RUHR-UNIVERSITÄT BOCHUM

Cognitive
Signal Processing

RUB

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

Wentao Yu, Steffen Zeiler, Dorothea Kolossa

Cognitive Signal Processing Group, Institute of Communication Acoustics, Faculty of Electrical Engineering and Information Technology, Ruhr University Bochum, Germany

Introduction

In noisy conditions, large-vocabulary end-to-end speech recognition remains difficult. In this paper, we address the question of how to optimally inform the end-to-end transformer/CTC model of any time-variant reliability of the acoustic and visual information streams. Our proposed decision fusion net (DFN) yields significant improvements compared to a state-of-the-art baseline model.

System Overview

Model structure: sequence-to-sequence transformer model with connectionist temporal classification—TM-CTC

Training:

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

Decoding:

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

Fusion Strategie

Encoder:

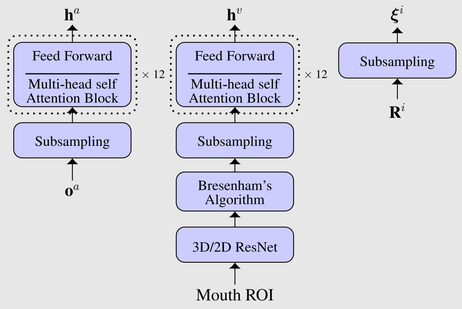

FIGURE 1: Audio encoder (left), video encoder (middle) and reliability measures encoder (right) for both modalities $i \in a, v$. $\mathbf{h}^a$, $\mathbf{h}^v$ and ξ^i are in the same length $N_F/4$, where N_F is the number of frames

Reliability Measures:

- acoustic ($\mathbf{R}^a$): MFCCs, SNR, f_0, Δf_0, Probability of voicing
- visual ($\mathbf{R}^v$): Confidence from OpenFace [1] and Action Units (AU12, AU15, AU17, AU23, AU25, AU26)

- Re-alignment in decoder

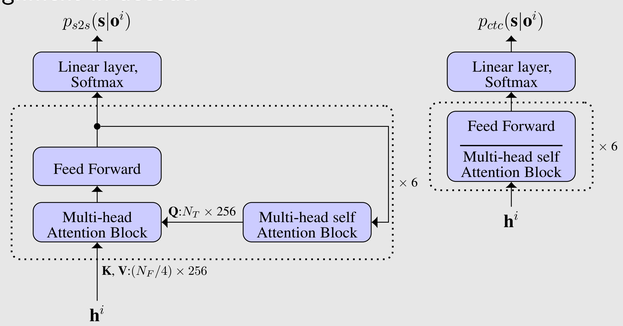

FIGURE 2: Transformer decoder (left) and CTC decoder (right) for both modalities $i \in a, v$

- Multi-head attention:

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

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

- Problem: Reliability measures ξ^i have length $N_F/4$, the token-by-token log-posteriors $p_{s2s}(\mathbf{s}|\mathbf{o}^i)$ have length N_T
- Solution: re-use the attention transform matrix of each head

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

- Proposed fusion architecture

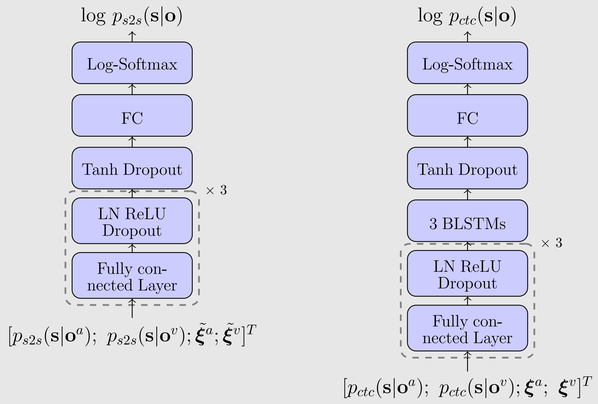

FIGURE 3: Topology of the Decision Fusion Networks (DFN)

Training Setup

ESPnet [2] is the ASR toolkit. All models are pretrained on LRS2 and LRS3 pre-train data [3, 4], then fine-tuned with the LRS2 training set. Acoustic model uses 80 log Mel features together with f_0, Δf_0, and the probability of voicing. Training set augmented with ambient noise, SNRs from -9 dB to 9 dB. Visual model uses 96×96 pixel grayscale mouth region of interest (ROI) at 25 fps fed into a pre-trained 3D/2D ResNet [5].

Results and Conclusion

The proposed DFN fusion shows best performance in all conditions. On average, the new system 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 [3].

models \ dB	-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**

FIGURE 4: Performance of audio-visual and uni-modal speech recognition (WER [%]). **AO**: audio only. **VO**: video only. **AV**: AV baseline [3]. **DFN**: proposed DFN fusion. **m**: music noise. **a**: ambient noise. **vc**: clean visual data. **gb**: visual Gaussian blur. **sp**: visual salt-and-pepper noise.

References

[1] T. Baltrušaitis, P. Robinson, and L. Morency, "Openface: an open source facial behavior analysis toolkit," in *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2016, pp. 1–10.

[2] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[3] T. Afouras, J. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *arXiv:1809.02108*, 2018.

[4] T. Afouras, J. Chung, and A. Zisserman, "LRS3-TED: a large-scale dataset for visual speech recognition," *arXiv:1809.00496*, 2018.

[5] T. Stafylakis and G. Tzimiropoulos, "Combining residual networks with LSTMs for lipreading," *arXiv preprint arXiv:1703.04105*, 2017.