RUB

Fusing Information Streams in End-to-End Audio-Visual Speech Recognition

Wentao Yu, Steffen Zeiler, Dorothea Kolossa

April 19, 2021

Cognitive Signal Processing Group
Institute of Communication Acoustics

DFG
Deutsche Forschungsgemeinschaft

Introduction

End-to-end automatic speech recognition (E2E ASR)

+ *simpler* structure than conventional hybrid ASR

- most E2E models *do not* use an explicit language model
- more likely to *overfit* [1]

Addressing both disadvantages: end-to-end speech processing toolkit ESPnet [2]

[1] A. Vaswani et al., "Attention is all you need," in *Advances in neural information processing systems*, pp. 5998-6008, 2017.

[2] S. Watanabe et al., "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech*, pp. 2207-2211, 2018.

Introduction

- Audio-visual speech recognition (AVSR): can significantly improve word error rates (WERs), but in noisy conditions, WERs are still clearly diminished [3] [4]
- Our contribution:
  - Combine the advantages of the end-to-end model and AVSR.
  - Additional reliability indicators help the integration of acoustic and visual information.

[3] H. Meutzner et al., "Improving audio-visual speech recognition using deep neural networks with dynamic stream reliability estimates," in *Proc. ICASSP*, pp. 5320-5324, IEEE, 2017.

[4] T. Afouras et al., "Deep audio-visual speech recognition," *arXiv:1809.02108*, 2018.

Introduction

Proposed structure: sequence-to-sequence (S2S) transformer model with connectionist temporal classification—TM-CTC, shows high performance in many tasks [5] [6]

- Training:

$$L = \alpha \cdot \log p_{ctc}(\mathbf{s}|\mathbf{o}) + (1-\alpha) \cdot \log p_{s2s}(\mathbf{s}|\mathbf{o})$$

- Decoding:

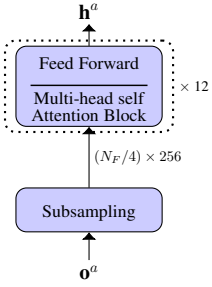
$$\log p^*(\mathbf{s}|\mathbf{o}) = \alpha \cdot \log p_{ctc}(\mathbf{s}|\mathbf{o}) + (1-\alpha) \cdot \log p_{s2s}(\mathbf{s}|\mathbf{o}) + \theta \cdot \log p_{LM}(\mathbf{s})$$

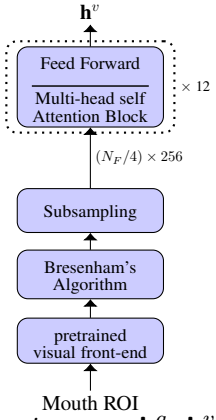
[5] T. Afouras et al., "Deep audio-visual speech recognition," *arXiv:1809.02108*, 2018.

[6] A. Baevskiet al., "wav2vec 2.0: A framework for self-supervised learning of speech representations," *arXiv preprint arXiv:2006.11477*, 2020.

Audio-visual TM-CTC model

Encoder:

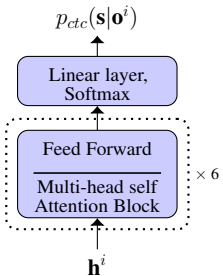


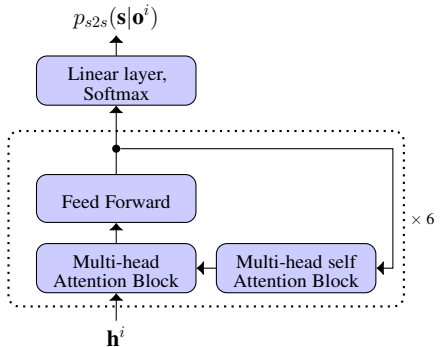


Audio encoder (left), video encoder (right) for both modalities $i \in a, v$. $\mathbf{h}^a$, $\mathbf{h}^v$ are in the same length $N_F/4$, where N_F is the number of frames

Audio-visual TM-CTC model

Decoder:

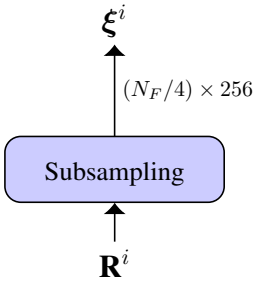




CTC decoder (left) and Transformer decoder (right) for both modalities $i \in a, v$

