modal attention to weight between the audio and visual representations for the AVSR, which avoids pretraining strategies and instead rely on the audio-visual data. However, it suffers from the convergence problems with the visual front end. Therefore, the regress action units (AUs) from the visual representations are used in [42] as an auxiliary loss in the transformer framework to overcome this problem. The parameters and models are available online.[2]

c) EG-seq2seq: The model in [41] is a two-stage model, which needs to obtain clean audio modality by separating the background noises with the help of visual information first. Then, the enhanced audio modality and visual modality are fed into elementwise-attention gated recurrent unit (EleAtt-GRU) to capture context dependence. The P3D-based front end is employed to extract features and pretrained on the LRW dataset. We use the method with single visual modality awareness as a baseline. The parameters and models are available online.[3]

V. Quantitative Analysis and Ablation Study

Extensive experiments are conducted to demonstrate the effectiveness of our proposed method on three datasets. Among them, we use the LRW dataset to compare with the baselines reported in the literature, focusing on the more current LRS2 and LRS3 datasets to evaluate our method. First, we compare some word-level lip-reading methods on the LRW dataset to pretrain the visual front end and evaluate the effectiveness of motion modality in the lip-reading task closely correlated with the AVSR. Second, we make a systematic comparison on the LRS2-dataset. A large number of ablation studies are performed to prove the availability of each module in MMST. Third, the extra experiment is conducted on the LRS3 dataset to further validate the performance of the proposed method. Fourth, the model parameter on the performance of the proposed method is explored. Finally, we investigate the robustness of the proposed method on different noise types. In the experiments, our methods do not use a separate language model; however, linguistic structure is learned implicitly from the video input.

A. Results on LRW Dataset

The pretraining of the visual-front-end C3D-P3D module on the LRW dataset is described in Section IV-C, which is a two-stream network consisting of visual and motion modalities. To demonstrate the effectiveness of the motion-modality input, we conduct the experimental comparison of the proposed two-stream method by implementing a word-level lip-reading task related to AVSR on the LRW dataset. We used two-layer Bi-LSTM as backend for modeling the temporal dependence similar to [41] that is a single stream using visual-modality input. Compared with the fifth row and the sixth row in Table III, the WER of the proposed method decreases by 2.1%. It demonstrates that the motion-modality input improves the recognition performance of single stream with visual modality. In addition, the results from the sixth row in Table III show the two-stream network better than different baselines.

