Jointly Learning Visual and Auditory Speech Representations from Raw Data - Extended Abstract

Alexandros Haliassos[1] Pingchuan Ma[1] Rodrigo Mira[1] Stavros Petridis[1,2] Maja Pantic[1,2]
[1]Imperial College London [2]Meta AI
alexandros.haliassos14@imperial.ac.uk

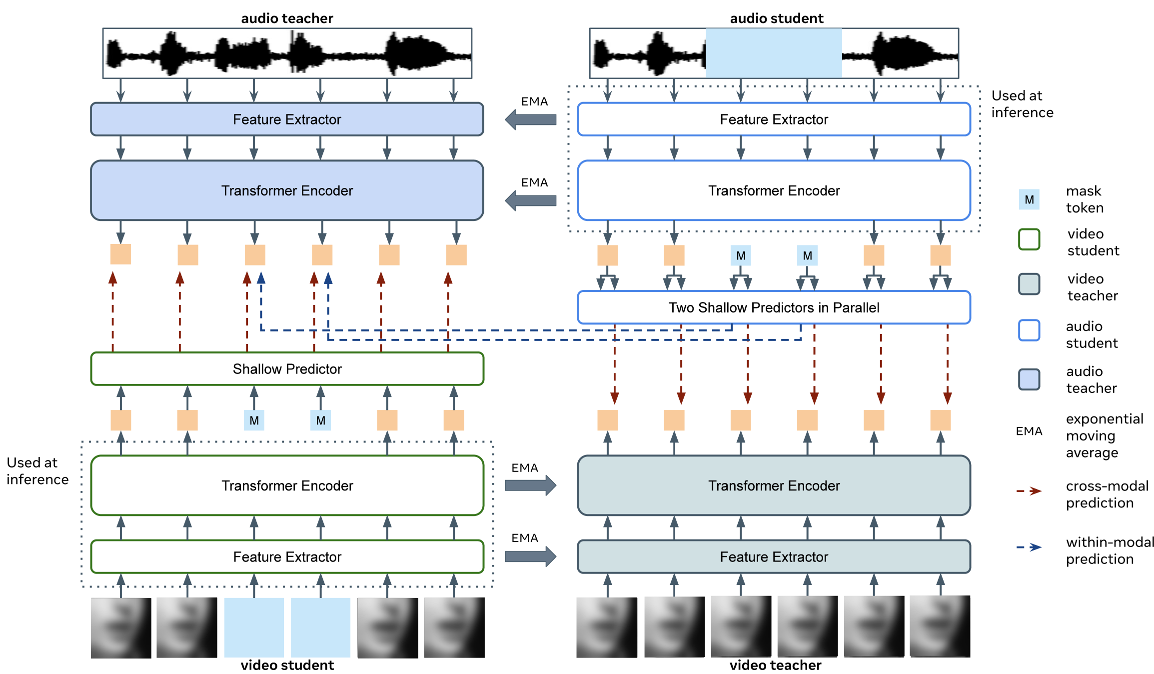

Figure 1. **RAVEn overview.** Given masked video and audio, students predict outputs of unmasked momentum teachers, via shallow Transformer predictors that intake mask tokens.

1. Introduction

Auditory (or automatic) speech recognition (ASR) and visual speech recognition (VSR; also known as lipreading) benefit greatly from the combination of high-capacity neural networks and large datasets, but the effort required for transcription hinders the scaling of labelled data.

A solution is to first learn, in a self-supervised way, general representations from large corpora of unlabelled data, and then fine-tune them on smaller labelled datasets [13]. The fine-grained correspondence between the (synchronised) visual and auditory modalities provides a natural source of self-supervision. However, approaches leveraging this correspondence either (1) only work for word-level samples rather than continuous speech [4–6]; (2) use handcrafted features (*e.g.,* spectrograms or MFCCs) as their inputs or targets [12, 17], which contain inductive biases that may influence the learned representations; (3) use multi-stage pre-training procedures [12, 14, 17]; and/or (4) use separate pre-training strategies for VSR and ASR [17], complicating the process of obtaining representations suitable for both tasks.

In this work, we present a single-stage self-supervised approach that jointly learns visual and auditory speech representations from raw video and audio only: RAVEn (**R**aw **A**udio-**V**isual Speech **En**coders). It involves a pair of student-teacher networks for each modality, whereby the students encode temporally-masked inputs, and, through the use of lightweight Transformer-based predictors, regress outputs of momentum-based teachers [2, 8] that are presented with unmasked inputs. Our experiments demonstrate state-of-the-art performance for self-supervised methods on

LRS3 [1] in low- and high-resource labelled data settings.

