北京大学
PEKING UNIVERSITY

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

Mutual Alignment between Audiovisual Features for End-to-End Audiovisual Speech Recognition

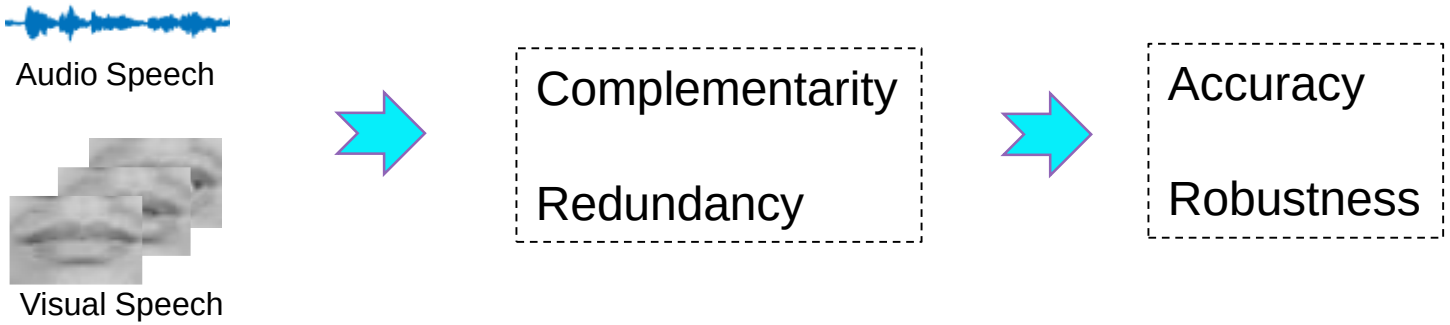
Hong Liu hongliu@pku.edu.cn

Yawei Wang Wongyawei@pku.edu.cn

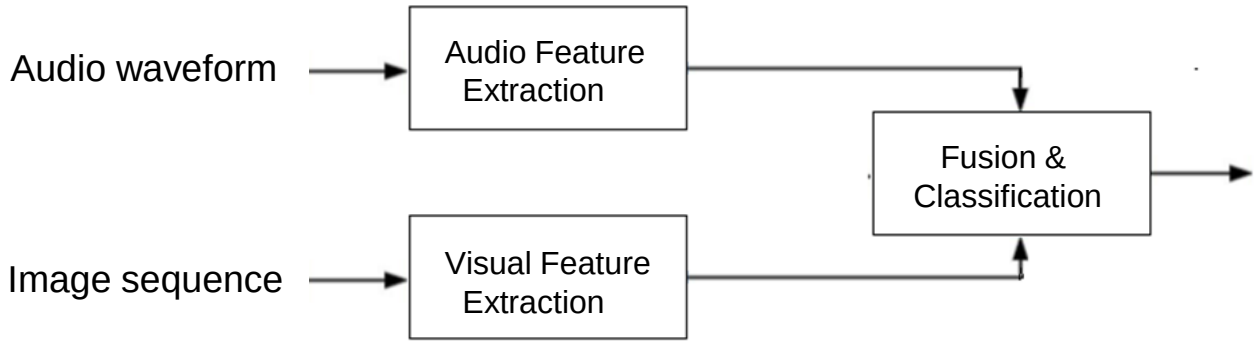
Bing Yang bingyang@sz.pku.edu.cn

Key Laboratory of Machine Perception, Shenzhen Graduate School, Peking University