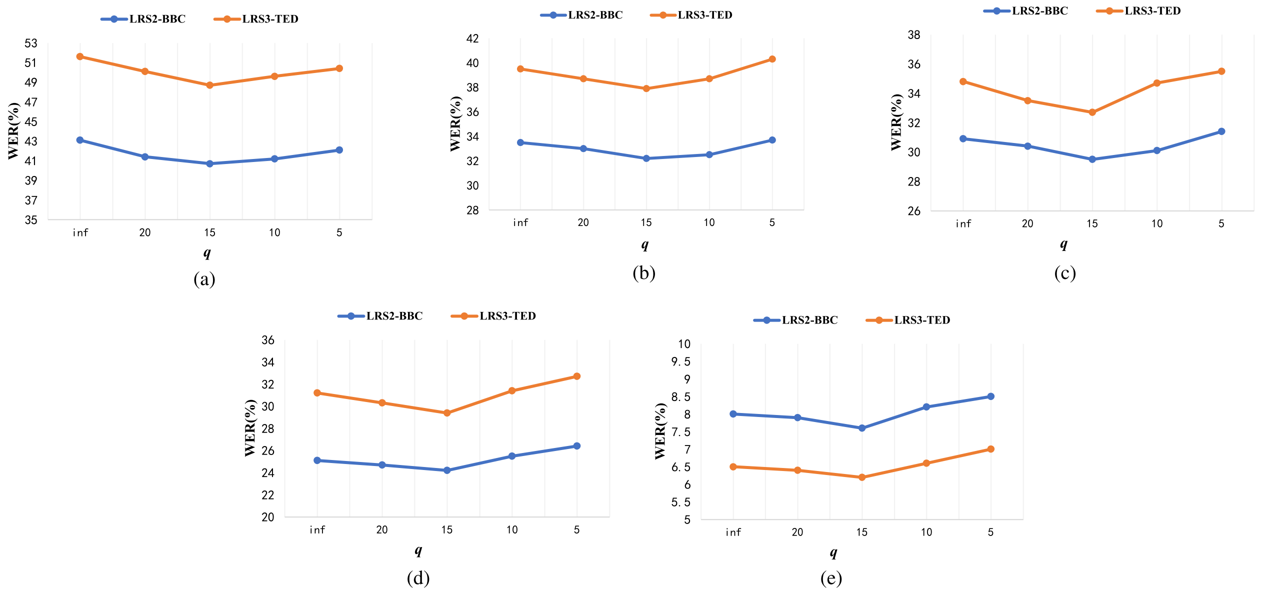
B. Results on LRS2 Dataset

In our network, the motion modality, cross-modal attention fusion, and sparse transformer are employed to obtain better performance for AVSR. To study the effectiveness of each module in MMST, we conduct the ablation experiments and compare the proposed method with several SOTA methods on the LRS2 dataset. The numerical results are shown that our network is more advantageous in speech recognition at different noise levels.

1) Effect of Sparse Transformer on LRS2 Dataset: To verify the effectiveness of sparse transformer module, we first experiment with only visual modality and the WER of 48.6% as described in Table IV from Rows V. Compared with the previous methods of TM-seq2seq, the WER is reduced by 1.6%. The results of Rows A and Rows AV show that our method surpasses the baselines at different noise levels. It indicates that the sparsity used in our method can achieve a better performance. Compared with Rows A in the different methods, the results show that all methods have a poor recognition performance with only audio modality under the low SNR conditions. However, the visual modality is not affected by the environment noises and the WER is constant on different levels. Comparing Rows A and Rows AV in Table IV demonstrates that the visual modality brings a significant improvement for the speech recognition accuracy in the low SNR conditions. Thereby, the motion modality is introduced to enhance the visual features by the CMAF.

[2] https://github.com/georgesterpu

[3] https://github.com/JackSyu/Discriminative-Multi-modality-Speech-Recognition



(a)

(b)

(c)

(d)

(e)

Fig. 7. Evaluation of the proposed method with the sparsity q of the attention mechanism on LRS2 and LRS3 datasets with the Babble noise. **inf** indicates that all features in the sequence are selected. **MMST**(·) represents that our method (TM + sparse + CMAF) is performed in different noise levels. (a) MMST(−5 dB). (b) MMST(0 dB). (c) MMST(5 dB). (d) MMST(10 dB). (e) MMST(clean).

TABLE IV

Quantitative Metrics of Different Methods on the LRS2-BBC Dataset in Terms of WER(%) With Different Modalities and Babble-Noise Levels. The Second Column M Represents the Input Modalities. A, V, and O Denote the Audio Modality, the Visual Modality, and the Motion Modality, Respectively

Methods	M	SNR(dB)				
		clean	10	5	0	-5
Google S2T	A	20.9	-	-	86.3	-
AV-transformer [42]	AV	12.5	28.7	33.9	37.5	48.9
TM-seq2seq [15]	A	10.8	28.6	42.6	58.5	78.4
TM-seq2seq [15]	V	50.2	-	-	-	-
TM-seq2seq [15]	AV	9.6	26.8	32.7	36.2	46.7
TM +sparse	A	9.1	28.1	40.7	56.3	76.8
TM +sparse	V	48.6	-	-	-	-
TM +sparse	AV	8.3	25.6	31.2	34.3	44.3
TM +sparse+concat fusion	VO	47.8	-	-	-	-
TM +sparse+concat fusion	AVO	8.0	25.1	30.6	33.5	43.1
TM +sparse+CMAF (MMST)	VO	**46.1**	-	-	-	-
TM +sparse+CMAF (MMST)	AVO	**7.6**	**24.2**	**29.5**	**32.2**	**40.7**

TABLE V

Quantitative Metrics of Different Methods on the LRS3-TED Dataset in Terms of WER(%) With Different Modalities and Babble-Noise Levels. The Second Column M Represents the Input Modalities. A, V, and O Denote the Audio Modality, the Visual Modality, and the Motion Modality, Respectively

