AUTO-AVSR: AUDIO-VISUAL SPEECH RECOGNITION WITH AUTOMATIC LABELS

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, UK
[2]Meta AI, UK

ABSTRACT

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. For this purpose, we use publicly-available pre-trained ASR models to automatically transcribe unlabelled datasets such as AVSpeech and VoxCeleb2. Then, we train ASR, VSR and AV-ASR models on the augmented training set, which consists of the LRS2 and LRS3 datasets as well as the additional automatically-transcribed data. We demonstrate that increasing the size of the training set, a recent trend in the literature, leads to reduced WER despite using noisy transcriptions. The proposed model achieves new state-of-the-art performance on AV-ASR on LRS2 and LRS3. In particular, it achieves a WER of 0.9 % on LRS3, a relative improvement of 30 % over the current state-of-the-art approach, and outperforms methods that have been trained on non-publicly available datasets with 26 times more training data.

Index Terms— audio-visual speech recognition, unlabelled audio-visual data, automatically generated transcriptions

1. INTRODUCTION

In the human perceptual system, the visual and audio streams often complement each other, yielding a unified robust response. It is also known that using visual signals along with audio signals leads to higher model robustness than using a single modality, especially in the presence of high levels of acoustic noise [1–4].

In this paper, we tackle Audio-Visual Automatic Speech Recognition (AV-ASR, or AVSR), which aims to transcribe continuous spoken sentences from both audio and visual streams. Recent audio- (ASR), video- (VSR) and audio-visual-based models have relied heavily on large-scale and well-labelled transcriptions to achieve convincing performance. However, accurate transcriptions require manual labelling, which is time-consuming and prohibitively expensive. In order to address this issue, several works have been proposed to build advanced ASR and VSR models by leveraging large-scale unlabelled audio-visual datasets. Popular approaches pre-train ASR and VSR models using self-supervised learning, where the goal is to learn audio and visual representations from

large unlabelled datasets [5–8]. The pre-trained models are then fine-tuned on smaller labelled datasets. Alternatively, another line of work solves the task using knowledge distillation [9, 10]. For example, Afouras *et al.* [9] use a pre-trained ASR model - acting as a teacher - to provide an extra supervisory signal to the target VSR model, where the goal is to force the posterior distribution of the VSR network to match the teacher. Similarly, Ren *et al.* [10] further improve the performance of visual-only models by distilling knowledge from pre-trained models with multiple modalities.

In this work, we propose a different approach to leverage large unlabelled datasets which does not require a two-step training approach used in self-supervised learning (SSL) (but it can easily be combined with any state-of-the-art SSL approach). In particular, we take advantage of the availability of good publicly-available pre-trained ASR models [5, 8, 11, 12] to automatically annotate large-scale audio-visual datasets. This approach is broadly related to self-training [13–15], where a model is first trained on annotated data and then used to generate pseudo-labels for the unlabelled data. A new model is trained with all the annotated data, and this process is repeated for a few iterations. This iterative process might be necessary in other domains for which a high-quality pre-trained model does not exist, but this is not the case for ASR, where accurate pre-trained models are relatively abundant. Thus, we sidestep the need for a costly iterative procedure. Moreover, we incorporate the automatically-transcribed unlabelled data into the training set rather than using the pre-trained ASR model for distillation [9]. As a result, we can easily train a large-scale AV-ASR system by simplifying the implementation and reducing both the computational and memory costs of using a teacher during training. We also find similarities with [3, 16], but instead of using owner-uploaded transcriptions or a production-quality ASR system, all models and datasets we use are publicly accessible.

Our main contributions can be summarised as follows: 1) 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 the LRS2 and LRS3 datasets. Concretely, the proposed approach leads to a WER of 0.9% for AV-ASR on the LRS3 dataset, which outperforms models trained on much larger training sets; 2) 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; 3) We observe that an increase in the number of hours of automatically-transcribed data used in the training set results in reduced WER, especially for the VSR models. On the other hand, the performance of the ASR models seems to saturate beyond 1500 hours.

