# AUDIO-VISUAL SPEECH RECOGNITION USING A TWO-STEP FEATURE FUSION STRATEGY

Hong Liu, Wanlu Xu, Bing Yang
Key Laboratory of Machine Perception, Peking University, Shenzhen Graduate School, China
Email: hongliu@pku.edu.cn, xuwanlu@pku.edu.cn, bingyang@sz.pku.edu.cn

## Ⅰ INTRODUCTION

- Audio-visual speech recognition (AVSR) aims at combining visual information with the audio information to effectively improve the recognition accuracy in noisy environment.
- Most approaches involve two separate audio and visual streams with early or late fusion strategies. Such a single-stage fusion method may fail to guarantee the integrity and representativeness of fusion information simultaneously.
- This paper extends a traditional single-stage fusion network to a two-step feature fusion network by adding an audio-visual early feature fusion (AVEFF) stream to the baseline model.

## Ⅱ MOTIVATION

- The visual information is particularly important when the audio information is contaminated severely in a noisy environment. most approaches usually extract the spatio-temporal feature of video sequences by local convolutional operation, which may lose some information between distant frames.

How to capture the long-range dependencies of sequential data?

- The way to fuse the visual and audio information is another point of audio-visual speech recognition task. most methods only consider the audio-visual fusion in a single stage of the network, which may not be able to balance the integrity and representativeness of audio and visual information.

How to design a fusion method to better integrate the two features?

## Ⅲ CONTRIBUTIONS

- A non-local block is inserted in the feature extraction part of the visual stream (NL-Visual) to capture long-range dependencies by calculating the distance of all positions.
- An audio-visual early feature fusion (AV-EFF) stream is added to form a two-step feature fusion strategy that can guarantee integrity and representativeness of features simultaneously.
- The experimental results show that our method can improve the fusion performance in strong noise environment greatly.

## Ⅳ THE PROPOSED METHOD

### NL-Visual Stream

the output of the 3D CNN is given by:

$$v_{ij}^{xyz} = tanh\left(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} \sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right),$$

the output of non-local block is:

$$Out_{nl} = W_z \frac{1}{\sum_{\forall j} f(x_i, x_j)} \sum_{\forall j} e^{W_\theta v_i^T W_\phi v_j} W_g v_j + v_i.$$

visual features can be formulated as:

$$V_{early} = ResNet34(Out_{nl}),$$
$$V_{late} = BGRU(V_{early}),$$

### Audio Stream

take the fast fourier transform (FFT) of x(n) to get the linear spectrum x(k):

$$x(k) = \sum_{n=0}^{N-1} x(n) e^{\frac{j2\pi nk}{N}}, \quad 0 \le n, K \le n-1.$$

the output of the 1D CNN can be expressed as:

$$a_{ij}^{x} = tanh\left(b_{ij} + \sum_m \sum_{p=0}^{P_i-1} w_{ijm}^{p} a_{(i-1)m}^{(x+p)}\right),$$

audio features can be represented as:

$$A_{early} = ResNet18(a),$$
$$A_{late} = BGRU(A_{early}),$$

### Audio-Visual Early Feature Fusion Stream

late audio-visual feature is obtained:

$$F_{early} = Concat(V_{early}, A_{early}),$$
$$F_{late} = BGRU(F_{early}),$$

### Audio-Visual Late Feature Fusion

the feature obtained here is: $F_{final} = BGRU(Concat(V_{late}, A_{late}, F_{late}))$,

the final fusion classification result can be obtained:

$$L_{final} = \arg\max(softmax(F_{final}))$$
$$= \arg\max_{j \in 1,\dots,K}\left(\frac{e^{F_{final}^{j}}}{\sum_{k=1}^{K} e^{F_{final}^{k}}}\right),$$

## Ⅴ THE PROPOSED METHOD

- ***Datasets***: LRW dataset, LRW-1000 dataset.

