Auto-AVSR: Audio-Visual Speech Recognition with Automatic Labels - Extended Abstract

Pingchuan Ma[1] Alexandros Haliassos[1] Adriana Fernandez-Lopez[2] Honglie Chen[2]
Stavros Petridis[1,2] Maja Pantic[1,2] *

[1]Imperial College London [2]Meta AI

1. Introduction

Audio-visual speech recognition has received a lot of attention due to its robustness against acoustic noise. Recently, the performance of automatic, visual, and audio-visual speech recognition (ASR, VSR, and AV-ASR, respectively) has been substantially improved, mainly due to the use of larger models and training sets. However, accurate labelling of datasets is time-consuming and expensive. Hence, in this work, we investigate the use of automatically-generated transcriptions of unlabelled datasets to increase the training set size. In this work, we automatically generate transcriptions for more than 2 000 hours of videos by utilising publicly-available ASR models. We then train ASR, VSR and AV-ASR models with these transcriptions and achieve state-of-the-art performance on LRS3. We show that the accuracy of the pre-trained ASR models used to automatically transcribe the unlabelled datasets is not highly correlated with the performance of the ASR and VSR models trained with these transcriptions. Furthermore, we show that our audio-visual model is more robust against different levels of noise than its audio-only counterpart.

2. Methodology

An overview of our label generation pipeline can be found at the top of Fig. 1. To be specific, audio waveforms from the unlabelled audio-visual datasets are fed into a pre-trained ASR model to produce automatic transcriptions. For the purpose of this study, we use two unlabelled datasets: VoxCeleb2 [5] and AVSpeech [7]. We are interested in training models in English. Thus, we use the VoxLingua107 language classifier [21] to filter the AVSpeech dataset, resulting in a total of 1 323 hours; the list of English data we use for VoxCeleb2 is obtained from [19], and comprises 1 307 data hours. Next, we leverage publicly-available ASR models to produce automatically generated transcriptions. It is worth pointing out that our work facilitates reproduction and comparison since all datasets and models used are publicly accessible.

Method	WER [%]			
	$A^{\dagger}$	$A^{\dagger\dagger}$	V	A
CM-Transducer [10]	1.62	3.31	19.1	0.99
HuBERT [9]	1.90	6.87	19.8	1.12
Wav2vec 2.0 [4]	3.40	11.22	19.1	1.06
Whisper [16]	4.10	1.81	19.0	1.04

Table 1. Impact of the pre-trained ASR models used to generate automatic transcriptions from unlabelled data on the performance of VSR/ASR models on the LRS3 dataset. † and †† denote the word error rate (WER) reported on Librispeech test-clean set [14] and LRS3 test set [2], respectively. "CM" denotes Conformer.

We investigate the impact of the automatic transcriptions given by four different ASR models on the performance of audio-only and visual-only models, i.e. Whisper [16], wav2vec2.0 [4], Hidden unit BERT (HuBERT) [9] and Conformer-Transducer [8, 10]. We adopt the off-the-shelf architecture presented in [11], which has achieved state-of-the-art performance on LRS3 without the use of external data. The architecture is shown at the bottom of Fig. 1.

For the purposes of this study, we use the LRS3 dataset [2] consisting of 151 819 video clips from TED talks with a total of 439 hours. For training, we also use the LRS2 dataset (223 hours) [6], English-speaking videos from AVSpeech (1 323 hours) [7] and VoxCeleb2 (1 307 hours) [5] as the additional training data together with automatically-generated transcriptions.