Stage 1: Leverage pre-trained ASR for automated transcripts of unlabelled data and combination with LRS2 and LRS3
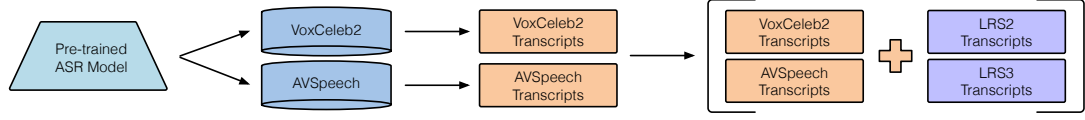
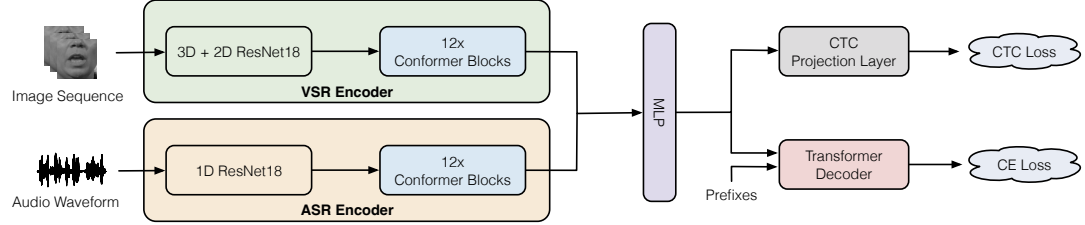**Stage 2: Training an audio-visual automatic speech recognition model**

Fig. 1. AV-ASR architecture overview. In the first stage, a pre-trained ASR model is leveraged to produce automatically-generated transcriptions for unlabelled audio-visual datasets. And then these unlabelled datasets are combined with the labelled training sets, including LRS2 and LRS3, for training. The frame rate of audio and visual features from the ASR and VSR encoders is 25 frames per second (fps).

2. AUTO-AVSR

2.1. Leveraging pre-trained models to produce automatically-generated transcriptions for unlabelled audio-visual datasets

In order to investigate the impact of the size of training data on the performance of audio-only, visual-only and audio-visual models, we scale the training sources by including publicly-available audio-visual clips into the training set. 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 [17] and AVSpeech [18]. In particular, AVSpeech [18] contains 4 700 hours of YouTube video segments in multiple languages, and VoxCeleb2 [17] consists of 2 300 hours of video segments from more than 6 000 people. However, we are interested in training models in English. Thus, we use the VoxLingua107 language classifier [19] 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 [7], 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.

2.2. Automatic speech recognition models

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 [11], wav2vec2.0 [5], Hidden unit BERT (HuBERT) [8] and Conformer-Transducer [12, 20]. In particular, wav2vec 2.0 [5] and HuBERT [8] are self-supervised learning methods for learning speech representations from unlabelled audio. In contrast, Conformer-Transducer [12] is a conformer-based model trained with recurrent neural network transducers (RNN-T) loss that uses the NeMo ASRSET dataset, consisting of 12 000 hours of English speech. Finally, Whisper [11] is a transformer-based model [21] trained with a total of 680 000 hours of labelled audio. In this work, we access the ASR models from the Hugging Face community, "nvidia/stt_en_conformer_transducer_xlarge", "facebook/hubert-large-ls960-ft", "facebook/wav2vec2-base-960h", and "openai/whisper-medium.en", respectively.

2.3. Architecture

We adopt the off-the-shelf architecture presented in [22], which has achieved state-of-the-art performance on the LRS2 and LRS3 datasets without the use of external data. The architecture is shown at the bottom of Fig. 1. In particular, the VSR front-end is based on a modified ResNet-18 [23, 24], where the first layer is a spatio-temporal convolutional layer with a kernel size of $5 \times 7 \times 7$ and a stride of $1 \times 2 \times 2$. The temporal back-end, which follows the front-end, is a Conformer [20]. Similarly, the ASR encoder consists of a 1D ResNet-18 [25] followed by a Conformer. The ASR and VSR encoder outputs are fused via a multi-layer perceptron (MLP). The rest of the network consists of a projection layer and a Transformer decoder for joint CTC/attention training [26].

