CATNet: Cross-modal fusion for audio–visual speech recognition

Xingmei Wang[a], Jianchen Mi[a], Boquan Li[ab], Yixu Zhao[a], Jiaxiang Meng[a]

a College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China

b School of Computing and Information Systems, Singapore Management University, 178902, Singapore

Published in Pattern Recognition Letters 178 (2024), 216-222. DOI: 10.1016/j.patrec.2024.01.002

Abstract:

Automatic speech recognition (ASR) is a typical pattern recognition technology that converts human speeches into texts. With the aid of advanced deep learning models, the performance of speech recognition is significantly improved. Especially, the emerging Audio–Visual Speech Recognition (AVSR) methods achieve satisfactory performance by combining audio-modal and visual-modal information. However, various complex environments, especially noises, limit the effectiveness of existing methods. In response to the noisy problem, in this paper, we propose a novel cross-modal audio–visual speech recognition model, named CATNet. First, we devise a cross-modal bidirectional fusion model to analyze the close relationship between audio and visual modalities. Second, we propose an audio–visual dual-modal network to preprocess audio and visual information, extract significant features and filter redundant noises. The experimental results demonstrate the effectiveness of CATNet, which achieves excellent WER, CER and converges speeds, outperforms other benchmark models and overcomes the challenge posed by noisy environments.

Keywords: Audio–visual speech recognition, Cross-modal fusion, Attention mechanism, Deep learning

1. Introduction

Automatic Speech Recognition (ASR) [1,2] is a critical branch in pattern recognition research. Such technologies focus on videos with speeches and recognize speech contents into texts, and have been ap□plied to multiple applications [3] such as Apple Siri. Previously, numerous conventional speech recognition models are broadly researched [4]. With the rapid development of deep learning models, the performance of speech recognition has been improved [5] to a new level. However, undesirable noises pose serious challenges to deep learning meth□ods [6]. Once a noisy environment is encountered by these models, their performance will be significantly degraded in practice.

As an alternative strategy, researchers propose that facial image features are observed in audio-noisy environments [7] and can be used to recognize speech contents. Such methods provide feasible solutions to reduce the impact of noises on speech recognition [8]. By applying an Automatic Lip-Reading (ALR) system [4], these models recognize speeches through vocal organs of mouths and lips [1,9], such as the typ□ical LipNet [10]. However, various complex environments pose great challenges to these models [11], such as optical noises or illumination conditions, which significantly degrade their performance.

In recent years, various typical methods for speech recognition are proposed, especially Audio–Visual Speech Recognition (AVSR). Motivated from that visual information is hard to be affected by audio noises and vice versa, AVSR utilizes the information from audio and visual modalities jointly [12,13], so as to circumvent the impact of complex environments. Typically, Saudi et al. [14] propose a Long and Short Term Memory Bidirectional Recurrent Neural Network (LSTM-BRNN). Zhang et al. [15] propose a Bimodal-DFSMN to enhance the robustness of AVSR against visual modal deficits. Vakhshiteh et al. [16] propose an AV-DBN-HMM that integrates visual and audio information based on the entropy of different layers in a Dynamic Bayesian Network (DBN). In addition, AVSR reveals limited performance as the visual information is insufficient, and thus Su et al. [17] propose an audio–visual inversion technique, Acoustic-to-Visual (A2V), to generate visual features for speech recognition. Although the emergence of AVSR improves the performance of speech recognition, the deficiencies in modality combination and feature-dimension splicing limit the performance of existing methods in noisy environments, due to the information lack caused by noise data.

In this paper, we propose a novel cross-modal audio–visual speech recognition network, named CATNet (Cross-modal Audio–visual Time–space-channel Attention Module LipNet). In CATNet, we devise (1) a cross-modal bidirectional fusion model based on a Feature Pyramid Network (FPN) [18] and (2) an audio–visual dual-modal speech recognition network consisting of an audio information processing network based on a fundamental LipNet, a Convolutional Block Attention Module (CBAM) [19] and a Temporal Attention Mechanism Module. CATNet is expected to be effective as well as robust against