3. Results

Do better Librispeech ASR models provide better transcriptions for VSR and ASR? Results of the ASR and VSR models trained with the automatically-generated

*Only non-Meta co-authors downloaded, accessed, and used the datasets. Only non-Meta authors conducted any of the dataset pre-processing (no dataset pre-processing took place on Meta's servers or facilities). Code and trained models are available at: https://github.com/mpc001/auto_avsr

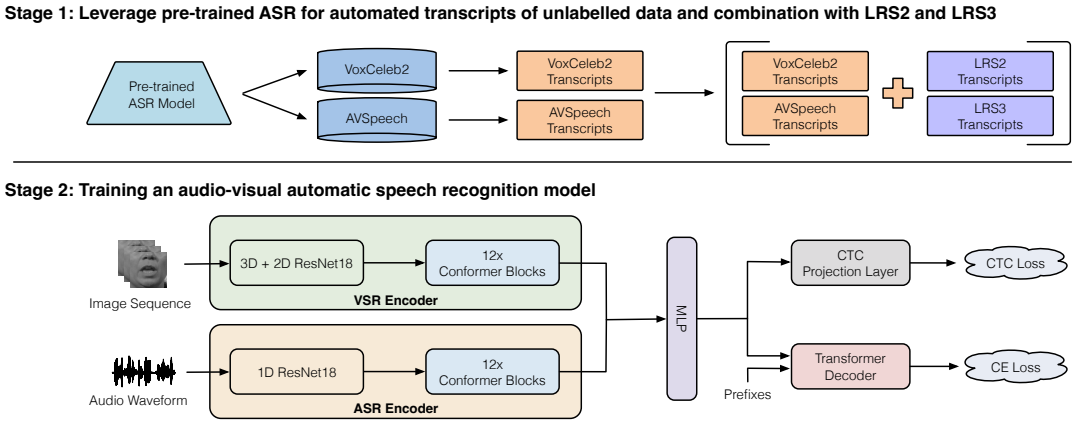

Figure 1. AV-ASR architecture overview.

Method	Type	Extra Data	Total Hours‡	WER (%)
CM-seq2seq [11]	V	✗	438	46.9
CM-aux [12]				37.9
Ours				**36.3**
KD + CTC [3]		✓	772	59.8
KD-seq2seq [17]			818	59.0
TM-seq2seq [1]			1 362	58.9
AVHuBERT [19]			1 759	26.9
RNN-T [13]			31 000	33.6
VTP [15]			2 676	30.7
ViT3D-CM [18]			90 000	17.0
Ours			818	**33.0**
Ours			1 902	**23.5**
Ours			3 448	**19.1**
CM-seq2seq [11]	A	✗	438	2.3
RNN-T [13]		✓	31 000	4.5
AV-HuBERT [19]			1 759	1.5
Ours			818	**1.5**
Ours			1 902	**1.0**
Ours			3 448	**1.0**
CM-seq2seq [11]	A+V	✗	438	2.3
RNN-T [13]		✓	31 000	4.8
AV-HuBERT [19]			1 759	1.3
ViT3D-CM [18]			90 000	1.6
Ours			1 902	**1.0**
Ours			3 448	**0.9**

Table 2. WER (%) of our audio-only, visual-only and audio-visual models on the LRS3 dataset. ‡ The total hours are counted by including the datasets used for both pre-training and training.

transcriptions on the LRS3 dataset are shown in the third and fourth columns, respectively, of Table 1. We observe that overall the WER on Librispeech is not highly correlated with the performance of the ASR and VSR models trained with the automatically-generated transcriptions from the corresponding pre-trained ASR models. The same conclusion is also true when we measure the WER on the LRS3 test.

Comparison with the state-of-the-art. Results on LRS3 are presented in 2. The best visual-only model has a WER of 19.1 %, which is outperformed only by [18] (17.0 % WER) which uses 26× more training data. Similarly, our audio-only model establishes a new state-of-the-art [19] by achieving a WER of 1.0 % when using 1 921 hours of training data from LRW, LRS3 and VoxCeleb2 datasets. However, when further introducing AVSpeech for training, no further improvement is observed, suggesting that the ASR performance may have reached saturation. State-of-the-art performance is also achieved for AV-ASR with a WER of 0.9 %.