- Audiovisual Speech Recognition

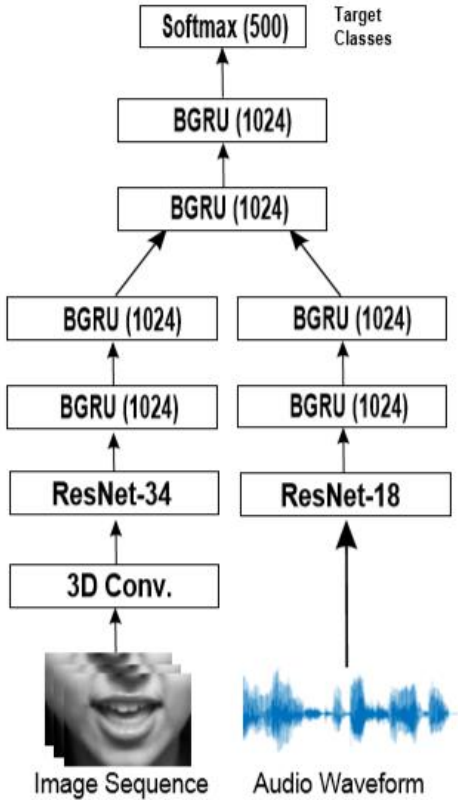


- Basic Framework



Asynchrony Modelling in Audiovisual Fusion

Concatenate Audio and Visual Features Directly



Overview of the end-to-end audiovisual speech recognition system[1]

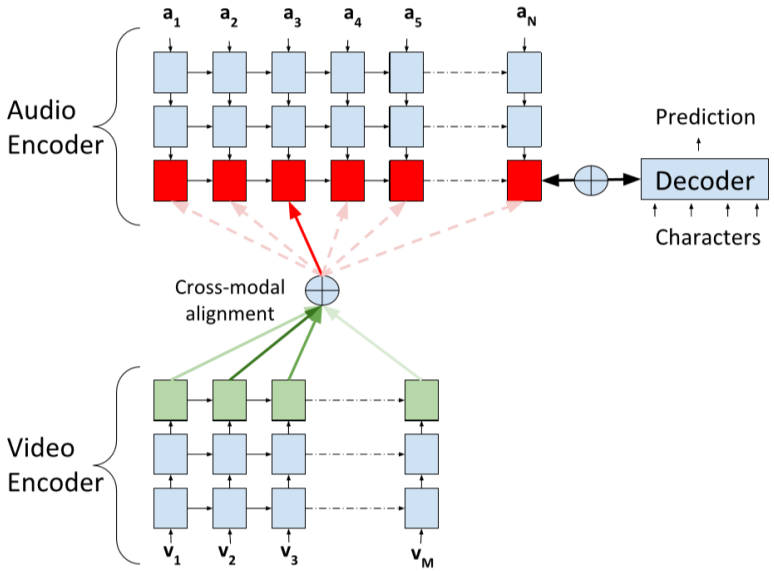
- Up sampling of video or down sampling of audio is performed to make the feature lengths identical as an alignment process

oversimplify the synchronization issue

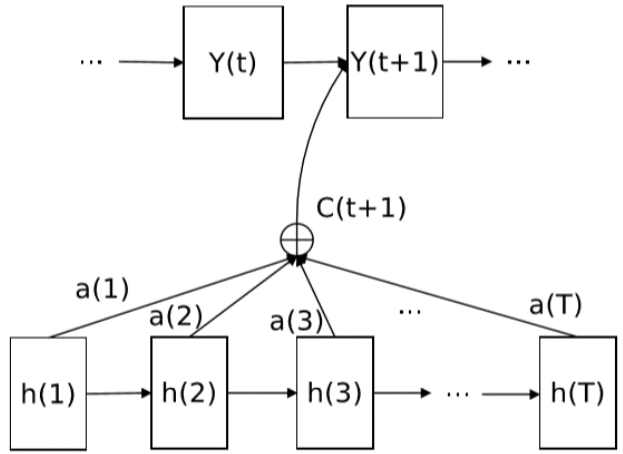
[1] Petridis S, Stafylakis T, Ma P, et al. End-to-end audiovisual speech recognition. *In ICASSP*, 2018: 6548-6552.