3. EXPERIMENTAL SETUP

3.1. Datasets

We conduct experiments on LRS2 [27] and LRS3 [28], which are the two largest publicly available datasets for audio-visual speech recognition in English. LRS2, collected from BBC programs, contains 144 482 video clips with a total of 225 hours. Specifically, the pre-training, training, validation and test set contains 96 318 (195 hours), 45 839 (28 hours), 1 082 (0.6 hours) and 1 243 (0.5 hours) video clips, respectively. LRS3 consists of 151 819 video clips from TED talks with a total of 439 hours. It contains 118 516 (408 hours), 31 982 (30 hours) and 1 321 clips (0.9 hours) in the pre-training, training-validation, and test set, respectively. For training, we also use the English-speaking videos from AVSpeech (1 323 hours) and VoxCeleb2 (1 307 hours) as the additional training data together with automatically-generated transcriptions.

3.2. Pre-processing

For the visual stream, we follow previous work [22] to pre-process the datasets. We crop the mouth region of interests (ROIs) using a bounding box of 96×96. Each frame is normalised by subtracting the mean and dividing by the standard deviation of the training set. For audio streams, we only perform z-normalisation per utterance.

Method	WER [%]			
	A[†]	A[††]	V	A
CM-Transducer [12]	1.62	3.31	19.1	0.99
HuBERT [8]	1.90	6.87	19.8	1.12
Wav2vec 2.0 [5]	3.40	11.22	19.1	1.06
Whisper [11]	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 [33] and LRS3 test set [28], respectively. "CM" denotes Conformer. "V" and "A" denote the visual-only and audio-only models trained on LRW, LRS2, LRS3, VoxCeleb2 and AVSpeech (using the automatically-generated transcriptions from the corresponding pre-trained ASR model), with a total of 3 448 hours.

3.3. Implementation details

For our audio- and visual-only ASR models, we use a ResNet-based front-end module pre-trained on LRW [29], followed by a Conformer encoder with 12 layers, 768 input dimensions, 3 072 feed-forward dimensions, and 16 attention heads. The decoder is a 6-layer Transformer with the same dimensions and number of heads as the encoder, resulting in a total of 243.1 M and 250.4 M parameters for the audio- and visual-only models, respectively. More specifically, the ASR front-end, VSR front-end, Conformer back-end, Transformer decoder and the projection layer of the CTC have 3.9 M, 11.2 M, 170.9 M, 64.5 M and 3.9 M parameters, respectively. For the audio-visual models, we concatenate the audio-visual encoder outputs and feed them to a 2-layer multi-layer perceptron (MLP) with hidden and output sizes of 8 192 and 768, respectively.

For data augmentation, we apply horizontal flipping, random cropping, and adaptive time masking [30] to the visual inputs, while we only use adaptive time masking for the audio stream. For both streams, we choose a number of masks that is proportional to the utterance length and a maximum masking length of up to 0.4 seconds. For the target vocabulary, we use SentencePiece [31] subword units with a vocabulary size of 5 000. We train the model for 75 epochs with the AdamW [32] optimizer, a cosine learning rate scheduler, and a warm-up of 5 epochs. The peak learning rate is 1e-3. The maximal number of frames in each batch is 1 800 frames. Following [30], our visual-only models are incorporated with a transformer-based language model trained on a corpus of 166 million characters [†] whereas language models are not included for ASR and AV-ASR models since no further improvements are observed.

4. RESULTS

4.1. Do better Librispeech ASR models provide better transcriptions for VSR?

Given that several publicly-available ASR models are available, we use performance on Librispeech as a criterion for model selection. We use models that have achieved state-of-the-art performance on the test-clean set of Librispeech, i.e., Conformer-Transducer [12] and HuBERT [8]. We also use ASR models that are widely used

---

| P | 0 % | 20 % | 40 % | 60 % | 80 % | 100 % |
| U | 0 h | 526 h | 1 052 h | 1 578 h | 2 104 h | 2 630 h |
T	818 h	1 344 h	1 870 h	2 396 h	2 922 h	3 448 h
A	1.5	1.3	1.3	1.1	1.0	1.0
V	33.0	26.6	23.6	21.9	20.0	19.1