2. Method

Masking. We employ masking to encourage the students to take context into account when solving the task. Given a grayscale video and an audio sample, we randomly sample with probability 0.2 each video frame to be the starting mask index, and if selected, then the consecutive three frames are zeroed out. A similar mask is applied to the auditory input, except that it is enlarged by a factor of 640.

Encoders. The masked video and audio are fed to their corresponding student encoders, each consisting of a modality-specific, convolutional feature extractor followed by a Transformer encoder. The video feature extractor is a 2D ResNet18 [9] with a 3D convolutional stem [15], outputting an embedding per frame. On the audio side, we use a 1D ResNet18 which produces features at 25 fps, to match the video sampling rate.

Predictors. The students contain lightweight 2-block, 512-D Transformer predictors , which regress targets given 1) the encoder outputs corresponding to the unmasked portions of the inputs and 2) mask tokens associated with the masked portions. Unlike other works which output global representations and thus use MLPs as predictors [3, 8], we use Transformers to allow modelling temporal dynamics.

Targets. The targets are the outputs of momentum-based teachers [2,8], which are given as input the unmasked video or audio, in order to force the students to predict the missing information. The momentum parameters of the teachers follow a cosine schedule from 0.999 to 1. The use of momentum-based teachers obviates the need for handcrafted targets or multi-stage training.

Prediction tasks. We propose a loss structure that reflects the asymmetry between the visual and auditory modalities w.r.t. speech information. The audio student predicts the targets from both the video and audio teacher, thus benefiting from the ability of cross-modal learning to induce semantic representations, while at the same time being encouraged to retain information from the auditory input that is absent from the visual one. As a result, two predictors are associated with the audio student, one for each target type. On the other hand, the video student only predicts the auditory targets, which are inevitably of higher quality.

Losses. The loss function is the negative cosine similarity [8] between pairs of (aligned) corresponding features, and then summed across the time dimension. For audio-to-audio prediction, the loss is applied only to targets corresponding to masked portions of the input [7]. For the cross-modal tasks, the loss is applied to all targets.

Method	Encoder	LM	Unlab hrs	Lab hrs	WER (%)	
					VSR	ASR
supervised						
Shillingford et al. (2018)	RNN	✓	-	3,886*	55.1	-
Makino et al. (2019)	RNN	✗	-	31,000*	33.6	4.8
Serdyuk et al. (2021)	Transf	✗	-	90,000*	25.9	2.3
Serdyuk et al. (2022)	Conf	✗	-	90,000*	**19.3**	**1.6**
self-supervised						
Base models, less data						
Ma et al. (2021)	Transf	✗	433	30	71.9†	-
Hsu et al. (2021)	Transf	✗	433	30	-	5.4
Shi et al. (2022)	Transf	✗	433	30	51.8	4.9
RAVEn	Transf	✗	433	30	**47.0**	**4.7**
Large models, more data						
Hsu et al. (2021)	Transf	✗	1,759	30	-	3.2
Shi et al. (2022)	Transf	✗	1,759	30	32.5	2.9
Shi et al. (2022) w/ self-training	Transf	✗	1,759	30	28.6	-
RAVEn	Transf	✗	1,759	30	32.5	2.7
RAVEn w/ self-training	Transf	✗	1,759	30	24.8	2.3
RAVEn w/ self-training	Transf	✓	1,759	30	**23.8**	**1.9**

Table 1. **LRS3 low-resource setting**.

Method	Encoder	LM	Unlab hrs	Lab hrs	WER (%)	
					VSR	ASR
self-supervised						
Base models, less data						
Shi et al. (2022)	Transf	✗	433	433	44.0	-
RAVEn	Transf	✗	433	433	**39.1**	**2.2**
Large models, more data						
Hsu et al. (2021)	Transf	✗	1,759	433	-	1.5
Shi et al. (2022)	Transf	✗	1,759	433	28.6	**1.3**
Shi et al. (2022) w/ self-training	Transf	✗	1,759	433	26.9	-
RAVEn	Transf	✗	1,759	433	27.8	1.4
RAVEn w/ self-training	Transf	✗	1,759	433	24.4	1.4
RAVEn w/ self-training	Transf	✓	1,759	433	**23.1**	1.4

Table 2. **LRS3 high-resource setting**.

