Speaker detection based on attention mechanism

Haozhen Song [#1,*1], Mintao Dong [#1], Kewei Chen [#1], Fangyan Dong [#1,*2]

#1 Faculty of Mechanical Engineering & Mechanics, Ningbo University, Ningbo, 315211,China
*1 E-mail: 2011081152@nbu.edu.cn
*2 E-mail: dongfangyan@nbu.edu.cn

Abstract: Speaker detection is a necessary pre-processing step for speech recognition and its downstream applications. Unimodal design architectures are currently widely used, but still have limitations in specific environments. Recent research has shown that bimodal fusion learning of audio and visual is a promising solution to the speaker detection task. Thus, in this paper, a bimodal model SdNet is proposed to implement speaker detection by introducing a cross-attention mechanism. Moreover, in order to optimize the different feature sizes caused by different attention levels at different locations of the network, a large number of residual connections are used to map to higher dimensional spaces. After the model is built，the largest AVA-Speech dataset is used for training, the mean average precision (mAP) reaches 80.1%, which is a certain improvement compared with the previously published models, and the experimental results show that the method can improve the accuracy of speaker recognition.

Keywords: Speaker detection; multi model; ResNet; Vision Transformer

1. INTRODUCTION

Research on audiovisual analysis of video speech content is gaining increasing attention for its academic and practical value. Some important topics include visual speech recognition and lip reading, as well as audiovisual speech separation and enhancement. However, while a great deal of work has been done in these areas, very little work has been explored in the area of active speaker detection[1]. Speaker detection, also known as "speech activity detection", has become an important part of the task processing for speech recognition and language recognition, and speaker detection aims to identify the current speaker from the candidate face detection in an arbitrary video stream, which is a necessary and important pre-processing step for visual speech recognition and other downstream applications. preprocessing step and is becoming increasingly important with the growth of audiovisual-based activities[2,3].

Research related to speaker recognition is a research hotspot that has gradually emerged in the last decade, and early studies were mostly done by sociologists, who proposed character features and research methods that can characterize individual influence by analyzing a large number of meetings and conversation processes, laying the foundation for computer- and mathematics-based systems research. in 2016, Punarjay Chakravarty[4] of Belgium presented at the ACM conference proposed to train audio detection in a weakly supervised manner and video classifiers in turn to train speech models.In 2019, JooSon, South Korea, proposed a 3D-CNN[5] front-end and LSTM[6] back-end integrated model for speaker prediction. In the same year, the Institute of Technology of the Chinese Academy of Sciences proposed training with contrast loss and conventional cross-loss in a multi-task learning approach after conventional feature extraction.In 2020, Juan[7] et al. proposed a contextual model for detecting active speakers that uses a self-attention mechanism to model the relationship between candidates so as to combine all the candidates by combining clipped facial trajectories with audio.In 2021, Tao[8] et al. made decisions by fusing long- and short-term features with audiovisual cross-attention by fusing at the decision level.

These studies show that audiovisual fusion-based schemes are feasible and promising for active speaker detection. Considering the temporality between video and audio, more and more researchers are designing models using structures that memorize temporal features, such as recurrent neural networks, gated recurrent units, long- and short-term memory networks, and visual converters. For visual information, some studies have started to extract visual information and visual dependencies by using 3D Convolutional Neural Networks (3D-CNN). For audio, good results have been achieved using Mel Frequency Cepstrum Coefficient (MFCC) or Short-time Fourier Transform (STFT) spectrograms as the input to convolutional neural networks.

Most prediction models, which encode modal features of audio and video separately, do not take full advantage of cross-modal synchronization information. As an inspiration, cross-attention mechanism is supposed to make full use of the information between different modalities.

2. RELATED WORK

2.1. Encoder-Decoder

The Encoder-Decoder model[9] is a common end-to-end model in the field of deep learning that has been widely used and achieved good results on various types of sequence transformation tasks.The Encoder-Decoder model usually employs a convolutional neural network or recurrent neural network to encode the input image, and the model subsequently implements the specified task by using the associated The model then decodes the feature vectors extracted by the encoder through the associated neural network to achieve the specified task.