Table 2. Impact of the size of additional training data (from AVSpeech and VoxCeleb2) on the WER (%) of audio-only and visual-only models evaluated on LRS3. All models are initialised from a model pre-trained on LRW and trained on LRS2, LRS3 plus X % hours of VoxCeleb2 and AVSpeech. "P" and "U" denote the amount of additional data in percentages and in hours, respectively. "T" denotes the total amount of training data (hours).

in the speech community, wav2vec 2.0 [5] and Whisper [11]. Performance of ASR models on Librispeech clean-test set is shown in the first column of Table 1. Results of the ASR and VSR models trained with the automatically-generated 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. We show that using the transcriptions from most ASR models (i.e., wav2vec 2.0 [5], Whisper [11], and Conformer-Transducer [12]) results in very similar WER for both audio-only and visual-only models. The only exception is the use of automatically-generated transcriptions from HuBERT [8] which results in slightly worse performance despite being one of the best performing models on Librispeech. In this work, we rely on the automatically-generated transcriptions from the Conformer-Transducer [12] since on average it leads to the best performance for both ASR and VSR models.

4.2. Impact of the number of hours of unlabelled data

Table 2 shows the impact of varying the numbers of hours of unlabelled data on the performance of ASR and VSR models on LRS3. An absolute improvement of 1.7 % in WER is observed for VSR by using only labelled data from LRS2 and LRS3 (818 hours) compared to [30]. This gain is likely due to the increase in model capacity. When including 20 % (526 hours) of AVSpeech [18] and VoxCeleb2 [17], the performance of audio- and visual-only models can be further improved to 1.3 % and 26.6 % WER, respectively. Increasing further the number of training hours leads to a further reduction of the WER especially for the VSR model. This is in line with the recent trend observed in the literature [30], where using larger training sets substantially improves performance. In this experiment, we also show that the WER can be improved even by adding data that have been automatically transcribed and inevitably have noisy labels. We also notice that the improvement for the ASR model is marginal when using more than 1 578 hours of unlabelled training data, indicating that the ASR performance may have saturated.

4.3. Comparison with the state-of-the-art

Results on LRS2 and LRS3 are presented in Tables 3 and 4, respectively. For LRS2, it is clear that our visual-only, audio-only and audio-visual models further push the state-of-the-art performance to a WER of 14.6 %, 1.5 % and 1.5 % respectively. For LRS3, the best visual-only model has a WER of 19.1 %, which is outperformed only by [16] (17.0 % WER) which uses 26× more training data. Similarly, our audio-only model establishes a new state-of-the-art [7] by

Method	Type	Extra Data	Total Hours†	WER (%)
MV-WAS [27]				70.4
TDNN [36]		✗	223	48.9
CM-seq2seq [22]				39.1
CM-aux [30]				32.9
CTC/Attention [2]			380	63.5
KD+CTC [9]			995	51.3
KD-seq2seq [10]			818	49.2
TM-seq2seq [1]	V		1 391	48.3
CTC/Attention [37]		✓	60 000	43.2
CM-aux [30]			1 459	25.5
VTP [38]			2 676	22.6
Ours			818	**27.9**
Ours			3 448	**14.6**
TDNN [36]		✗	223	6.7
CM-seq2seq [22]				4.3
CTC/Attention [37]	A		60 000	2.7
Ours		✓	818	**2.6**
Ours			3 448	**1.5**
TDNN [36]		✗	223	5.9
CM-seq2seq [22]				4.2
TM-seq2seq [1]	A+V		1 391	8.3
CTC/Attention [2]		✓	380	7.0
CM-seq2seq [22]			380	3.9
Ours			3 448	**1.5**

Table 3. WER (%) of our audio-only, visual-only and audio-visual models on the LRS2 dataset. † The total hours are counted by including the datasets used for both pre-training and training. Our model trained on 818 hours uses LRW, LRS2 and LRS3. Our model trained on 3 448 hours uses LRW, LRS2, LRS3, VoxCeleb2 and AVSpeech.

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 %.

