Robust Audio-Visual Speech Recognition Based on Hybrid Fusion

ICPR 2020

Hong Liu, Wenhao Li, Bing Yang
Key Laboratory of Machine Perception
Shenzhen Graduate School, Peking University

Presenter: Wenhao Li

Advisor: Prof. Hong Liu

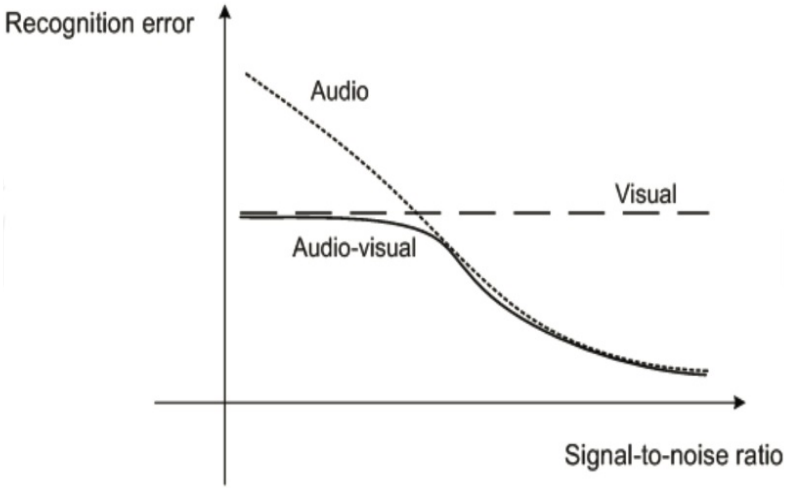
Date: 2020.12

Introduction

- **Audio-Visual Speech Recognition (AVSR)**
  - The approach of AVSR system is to leverage the extracted information from one modality to improve the recognition accuracy of the other modality by complementing the missing information.
  - Visual information, which is not affected by acoustic noise, can significantly improve the performance of speech recognition in noisy environments.

Traditional speech recognition device

Affected by acoustic noise

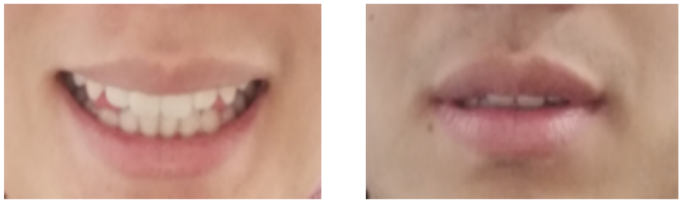


AVSR recognition error

Improve the noise-robustness

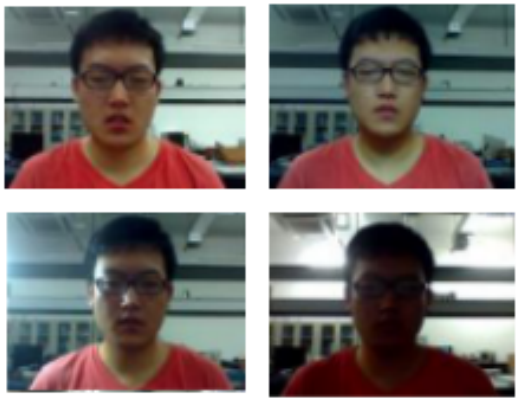
Introduction

- **Challenges**
  - Intra-Class variations
  - Inter-Class similarities
  - Various lighting
  - Complex backgrounds



Two different individuals pronounce "you"

- **Major Challenges**
  - Fuse the two modalities more effectively and robustly due to the intrinsically ambiguous nature of homophones, especially in noisy environments