2.2. Transformer

For long sequences, the recurrent neural network[10] (RNN) suffers from the gradient disappearance problem during training, and the long short-term memory network (LSTM) is slow to train because of its inability to perform parallel computation, although a gating mechanism is introduced to enhance the memory capability of the network. In response to the above problems, the team at Google proposed Transformer[11] model in 2017, which has no structure similar to recurrent neural network and constructs the dependency between input and output sequences entirely through the attention mechanism, which greatly increases the memory time of the model and realizes the parallelization of data in the model for training, which makes the training speed greatly improved. Its structure diagram is shown in Figure 1. The Transformer follows this overall architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder, shown in the left and right halves of Figure 1, respectively.

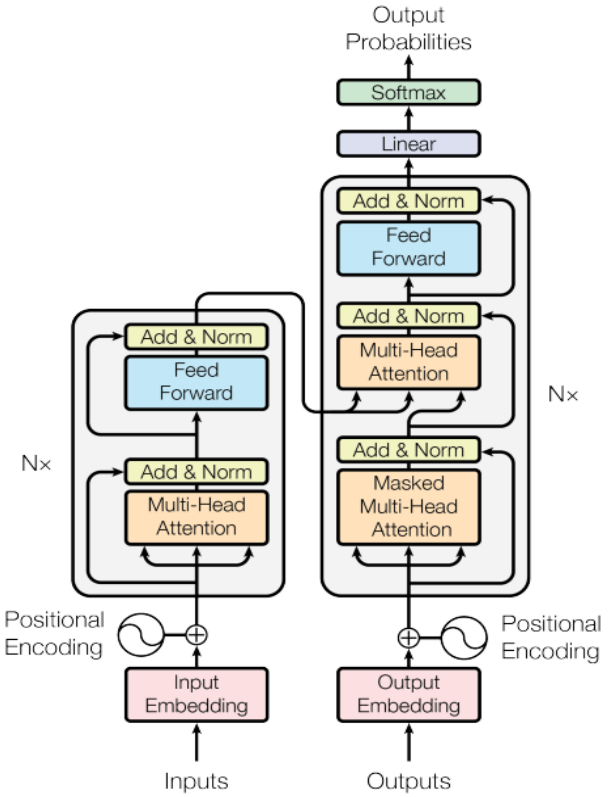


Fig.1 The Transformer-model architecture

In the classical Transformer structure, the encoder and decoder are built with N layers respectively, the mask component makes the current information not directly visible to the future information, the two parts are aligned using attention connections, and the output part contains transformations mapped to higher dimensional space and normalized.

2.3. Attention

The self-attentive layer mainly consists of Scaled Dot-Product Attention and Multi-Headed Attention mechanisms together, and the structure diagram of Scaled Dot-Product Attention is shown in Figure 2.

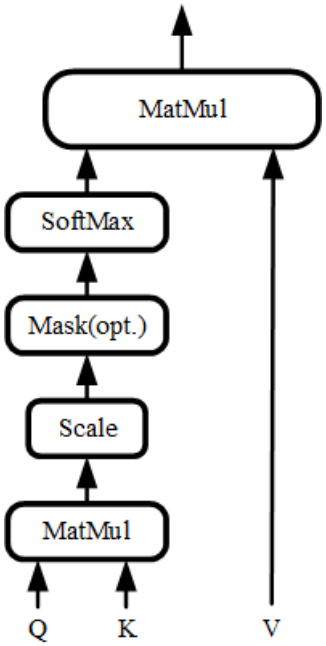


Fig.2 Scaled Dot-Product Attention

The output matrix is calculated as:

$$\text{attention}(\mathrm{Q},\mathrm{K},\mathrm{V}) = \text{Softmax}\left(\frac{\mathrm{Q}K^T}{\sqrt{d}}\right)\mathrm{V} \qquad (1)$$

The attention mechanism is mainly implemented by calculating the mapping relationship between three vector matrices of query (query,Q) and key-value pair (key-value,K-V), where query and key vectors have the same dimension, both of d_k dimension, and value vector is d_v dimension, first calculate the dot product matrix of matrix Q and K^T, divide by $\sqrt{d_k}$ to prevent the matrix value from being too large affect the results of the subsequent normalized exponential function, followed by softmax to obtain the correlation coefficients between Q and K. Finally, the results are multiplied by the vector V to obtain the final output vector, whose dimension is d_v dimension.