4.4. Noise experiments

Results of ASR and AV-ASR models, when tested with different acoustic noise levels, are shown in Table. 5. During training we use the babble noise from the NOISEX dataset [39], while the SNR level is selected from [-5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, ∞ dB] with a uniform distribution. For evaluation, we test three types of noise: babble noise [39], pink and white noise from the Speech Commands dataset [40]. We show that, overall, the results are consistent with those presented in [1, 3, 4, 22], 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. We notice that when using babble noise for evaluation, the performance of either audio-only or audio-visual models has a WER lower than 10 % at -7.5 dB. This is likely mainly a consequence of the overlapping noise type in the training and testing phases (despite mismatched levels of noise).

5. CONCLUSIONS

In this work, we propose a simple and efficient method for scaling up audio-visual data for speech recognition. We present a detailed study

Method	Type	Extra Data	Total Hours‡	WER (%)
CM-seq2seq [22]		✗	438	46.9
CM-aux [30]				37.9
Ours				**36.3**
KD+CTC [9]			772	59.8
KD-seq2seq [10]			818	59.0
TM-seq2seq [1]			1 362	58.9
AVHuBERT [7]			1 759	26.9
RNN-T [3]	V		31 000	33.6
VTP [38]		✓	2 676	30.7
ViT3D-CM [16]			90 000	17.0
Ours			818	**33.0**
Ours			1 902	**23.5**
Ours			3 448	**19.1**
CM-seq2seq [22]		✗	438	2.3
RNN-T [3]			31 000	4.5
AV-HuBERT [7]	A		1 759	1.5
Ours		✓	818	**1.5**
Ours			1 902	**1.0**
Ours			3 448	**1.0**
CM-seq2seq [22]		✗	438	2.3
RNN-T [3]			31 000	4.8
AV-HuBERT [7]	A+V		1 759	1.3
ViT3D-CM [16]		✓	90 000	1.6
Ours			1 902	**1.0**
Ours			3 448	**0.9**

Table 4. 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. Our model trained on 818 hours uses LRW, LRS2 and LRS3. Our model trained on 1 902 hours uses LRW, LRS3 and VoxCeleb2. Our model trained on 3 448 hours uses LRW, LRS2, LRS3, VoxCeleb2 and AVSpeech.

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 5. WER (%) of our audio-only and audio-visual models as a function of the noise levels on the LRS3 dataset. The babble noise from the NOISEX dataset [39] is used for training while one of SNR levels from [-5 dB, 0 dB, 5 dB, 10 dB, 15 dB, 20 dB, ∞ dB] is selected with a uniform distribution. For testing, the pink and white noise from the Speech Commands dataset [40] is added to the raw audio waveforms with a specific SNR level. ‡ denotes the noise type used in both training and test set.

on the performance of LRS3 in terms of the amount of unlabelled training data. By leveraging publicly-available ASR models to produce automatically-generated transcriptions, we train an AV-ASR system and achieve state-of-the-art performance on both publicly-available audio-visual benchmarks, LRS2 and LRS3. Furthermore, we show that our audio-visual model is more robust against different levels of noise than its audio-only counterpart.

6. REFERENCES

[1] T. Afouras *et al.*, "Deep audio-visual speech recognition," *IEEE TPAMI*, 2018.

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

[3] T. Makino *et al.*, "Recurrent neural network transducer for audio-visual speech recognition," in *ASRU*, 2019, pp. 905–912.

[4] B. Shi *et al.*, "Robust self-supervised audio-visual speech recognition," in *Interspeech*, 2022, pp. 2118–2122.

[5] A. Baevski *et al.*, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NIPS*, vol. 33, 2020, pp. 12 449–12 460.

[6] P. Ma *et al.*, "LiRA: Learning visual speech representations from audio through self-supervision," in *Interspeech*, 2021, pp. 3011–3015.

[7] B. Shi *et al.*, "Learning audio-visual speech representation by masked multimodal cluster prediction," in *ICLR*, 2022.

[8] W.-N. Hsu *et al.*, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 29, pp. 3451–3460, 2021.

[9] T. Afouras *et al.*, "ASR is all you need: Cross-modal distillation for lip reading," in *ICASSP*, 2020, pp. 2143–2147.