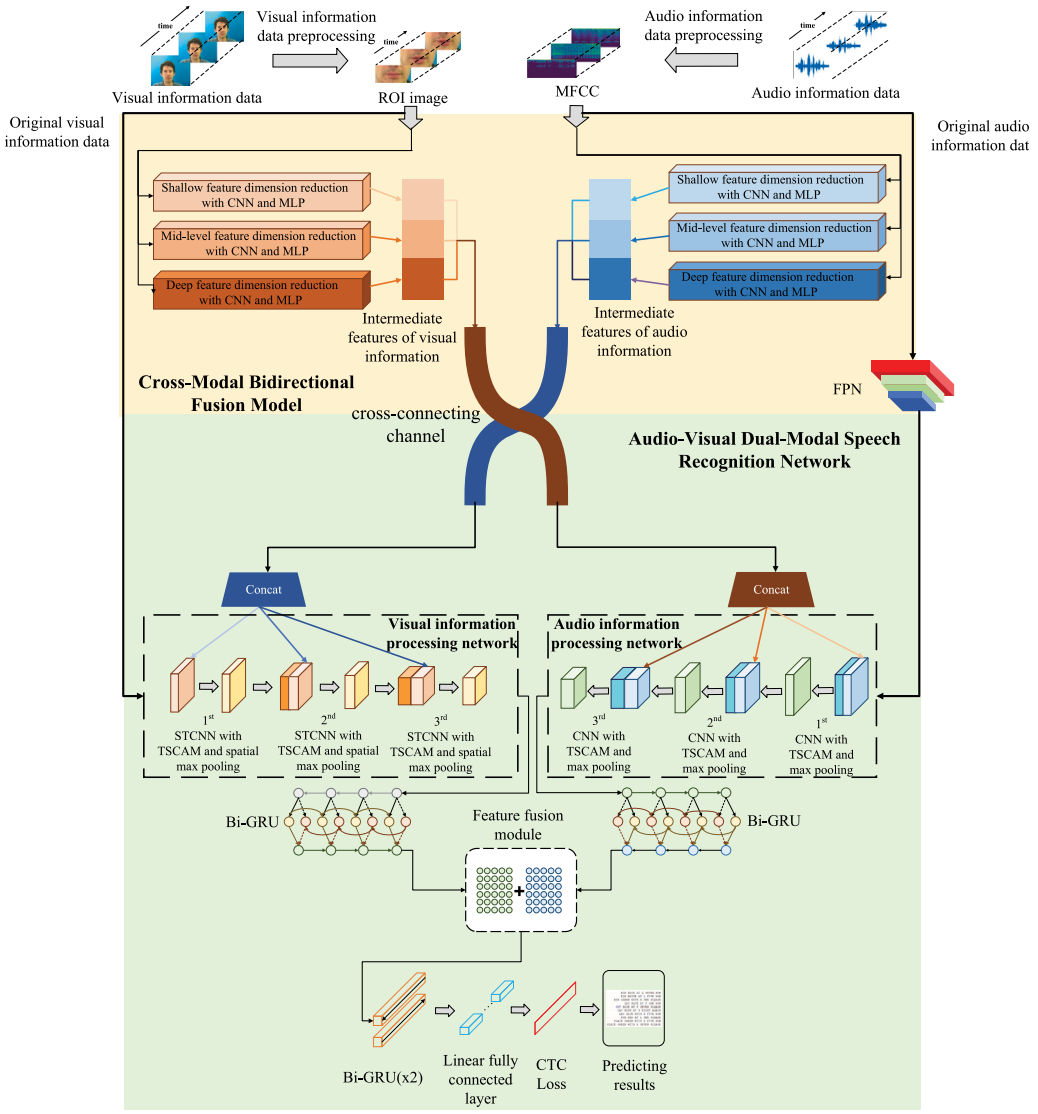


Fig. 1. Architecture of CATNet.

noises, i.e., to overcome the challenge of speech recognition in noisy environments.

Overall, the main contributions of this paper are summarized as follows.

- We devise a cross-modal bidirectional fusion model that fuses information from audio as well as visual modalities, so as to absorb the advantages of audio-based and visual-based speech recognition methods.
- We devise an audio–visual dual-modal speech recognition network, where an audio information processing network extracts significant features, a Convolutional Block Attention Module (CBAM) filters out redundant noises, and a Temporal Attention Mechanism Module overcomes the problem brought by feature similarity between adjacent frames.
- The effectiveness of CATNet is evaluated by conducting systematic experiments. The results demonstrate that CATNet outperforms other benchmarks in terms of its promising Word Error Rate (WER) and Character Error Rate (CER), efficient convergence speed, and excellent robustness against noises.

The rest of this paper is organized as follows. Section 2 presents the methodology of CATNet in detail, which is further evaluated in Section 3. Finally, we conclude this work in Section 4.

2. Methodology

In this section, we present the methodology of CATNet. As the architecture illustrated in Fig. 1, CATNet is composed of three modules, i.e., data preprocessing module, cross-modal bidirectional fusion model, and audio–visual dual-modal speech recognition network.

2.1. Data preprocessing module

A processing step is first performed to align the visual and audio information in temporal dimensions.

In visual data, motivated from that the changes in mouth regions are more imperative than other ones such as eyes and noses [20,21], the mouth regions in images are extracted as Regions of Interest (ROI). Specifically, we utilize a DLib toolkit [22] to locate and extract salient mouth regions, where the size of each ROI image frame is 160 × 80, and the ROI image of each frame is normalized as:

$$X_v = X_v/255.0, \tag{1}$$

where X_v is the pixel in each ROI image. To further extract enough meaningful information from mouth regions, the extracted ROI image sequences are flipped horizontally into units of video:

$$X_v^{(i,j)} = X_v^{(i,160-j)} if random \geq p, \tag{2}$$

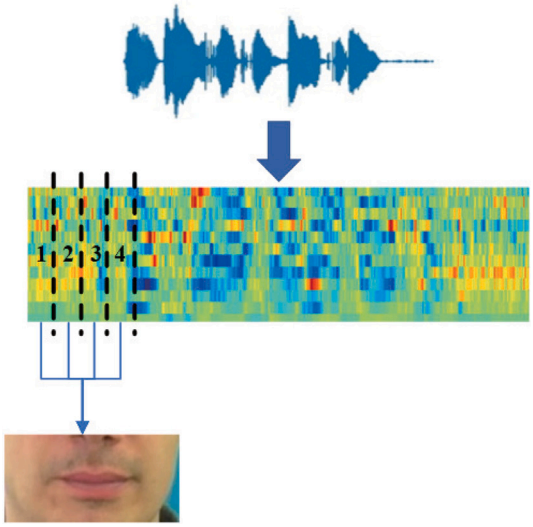


Fig. 2. Alignment between Audio and Visual data.

where *random* is the random number. p is a hyperparameter that refers to a probability threshold to control the horizontal-flipping operations.