3. SDNET(SPEAKER DETECT NETWORK)

3.1. Construct Ideas

An entire sentence spoken by an active speaker may span hundreds of video frames or may be a single tone word. Simply dividing 200ms-600ms video clips to extract

features is not reliable; for example, a 300ms video clip cannot cover a single word and it is difficult to determine speech activity, let alone audiovisual synchronization. Also unimodal input is not reliable, and there may be misjudgment of voices other than the target or behaviors such as eating or laughing. Therefore, I think a better result can be achieved by fusing two modalities on a sentence interval.In order to further improve the performance, instead of using Recurrent Neural Network (RNN), which has a short memory time, and Long Short-Term Memory (LSTM), which cannot be computed in parallel, Transformer was choosed, which has a faster training speed and longer memory time. model.

In this paper, a bimodal fusion recognition model will be build with a face detection frame, and the main backbone feature extraction network of the model mainly uses Residual Neural Network (ResNet). The video feature extraction part abandoned the standard recognition method based on 3D Convolutional Neural Networks (3D ConvNets) and used the Video Transformer Network[12] (VTN) based on Transformer to extract spatial information using 2D structure to avoid the problem of large computation of traditional 3D convolution, and the audio feature extraction network is modified from ResNet18. In order to realize the interaction between bimodal information, a cross-attention mechanism is introduced to improve the fine-grainedness of matching between detecting lips and speech, and finally a cross-entropy function is chosen to perform biclassification.

3.2. Model Structure

A new end-to-end model is constructed, consisting mainly of a feature extraction front-end and a detector back-end classifier, which is named SdNet in this paper. the visual front-end aims to explore the spatial information in each video frame and learn the dynamic features of facial expressions. The back-end classifier adds a cross-attention mechanism between bimodal states, which is used to learn speaking activities from discourse level in time. The overall structure of the model is shown in Figure 3.

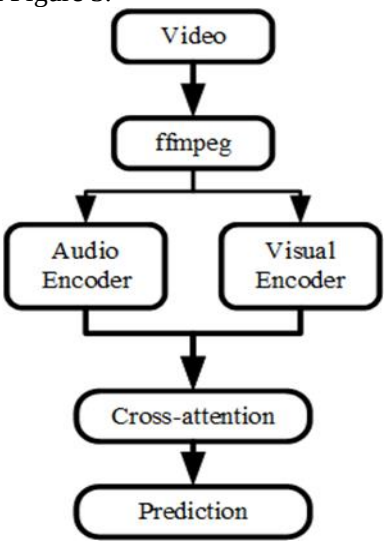
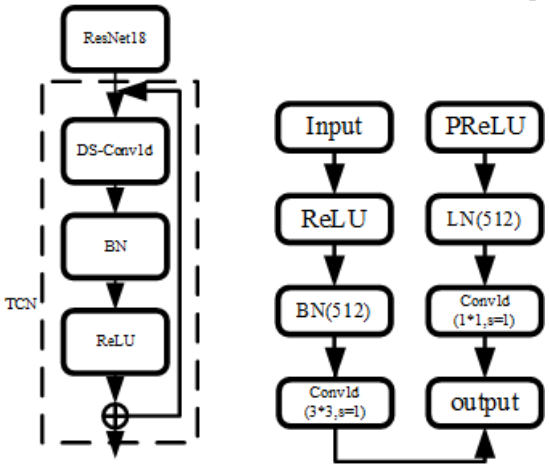


Fig.3 Schematic diagram of SdNet model structure

3.3. Feature Extraction Frontend

The first part of the video side is composed of ResNet18 residual neural network to extract the information of each frame, after which a video time-domain convolution block (TCN) is connected to extract the information between frames. The video time-domain convolution block contains five linear rectification functions (ReLU), a batch normalization layer (BN), a depth-separable convolution layer (DS-Conv1d) and a one-dimensional convolution layer (Conv1d), whose structure is shown in Figure 4. This structure is able to achieve a perceptual field of 840ms when the video frame rate is 25fps.