[10] S. Ren *et al.*, "Learning from the master: Distilling cross-modal advanced knowledge for lip reading," in *CVPR*, 2021, pp. 13 325–13 333.

[11] A. Radford *et al.*, *Introducing whisper*, https://openai.com/blog/whisper/, [Online; accessed 18-October-2022], 2022.

[12] O. Kuchaiev *et al.*, "Nemo: A toolkit for building ai applications using neural modules," *arXiv preprint arXiv:1909.09577*, 2019.

[13] Q. Xie *et al.*, "Self-training with noisy student improves imagenet classification," in *CVPR*, 2020, pp. 10 684–10 695.

[14] J. Kahn *et al.*, "Self-training for end-to-end speech recognition," in *ICASSP*, 2020, pp. 7084–7088.

[15] D. S. Park *et al.*, "Improved noisy student training for automatic speech recognition," in *Interspeech*, 2020, pp. 2817–2821.

[16] D. Serdyuk *et al.*, "Transformer-based video front-ends for audio-visual speech recognition for single and multi-person video," in *Interspeech*, 2022, pp. 2833–2837.

[17] J. S. Chung *et al.*, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018, pp. 1086–1090.

[18] A. Ephrat *et al.*, "Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation," *ACM Transactions on Graphics*, vol. 37, no. 4, 112:1–112:11, 2018.

[19] J. Valk *et al.*, "VoxLingua107: A dataset for spoken language recognition," in *SLT*, 2021.

[20] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, 2020, pp. 5036–5040.

[21] A. Vaswani *et al.*, "Attention is all you need," in *NIPS*, 2017, pp. 6000–6010.

[22] P. Ma *et al.*, "End-to-end audio-visual speech recognition with conformers," in *ICASSP*, 2021, pp. 7613–7617.

[23] K. He *et al.*, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[24] T. Stafylakis *et al.*, "Combining residual networks with LSTMs for lipreading," in *Interspeech*, vol. 9, 2017, pp. 3652–3656.

[25] S. Petridis *et al.*, "End-to-end audiovisual speech recognition," in *ICASSP*, 2018, pp. 6548–6552.

[26] S. Watanabe *et al.*, "Hybrid CTC/attention architecture for end-to-end speech recognition," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 8, pp. 1240–1253, 2017.

[27] J. S. Chung *et al.*, "Lip reading sentences in the wild," in *CVPR*, 2017, pp. 3444–3453.

[28] T. Afouras *et al.*, "LRS3-TED: A large-scale dataset for visual speech recognition," *arXiv preprint arXiv:1809.00496*, 2018.

[29] P. Ma *et al.*, "Towards practical lipreading with distilled and efficient models," in *ICASSP*, 2021, pp. 7608–7612.

[30] P. Ma *et al.*, "Visual Speech Recognition for Multiple Languages in the Wild," *Nature Machine Intelligence*, pp. 930–939, 2022.

[31] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *ACL*, 2018, pp. 66–75.

[32] I. Loshchilov *et al.*, "Decoupled weight decay regularization," in *ICLR*, 2019.

[33] V. Panayotov *et al.*, "Librispeech: An ASR corpus based on public domain audio books," in *ICASSP*, 2015, pp. 5206–5210.

[34] F. Hernandez *et al.*, "TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation," in *SPECOM*, vol. 11096, 2018, pp. 198–208.

[35] R. Ardila *et al.*, "Common voice: A massively-multilingual speech corpus," English, in *LREC*, 2020, pp. 4218–4222.

[36] J. Yu *et al.*, "Audio-visual recognition of overlapped speech for the LRS2 dataset," in *ICASSP*, 2020, pp. 6984–6988.

[37] X. Pan *et al.*, "Leveraging uni-modal self-supervised learning for multimodal audio-visual speech recognition," in *ACL*, 2022, 4491–4503.

[38] K. Prajwal *et al.*, "Sub-word level lip reading with visual attention," in *CVPR*, 2022, pp. 5162–5172.

[39] A. Varga *et al.*, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, 1993.

[40] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *CoRR*, vol. abs/1804.03209, 2018.