In audio data, to maintain the time consistency with visual data, the frame ratio of audio to images is set to 4:1 on Mel Frequency Cepstrum Coefficient (MFCC) [23], as illustrated in Fig. 2, and the visual feature is then obtained by Min–Max Normalization on MFCC:

$$X'_a = \frac{X_a - X_{a\_\min}}{X_{a\_\max} - X_{a\_\min}}. \tag{3}$$

2.2. Cross-modal bidirectional fusion model

In the yellow part of Fig. 1, a cross-modal bidirectional fusion model is divided into visual and audio flows, which respectively extract visual and audio information from the image and voice of a speaker.

The visual and audio flows are first input into parallel but different Convolutional Neural Networks (CNN), so as to extract shallow, middle and deep levels of features from different dimensions. Specifically, the intermediate features including shallow feature f_{a_s}, middle feature f_{a_m} and deep feature f_{a_d} in audio flow, and shallow feature f_{v_s}, middle feature f_{v_m} and deep feature f_{v_d} in visual flow. These features are extracted by a Multilayer Perceptron (MLP) [24]-based dimensional transformation to match the features in the following dual-modal speech recognition network, and are then input into a cross-connecting channel, so as to concatenate features of different layers in another modality.

In addition, a Feature Pyramid Network (FPN) [18] is appended to the visual flow that enables the model to focus on overall features instead of any detailed one, and to perform multi-scale feature extraction. Ultimately, the network extracts the last-layer feature map in the top-down feature enhancement channel in FPN.

2.3. Audio–visual dual-modal speech recognition network

In the green part of Fig. 1, an audio–visual dual-modal speech recognition network, ATNet, is divided into three modules, i.e., visual information processing network, audio information processing network and feature fusion network, as detailed in Fig. 3.

ATNet is constructed based on LipNet [10], which processes visual and audio information simultaneously to fill the gap in single-modal networks. Note that the visual and audio information processing networks receive intermediate features from cross-connecting channels, and transmit shallow, medium and deep features between different modalities. Such strategies could improve the effectiveness of information fusion between visual and audio data, and thereby enabling the model to be robust against noisy environments.

The audio information processing network contains a three-layer CNN as well as max pooling layers. Differently, the visual information processing network consists of a three-layer Spatio-Temporal Convolutional Neural Network (STCNN) [25] as well as spatial max pooling layers which is formulated as:

$$[stconv(x,w)]_{d'tij} = \sum_{d=1}^{D}\sum_{t'=1}^{k_t}\sum_{i'=1}^{k_w}\sum_{j'=1}^{k_h} w_{d'dt'i'j'} x_{d,t+t',i+i',j+j'}, \tag{4}$$

where k_t is the length of a time step, k_w and k_h are kernel sizes. d is the number of STCNN layers and w represents its parameters Moreover, x refers to the image representation.

The convolution operation mixes cross-channel information and spatial information to extract features. To enable the network to focus on meaningful features and suppress unimportant ones, a Convolutional Block Attention Module (CBAM) [19] is appended to each network, so as to emphasize meaningful features in channel and spatial dimensions.

Intuitively, it is observe that features are similar between adjacent frames in the audio and visual information processing networks, which indicates that some necessary features are extracted from such adjacent similar frames. Thus, a Temporal Attention Mechanism [26] module is further adopted to extract meaningful information in the temporal dimension as well as filter out redundant features, as follows:

$$\begin{aligned} M_t(F) &= \sigma_2\left(MLP\left(\sigma_1\left(MLP\left(MaxPool\left(F\right)\right)\right)\right)\right. \\ &\quad \left.+MLP\left(\sigma_1\left(MLP\left(AvgPool\left(F\right)\right)\right)\right)\right) \\ &= \sigma_2\left(W_1\left(\sigma_1\left(W_0\left(F^t_{\max}\right)\right)\right)+W_1\left(\sigma_1\left(W_0\left(F^t_{avg}\right)\right)\right)\right), \end{aligned} \tag{5}$$

where F is the input feature, $F^t_{\max}$ and F^t_{avg} are respectively features with significant information and with background information. MLP is the convolutional layer and W_0 and W_1 are parameters in MLP. $\sigma_1(\cdot)$ refers to a ReLU function [27] and $\sigma_2(\cdot)$ represents a Sigmoid function.

Fig. 4 illustrates the Temporal Attention Mechanism module utilizes max pooling and average pooling layers to reduce the temporal dimension, so as to obtain features F^t_{max} and F^t_{avg} that severally highlight the distinguishing features and background information, and utilizes a convolutional layer to filter out redundant information.

Note that the STCNN (or CNN) is combined with CBAM to extract features in channel and space dimensions respectively, and the Temporal Attention Mechanism module also extracts features in the time dimension. CBAM and the Temporal Attention Mechanism module are combined as TSCAM, and these product results are added and then divided by 2.

Specifically, the operations to concatenate the intermediate features from cross-connecting channels, and different levels of features in CNN, are formulated as:

$$\begin{aligned} F_{a\_fianl} =& CNN\_TSCAM_3\left(Concat\left(CNN\_TSCAM_2\left(\right.\right.\right. \\ & Concat\left(CNN\_TSCAM_1\left(Concat\left(F_{a\_o}, f_{v\_s}\right)\right),\right. \\ & \left.\left.\left. f_{v\_m}\right)\right), f_{v\_d}\right)\right), \end{aligned} \tag{6}$$

where F_{a_o} represents the MFCC of audio features, and $CNN\_TSCAM_i(\cdot)$ is the ith layer of CNN with TSCAM. Similarly, such concatenation operations in STCNN are formulated as:

$$\begin{aligned} F_{v\_fianl} =& STCNN\_TSCAM_3\left(Concat\left(STCNN\_TSCAM_2\left(\right.\right.\right. \\ & Concat\left(STCNN\_TSCAM_1\left(Concat\left(F_{v\_o}, f_{a\_s}\right)\right),\right. \\ & \left.\left.\left. f_{a\_m}\right)\right), f_{a\_d}\right)\right), \end{aligned} \tag{7}$$

where F_{v_o} is the last-layer visual features based on the top-down feature enhancements in FPN. $STCNN\_TSCAM_i(\cdot)$ is the ith layer of STCNN with TSCAM. In this way, the shallow, middle and deep features from the cross-modal bidirectional fusion model are extracted as single modalities, and are then transferred to the different levels in another modality, so as to learn dual-modal information in a fused manner.

