BRAVEn: Improving Self-Supervised Pre-training for Visual and Auditory Speech Recognition - Extended Abstract

Alexandros Haliassos* Andreas Zinonos* Rodrigo Mira Stavros Petridis Maja Pantic

Imperial College London

*Equal contribution

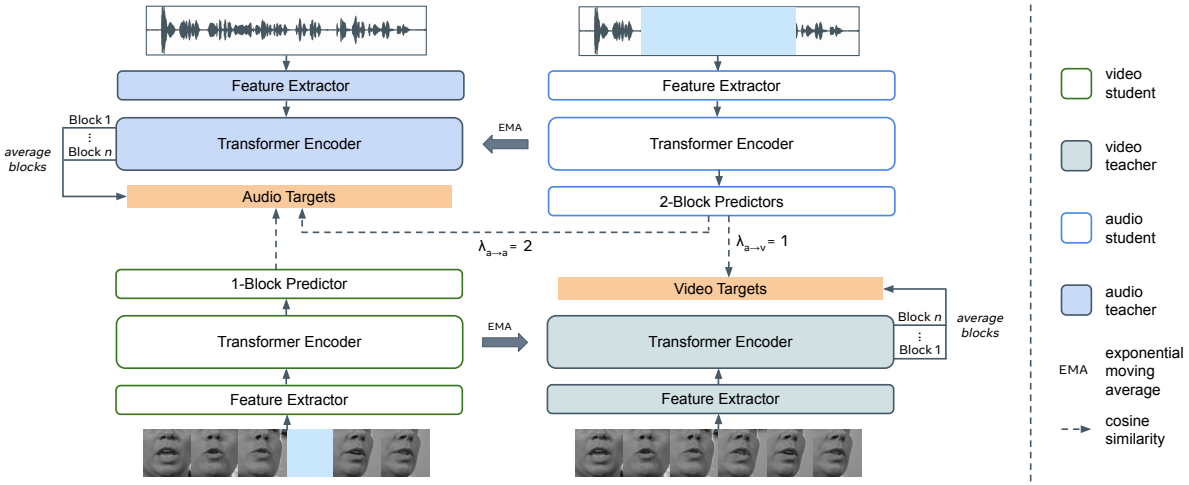

Figure 1. **BRAVEn overview.**

1. Introduction

Visual and auditory speech recognition (VSR / ASR) scale very well with the amount of transcribed data [10–13]. However, the cost of accurately annotating large-scale datasets can be prohibitive, spurring significant research into bypassing annotated data requirements. One line of work along this direction is audio-visual self-supervised learning, where representations are first learned by exploiting the fine-grained correspondence between the visual and auditory modalities, and then the resulting encoders are fine-tuned on potentially less labelled data for speech recognition [2, 7, 9, 14, 16]. In particular, RAVEn [7] uses separate video and audio students which regress the outputs of the teacher networks through lightweight Transformer predictors, and learns strong visual and auditory speech representations entirely from raw data.

In this work, we extend RAVEn through modifications that yield better representations by acknowledging the semantic asymmetries between audio and video. We dub our approach Better RAVEn, or BRAVEn. Our enhancements are as follows: (1) We use the average of the outputs of each Transformer encoder block as our targets [8], rather than the output of the last block, in order to create smoother targets; (2) We use a shallower predictor for the video student, which encourages the video encoder to better capture the information embedded in the predicted audio targets; (3) We use stronger masking for the audio inputs to address the difference in relative difficulty between VSR and ASR; (4) Finally, we use different loss weights for the audio predictors, which empirically benefits ASR performance.

We find that BRAVEn scales well both with the model size as well as the amount of unlabelled data, consistently achieving state-of-the-art performance across self-supervised methods in comparable settings. Notably, BRAVEn-Large trained with around 3,000 hours of unlabelled data and only 30 hours of annotated data achieves 20.0 % / 1.7 % word error rate (WER) for VSR / ASR on the LRS3 test set, making it competitive with methods trained on orders of magnitude more transcribed data [10, 12].

2. Method

In the following, we describe in detail each of BRAVEn's design improvements over RAVEn (see Figure 1).