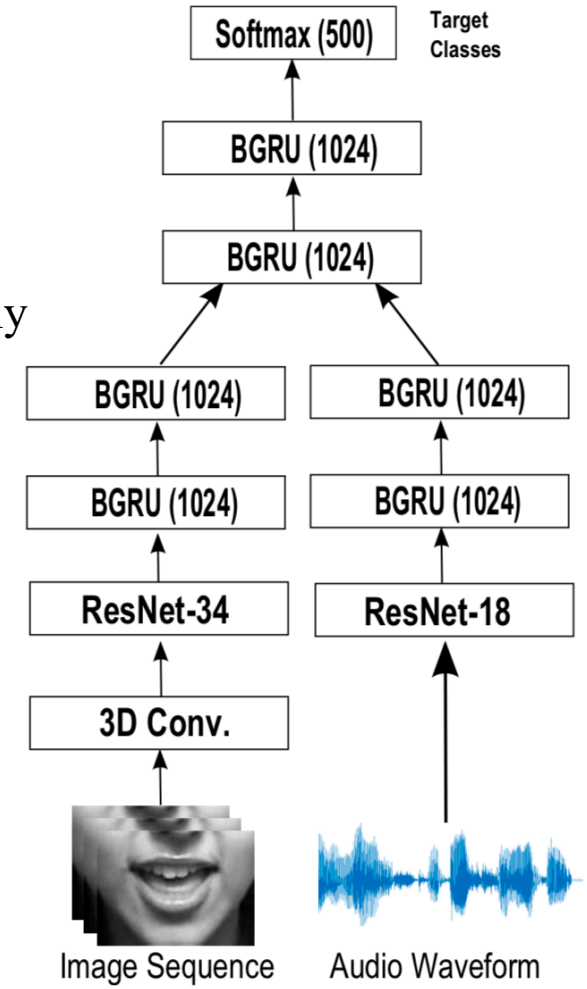


Various lighting

Complex backgrounds

Introduction

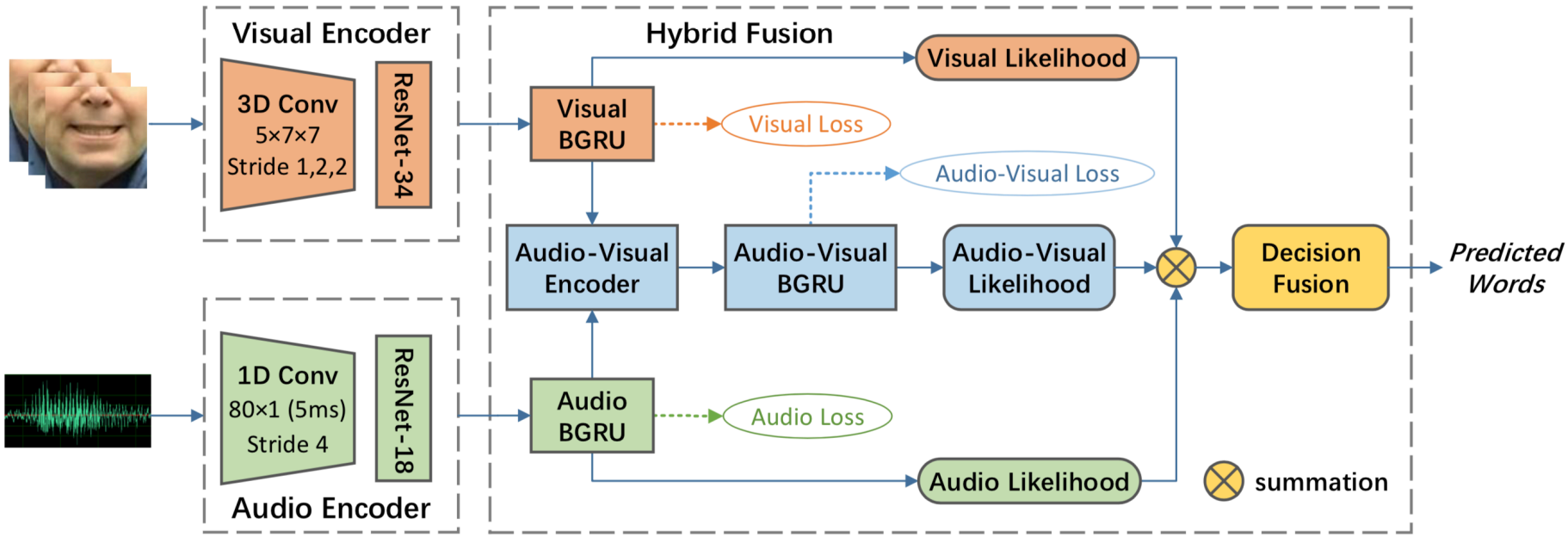
- **Motivation**
  - **Baseline**
    - Only use feature fusion
    - Audio and visual features are concatenated directly
    - Only use one audio-visual loss
  - **Solution**
    - Introduce decision fusion
    - Introduce audio-visual encoder
    - Introduce combined loss



End-to-end audiovisual speech recognition, ICASSP 2018.