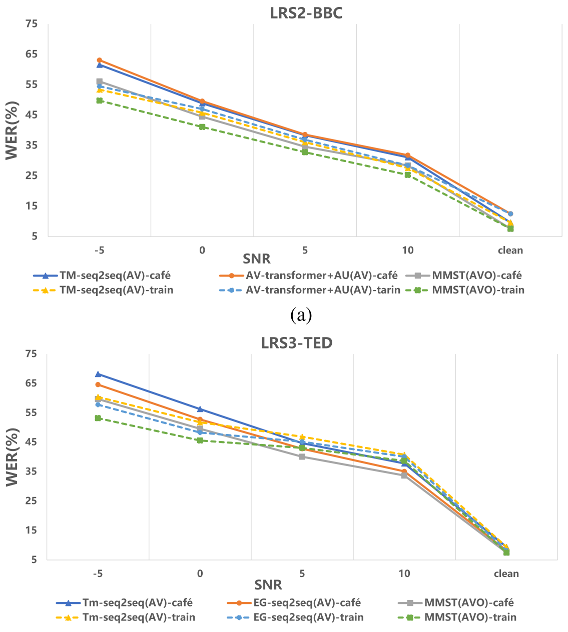
Methods	M	SNR(dB)				
		clean	10	5	0	-5
TM-seq2seq [15]	V	60.5	-	-	-	-
TM-seq2seq [15]	AV	8.2	33.4	38.3	44.5	53.9
EG-seq2seq [41]	V	59.1	-	-	-	-
EG-seq2seq [41]	AV	7.1	32.6	37.1	42.4	53.1
TM +sparse	V	57.4	-	-	-	-
TM +sparse	AV	6.7	31.7	36.1	41.9	51.1
TM +sparse+concat fusion	VO	56.5	-	-	-	-
TM +sparse+concat fusion	AVO	6.5	31.2	34.9	41.1	49.6
TM +sparse+CMAF (MMST)	VO	**55.1**	-	-	-	-
TM +sparse+CMAF (MMST)	AVO	**5.5**	**29.2**	**33.2**	**39.4**	**47.6**

2) Effect of Motion Modality on LRS2 Dataset: In our work, the motion modality is introduced to enhance visual features to further improve the performance of the AVSR. We utilize two fusion methods to combine visual modality and motion modality to better utilize visual information. The results from Rows VO and Rows AVO in Table IV show that the TM + sparse + concat fusion and TM + sparse + CMAF methods achieve a better performance than the Tm + spare and the TM-seq2seq methods. It demonstrates the availability of introducing the motion modality to our method.

3) Effect of CMAF Module on LRS2 Dataset: To prove the role of the CMAF module, we perform a comparative experiment with a common fusion strategy by concatenating the visual and motion features (TM + sparse + concat fusion). Compared with Rows VO in Table IV, the WER of the TM + sparse + CMAF (MMST) method is reduced by 1.4%. In addition, the results of Rows **AVO** validate that the MMST method is also superior to the TM + sparse + concat fusion method at different noise levels, which demonstrates the availability of the CMAF module. From Row AVO in Table IV, the MMST method obtains the lower WER in the different noise levels compared with other methods.

C. More Results on LRS3 Dataset

To further validate the effectiveness of our method, we perform experiments and compare with existing superior works



(a)

(b)

Fig. 8. Performance comparison with some state-of-the-art methods under different noise types on the LRS2 and LRS3 datasets (lower is better). Each noise type has a different signal noise rate (SNR) from -5 to 10 dB. The combination of A, V, and O indicates the input modality of this method. **MMST**$(\cdot)$ represents that our method (TM + sparse + CMAF) is performed in different noise levels. (a) WER(%) on LRS2-BBC. (b) WER(%) on LRS3-TED.

on the LRS3-TED dataset, and the results are also clearly shown in Table V. Compared with the methods of the EG-seq2seq [41] and the TM-seq2seq, our method achieves a better performance for speech recognition at different noise levels. It indicates the high performance of integrating the visual-enhanced component and sparse transformer component.

D. Analysis on Model Parameters

The parameter q is the sparsity of the attention mechanism. To further investigate the effect of q on the performance of the proposed method, we evaluate the results of the proposed method with different values of q. We set different initializations $q \in \{5, 10, 15, 20\}$ as a parameter set as shown in Fig. 7. $q = \inf$ represents the traditional global attention. The results show that the global attention has a cutoff point. The WER decreases as q varies from inf to the cutoff point. It demonstrates that the sparse attention can improve the concentration of the global attention by discarding some irrelevant information. When q is lower than the cutoff point, the model cannot capture sufficient information of the sequences, resulting in a poor performance. It demonstrates that the suitable sparsity in attention can establish the correlation better between the sequences. The model achieves its best performance at $q = 15$ for all noise levels and different datasets. This proves that the sparse attention by the top-q selection is effective to fuse visual and audio features in the AVSR.

