SynthVSR: Scaling Up Visual Speech Recognition With Synthetic Supervision

Xubo Liu[1*], Egor Lakomkin[2], Konstantinos Vougioukas[2], Pingchuan Ma[2], Honglie Chen[2], Ruiming Xie[2], Morrie Doulaty[2], Niko Moritz[2], Jáchym Kolář[2], Stavros Petridis[2], Maja Pantic[2], Christian Fuegen[2]

[1]University of Surrey [2]Meta AI

Abstract

Recently reported state-of-the-art results in visual speech recognition (VSR) often rely on increasingly large amounts of video data, while the publicly available transcribed video datasets are limited in size. In this paper, for the first time, we study the potential of leveraging synthetic visual data for VSR. Our method, termed SynthVSR, substantially improves the performance of VSR systems with synthetic lip movements. The key idea behind SynthVSR is to leverage a speech-driven lip animation model that generates lip movements conditioned on the input speech. As plenty of transcribed acoustic data and face images are available, we are able to generate large-scale synthetic data using the proposed lip animation model for semi-supervised VSR training. We evaluate the performance of our approach on the largest public VSR benchmark - Lip Reading Sentences 3 (LRS3). SynthVSR achieves a WER of 27.9% when using 438 hours of labeled data from LRS3, which is on par with the state-of-the-art self-supervised AV-HuBERT method. Furthermore, when combined with large-scale pseudo-labeled audio-visual data SynthVSR yields a new state-of-the-art VSR WER of 16.9% using publicly available data only, surpassing the recent state-of-the-art approaches trained with 29 times more non-public machine-transcribed video data (90,000 hours). Project page: `https://liuxubo717.github.io/SynthVSR/`

1. Introduction

Visual speech recognition (VSR) is the task of recognizing speech content based on visual lip movements. VSR has a wide range of applications in real-world scenarios such as improving automatic speech recognition (ASR) in noisy environments. Recently, with the release of large-scale transcribed audio-visual datasets such as LRS2 [1] and LRS3 [2], deep neural networks have become the mainstream approach for VSR. However, even the largest public dataset for English VSR, LRS3, does not exceed 500 hours of transcribed video data. The lack of large-scale transcribed audio-visual datasets potentially results in VSR models that could only work in a laboratory environment [15].

A common solution to this issue is to collect and annotate large-scale audio-visual datasets. For example, [19, 20] collected 90,000 hours of YouTube videos with user-uploaded transcriptions to achieve state-of-the-art performance on standard benchmarks. However, such a procedure is expensive and time-consuming, especially for most of the world's 7,000 languages [21]. If annotations are missing, the ASR can be used to generate the transcriptions automatically and this has been shown to be an effective approach to significantly improve VSR performance [15]. The other promising direction is to learn audio-visual speech representations from large amounts of parallel unlabeled audio-visual data in a self-supervised approach, and then fine-tune them on the limited labeled video dataset [21]. Nevertheless, publicly available video datasets are also limited and their usage may raise license-related[1] concerns, barring their use in commercial applications.

Human perception of speech is inherently multimodal, involving both audition and vision [21]. ASR, which is a complementary task to VSR, has achieved impressive performance in recent years, with tens of thousands of hours of annotated speech datasets [17, 10, 3] available for large-scale training. It is intuitive to ask: *Can we improve VSR with large amounts of transcribed acoustic-only ASR training data?* In this work, we present SynthVSR, a novel semi-supervised framework for VSR. In particular, we first propose a speech-driven lip animation model that can generate synthetic lip movements video conditioned on the speech content. Next, the proposed lip animation model is used to generate synthetic video clips from transcribed speech datasets (e.g., Librispeech [17]) and human face datasets (e.g., CelebA [12]). Then, the synthetic videos together with the corresponding transcriptions are used in combination with the real video-text pairs (e.g., LRS3 [2]) for large-scale semi-supervised VSR training. Synthetic videos provide advantages such as having control over the target text

*Work done during an internship at Meta AI.

[1]Such as LRW [6] and LRS2 [1] datasets which are only permitted for the purpose of academic research.

