End-to-End Audio-Visual Speech Recognition with Conformers

Pingchuan Ma, Stavros Petridis, Maja Pantic

Imperial College London, UK

Motivation

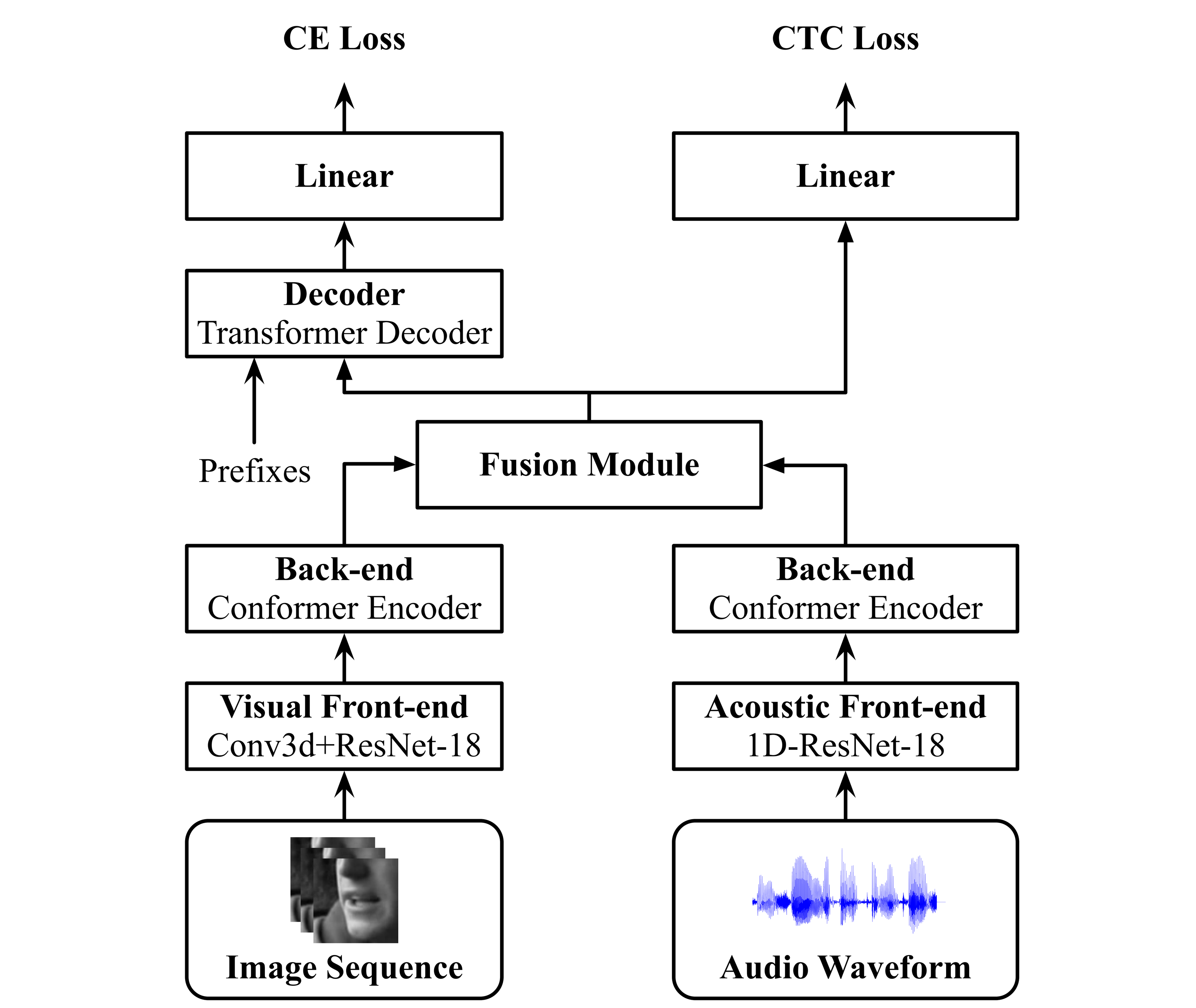
- Audio-Visual Speech Recognition is useful in noisy environments where the audio signal is corrupted.
- Limitations
  - It is common in the literature to have a two-step approach where they first extract visual/audio features and then do recognition.