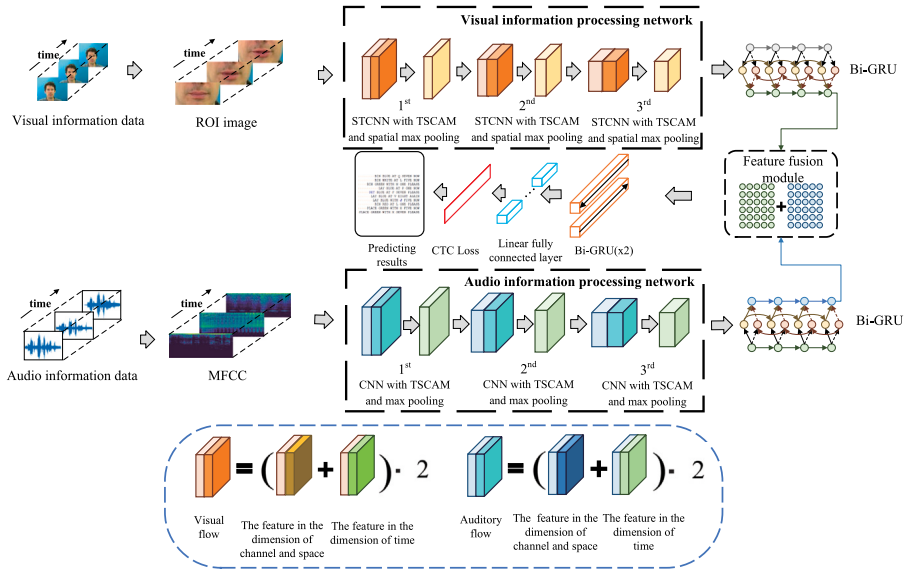


Fig. 3. Architecture of ATNet.

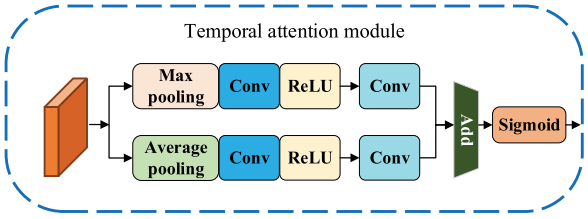


Fig. 4. Architecture of temporal attention mechanism module.

Subsequently, the features from the visual and audio information processing networks are input into a Bidirectional Gate Recurrent Unit (Bi-GRU) layer to obtain the visual feature X_v and audio feature X_a, and a feature fusion network concatenates X_v and X_a to fuse features as follows:

$$X = X_v \oplus X_a, \tag{8}$$

where X is the fused feature, and $\oplus$ is the concatenation equation of the feature matrix.

Finally, based on the fused features, an additional two-layer Bi-GRU with a Connectionist temporal classification (CTC) [28] loss function is utilized to analyze temporal information. Formally, the objective of the optimization is:

$$\begin{aligned} \arg\max_{\theta} L(D) &= -\ln\left(\prod_{(X,l)\in D} p(l\,|X)\right) \\ &= -\sum_{(X,l)\in D} \ln\left(\sum_{\pi\in B^{-1}(l)} p(\pi|X)\right), \end{aligned} \tag{9}$$

where D is the training set, X is the fused feature, l is the ground-truth label and θ is the parameter of CATNet.

3. Experiments

In this section, we introduce our experimental settings, conduct comparative experiments on the Temporal Attention Mechanism module, and evaluate CATNet in both conventional as well as noisy environments.

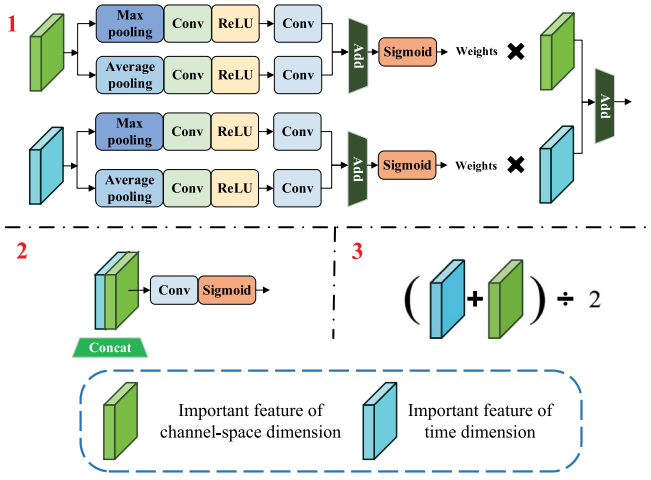


Fig. 5. Combination schemes between temporal attention mechanism and CBAM.

3.1. Experimental settings

We start with introducing the datasets and metrics that are utilized in our experiments.