Asynchrony Modelling in Audiovisual Fusion

Conventional Additive Attention Mechanism



Overview of the AV Align[2] model



The dynamic part of the AliNN[3] model

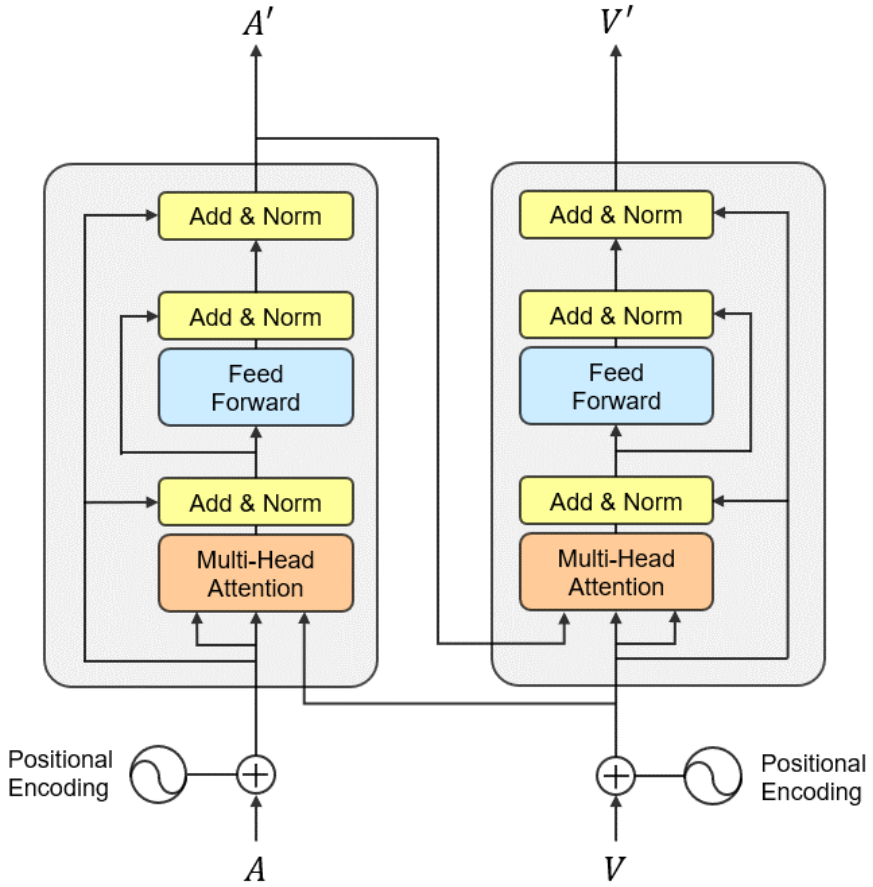
Overly rely on the audio modality

Do harm to performance under some noise conditions

[2] Sterpu G, Saam C, Harte N. Attention-based audio-visual fusion for robust automatic speech recognition. *In ICMI*, 2018: 111-115.

[3] Tao F, Busso C. Aligning audiovisual features for audiovisual speech recognition. *In ICME*, 2018: 1-6.

Mutual Attention



- Positional Encoding

$$PE_{(pos,2i)} = \sin\left(pos/10000^{2i/d_{\text{input}}}\right),$$

$$PE_{(pos,2i+1)} = \cos\left(pos/10000^{2i/d_{\text{input}}}\right)$$

- Scaled Dot-Product Attention

$$\text{Att}_i(Q,S) = \text{softmax}\left(\frac{QW_i^Q\left(SW_i^K\right)^\mathsf{T}}{\sqrt{d_k}}\right)SW_i^V$$

- Multi-Head Attention

$$\text{MultiHead}(Q,S) = \left[\text{Att}_1(Q,S),\ldots,\text{Att}_k(Q,S)\right]W^O$$