Contributions

CTC/Attention[3]	Ours
Hand-crafted acoustic features	**Raw** audio waveforms
LSTM encoder and decoder	**Conformer** encoder and **Transformer** decoder
Two-step approach	Joint **end-to-end training**
RNN-based Language Model	**Transformer**-based Language Model

Methodology

Audio-Visual Framework



Experiments

Datasets

- LRS2: 144 482 video clips from BBC programs (224.1 hours)
- LRS3: 151 819 video clips from TED talks (438.9 hours)

Setup

- **Data augmentation**
  - Visual Stream: Horizontal Flipping, Random Cropping
  - Audio Stream: Additive Noise, Time Mask, Frequency Mask
- **Network settings** ($e=12$, $d^{\text{ff}}=2048$, $d^{\text{k}}=256$, $d^{\text{v}}=256$), where e denotes the number of conformer blocks, d^{ff} denotes the dimension of linear layer in the feed-forward module, d_k and d_v are the dimensions for queries/keys and values, respectively.
- **Experimental settings** We train the model for **50** epochs. The learning rate increases linearly with the first **25 000** steps, yielding a peak learning rate of **0.0004** and thereafter decreases proportionally to the inverse square root of the step number.
- **Language model corpus** the training transcriptions of LibriSpeech (960 h), pre-training and training sets of LRS2 and LRS3, with a total of **16.2** million words.

References

[1] T. Afouras et al. "Deep audio-visual speech recognition". In: *IEEE PAMI* (2018). DOI: 10.1109/TPAMI.2018.2889052.

[2] T. Makino et al. "Recurrent neural network transducer for audio-visual speech recognition". In: *ASRU*. 2019, pp. 905–912. DOI: 10.1109/ASRU46091.2019.9004036.

[3] S. Petridis et al. "Audio-Visual Speech Recognition with a Hybrid CTC/Attention Architecture". In: *SLT*. 2018, pp. 513–520. DOI: 10.1109/SLT.2018.8639643.

[4] S. Petridis et al. "End-to-End Audiovisual Speech Recognition". In: *ICASSP*. 2018, pp. 6548–6552. DOI: 10.1109/ICASSP.2018.8461326.

[5] J. Yu et al. "Audio-Visual Recognition of Overlapped Speech for the LRS2 Dataset". In: *ICASSP*. 2020, pp. 6984–6988. DOI: 10.1109/ICASSP40776.2020.9054127.

Results

Ablation Study

Method	WER
Baseline [4]	63.5
+ E2E	50.9
+ LRW pre-training	46.2
+ Conformer encoder	42.4
+ Transformer LM	37.9