Reliability measures

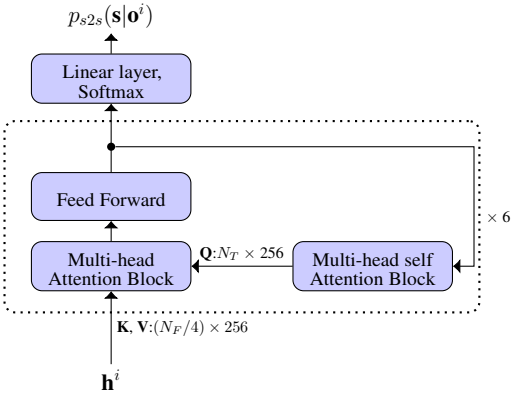
- acoustic ($\mathbf{R}^a$): MFCCs, SNR, f_0, Δf_0, Probability of voicing
- visual ($\mathbf{R}^v$): Face detector confidence (OpenFace [7]) and Action Units (AU12, AU15, AU17, AU23, AU25, AU26 [8])
- Reliability measures encoder: without self-attention block. $\boldsymbol{\xi}^i$ are in the length $N_F/4$.



[7] T. Baltrušaitis et al., "Openface: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision*, pp. 1-10, IEEE, 2016.

[8] G. Sterpu et al., "How to teach DNNs to pay attention to the visual modality in speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1052-1064, 2020.

Re-alignment in the decoder



$$\mathbf{T}_j = \text{softmax}\left(\frac{\left(\mathbf{W}_j^Q\mathbf{Q}^T\right)^T\left(\mathbf{W}_j^K\mathbf{K}^T\right)}{\sqrt{d_k}}\right)$$

$$\boldsymbol{\alpha}_j = \text{attention}_j(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathbf{T}_j\left(\mathbf{W}_j^V\mathbf{V}^T\right)^T$$

j: attention head index, $j \in \{0,\cdots,J\}$

$$\boldsymbol{\alpha} = \text{fully-connect}(\boldsymbol{\alpha}_0,\boldsymbol{\alpha}_1,\cdots,\boldsymbol{\alpha}_J)$$

Re-alignment in decoder

- Problems:
  - Reliability measures $\boldsymbol{\xi}^i$ have length $N_F/4$
  - Reliability measures $\boldsymbol{\xi}^i$ do not contain speech information. We should not use attention mechanism to re-align them.
  - For decision fusion, the token-by-token log-posteriors $p_{s2s}(\mathbf{s}|\mathbf{o}^i)$ have length N_T, where N_T is the length of tokens, of the transcription.
- Solution: re-use the attention transform matrix of each head for re-alignment.

$$\tilde{\boldsymbol{\xi}}^i_j = \mathbf{T}^i_j \cdot \left(\mathbf{W}^{i\xi}_j (\boldsymbol{\xi}^i)^T\right)^T,$$

$$\tilde{\boldsymbol{\xi}}^i = \text{fully-connect}(\tilde{\boldsymbol{\xi}}^i_0, \tilde{\boldsymbol{\xi}}^i_1, \cdots, \tilde{\boldsymbol{\xi}}^i_J)$$

Fusion strategy: Decision Fusion Network

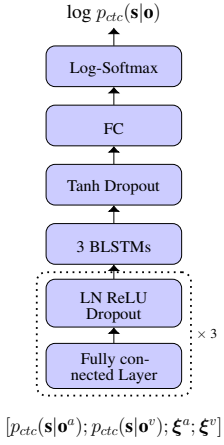


Figure 1 CTC Decision Fusion Network

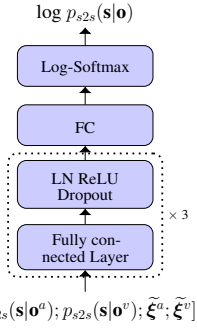


Figure 2 S2S Decision Fusion Network

Experimental Setup

- ESPnet is trained on LRS2 corpus and LRS3 pre-train set
- Features:
  - acoustic: 80 log Mel features with pitch, delta pitch, probability of voicing. 25 ms frame size and 10 ms frameshift.
  - visual: 96×96 pixel grayscale mouth region of interest (ROI) detected via OpenFace at 25 fps, fed into pre-trained 3D/2D ResNet [9].
- Augmentation:
  - acoustic: training set augmented with ambient noise, SNRs from -9 dB to 9 dB. Test set augmented with ambient & music noise, SNRs from -12 dB to 12 dB.
  - visual: Gaussian blur & salt-and-pepper noise for test set.