- Feed Forward Network

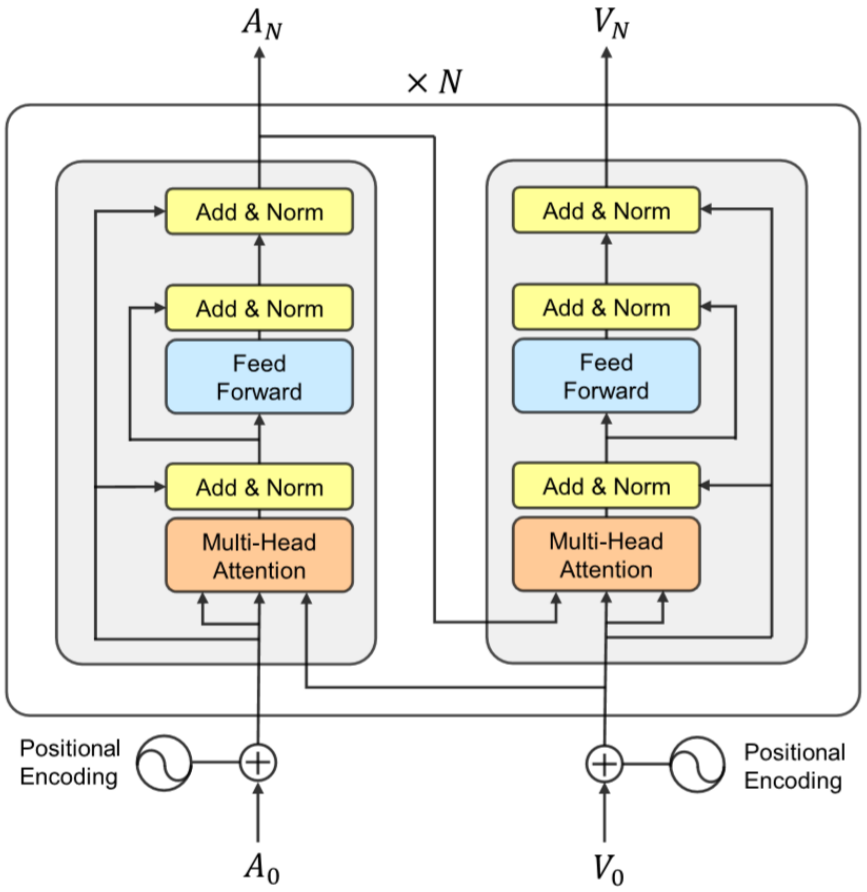
$$\text{FFN}(X) = \max\left(0, XW_1 + b_1\right)W_2 + b_2$$