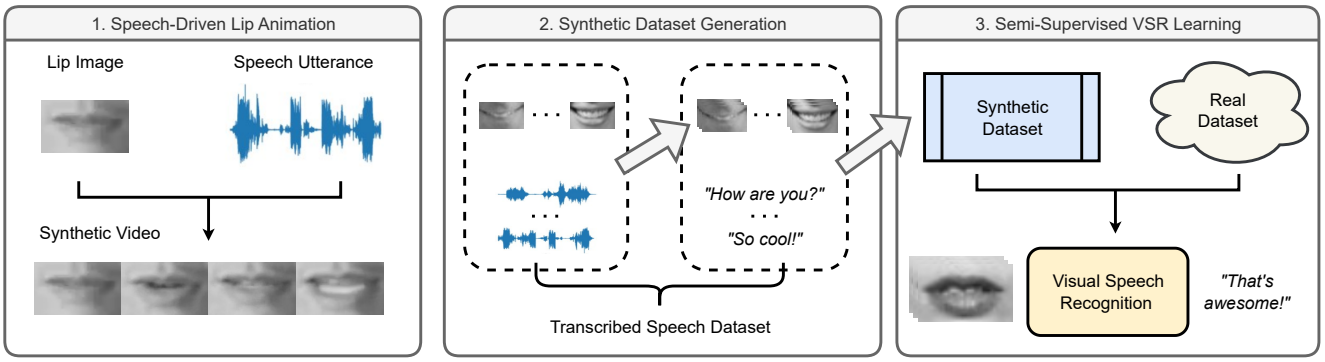

Figure 1. **Scaling up visual speech recognition with synthetic supervision (SynthVSR):** we propose SynthVSR, a semi-supervised framework that can substantially improve the performance of VSR models by using synthetic lip movements. Firstly, we introduce a speech-driven lip animation model that generates lip movement videos conditioned on input lip images and speech utterances (left). Secondly, we generate large-scale synthetic videos using transcribed speech datasets and lip images. The combination of synthetic videos and their corresponding speech transcriptions constitutes the synthetic dataset (centre). Finally, we conduct semi-supervised VSR training with synthetic and real datasets. Our method substantially improves the performance of VSR models with large-scale synthetic data (right).

and lip image as well as the duration of a generated utterance. To the best of our knowledge, the potential of leveraging synthetic visual data for improving VSR has never been studied in the literature.

SynthVSR achieves remarkable performance gains with labeled video data at different scales. We evaluate the performance of SynthVSR on LRS3 with a Conformer-Transformer encoder-decoder VSR model [15]. Using the complete 438 hours from LRS3, SynthVSR achieves a VSR WER of 27.9% which is on par with the state-of-the-art self-supervised method AV-HuBERT-LARGE [21] that uses external 1,759 hours of unlabeled audio-visual data, but with fewer model parameters. Furthermore, following a recent high-resource setup [13] which uses additional 2,630 hours of ASR pseudo-labeled publicly available audio-visual data, our proposed method yields a new state-of-the-art VSR WER of 16.9%, surpassing the former state-of-the-art approaches [19, 20] trained on 90,000 hours of non-public machine-transcribed data.

2. SynthVSR

In this section, we introduce SynthVSR, a novel semi-supervised VSR framework to improve the performance of VSR models using synthetic data, as shown in Figure 1. We will introduce the VSR model, the speech-driven lip animation model next. More details are described in the published CVPR paper [11].

2.1. VSR Model

The VSR model we use in this work is based on [15, 14], which is an encoder-decoder architecture. The encoder is comprised of two components, the visual front-end [9, 23]) and a Conformer [8] encoder. The decoder is based on the transformer architecture [24]. The baseline VSR model is trained end-to-end using a combination of the CTC loss [4, 22] with an attention-based Cross-Entropy (CE) loss.

2.2. Speech-Driven Lip Animation

Inspired by the recent advances in speech-driven facial animation [28, 29, 26, 27], we propose an approach for speech-driven lip animation that generates videos of talking mouth regions conditioned on speech utterances. The output space of the lip animation model is the same as the VSR input space. The proposed lip animation model is based on a temporal GAN [26, 27] with two discriminators. We further propose a VSR perceptual loss when labeled video data is available. The architecture of the speech-driven lip animation model is illustrated in the left part of Figure 2. We will introduce each component in the next sections.

Generator. The generator G is an encoder-decoder structure, as shown in the right part of Figure 2. The generator uses the first frame of a video clip, a speech clip, and a head rotation sequence as inputs. The speech clip is divided into overlapping chunks. The generator produces the corresponding video frame for each speech chunk. Specifically, an image encoder E_i and a speech encoder E_s are used to capture the visual information and speech context into latent embeddings z_i and z_s. The head rotations are provided in the form of sequences of 3D rotation matrices [29] $z_r \in \mathbb{R}^{3\times3}$ with respect to the first frame. The three embeddings z_i, z_s and z_r are concatenated and used to modulate the convolutional layers in the frame decoder D_{frame}, which is similar to StyleGAN2 [25].