3.1.1. Dataset

To evaluate CATNet, we adopt a typical audio–visual speech recognition dataset, GRID [29], to our experiments. GRID is adopted based on the following reasons: (1) It is representative and is extensively researched by numerous work (over 1200 citations). (2) It is broadly applied to evaluate various speech recognition methods [10,30,31]. (3) It covers a variety of lexical properties in English sentences collected in complex speech-recognition environments, which has constructed challenging conditions for speaker recognition, and is in accordance with the challenge that is aimed to be addressed by this work.

Specifically, GRID contains 34,000 videos from 16 female and 18 male speakers, and each speaker records 1000 videos. Each video in the dataset consists of 75 frames in the resolution of 360 * 288, segment of 3 s, and frame rate of 25 fps. Moreover, each sentence in GRID is composed of six word categories following the grammar: command (4) + color (4) + preposition (4) + letter (25) + digit (10) + adverb (4),

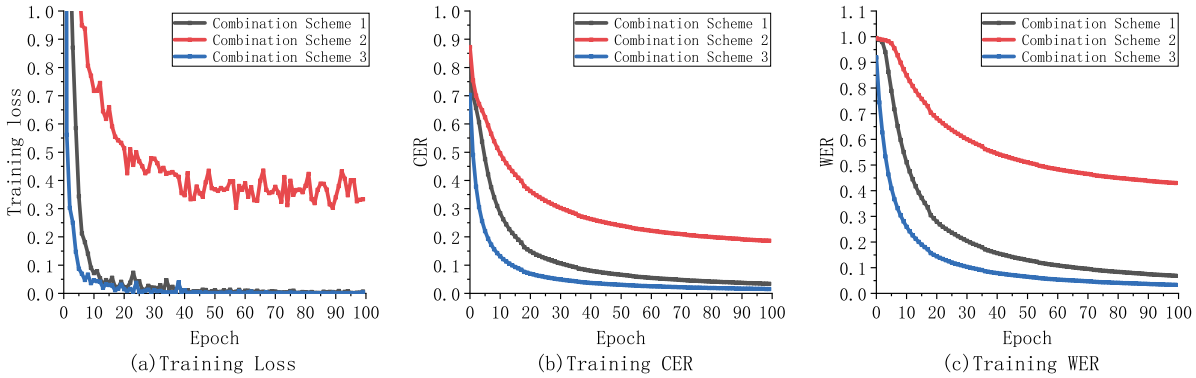


Fig. 6. Experimental results on comparative combination schemes.

Table 1
Sentence structures of GRID.

Command statement	Color	Preposition	Letter	Number	Adverb
bin	blue	at	A-Z (Excluding W)	1–9, zero	again
lay	green	by	/	/	now
place	red	in	/	/	please
set	while	with	/	/	soon

Table 2
Evaluation results on conventional environments.

Speech recognition model [Modality]	WER (%)	CER (%)
A-LipNet [audio]	5.68	2.57
Asymmetric BLSTM [audio]	70.35	35.04
LipNet [visual]	26.18	11.56
TM-seq2seq [visual]	93.11	48.17
XFLow [visual]	5.05	2.38
AV-LipNet [audio–visual]	3.52	1.40
AV-CBAM-LipNet [audio–visual]	2.97	1.32
ATNet [audio–visual]	2.03	1.13
CATNet [audio–visual]	**1.60**	**0.68**

such as 'SET WHITE WITH M 6 PLEASE'. The numbers in parentheses refer to the candidates for these word categories, which could produce over 60,000 potential sentences that are sufficient for evaluation.

Table 1 presents the sentence structure of GRID, i.e., the possible words in each position. In these positions, the letter 'W' is excluded, and the number '0' is pronounced 'Zero' to avoid potential pronunciation problems.

3.1.2. Evaluation metrics

We adopt two typical metrics to evaluate CATNet, i.e., WER and CER, which are extensively utilized to metric the effectiveness of speech recognition methods [32].

Specifically, WER (%) is formulated as:

$$WER = 100 \times \frac{WS + WD + WI}{WN}, \tag{10}$$

where WS, WD and WI are the numbers of words to be replaced, deleted, and inserted respectively, and WN is the number of words with ground-truth labels.

Similarly, CER (%) is formulated as:

$$CER = 100 \times \frac{CS + CD + CI}{CN}, \tag{11}$$

where CI, CD, and CS are the numbers of characters to be replaced, deleted, and inserted respectively, and CN is the number of characters with ground-truth labels.

Note that WER and CER results are possible over 100% depending on the number of inserted words and characters.