- An Extra Shortcut Connection
- Mutual Attention

$$A' = \text{FFN}\left(\text{MultiHead}\left(V,A\right)\right),$$

$$V' = \text{FFN}\left(\text{MultiHead}\left(A',V\right)\right)$$

Mutual Iterative Attention



- First Round

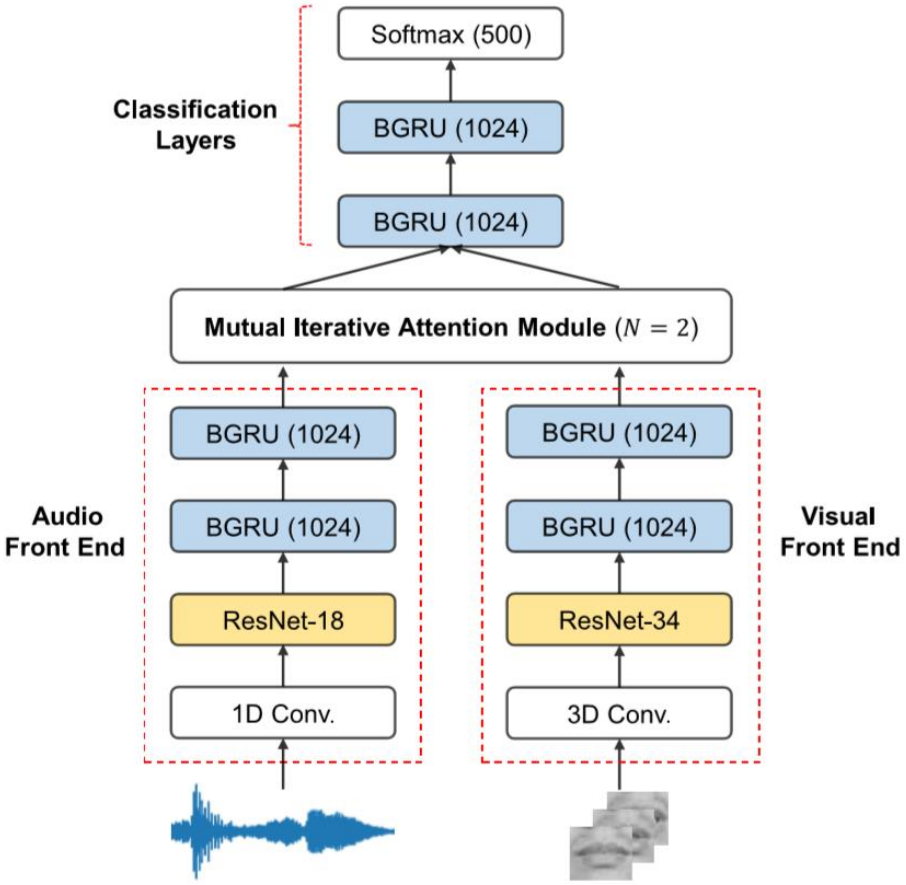
$$A_1 = \mathrm{FFN}\left(\mathrm{MultiHead}\left(V_0, A_0\right)\right),$$
$$V_1 = \mathrm{FFN}\left(\mathrm{MultiHead}\left(A_1, V_0\right)\right)$$

- After N Iterations

$$A_N = \mathrm{FFN}\left(\mathrm{MultiHead}\left(V_{N-1}, A_{N-1}\right)\right),$$
$$V_N = \mathrm{FFN}\left(\mathrm{MultiHead}\left(A_N, V_{N-1}\right)\right)$$

The parameters are shared in each iteration

Overall Framework



- Audio and Visual Front End
  - A convolutional layer
  - A residual network
  - A 2-layer BGRU
- Mutual Iterative Attention Module
  - A single iteration is not enough
  - To avoid an over-smoothing problem
- Classification Layers
  - A 2-layer BGRU
  - A softmax layer

- Lip Reading in the Wild (LRW) dataset

  - a publicly available audiovisual speech recognition dataset collected from in-the-wild videos.

  - 500 different words, each word contains 900~1100 utterances, spoken by over 1000 different speakers

  - More than 500000 speech instances, each video contains 29 frames

Some examples of LRW dataset

Experimental Results and Discussion

Table. Recognition performance in word classification rate [%] of various models on LRW dataset at different SNR levels.

Model	Word Classification Rate (%)							
	clean	20dB	15dB	10dB	5dB	0dB	-5dB	AVG
audio-only	96.74	96.68	96.48	95.85	94.07	88.07	68.90	90.97
visual-only	77.24	77.24	77.24	77.24	77.24	77.24	77.24	77.24
AV_baseline[1]	97.42	97.38	97.36	97.12	96.49	94.22	87.17	95.31
AV_MIA(Ours)	**97.55**	**97.54**	**97.48**	**97.27**	**96.82**	**94.92**	**89.32**	**95.84**

- The performance of audio-only drops significantly along with the descent of SNR
- The performance of visual-only maintains a constant over all noisy conditions
- AV_MIA achieves higher WCR in clean condition than AV_baseline although audio and visual features are still equally treated

Summary

- In this work, a mutual feature alignment method is proposed to address the asynchronization issue in audiovisual speech recognition
- We introduce Mutual Iterative Attention mechanism to align the audio and visual features by performing mutual attention over the two modalities iteratively to make full use of cross modality information
- Our proposed method outperforms the feature concatenation based AVSR system over all noisy conditions

Thanks! Q&A