Type	Noise	SNR levels [dB] 12.5	7.5	2.5	-2.5	-7.5
A	Babble‡	1.1	1.2	1.6	2.7	8.3
A+V		1.0	1.0	1.5	2.2	5.6
A	Pink	1.4	1.9	4.3	13.1	56.8
A+V		1.2	1.4	2.3	6.0	16.2
A	White	2.1	4.0	10.4	30.2	88.9
A+V		1.4	2.3	4.3	9.5	24.2

Table 3. WER (%) of our audio-only and audio-visual models as a function of the noise levels on the LRS3 dataset. ‡ denotes the noise type used in both training and test set.

Noise experiments. Results of ASR and AV-ASR models, when tested with different acoustic noise levels, are shown in Table. 3. We show that, overall, the results are consistent with those presented in [1, 11, 13, 20], i.e. the performance of audio-only models is closer to the audio-visual counterpart in the presence of low levels of noise, whereas the performance gap becomes larger as the noise levels increase.

4. Conclusion

In this work, we propose a new strategy to scaling up audio-visual data for speech recognition, which takes advantage of well-known ASR models to annotate audio-visual data. In this way, our AV-ASR system achieves state-of-the-art on LRS3 with less than 1% of WER.

References

[1] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Deep audio-visual speech recognition. *IEEE TPAMI*, 2018. 2

[2] T. Afouras, J. S. Chung, and A. Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496*, 2018. 1

[3] T. Afouras, J. S. Chung, and A. Zisserman. ASR is all you need: Cross-modal distillation for lip reading. In *ICASSP*, pages 2143–2147, 2020. 2

[4] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NIPS*, volume 33, pages 12449–12460, 2020. 1

[5] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech*, pages 1086–1090, 2018. 1

[6] J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *CVPR*, pages 3444–3453, 2017. 1

[7] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics*, 37(4):112:1–112:11, 2018. 1

[8] A. Gulati, J. Qin, C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *Interspeech*, pages 5036–5040, 2020. 1

[9] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE Trans. Audio, Speech, Lang. Process.*, 29:3451–3460, 2021. 1

[10] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook, et al. Nemo: a toolkit for building ai applications using neural modules. *arXiv preprint arXiv:1909.09577*, 2019. 1

[11] P. Ma, S. Petridis, and M. Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP*, pages 7613–7617, 2021. 1, 2

[12] P. Ma, S. Petridis, and M. Pantic. Visual Speech Recognition for Multiple Languages in the Wild. *Nature Machine Intelligence*, pages 930–939, 2022. 2

[13] T. Makino, H. Liao, Y. Assael, B. Shillingford, B. Garcia, O. Braga, and O. Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *ASRU*, pages 905–912, 2019. 2

[14] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *ICASSP*, pages 5206–5210, 2015. 1

[15] K. Prajwal, T. Afouras, and A. Zisserman. Sub-word level lip reading with visual attention. In *CVPR*, pages 5162–5172, 2022. 2

[16] A. Radford, J. W. Kim, C. M. Payne, P. Mishkin, T. Xu, G. Brockman, and I. Sutskever. Introducing whisper. https://openai.com/blog/whisper/, 2022. [Online; accessed 18-October-2022]. 1

[17] S. Ren, Y. Du, J. Lv, G. Han, and S. He. Learning from the master: Distilling cross-modal advanced knowledge for lip reading. In *CVPR*, pages 13325–13333, 2021. 2

[18] D. Serdyuk, O. Braga, and O. Siohan. Transformer-based video front-ends for audio-visual speech recognition for single and multi-person video. In *Interspeech*, pages 2833–2837, 2022. 2

[19] B. Shi, W. Hsu, K. Lakhotia, and A. Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *ICLR*, 2022. 1, 2

[20] B. Shi, W. Hsu, and A. Mohamed. Robust self-supervised audio-visual speech recognition. In *Interspeech*, pages 2118–2122, 2022. 2

[21] J. Valk and T. Alumäe. VoxLingua107: a dataset for spoken language recognition. In *SLT*, 2021. 1