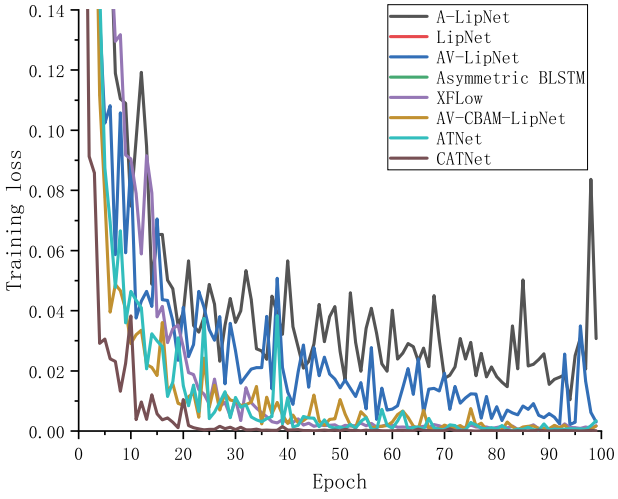


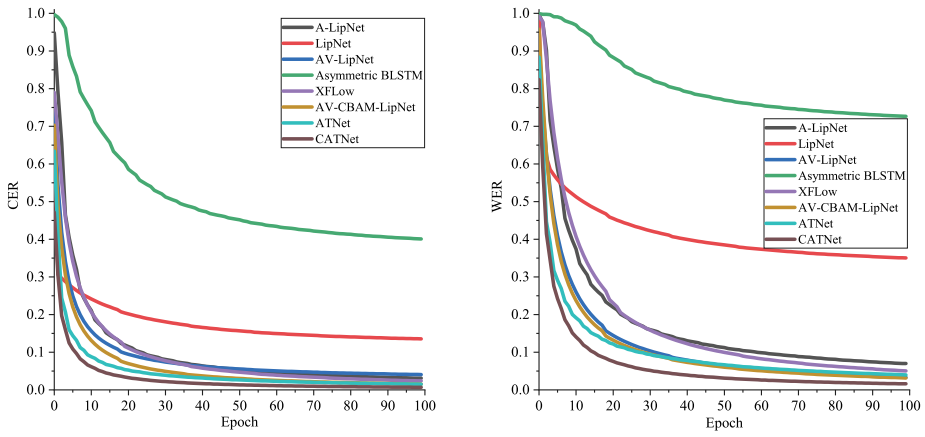
Fig. 7. Training Losses of CATNet and other benchmark models.

3.2. Comparative experiments on temporal attention mechanism

Recall our methodology in Section 2, we propose to combine the Temporal Attention Mechanism module with CBAM. To explore the most effective combination solution beforehand, we perform comparative experiments to evaluate the effect of significant features in channel-spatial and temporal dimensions.

Specifically, as illustrated in Fig. 5, three combination strategies are set as candidates: (1) For the features in channel-spatial and temporal dimensions, the weight distributions are severally calculated for the channel attention mechanism, and their sum weights are then utilized to extract the features of the temporal–spatial-channel dimension. (2) The features in channel-spatial and temporal dimensions are concatenated, and are then input into CNN to extract the features of three dimensions. (3) The features in channel-spatial and temporal dimensions are superimposed, and the result is divided by 2.

Further, we conduct speech recognition experiments based on these combination schemes. As the results illustrated in Fig. 6(a), Scheme 3 achieves the best convergence speed, and the loss drops to a small range in the first five rounds. Moreover, as results illustrated in Fig. 6(b) and



(a)CER of CATNet and Other Benchmark Models (b)WER of CATNet and Other Benchmark Models

Fig. 8. (a) CER and (b) WER of CATNet and Other Benchmark Models.

Table 3
Evaluation results on anti-noise robustness.

Speech recognition model [Modality]	SNR					
	−5 dB		10 dB		20 dB	
	WER (%)	CER (%)	WER (%)	CER (%)	WER (%)	CER (%)
A-LipNet [audio]	129.23	80.36	37.28	22.39	7.16	4.63
Asymmetric BLSTM [audio]	94.74	65.21	81.06	51.21	75.18	47.64
LipNet [visual]	26.19	9.09	26.19	9.09	26.19	9.09
XFLow [audio–visual]	88.04	58.04	73.62	46.69	37.05	25.17
AV-LipNet [audio–visual]	99.47	69.18	14.34	6.99	2.11	0.64
AV-CBAM-LipNet [audio–visual]	107.17	66.95	13.81	9.59	1.14	0.48
ATNet [audio–visual]	80.61	61.77	10.25	5.31	7.61	4.18
CATNet [audio–visual]	54.92	32.08	**3.08**	**1.22**	**0.88**	**0.29**

Fig. 6(c), Scheme 3 achieves the best WER and CER. Based on such results, Scheme 3 is adopted to perform the following experiments.

3.3. Evaluation experiments on catnet

Next, we evaluate the complete CATNet in terms of: (1) WER and CER, (2) convergence rate and (3) anti-noise robustness.

Specifically, based on GRID, we perform speech recognition experiments to CATNet as well as multiple advanced benchmark models, including Audio-LipNet (A-LipNet), Asymmetric BLSTM [33], LipNet [10], TM-seq2seq [34], XFlow [35], AV-LipNet, AV-CBAM-LipNet and ATNet. Note that A-LipNet is constructed based on audio features only.

3.3.1. Evaluation experiments on conventional environments

These models are first evaluated based on GRID S8/bbaefn (BIN BLUE AT E FIVE NOW). As the results in Table 2, most of these models successfully predict the speaking contents, and CATNet outperforms the other models. Especially, compared with the sub-optimal ATNet, CATNet reduces WER and CER by 0.43% and 0.45% respectively. Such results demonstrate the promising speech recognition ability of CATNet, which is capable of leveraging audio and visual information to establish potential associations.

Moreover, we evaluate the convergence performance of these models based on GRID. As the training loss results illustrated in Fig. 7, compared with other benchmark models, CATNet shows substantial drops in the training loss of the first-ten epochs, and the training loss curve becomes stable at around the 25th epoch. For comprehensive evaluations, we present the trends of WER and CER during training in Fig. 8. Compared with other benchmark models, CATNet also presents promising performance in dropping both WER and CER. In general, CATNet achieves the best convergence speeds among these benchmark models, which demonstrates its excellent fitting effects and is prospective to be applied in practice.

3.3.2. Evaluation experiments on anti-noise robustness

In the following, we evaluate the anti-robustness performance of CATNet. Specifically, we pollute the data in GRID based on different intensities of Gaussian white noises, i.e., under Signal to Interference plus Noise Ratio (SNR) [36] of 20 dB, 10 dB and −5 dB.