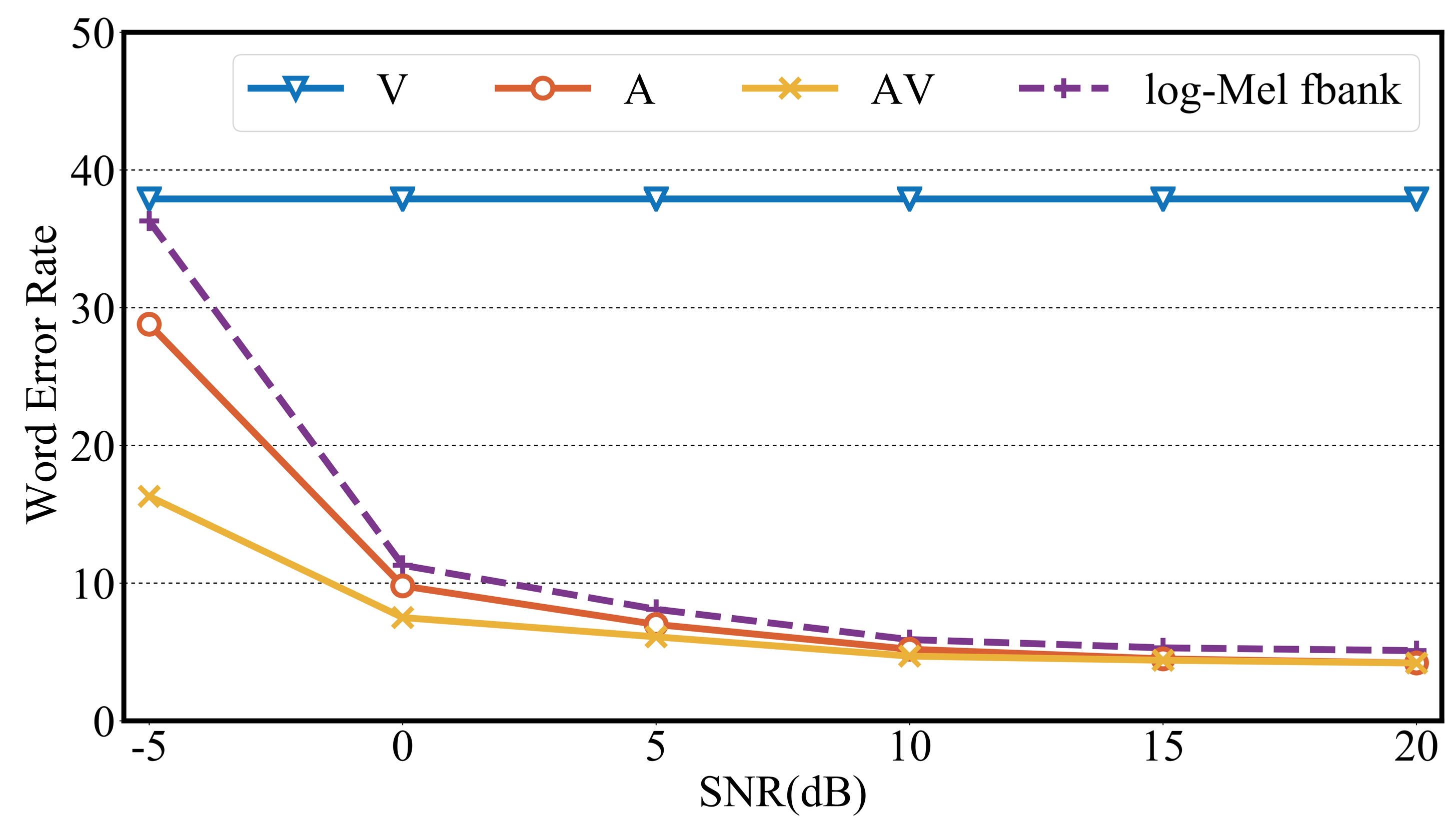
Performance on LRS2

Method	Training Data (Hours)	WER
Visual-only (↓)		
TDNN [5]	LRS2 (224)	48.9
TM-seq2seq [1]	MVLRS (730) + LRS2&3$^{v0.4}$ (632)	48.3
Ours (V)	LRS2 (224)	**39.1**
Ours (V)	LRW (157) + LRS2 (224)	**37.9**
Audio-only (↓)		
TM-seq2seq [1]	MVLRS (730) + LRS2&3$^{v0.4}$ (632)	9.7
TDNN [5]	LRS2 (224)	6.7
Ours (filter-bank)	LRS2 (224)	**4.3**
Ours (raw A)	LRS2 (224)	**4.3**
Ours (raw A)	LRW (157) + LRS2 (224)	**3.9**
Audio-visual (↓)		
TM-seq2seq [1]	MVLRS (730) + LRS2&3$^{v0.4}$ (632)	8.5
TDNN [5]	LRS2 (224)	5.9
Ours (raw A + V)	LRS2 (224)	**4.2**
Ours (raw A + V)	LRW (157) + LRS2 (224)	**3.7**

Performance on LRS3$^{v0.4}$

Method	Training Data (Hours)	WER
Visual-only (↓)		
TM-seq2seq [1]	MVLRS (730) + LRS2&3$^{v0.4}$ (632)	58.9
RNN-T [2]	YT (31 000)	33.6
Ours (V)	LRS3$^{v0.4}$ (438)	**46.9**
Ours (V)	LRW (157) + LRS3$^{v0.4}$ (438)	**43.3**
Audio-only (↓)		
TM-seq2seq [1]	MVLRS (730) + LRS2&3$^{v0.4}$ (632)	8.3
RNN-T [2]	YT (31 000)	4.8
Ours (filter-bank)	LRS3$^{v0.4}$ (438)	**2.3**
Ours (raw A)	LRS3$^{v0.4}$ (438)	**2.3**
Ours (raw A)	LRW (157) +LRS3$^{v0.4}$ (438)	**2.3**
Audio-visual (↓)		
TM-seq2seq [1]	MVLRS (730) + LRS2&3$^{v0.4}$ (632)	7.2
RNN-T [2]	YT (31 000)	4.5
Ours (raw A + V)	LRW (157) + LRS3$^{v0.4}$ (438)	**2.3**

Performance in a noisy scenario on LRS2



Conclusions

- We present end-to-end speech recognition models that improve audio-only, visual-only and audio-visual performance on LRS2 and LRS3$^{v0.4}$.
- On LRS3$^{v0.4}$, our audio-visual model is trained on a dataset which is 52× smaller than the state-of-the-art audio-visual model, 595 vs 31000 hours.
- We propose a convolutional neural network based backbone for acoustic modeling, showing that deep speech representations are more robust to audio noise than log-Mel filter-bank features.