(a) Overall structure diagram
(b) DS-Conv1d
Fig.4 Visual Encoder Modules

where a parametric linear rectification unit PReLU is used to improve the fit of the model with a very small computational effort enhancement, here are the definition of PReLU:

$$PReLU(x_i) = f(x) = \begin{cases} x_i, & if\, x_i > 0 \\ a_i x_i, & if\, x_i \le 0 \end{cases} \quad (2)$$

i denotes different channels and a_i is a fixed value set in advance. Due to the inconsistent length of the videos identified each time, and also to reduce the computational effort during training, one of the BN layers is replaced with an LN layer, making it possible to achieve normalization without mini-batch and within a single piece of data, and the LN layer is represented as follows:

$$y = \frac{x - E[x]}{\sqrt{Var[x] + \in}} * \gamma + \beta \quad (3)$$

$E[x]$ is the mean, $Var[x]$ is the variance, γ is initially 1, and β is initially 0, both of which can vary with training. For the use of MFCC features, the sampling rate were set to 16000, the number of cepstrum returned to 13, the length of the analysis window to 0.025s, and a preset value of 0.01s for the interval between consecutive windows (the above parameters were set to default video at 25 frames/second). ResNet34 was chosen as the feature extraction network for audio, and the temporal resolution was matched with the video side by designing the null convolution kernel for the subsequent cross-

attention mechanism. When the MFCC window step is 10ms, the perceptual field can reach 1890ms.

3.4. Attentional Mechanisms

At the speaker detection backend, since the video and audio sides are not perfectly aligned in time and the actual audiovisual activity depends on the immediate speech content and the speaker's face movements, a structure like cross-attention from two temporal dimensions of audio-video is proposed. This structure is modified from the Scaled Dot-Product Attention module in Transformer, as shown in Figure 2.

F_a and F_v were used to describe the events associated with audio and video active speech activities, respectively. The cross-attention network is 2 Dot-Product Attention with the same structure, and the core part is shown in Figure 5.

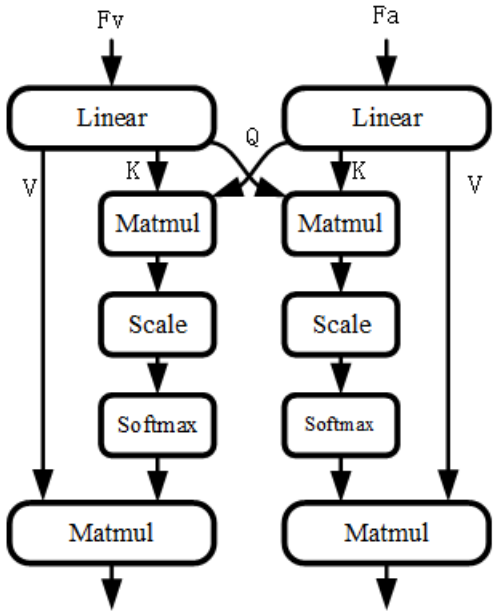


Fig.5 Cross-attention Modules

The cross-attention network is 2 structurally identical Dot-Product Attention, the core part of which is shown in Figure 3. Its inputs are audio and video features F_a and F_v, which are reduced to lower dimensional vectors X_a and X_v by a fully connected layer. multiplying X_a and X_v with parameter matrices W^q, W^k, W^v to get three vectors query(Q_a, Q_v), key(K_a, K_v) and value(V_a, V_v), respectively. Using audio embedding as the source and visual feature F_v as the target, the vector dot product of Q_a and K_v is performed, and the result obtained is used as the weight of V_v that affects the video feature. the function expression of the Attention process is as follows:

$$[Q, K, V] = X * [W^Q, W^K, W^V] \quad (4)$$

$$\text{attention}(\mathrm{Q}, \mathrm{K}, \mathrm{V}) = \text{Softmax}\left(\frac{\mathrm{Q}K^T}{\sqrt{d}}\right)\mathrm{V} \quad (5)$$

Similarly, audio features will also interact with video features $F_v \rightarrow a$. The attention layer is followed by a feedforward layer, and after these two layers, the entire cross-modal attention network is also generated using residual connectivity and layer normalization methods, with the outputs connected together along the temporal direction.