We present the evaluation results in Table 3. To first observe the results of those benchmark models, it is observed that LipNet achieves promising results under 10-dB and 20-dB noises. Especially, under −5-dB noises (the loudest noise level), LipNet performs best in terms of both WER and CER. Recall that LipNet is a visual model that performs speech recognition by mapping the sequence of a speaker's mouth, so as to recognize the entire sentences. Moreover, the added Gaussian noise mainly affects the audio information in the video, and thus will not affect the performance of such visual-based models. Similarly, A-LipNet is an audio model that shows poor anti-noise robustness. Such results further reveal the superiority of selecting LipNet as a basic component in our CATNet.

To further observe the results in Table 3, under high-intensity noises, Asymmetric BLSTM and XFLow achieve better results than AV-LipNet and AV-CBAM-LipNet, since the latter models are affected by the inclusion of susceptible audio modalities. However, the advantages of Asymmetric BLSTM and XFLow cease to exist under low-intensity noises, since AV-LipNet and AV-CBAM-LipNet introduce visual modalities where the feature fusion modules are able to reject noises so as to increase the noise robustness of the model. Finally, AV-CBAM-LipNet performs better than AV-LipNet since CBAM enables the network to pay meticulous attention to subtle features.

Finally, it is observed that CATNet achieves the best performance among all the benchmark models, which indicates our designed network has integrated the above advantages of information fusion, noise filtering and significant feature extraction. Overall, CATNet overcomes the challenge posed by noisy environments and is with outstanding anti-noise robustness.

4. Conclusion

In this paper, we propose CATNet, a novel speech recognition model. In CATNet, (1) a cross-modal bidirectional fusion model analyzes the close relationship between audio and visual modalities, and (2) an audio–visual dual-modal network preprocesses audio and visual information, extracts significant features and filters redundant noises. Our evaluation results demonstrate CATNet achieves promising WER, CER and converges speed, overcomes the noisy-environment challenges, and outperforms other benchmark models.

In the future, we attempt to introduce more meaningful visual information, such as human gestures, into our models, so as to explore potential improvements.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data used in this paper is open.

Acknowledgments

This work was supported by Key Laboratory of Avionics System Integrated Technology, and Fundamental Research Funds for the Central Universities in China [Grant number: 3072022JC0601].

References

[1] S. Petridis, Y. Wang, P. Ma, Z. Li, M. Pantic, End-to-end visual speech recognition for small-scale datasets, Pattern Recognit. Lett. 131 (2020) 421–427, http://dx.doi.org/10.1016/j.patrec.2020.01.022.

[2] N. Radha, A. Shahina, P. Prabha, B.P. Sri, A.N. Khan, An analysis of the effect of combining standard and alternate sensor signals on recognition of syllabic units for multimodal speech recognition, Pattern Recognit. Lett. 115 (2018) 39–49, http://dx.doi.org/10.1016/j.patrec.2017.10.011.

[3] V. Kepuska, G. Bohouta, Next-generation of virtual personal assistants (microsoft cortana, apple siri, amazon alexa and google home), in: 2018 IEEE 8th Annual Computing and Communication Workshop and Conference, CCWC, IEEE, 2018, pp. 99–103, http://dx.doi.org/10.1109/CCWC.2018.8301638.

[4] S. Bhaskar, T. Thasleema, Scope for deep learning: A study in audio-visual speech recognition, in: 2019 International Conference on Computational Intelligence and Knowledge Economy, ICCIKE, IEEE, 2019, pp. 72–77, http://dx.doi.org/10.1109/ICCIKE47802.2019.9004287.

[5] L. Deng, J. Platt, Ensemble deep learning for speech recognition, in: Proc. Interspeech, 2014.

[6] P. Swietojanski, A. Ghoshal, S. Renals, Revisiting hybrid and GMM-HMM system combination techniques, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 6744–6748, http://dx.doi.org/10.1109/ICASSP.2013.6638967.

[7] S. Dupont, J. Luettin, Audio-visual speech modeling for continuous speech recognition, IEEE Trans. Multimed. 2 (3) (2000) 141–151, http://dx.doi.org/10.1109/6046.865479.

[8] K. Noda, Y. Yamaguchi, K. Nakadai, H.G. Okuno, T. Ogata, Audio-visual speech recognition using deep learning, Appl. Intell. 42 (4) (2015) 722–737, http://dx.doi.org/10.1007/s10489-014-0629-7.

[9] U. Sharma, S. Maheshkar, A.N. Mishra, R. Kaushik, Visual speech recognition using optical flow and hidden Markov model, Wirel. Pers. Commun. 106 (4) (2019) 2129–2147, http://dx.doi.org/10.1007/s11277-018-5930-z.

[10] Y.M. Assael, B. Shillingford, S. Whiteson, N. De Freitas, Lipnet: End-to-end sentence-level lipreading, 2016, http://dx.doi.org/10.48550/arXiv.1611.01599, arXiv preprint arXiv:1611.01599.

[11] J. Huang, B. Kingsbury, Audio-visual deep learning for noise robust speech recognition, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 7596–7599, http://dx.doi.org/10.1109/ICASSP.2013.6639140.

[12] M. Wand, J. Schmidhuber, N.T. Vu, Investigations on end-to-end audiovisual fusion, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2018, pp. 3041–3045, http://dx.doi.org/10.1109/ICASSP.2018.8461900.

[13] L. Schoneveld, A. Othmani, H. Abdelkawy, Leveraging recent advances in deep learning for audio-visual emotion recognition, Pattern Recognit. Lett. 146 (2021) 1–7, http://dx.doi.org/10.1016/j.patrec.2021.03.007.