### Comparisons with the state-of-the-art methods

| Task | Method | LRW Accuracy(%) | LRW1000 Accuracy(%) |
|---|---|---|---|
| Lip-reading | LSTM-5 [30] | 71.50 | 25.76 |
| | D3D [31] | 78.02 | 34.76 |
| | 3D+2D [21] | 83.00 | **38.19** |
| | Multi-Grained [32] | 83.34 | 36.91 |
| | ResNet34+BGRU(Baseline) [8] | 82.80 | 36.72 |
| | NL-Visual(Ours) | **83.41** | 37.03 |
| AVSR (clean) | MCNN [33] | 96.98 | 39.60 |
| | ETE-AVSR(Baseline) [8] | 97.60 | 37.52 |
| | Two-Step(Ours) | **98.26** | **41.57** |

### Ablation study

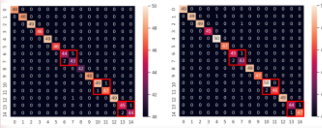
| Baseline [8] | NL-Visual | AV-EFF | -5 | 0 | 5 | 10 | 15 | 20 | clean |
|---|---|---|---|---|---|---|---|---|---|
| ✓ | | | 86.66 | 94.13 | 96.29 | 96.70 | 97.00 | 97.50 | 97.90 |
| ✓ | ✓ | | 88.21 | 95.01 | 97.18 | 97.22 | 97.53 | 97.86 | 98.10 |
| ✓ | | ✓ | 90.65 | 95.56 | 97.28 | 97.74 | 98.04 | 98.08 | 98.14 |
| ✓ | ✓ | ✓ | **92.10** | **96.19** | **97.35** | **97.86** | **98.08** | **98.15** | **98.26** |

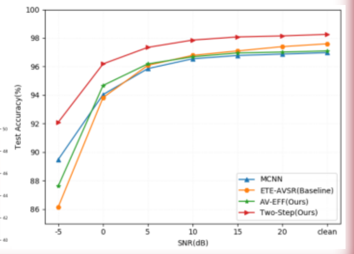
### Evaluation of two-step feature fusion method

| Modality | Method | -5 | 0 | 5 | 10 | 15 | 20 | clean |
|---|---|---|---|---|---|---|---|---|
| Single | Audio only | 71.60 | 90.55 | 95.34 | 96.89 | 97.32 | 97.58 | 97.70 |
| | Visual only | 83.41 | 83.41 | 83.41 | 83.41 | 83.41 | 83.41 | 83.41 |
| | AV-EFF only | 87.63 | 94.68 | 96.19 | 96.69 | 96.96 | 97.02 | 97.10 |
| Fusion | Two-step(Ours) | **92.10** | **96.19** | **97.35** | **97.86** | **98.08** | **98.15** | **98.26** |

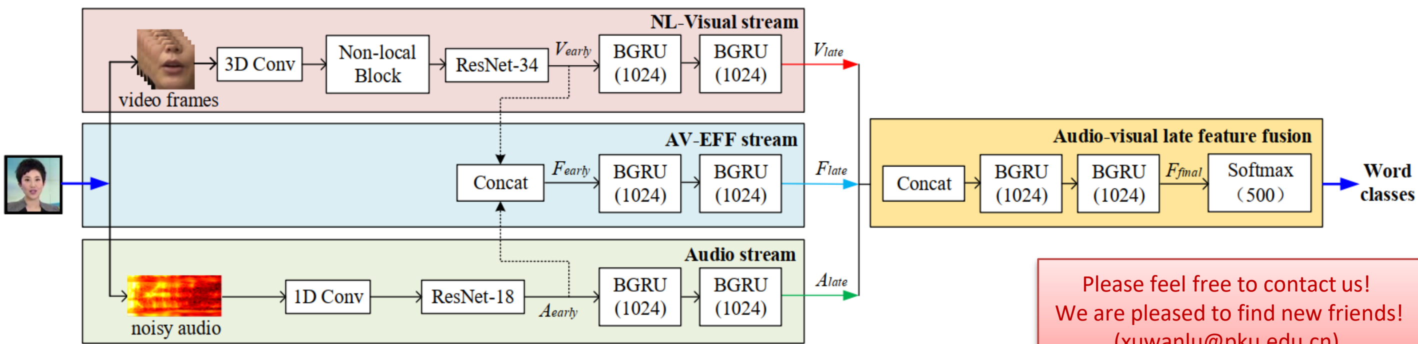
### Visualization

1. Confusion matrices of baseline model and our two-step feature fusion network at -5dB SNR.
2. Classification accuracy of different fusion methods under different SNR.






## Ⅵ PIPELINE OF PROPOSED TWO-STEP FEATURE FUSION NETWORK




Please feel free to contact us!
We are pleased to find new friends!
(xuwanlu@pku.edu.cn)