Discriminators. The speech-driven lip animation system has two discriminators: frame discriminator D_{img} and sequence discriminator D_{seq}. The frame discriminator operates on the image frame level. and helps enforce visual consistency. The sequence discriminator operates on the sequence level to ensure the temporal consistency of synthetic lip movements. Specifically, the frame discriminator D_{img} is trained on frames that are uniformly sampled from

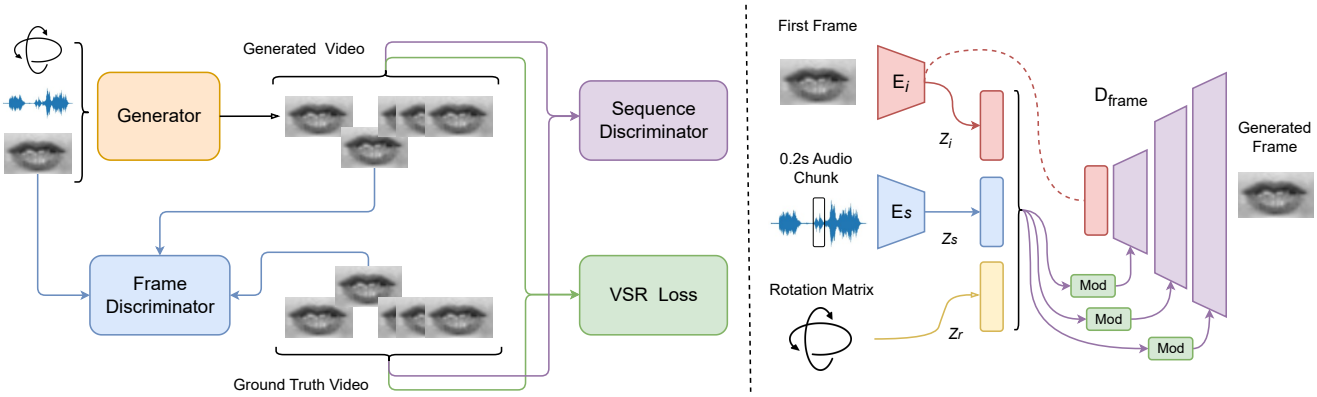

Figure 2. **Architecture of proposed speech-driven lip animation model.** Left: GAN-based speech-driven lip animation model generating lip movements given a lip image, a speech utterance, and a rotation sequence; Right: structure of the generator in the lip animation model.

a video v using a sampling function $S(v)$. The first frame v_1 is fed to the D_{img} as the condition. The input speech signal is s. The adversarial loss of the D_{img} is defined as:

$$\begin{aligned}\mathcal{L}_{Disc}^{img} = \mathbb{E}_v[\log D_{img}(S(v), v_1)] \\ +\mathbb{E}_{v,s}[\log(1 - D_{img}(S(G(s, v_1)), v_1)].\end{aligned} \tag{1}$$

D_{seq} operates on the entire sequence video v. The adversarial loss of the D_{seq} is defined as follows:

$$\begin{aligned}\mathcal{L}_{Disc}^{seq} = \mathbb{E}_v[\log D_{seq}(v)]+ \\ \mathbb{E}_{v,s}[\log(1 - D_{seq}(G(s, v_1))].\end{aligned} \tag{2}$$

VSR Perceptual Loss. We further propose to optimize the lip animation model towards a VSR perceptual loss if labeled video data is available. We first pre-train a VSR model as introduced in Section 2.1. The proposed VSR perceptual loss corresponds to a weighted sum of feature distances computed from the visual front-end and the Transformer decoder of the pre-trained VSR model for real and generated samples. We use L_1 norm to measure the visual embedding distance and Kullback–Leibler (KL) divergence to measure the logits distribution distance, respectively. The VSR perceptual loss is obtained by:

$$\mathcal{L}_{VSR} = \lambda_{visual} \left\| z_f^r - z_f^s \right\|_1 + \lambda_{logits} \, \mathrm{KL}(\hat{y}^r, \hat{y}^s), \tag{3}$$

where z_f^r and z_f^s are the VSR front-end visual features of the real and synthetic video, respectively, $\hat{y}^r$ and $\hat{y}^s$ is the VSR predicted logits distribution of real and synthetic video, respectively, λ_{visual} and λ_{logits} control the weights of these two perceptual losses. The VSR model is frozen during the lip animation model training.