3.5. Loss Function

A fully-connected layer is connected at the end of the model to project the output of the cross-attention network to the ASD label sequence and compare it with the sequence of truth labels. Since it is a binary classification task, the Adam optimizer and the cross-entropy loss function are used to perform the classification with a loss function of:

$$L = -\frac{1}{T}\sum_{i=1}^{T}(y_i * \log s_i + (1 - y_i) * \log(1 - s_i)) \quad (6)$$

where s_i and y_i are the predicted and true value ASD labels of the i video frame, $i \in [1, T]$, and T is the number of video frames.

4. EXPERIMENTS

4.1. Dataset

The AVA-Speech[13] dataset released by Google in 2018 was selected, which is 5 times larger than the largest video dataset currently available and contains clips of videos from 192 movies on YouTube, allowing the training to have a broader diversity of contexts. Moreover, this dataset is very different from the recently released video datasets (iQIYI-VID-2019[14], etc.).

4.2. Analysis

The detail of Hard Ware Platform is shown in Table 1.

Table.1 Hard Ware Platform

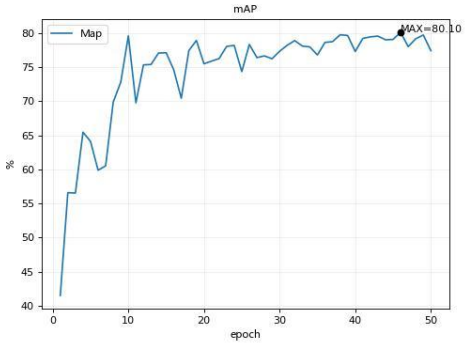
Name	Model
CPU	i5-11400F
GPU	RTX 3060(12GB)
RAM	16GB

The AVA-Speech dataset was selected as the training dataset, and the PyTorch framework was used to build the model. 21 video clips containing 416355 pictures and 4482 audio clips were selected, and the training, test and validation sets were divided in the ratio of 8:1:1, sample example is shown in Figure 6.

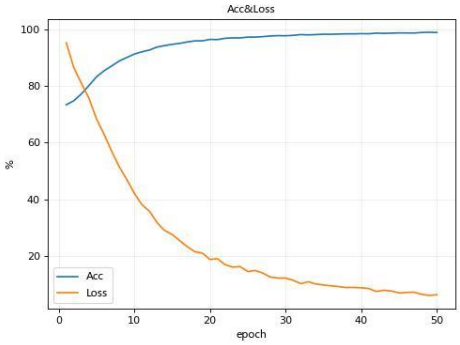


Fig.6 Example of AVA-Speech dataset sample

The input face image size is 128*128, the dimension of MFCC is 13, the initial learning rate is 0.0001, each round is reduced by 5%, the batchSize is set to 600, the training process is shown in Figure 7, and the model basically converges at the 50th round.



(a) mAP



(b) Acc and Loss

Fig.7 Visualization of the training process

Our model achieved a test score of 80.10% in AVA-Speech, and our results were compared with those of other teams trained using the same dataset, with higher scores than those teams as shown in Table 2.

Table.2 Comparison with other methods on AVA dataset

Method	mAP(%)
Arnia Software	43.886
Roth et al.[15,16]	79.2
SdNet	**80.1**

The results demonstrate that increasing the perceptual field at the audiovisual end is beneficial to improve the accuracy of the model in recognizing speaker detection, and introducing the cross-attention mechanism can improve the deep fusion of different modal features in the bimodal fusion model, and the model can have high robustness in video detection of different picture quality. The experimental results of the method in this paper are shown in Figure 8. The results of a movie clip on SdNet, with green boxes indicating people who are talking and red boxes indicating people who are not talking.