Proposed Method

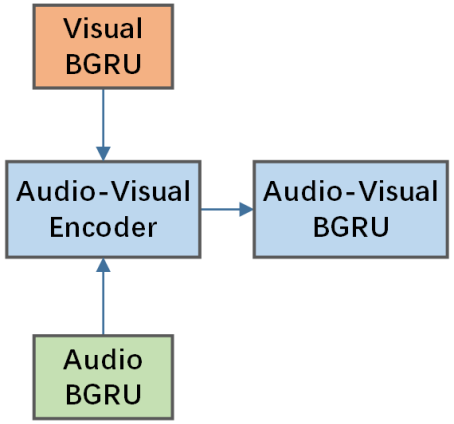
- **Overview of the Proposed Method**



Overview of the proposed end-to-end audio-visual speech recognition method based on hybrid fusion

Proposed Method

- **Audio-Visual Hybrid Fusion Network**
  - **Audio-Visual feature fusion**
    - **Audio-Visual encoder**
      - A simple yet effective and light-weight, two fully-connected layers, residual connection
      - Map audio and visual features into a shared latent space to capture more discriminative multi-modal feature
      - Learn more crucial information and find the internal correlation between spatial-temporal information from audio and visual networks effectively
    - **Audio-Visual BGRU**
      - 2-layer BGRU: consist of 1024 cells in each layer
      - Fuse the information from the two modalities



Audio-Visual Feature Fusion

Proposed Method

- **Audio-Visual Hybrid Fusion Network**
  - **Audio-Visual decision fusion**
    - The fusion likelihood

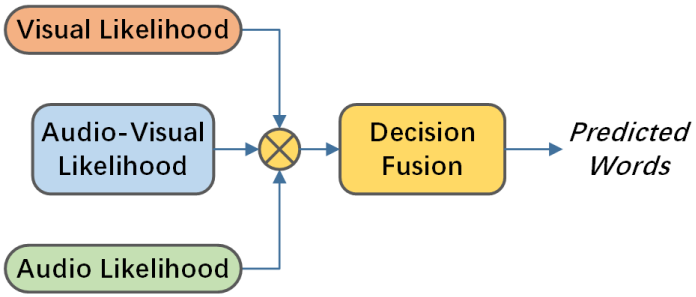
$$P\left(c_i|x\right) = \alpha P\left(c_i^a|x^a\right) + \beta P\left(c_i^v|x^v\right) + \gamma P\left(c_i^{av}|x^{av}\right)$$

    - Satisfy the constraints

$$\alpha + \beta + \gamma = 1, 0 \leq \alpha, \beta, \gamma \leq 1$$

    - The class label

$$z = \arg\max_i \left\{P\left(c_i|x\right)\right\}$$



Audio-Visual Decision Fusion

Proposed Method

- **Audio-Visual Hybrid Fusion Network**
  - **Audio-Visual training**
    - Cross-entropy loss for each network

$$L_k = -\sum_{i=1}^{C} y_i \log P\left(c_i^k | x^k\right)$$

      - k ∈ {a,v,av}: audio, visual and audio-visual stream
      - y: the true class label of each sequence
      - C: the number of target isolated words

    - Total objective

$$L = \lambda_a L_a + \lambda_v L_v + \lambda_{av} L_{av}$$

      - λ : the weighting factor
      - L: the combined loss of the audio-visual network

The model focuses on various modalities and shows its noise-robustness in learning the joint representation across audio- visual modalities.

Experiments

- **Datasets**
  - Lip Reading in the Wild (LRW)
    - The largest publicly available dataset for audio-visual speech recognition task, more than 1000 speakers, 500 different isolated words, such as ”YOUNG”, ”SOCIAL” and ”UNITED”
    - Short video clips with 29 frames (1.16 seconds) from BBC television
- **Experimental Setting**
  - Babble and white noises from the Noisex92 dataset, SNRs of 20 dB, 15 dB, 10 dB, 5 dB, 0 dB, -5dB and -10 dB to the traing and test sets.

Top: Example of video frames from Lip Reading in the Wild dataset. Bottom: Mouth ROI sequences for ‘about’ from two different speakers.

Experiments

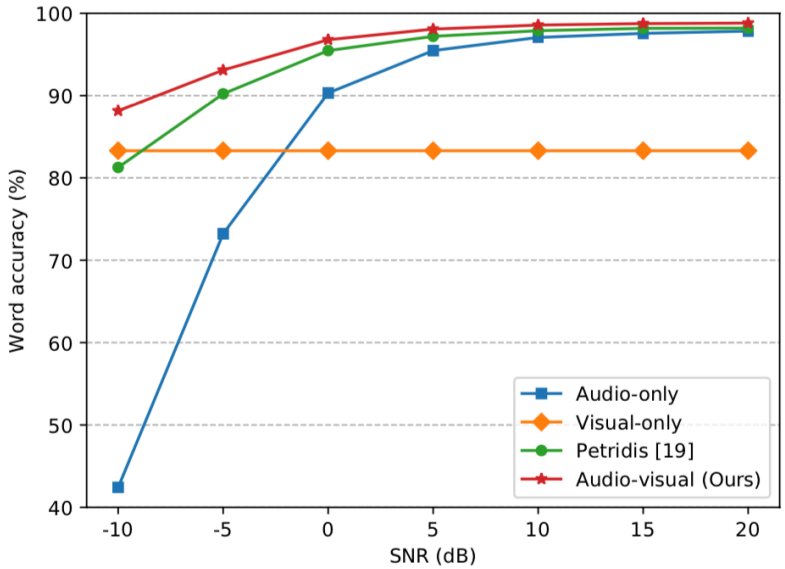
- **Comparison with the State-of-the-art**