E. Robustness on Different Noise Types

To evaluate the robustness of our method on different noise types, we conduct comparative experiments with another two types of noise, i.e., Cafeteria (caff) [42] and Train [45], on the LRS2-BBC and LRS3-TED datasets. The noise is added to the audio stream with different SNRs from -5 to 10 dB. As shown in Fig. 8, compared with the different methods, our method achieves a better performance for different types of noise on the LRS2-BBC and LRS3-TED datasets.

VI. Conclusion

In this article, we propose the multimodal sparse transformer network for the AVSR, which facilitates the visual information to improve the accuracy of speech recognition. Apart from the visual features used in the traditional AVSR methods, the proposed method also takes advantage of the motion feature represented by the optical flow extracted from the video. We also design the cross-modal attention-based fusion module to enhance the visual features with the motion modality. Moreover, the STN is employed to model the correlation of different features as well as avoid introducing too much redundant information. We have conducted extensive experiments to study the effectiveness of the proposed modules and demonstrate the superiority of the proposed method over the state-of-the-art methods under different noisy conditions.

Acknowledgment

The authors would like to thank the Visual Geometry Group from the University of Oxford for providing the three datasets for this work. They also would like to thank the editors and reviewers for their insightful comments and suggestions.

References

[1] C.-C. Chiu *et al.*, "State-of-the-art speech recognition with sequence-to-sequence models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Apr. 2018, pp. 4774–4778.

[2] J. Davila-Chacon, J. Liu, and S. Wermter, "Enhanced robot speech recognition using biomimetic binaural sound source localization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 138–150, Jan. 2019.

[3] Q. Yu, Y. Yao, L. Wang, H. Tang, J. Dang, and K. C. Tan, "Robust environmental sound recognition with sparse key-point encoding and efficient multispike learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 625–638, Feb. 2021.

[4] G. Potamianos, C. Neti, J. Luettin, and I. Matthews, "Audio-visual automatic speech recognition: An overview," *Issues Vis. Audio-Vis. Speech Process.*, vol. 22, pp. 23–62, Mar. 2004.

[5] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. W. Senior, "Recent advances in the automatic recognition of audiovisual speech," *Proc. IEEE*, vol. 91, no. 9, pp. 1306–1326, Sep. 2003.

[6] C. Cangea, P. Velickovic, and P. Lio, "XFlow: Cross-modal deep neural networks for audiovisual classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 9, pp. 3711–3720, Sep. 2020.

[7] Q. Summerfield, "Lipreading and audio-visual speech perception," *Philos. Trans. Roy. Soc. London B, Biol. Sci.*, vol. 335, no. 1273, pp. 71–78, 1992.

[8] J. S. Chung and A. Zisserman, "Out of time: Automated lip sync in the wild," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Taipei, Taiwan, 2016, pp. 251–263.

[9] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, no. 5588, pp. 746–748, 1976.

[10] G. Sterpu, C. Saam, and N. Harte, "Attention-based audio-visual fusion for robust automatic speech recognition," in *Proc. 20th ACM Int. Conf. Multimodal Interact.*, Oct. 2018, pp. 111–115.

[11] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proc. IEEE*, vol. 103, no. 9, pp. 1635–1653, Sep. 2015.

[12] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 2, pp. 206–219, Apr. 2019.

[13] A. Mohamed, D. Okhonko, and L. Zettlemoyer, "Transformers with convolutional context for ASR," 2019, *arXiv:1904.11660*.

[14] S. Petridis, Y. Wang, Z. Li, and M. Pantic, "End-to-end audiovisual fusion with LSTMs," 2017, *arXiv:1709.04343*.

[15] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Dec. 21, 2018, doi: 10.1109/TPAMI.2018.2889052.

[16] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[17] T. Afouras, J. Son Chung, and A. Zisserman, "The conversation: Deep audio-visual speech enhancement," 2018, *arXiv:1804.04121*.

[18] M. Riva, M. Wand, and J. Schmidhuber, "Motion dynamics improve speaker-independent lipreading," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 4407–4411.

[19] B. Martinez, P. Ma, S. Petridis, and M. Pantic, "Lipreading using temporal convolutional networks," in *Proc. ICASSP - IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 6319–6323.

[20] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Taipei, Taiwan, 2016, pp. 87–103.

[21] Y. M. Assael, B. Shillingford, S. Whiteson, and N. de Freitas, "LipNet: End-to-end sentence-level lipreading," 2016, *arXiv:1611.01599*.