Fine-tuning For fine-tuning, we append a linear layer and a Transformer decoder to the student encoders for joint CTC / attention decoding [18]. We use SentencePiece [11] subword units as our targets.

3. Main results

Low-resource setting. We pre-train our models on LRS3 or LRS3+Vox2-en [17] and then fine-tune on the 30-hour LRS3 subset to evaluate performance when labels are scarce (see Table 1). Our Base variant outperforms all related methods on VSR. The Large model provides significant boosts over the Base model (32.5% vs 40.2% WER) when using LRS3+Vox2-en for pre-training, keeping the number of labelled data points fixed. Using a language model with self-training leads to a WER of 23.8%, better than a method [16] trained on *90,000* hours of non-public data.

On ASR, RAVEn significantly outperforms the audio-only Hubert [10] model, and in all cases is better than or on par with AV-HuBERT. Our best ASR model without self-training achieves 2.7% WER vs AV-HuBERT's 2.9%, despite using the same pre-training for VSR and ASR.

High-resource setting. Table 2 reports results when fine-tuning on the full 433 hours of LRS3. RAVEn outperforms

AV-HuBERT under all configurations on VSR. Our best result is 23.1%, achieved using self-training and a language model. We are on par with the state-of-the-art for ASR in the high-resource setting, achieving a WER of 1.4% with the Large model. This is despite using raw audio as input (rather than spectrograms [17]).

4. Conclusion

We propose RAVEn, a single-stage method that jointly learns visual and auditory speech representations entirely from raw data, and achieves state-of-the-art results for VSR and ASR on LRS3 for self-supervised methods. Our pre-training methodology is general, and we hope it inspires future research extending beyond speech recognition.

Acknowledgements

Only non-Meta authors conducted any of the dataset pre-processing (no dataset pre-processing took place on Meta's servers or facilities).

References

[1] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 2

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9650–9660, 2021. 1, 2

[3] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the 18th IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9640–9649, 2021. 2

[4] Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In *Proceedings of the 13th Asian Conference on Computer Vision (ACCV)*, pages 251–263, 2016. 1

[5] S. Chung, J. Son Chung, and H. Kang. Perfect match: Improved cross-modal embeddings for audio-visual synchronisation. In *Proceedings of the 44th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3965–3969, 2019. 1

[6] Soo-Whan Chung, Hong-Goo Kang, and Joon Son Chung. Seeing voices and hearing voices: learning discriminative embeddings using cross-modal self-supervision. In *Proceedings of the 21st Annual Conference of International Speech Communication Association (INTERSPEECH)*, pages 3486–3490, 2020. 1

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT*, pages 4171–4186, 2019. 2

[8] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. In *Proceedings of the 33rd Advances in Neural Information Processing System (NIPS)*, volume 33, pages 21271–21284, 2020. 1, 2

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the 29th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[10] Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460, 2021. 2

[11] Taku Kudo and John Richardson. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 56th Conference of the Association for Computational Linguistics (ACL)*, pages 66–71, 2018. 2

[12] Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W. Schuller, and Maja Pantic. LiRA: Learning visual speech representations from audio through self-supervision. In *Proceedings of the 22nd Annual Conference of International Speech Communication Association (INTERSPEECH)*, pages 3011–3015, 2021. 1

[13] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al. Self-supervised speech representation learning: A review. *arXiv preprint arXiv:2205.10643*, 2022. 1

[14] Xichen Pan, Peiyu Chen, Yichen Gong, Helong Zhou, Xinbing Wang, and Zhouhan Lin. Leveraging uni-modal self-supervised learning for multimodal audio-visual speech recognition. In *Proceedings of the 60th Conference of the Association for Computational Linguistics (ACL)*, page 4491–4503, 2022. 1

[15] Stavros Petridis, Themos Stafylakis, Pingchuan Ma, Georgios Tzimiropoulos, and Maja Pantic. Audio-visual speech recognition with a hybrid ctc/attention architecture. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 513–520, 2018. 2

[16] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. Audio-visual speech recognition is worth 32x32x8 voxels. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding (ASRU) Workshop*, 2021. 2

[17] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022. 1, 2, 3

[18] Shinji Watanabe, Takaaki Hori, Suyoun Kim, John R Hershey, and Tomoki Hayashi. Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253, 2017. 2