Fig.8 Detection effect picture

5. CONCLUSION

In this study, speaker detection is explored as a pre-processing for future tasks based on audiovisual information.In order to propose a promising solution for the challenge of audio-visual multi-task learning.An end-to-end bimodal recognition model was proposed to be designed, called SdNet, by introducing an attention mechanism and a parametric correction linear rectifier. Finally, SdNet achieves 80.10% mAP on the AVA-Speech dataset. The results demonstrate that increasing the perceptual field at the audiovisual end is beneficial to improve the accuracy of the model in recognizing speaker detection, and introducing the cross-attention mechanism can improve the deep fusion of different modal features in the bimodal fusion model, and the model can have high robustness in video detection of different picture quality. In this paper, a bimodal model SdNet is proposed to implement speaker detection by introducing a cross-attention mechanism. Compared with the previously published models, and the experimental results show that the method can improve the accuracy of speaker recognition.

On the feature extraction side, ResNet18 is used to extract information from picture frames and video time-domain convolution block (TCN) to extract information between frames to enhance the perceptual field as much as possible.ReLU was modified to PReLU. On the audio side, MFCC frames is used for training. Secondly, the extracted features are also fed into the cross-attention mechanism module to achieve information interaction between different modalities.

In the experiments,the recognition accuracy of the model still needs to be improved when the video images are rapidly switching or when there are multiple people talking in the scene. In the future work, a larger video datasets with content closer to natural conditions will be used for training, and to further improve the robustness of the model. Noise will be used to add in the image features and audio features. In addition, the backbone extraction network at the feature extraction end will be adjust to reduce the parameters during training, improve the training speed, and reduce the hardware requirements.

REFERENCES:

[1] Alcazar J, Heilbron F, Thabet A, et al. MAAS: Multi-modal Assignation for Active Speaker Detection[M]. 2021: 265-274.

[2] Son Chung J. Naver at ActivityNet Challenge 2019 -- Task B Active Speaker Detection (AVA), 2019: arXiv:1906.10555.

[3] Alcazar J L, Arbeláez P. Universidad de los Andes at ActivityNet Challenge 2020-Task B Active Speaker Detection (AVA)[C], 2020.

[4] Chakravarty P, Zegers J, Tuytelaars T, et al. Active speaker detection with audio-visual co-training[M].

2016: 312-316.
[5] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation, 2013: arXiv:1311.2524.
[6] Hochreiter S, Schmidhuber J. Long Short-term Memory[J]. Neural computation, 1997, 9: 1735-80.
[7] Alcazar J L, Caba Heilbron F, Mai L, et al. Active Speakers in Context, 2020: arXiv:2005.09812.
[8] Tao R, Pan Z, Das R K, et al. Is Someone Speaking? Exploring Long-term Temporal Features for Audio-visual Active Speaker Detection, 2021: arXiv:2107.06592.
[9] Li C, Ma X, Jiang B, et al. Deep Speaker: an End-to-End Neural Speaker Embedding System, 2017: arXiv:1705.02304.
[10] Dipietro R, Hager G D: Chapter 21 - Deep learning: RNNs and LSTM, Zhou S K, Rueckert D, Fichtinger G, editor, Handbook of Medical Image Computing and Computer Assisted Intervention: Academic Press, 2020: 503-519.
[11] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need, 2017: arXiv:1706.03762.
[12] Neimark D, Bar O, Zohar M, et al. Video Transformer Network, 2021: arXiv:2102.00719.
[13] Roth J, Chaudhuri S, Klejch O, et al. AVA-ActiveSpeaker: An Audio-Visual Dataset for Active Speaker Detection, 2019: arXiv:1901.01342.
[14] Liu Y, Peng B, Shi P, et al. iQIYI-VID: A Large Dataset for Multi-modal Person Identification, 2018: arXiv:1811.07548.
[15] Roth J, Chaudhuri S, Klejch O, et al. Ava Active Speaker: An Audio-Visual Dataset for Active Speaker Detection[C]. ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020: 4492-4496.
[16] León-Alcázar J, Caba Heilbron F, Thabet A, et al. MAAS: Multi-modal Assignation for Active Speaker Detection, 2021: arXiv:2101.03682.