First, BRAVEn uses the mean of the outputs of all Trans-

Method	Encoder	LM	Unlab hrs	Lab hrs	WER (%)	
					VSR	ASR
Base models						
AV-HuBERT [14]	Transf	✗	403	30	51.8	4.9
RAVEn [7]	Transf	✗	403	30	47.0	4.7
AV-data2vec [8]	Transf	✗	403	30	45.2	4.4
BRAVEn	Transf	✗	403	30	**43.4**	**4.0**
Large models						
AV-HuBERT [14]	Transf	✗	1,729	30	32.5	2.9
RAVEn [7]	Transf	✗	1,729	30	32.5	2.7
AV-data2vec [8]	Transf	✗	1,729	30	30.8	2.7
RAVEn w/ ST [7]	Transf	✗	1,729	30	24.8	2.3
BRAVEn	Transf	✗	1,729	30	30.8	2.3
BRAVEn	Transf	✗	3,052	30	24.8	2.1
BRAVEn w/ ST	Transf	✗	3,052	30	21.3	1.9
BRAVEn w/ ST	Transf	✓	3,052	30	**20.0**	**1.7**

(a) **Low-resource.**

Method	Encoder	LM	Unlab hrs	Lab hrs	WER (%)	
					VSR	ASR
Base models						
AV-HuBERT [14]	Transf	✗	-	433	44.0	-
RAVEn [7]	Transf	✗	-	433	39.1	2.2
AV-data2vec [8]	Transf	✗	-	433	39.0	2.0
BRAVEn	Transf	✗	-	433	**36.0**	**1.9**
Large models						
AV-HuBERT [14]	Transf	✗	1,326	433	28.6	1.3
RAVEn [7]	Transf	✗	1,326	433	27.8	1.4
AV-data2vec [8]	Transf	✗	1,326	433	28.5	1.4
RAVEn w/ ST [7]	Transf	✗	1,326	433	24.4	1.4
BRAVEn	Transf	✗	1,326	433	26.6	1.2
BRAVEn	Transf	✗	2,649	433	23.6	1.2
BRAVEn w/ ST	Transf	✗	2,649	433	20.9	1.2
BRAVEn w/ ST	Transf	✓	2,649	433	**20.1**	**1.1**

(b) **High-resource.**

Table 1. **LRS3 results**. "Unlab hrs" / "lab hrs" denote unlabelled hours / labelled hours. Self-supervised methods use labelled data for pre-training (without labels) along with unlabelled data (if any). "LM" denotes language model. "ST" denotes self-training.

	WER (%)	
	VSR	ASR
RAVEn	47.0	4.7
+ average blocks	45.3 ↓1.7	4.6 ↓0.1
+ shallower video predictor	44.1 ↓1.2	4.6 =
+ stronger audio masking	43.5 ↓0.6	4.2 ↓0.4
+ different loss weights = **BRAVEn**	**43.4** ↓0.1	**4.0** ↓0.2

Table 2. **Ablations.** We show results for the low-resource setting using the Base model.

former blocks rather than only the final block's output [8]. This averaging operation likely results in smoother and higher-quality targets, which can aid the training dynamics [6]. Moreover, we apply instance normalisation [15] to the averaged targets to prevent representation collapse [3]. In RAVEn, the layer normalisation [1] that follows the final block in the Transformer architecture serves a similar purpose.

Second, BRAVEn uses asymmetric predictor depths. RAVEn uses two-block Transformer predictors (with attention dimension 512), shown to work optimally *when using the same design for both video and audio students*. BRAVEn instead uses a *one-block* Transformer predictor for the video student (which predicts the audio targets), while retaining the original design of the audio predictors. The intuition is that audio is more relevant to speech recognition than video, and thus a shallower predictor leads to visual representations that more closely match the information in the audio targets.

Third, BRAVEn applies stronger masking to the audio inputs. Each corresponding video frame index has a 40 % probability of being picked as the start of a mask for audio, while a probability of 20 % is used for video. In contrast, RAVEn uses a probability of 20 % for both modalities, which was chosen based on the assumption that *both modalities would equally benefit from the same masking strategy*. BRAVEn's masking asymmetry is beneficial likely due to the difference in difficulty between VSR and ASR: Stronger masking for audio leads to more context-aware representations, but for video, it may make the pretext task overly difficult.

Finally, BRAVEn uses asymmetric loss weights. In RAVEn, the weights of the two losses associated with the audio student are equal. BRAVEn instead uses a larger weight for the audio-to-audio ($\lambda_{a\to a} = 2$) than the audio-to-video loss ($\lambda_{a\to v} = 1$), which we empirically find leads to improvements for ASR.