[22] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 568–576.

[23] J. Yu and R. Ramamoorthi, "Learning video stabilization using optical flow," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8159–8167.

[24] X. Weng and K. Kitani, "Learning spatio-temporal features with two-stream deep 3D CNNs for lipreading," 2019, *arXiv:1905.02540*.

[25] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Lip reading sentences in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3444–3453.

[26] G. Zhao, J. Lin, Z. Zhang, X. Ren, Q. Su, and X. Sun, "Explicit sparse transformer: Concentrated attention through explicit selection," 2019, *arXiv:1912.11637*.

[27] P. Zhou, W. Yang, W. Chen, Y. Wang, and J. Jia, "Modality attention for end-to-end audio-visual speech recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 6565–6569.

[28] D. Castelli and P. Pagano, "OpenDLib: A digital library service system," in *Proc. Int. Conf. Theory Pract. Digit. Libraries*, Rome, Italy, 2002, pp. 292–308.

[29] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[30] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5533–5541.

[31] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," 2017, *arXiv:1703.04105*.

[32] X. Zhao, S. Yang, S. Shan, and X. Chen, "Mutual information maximization for effective lip reading," 2020, *arXiv:2003.06439*.

[33] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 8934–8943.

[34] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black, "A naturalistic open source movie for optical flow evaluation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Springer, 2012, pp. 611–625.

[35] J. Xiao, S. Yang, Y. Zhang, S. Shan, and X. Chen, "Deformation flow based two-stream network for lip reading," 2020, *arXiv:2003.05709*.

[36] P. Gao *et al.*, "Dynamic fusion with intra-and inter- modality attention flow for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6639–6648.

[37] Y.-H.-H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, p. 6558.

[38] T. Afouras, J. Son Chung, and A. Zisserman, "LRS3-TED: A large-scale dataset for visual speech recognition," 2018, *arXiv:1809.00496*.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[40] C. Wang, "Multi-grained spatio-temporal modeling for lip-reading," 2019, *arXiv:1908.11618*.

[41] B. Xu, C. Lu, Y. Guo, and J. Wang, "Discriminative multi-modality speech recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14433–14442.

[42] G. Sterpu, C. Saam, and N. Harte, "Should we hard-code the recurrence concept or learn it instead? Exploring the transformer architecture for audio-visual speech recognition," 2020, *arXiv:2005.09297*.

[43] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 27, 2014, pp. 3104–3112.

[44] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proc. 26th Annu. Int. Conf. Mach. Learn. (ICML)*, 2009, pp. 41–48.

[45] Y. Hu, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Commun.*, vol. 49, pp. 588–601, Jul./Aug. 2007.

Qiya Song received the B.S. degree from the Lanzhou University of Technology, Lanzhou, China, in 2018. He is currently pursuing the Ph.D. degree with the Laboratory of Vision and Image Processing, Hunan University, Changsha, China.

His research interests include multimodal information fusion, speech recognition, and human–robot natural interaction.

Bin Sun (Member, IEEE) received the B.S. and Ph.D. degrees in control science and engineering from Hunan University, Changsha, China, in 2010 and 2016, respectively.

From 2017 to 2019, he was a Post-Doctoral Researcher in electrical engineering with the College of Electrical and Information Engineering, Hunan University, where he is currently an Associate Professor. His research interests include computer vision and human–robot natural interaction.

Shutao Li (Fellow, IEEE) received the B.S., M.S., and Ph.D. degrees from Hunan University, Changsha, China, in 1995, 1997, and 2001, respectively.

In 2001, he joined the College of Electrical and Information Engineering, Hunan University, where he is currently a Full Professor. From May 2001 to October 2001, he was a Research Associate with the Department of Computer Science, The Hong Kong University of Science and Technology, Hong Kong. From November 2002 to November 2003, he was a Post-Doctoral Fellow with the Royal Holloway College, University of London, Egham, U.K. From April 2005 to June 2005, he was a Visiting Professor with the Department of Computer Science, The Hong Kong University of Science and Technology. He has authored or coauthored over 200 refereed articles. His current research interests include image processing, pattern recognition, and artificial intelligence.

Dr. Li received two Second-Grade State Scientific and Technological Progress Awards of China in 2004 and 2006. He is an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT. He is a member of the Editorial Board of the *Information Fusion* and the *Sensing and Imaging*.