[14] A.S. Saudi, M.I. Khalil, H.M. Abbas, Improving audio-visual speech recognition using gabor recurrent neural networks, in: IAPR Workshop on Multimodal Pattern Recognition of Social Signals in Human-Computer Interaction, Springer, 2018, pp. 71–83, http://dx.doi.org/10.1007/978-3-030-20984-1_7.

[15] S. Zhang, M. Lei, B. Ma, L. Xie, Robust audio-visual speech recognition using bimodal DFSMN with multi-condition training and dropout regularization, in: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2019, pp. 6570–6574, http://dx.doi.org/10.1109/ICASSP.2019.8682566.

[16] F. Vakhshiteh, F. Almasganj, Exploration of properly combined audiovisual representation with the entropy measure in audiovisual speech recognition, Circuits Systems Signal Process. 38 (6) (2019) 2523–2543, http://dx.doi.org/10.1007/s00034-018-0975-5.

[17] R. Su, X. Liu, L. Wang, J. Yang, Cross-domain deep visual feature generation for mandarin audio–visual speech recognition, IEEE/ACM Trans. Audio Speech Lang. Process. 28 (2019) 185–197, http://dx.doi.org/10.1109/TASLP.2019.2950602.

[18] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2117–2125.

[19] S. Woo, J. Park, J.-Y. Lee, I.S. Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 3–19.

[20] A. Nieto-Castanon, S.S. Ghosh, J.A. Tourville, F.H. Guenther, Region of interest based analysis of functional imaging data, Neuroimage 19 (4) (2003) 1303–1316, http://dx.doi.org/10.1016/S1053-8119(03)00188-5.

[21] R.A. Poldrack, Region of interest analysis for fMRI, Soc. Cogn. Affect. Neurosci. 2 (1) (2007) 67–70, http://dx.doi.org/10.1093/scan/nsm006.

[22] D.E. King, Dlib-ml: A machine learning toolkit, J. Mach. Learn. Res. 10 (2009) 1755–1758.

[23] V. Tiwari, MFCC and its applications in speaker recognition, Int. J. Emerg. Technol. 1 (1) (2010) 19–22.

[24] H. Taud, J. Mas, Multilayer perceptron (MLP), in: Geomatic Approaches for Modeling Land Change Scenarios, Springer, 2018, pp. 451–455, http://dx.doi.org/10.1007/978-3-319-60801-3_27.

[25] Z. He, C.-Y. Chow, J.-D. Zhang, STCNN: A spatio-temporal convolutional neural network for long-term traffic prediction, in: 2019 20th IEEE International Conference on Mobile Data Management, MDM, IEEE, 2019, pp. 226–233, http://dx.doi.org/10.1109/MDM.2019.00-53.

[26] C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, Q. Dai, STAT: Spatial-temporal attention mechanism for video captioning, IEEE Trans. Multimed. 22 (1) (2019) 229–241, http://dx.doi.org/10.1109/TMM.2019.2924576.

[27] A.F. Agarap, Deep learning using rectified linear units (relu), 2018, arXiv preprint arXiv:1803.08375.

[28] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 23rd International Conference on Machine Learning, 2006, pp. 369–376, http://dx.doi.org/10.1145/1143844.1143891.

[29] M. Cooke, J. Barker, S. Cunningham, X. Shao, An audio-visual corpus for speech perception and automatic speech recognition, J. Acoust. Soc. Am. 120 (5) (2006) 2421–2424, http://dx.doi.org/10.1121/1.2229005.

[30] S.T. Shivappa, B.D. Rao, M.M. Trivedi, Multimodal information fusion using the iterative decoding algorithm and its application to audio-visual speech recognition, in: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2008, pp. 2241–2244, http://dx.doi.org/10.1109/ICASSP.2008.4518091.

[31] A. Thanda, S.M. Venkatesan, Audio visual speech recognition using deep recurrent neural networks, in: Multimodal Pattern Recognition of Social Signals in Human-Computer-Interaction: 4th IAPR TC 9 Workshop, MPRSS 2016, Cancun, Mexico, December 4, 2016, Revised Selected Papers 4, Springer, 2017, pp. 98–109, http://dx.doi.org/10.1007/978-3-319-59259-6_9.

[32] Y. Zhou, C. Xiong, R. Socher, Improving end-to-end speech recognition with policy learning, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP, IEEE, 2018, pp. 5819–5823, http://dx.doi.org/10.1109/ICASSP.2018.8462361.

[33] X. Wang, F. Xue, W. Wang, A. Liu, A network model of speaker identification with new feature extraction methods and asymmetric BLSTM, Neurocomputing 403 (2020) 167–181, http://dx.doi.org/10.1016/j.neucom.2020.04.041.

[34] T. Afouras, J.S. Chung, A. Senior, O. Vinyals, A. Zisserman, Deep audio-visual speech recognition, IEEE Trans. Pattern Anal. Mach. Intell. (2018) http://dx.doi.org/10.1109/TPAMI.2018.2889052.

[35] C. Cangea, P. Veličković, P. Lio, Xflow: Cross-modal deep neural networks for audiovisual classification, IEEE Trans. Neural Netw. Learn. Syst. 31 (9) (2019) 3711–3720, http://dx.doi.org/10.1109/TNNLS.2019.2945992.

[36] K.A. Hamdi, On the statistics of signal-to-interference plus noise ratio in wireless communications, IEEE Trans. Commun. 57 (11) (2009) 3199–3204, http://dx.doi.org/10.1109/TCOMM.2009.11.060425.