3. Main results

3.1. Low-resource setting

We provide results for the low-resource setting in Table 1a, where we fine-tune on the 30-hour "trainval" LRS3 partition. Using BRAVEn-Base and LRS3 as the pre-training dataset, we achieve better VSR and ASR word error rate (WER) than the previous state-of-the-the-art obtained by AV-data2vec [8]. When also including an English-only version of VoxCeleb2 for pre-training, BRAVEn-Large achieves 30.8 % and 2.3 % WER for VSR and ASR respectively, demonstrating BRAVEn's strong scalability. We are also the first to experiment with pre-training on a combination of LRS3, VoxCeleb2 [4], and AVSpeech [5], consisting of 3,082 hours of unlabelled data. This results in significant WER improvement for VSR (30.8 %→24.8 %) and modest improvement for ASR (2.3 %→2.1 %). Finally, using self-training and a language model during inference leads to 20.0 % / 1.7 % WER for VSR / ASR.

3.2. High-resource setting

The results for the high-resource setting, where we fine-tune on the full 433-hour LRS3 dataset, are shown in Table 1b. BRAVEn achieves state-of-the-art performance among the self-supervised methods. Our best results for the high-resource setting are 20.1 % / 1.1 % for VSR / ASR. Interestingly, BRAVEn-Large coupled with self-training and a language model yields similar VSR results ($\sim$ 20 % WER) in both low- and high-resource settings, but better results are obtained in the high-resource setting for ASR. This suggests that high-quality transcriptions are more important for ASR than VSR.

3.3. Ablations

We conduct an ablation study to show the effect of each design choice on the VSR and ASR performance (see Table 2). Each new addition improves the VSR and / or ASR performance. We observe that averaging targets and using a shallower video predictor has a larger effect on VSR than ASR, while stronger audio masking and different loss weights have a more significant impact on ASR. Furthermore, using even stronger audio masking or using the average of the last 6 Transformer blocks as targets lead to worse performance.

4. Conclusion

We have proposed some design improvements to the recent RAVEn method, which cumulatively have a significant impact on the downstream VSR and ASR performance and achieve state-of-the-art results for audio-visual self-supervised methods in various settings. Furthermore, we experiment with doubling the amount of unlabelled data during pre-training as used in other self-supervised works and observe strong scaling behaviour, even when fine-tuning with only 30 hours of labelled data. Our work provides compelling evidence that readily accessible unlabelled audio-visual data can effectively substitute costly annotated samples.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 2

[2] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *ICML*, volume 162, pages 1298–1312. PMLR, 2022. 1

[3] 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. 2

[4] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech*, pages 1086–1090, 2018. 2

[5] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T. Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Trans. Graph.*, 37(4):112, 2018. 2

[6] 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 *NeurIPS*, volume 33, pages 21271–21284, 2020. 2

[7] Alexandros Haliassos, Pingchuan Ma, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Jointly learning visual and auditory speech representations from raw data. In *ICLR*. OpenReview.net, 2023. 1, 2

[8] Jiachen Lian, Alexei Baevski, Wei-Ning Hsu, and Michael Auli. Av-data2vec: Self-supervised learning of audio-visual speech representations with contextualized target representations. *arXiv*, 2023. 1, 2

[9] Pingchuan Ma, Rodrigo Mira, Stavros Petridis, Björn W. Schuller, and Maja Pantic. LiRA: Learning visual speech representations from audio through self-supervision. In *Interspeech*, pages 3011–3015, 2021. 1

[10] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *ASRU Workshop*, pages 905–912, 2019. 1

[11] K. R. Prajwal, Triantafyllos Afouras, and Andrew Zisserman. Sub-word level lip reading with visual attention. In *CVPR*, 2022. 1

[12] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. Audio-visual speech recognition is worth 32x32x8 voxels. In *ASRU Workshop*, 2021. 1

[13] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. Transformer-based video front-ends for audio-visual speech recognition for single and muti-person video. In Hanseok Ko and John H. L. Hansen, editors, *Interspeech*, pages 2833–2837. ISCA, 2022. 1

[14] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *ICLR*, 2022. 1, 2

[15] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 2

[16] Qiushi Zhu, Long Zhou, Ziqiang Zhang, Shujie Liu, Binxing Jiao, Jie Zhang, Lirong Dai, Daxin Jiang, Jinyu Li, and Furu Wei. Vatlm: Visual-audio-text pre-training with unified masked prediction for speech representation learning. *IEEE Trans. Multimed.*, 2023. 1