Modalities	Methods	Word accuracy (%)
Audio	Petridis *et al.* [19]	97.70
	Stafylakisa *et al.* [25]	97.96
Visual	Chung *et al.* [20]	61.10
	Chung *et al.* [12]	76.20
	Petridis *et al.* [19]	82.00
	Stafylakis *et al.* [23]	83.00
	Wang *et al.* [26]	83.34
	Zhao *et al.* [27]	84.41
Audio-Visual	Petridis *et al.* [19]	98.20
	Ours	**98.91**

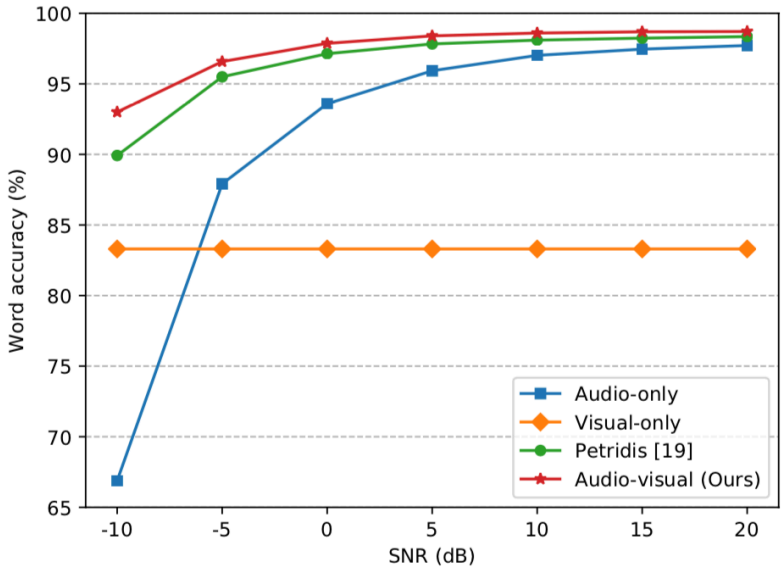
Word accuracy (%) of the audio-only, visual-only and audio-visual models on the LRW dataset in clean audio condition.

Experiments

- **Comparison with the State-of-the-art**



(a) Babble noise



(b) White noise

Comparisons of the word accuracy (%) with the state-of-the-art method under different levels of babble noise and white noise on LRW dataset.

Experiments

Ablation Studies

- w/ AV encoder: insert an audio-visual encoder into base- line.
- w/ combined loss: insert a combined loss into baseline.
- w/ AV decision: insert a decision fusion module into baseline.
- w/ A+V decision: insert a decision fusion strategy into our audio and visual networks.

SNR(dB)	-10	-5	0	5	10	15	20
Baseline [19]	81.27	90.22	95.47	97.20	97.89	98.17	98.20
w/ combined loss	86.08	92.58	96.29	97.50	97.97	98.23	98.29
w/ AV encoder	85.05	91.90	96.34	97.81	98.35	98.52	98.62
w/ AV decision	85.91	92.09	96.32	97.80	98.34	98.55	98.64
w/ A+V decision [24]	84.69	89.48	94.64	96.72	97.59	97.98	98.15
Ours	**88.15**	**93.10**	**96.79**	**98.08**	**98.57**	**98.75**	**98.81**

Ablation studies on different components of our audio-visual model at varying SNR levels of babble noise. Word accuracy (%) on the LRW dataset.

Conclusions

- We present a novel hybrid fusion based method for AVSR to address the challenge of inherent ambiguity, which is able to distinguish words with similar pronunciations and becomes robust to various noisy conditions.
- A simple yet effective audio-visual encoder is proposed to capture more discriminative multi-modal feature from both modalities; A decision fusion module is designed in order to complementarily utilize the reliability measures of audio and visual information; A combined loss to make the model learn the joint representation across audio-visual modalities robustly.
- Experiments on LRW dataset demonstrate that our method achieves superior performance compared to other state-of-the-art methods in both clean and noisy conditions.

Thank You! Q&A

Open Lab on Robot Vision System, Shenzhen Graduate School, Peking University