[9] T. Stafylakis et al., "Combining residual networks with LSTMs for lipreading," *arXiv preprint arXiv:1703.04105*, 2017.

Training Setup

- Acoustic & visual models: pretrained with LRS2 and LRS3 pre-train sets, fine-tuned with the LRS2 training set.
- Language model: trained with the LibriSpeech corpus [10] .

[10] V. Panayotov et al., "LibriSpeech: an ASR corpus based on public domain audio books," in *Proc. ICASSP*, pp. 5206-5210, IEEE, 2015.

Audio-only results

- Dataset: LRS2 corpus with clean acoustic data
- Baselines:
  - **Kaldi**-trained hybrid model with the nnet2 p-norm network recipe [11].
  - Hybrid **CTC/Attention** model without transformer [12]
  - **TM-CTC** model

Table 1 WER (%) for the LRS2 corpus.

	Kaldi	CTC/Attention	TM-CTC
WER	11.28	8.3	3.7

[11] X. Zhang et al., "Improving deep neural network acoustic models using generalized maxout networks," in *Proc. ICASSP*, pp. 215-219, 2014.

[12] S. Petridis et al., "Audio-visual speech recognition with a hybrid CTC/Attention architecture," in *IEEE SLT*, pp. 513-520, IEEE, 2018.

Results

Performance of audio-visual and uni-modal speech recognition (WER [%]). **AO**: audio only. **VO**: video only. **AV**: AV baseline [13]. **DFN**: proposed DFN fusion. **m**: music noise. **a**: ambient noise. **vc**: clean visual data. **gb**: visual Gaussian blur. **sp**: visual salt-and-pepper noise.

dB / models	-12	-9	-6	-3	0	3	6	9	12	clean	avg.
AO(m)	18.9	13.7	11.2	8.4	6.3	6.8	4.5	4.1	4.3	4.2	8.24
AO(a)	25.7	23.4	18.5	11.6	8.2	9.0	5.9	3.8	4.4	4.2	11.47
VO(vc)	58.7	61.0	61.7	69.6	69.6	63.5	64.6	63.6	66.6	61.9	64.08
VO(gb)	66.6	69.2	71.0	68.5	68.5	71.1	62.7	69.4	67.6	66.9	68.15
VO(sp)	68.5	72.5	73.7	70.1	70.1	70.6	68.3	69.1	73.1	67.9	70.39
AV(m.vc)	14.6	11.8	6.4	7.9	7.9	6.3	5.2	4.4	3.4	4.0	7.19
DFN(m.vc)	**11.1**	**8.7**	**5.5**	**4.8**	**4.8**	**4.5**	**3.6**	**3.3**	**2.2**	**2.4**	**5.09**
AV(a.vc)	19.1	19.0	14.3	7.3	6.3	6.0	5.7	4.5	4.9	4.0	9.11
DFN(a.vc)	**14.3**	**11.9**	**8.1**	**4.8**	**4.0**	**5.4**	**3.7**	**2.8**	**3.6**	**2.4**	**6.10**
AV(a.gb)	20.6	18.9	15.0	7.7	6.8	7.5	5.9	3.9	4.8	4.0	9.51
DFN(a.gb)	**14.9**	**12.8**	**9.4**	**5.2**	**4.2**	**5.5**	**3.8**	**3.0**	**4.1**	**2.6**	**6.55**
AV(a.sp)	19.5	19.9	15.3	7.7	7.2	6.3	5.6	4.4	4.6	4.3	9.48
DFN(a.sp)	**15.4**	**12.8**	**9.9**	**5.2**	**4.7**	**5.5**	**3.4**	**2.6**	**4.0**	**2.5**	**6.60**

[13] T. Afouras et al., "Deep audio-visual speech recognition," *arXiv:1809.02108*, 2018.

Conclusion and Future Work

Conclusion

- The proposed decision fusion net (DFN) optimally combines posterior token probabilities of acoustic and visual models based on time-variant reliability information.
- Our new model shows significant improvements on noisy as well as on clean data. On average, it achieves a relative word error rate reduction of 43% compared to the audio-only setup and 31% compared to the audio-visual end-to-end baseline.

Future Work: extend reliability-guided fusion to a general concept within attention models.

Thank you for your attention