Training Objectives. The speech-driven lip animation model is trained using a combination of a reconstruction loss, adversarial losses, and a VSR perceptual loss. The reconstruction loss is computed based on the L_1 distance between the generated video $\hat{v}$ and ground truth video v:

$$\mathcal{L}_{rec} = \left\| v - \hat{v} \right\|_1 . \tag{4}$$

The overall training loss for the lip animation model is:

$$\begin{aligned}\mathcal{L}_{Animation} = \lambda_{disc}^{img}\mathcal{L}_{disc}^{img} + \lambda_{disc}^{seq}\mathcal{L}_{disc}^{seq} \\ +\lambda_{rec}\mathcal{L}_{rec} + \mathcal{L}_{VSR},\end{aligned} \tag{5}$$

where λ_{disc}^{img}, λ_{disc}^{seq}, and λ_{rec} represent the coefficient of the adversarial loss of frame discriminator D_{img}, the adversarial loss of sequence discriminator D_{seq}, and the reconstruction loss, respectively.

3. Experiments

3.1. Dataset

VSR Benchmark. We conduct experiments on the LRS3 [2] dataset, which is the largest public benchmark for English VSR containing 438.9 hours of video clips from TED talks (408, 30, and 0.9 hours in the pre-training, training-validation, and test set, respectively).

Datasets for Speech-Driven Lip Animation Training. The lip animation model is trained on a combination of LRS3 (pre-training and training-validation splits) and the English subset (933 hours) of AVSpeech [7] datasets.

Datasets for Synthetic Data Generation. We use Librispeech [17], TED-LIUM 3 [10], Common Voice (English split) [3] datasets as the speech sources. We use the CelebA [12] dataset as a source of lip images. For each speech clip, we randomly sample one image from CelebA to generate one synthetic video. We use a static rotation matrix in the generation process. In total, 3,652 hours of synthetic video clips are generated for scaling up VSR training.

3.2. Implementation Details

We describe the implementation details of VSR models and speech-driven lip animation models here. More details are described in the published CVPR paper [11].

VSR Model. We consider two model configurations: (1) Conformer-BASE (250M) with 12-layer Conformer

Method	Backbone	LM	Labeled data (hrs)	Unlabeled data (hrs)	Synthetic data (hrs)	WER (%)
AV-HuBERT-BASE [21]	Transformer	✗	433	1,759	-	34.8
Makino et al. [16]	Transformer	✗	31,000†	-	-	33.6
Ma et al. [15]	Conformer	✓	1,459‡	-	-	31.5
Prajwal et al. [18]	Transformer	✓	2,676†	-	-	30.7
AV-HuBERT-LARGE [21]	Transformer	✗	433	1,759	-	28.6
AV-HuBERT-LARGE w. Self-Training [21]	Transformer	✗	433	1,759	-	26.9
Auto-AVSR [13]	Conformer	✓	3,448‡	-	-	19.1
Serdyuk et al. [20]	Transformer	✗	90,000†	-	-	25.9
Serdyuk et al. [19]	Transformer	✗	90,000†	-	-	17.0
SynthVSR	Conformer-BASE	✗	438	-	-	36.7
		✗	438	-	3,652	28.4
		✓	438	-	3,652	**27.9**
		✗	3,068	-	-	21.2
		✗	3,068	-	3,652	19.4
		✓	3,068	-	3,652	**18.7**
SynthVSR	Conformer-LARGE	✗	3,068	-	3,652	18.2
		✓	3,068	-	3,652	**16.9**

Table 1. Experimental results of LRS3 & high-resource labeled data setting on LRS3 (test). LM denotes whether or not a language model is used in the decoding. †Includes non-publicly available data. ‡Includes datasets that are only permitted for the purpose of academic research. hrs is an abbreviation for hours.

encoder, 6-layer Transformer decoder; (2) Conformer-LARGE (783M) with 24-layers Conformer encoder, 9-layer Transformer decoder.

Speech-Driven Lip Animation. We consider two lip animation model configurations. First, The lip animation model with the BASE VSR model trained on LRS3 is referred to as LAM-LRS3-VSR-VL. Second, the lip animation model with the BASE VSR model trained on LRS3 and 2,630 hours of pseudo-labeled AVSpeech and VoxCeleb2 is referred to as LAM-LRS3-AVoX-VSR.

3.3. LRS3 Labeled Data Setting

We report the results when using the full 438 hours of LRS3. Experiments are conducted on the BASE VSR model. The results are shown in Table 1. Our BASE model achieves WER 36.7% when using 438 hours of LRS3 data for training. We generate 3,652 hours of synthetic data using the LAM-LRS3-VSR-VL model. Using 3,652 hours of synthetic data and 438 hours of LRS3 labeled data, the BASE VSR model achieves the WER 27.9% (28.4% w/o language model, corresponding to a WER reduction of 8.3%). Our method outperforms three recent approaches using 31,000 (33.6%), 1,459 (31.5%), and 2,679 (30.7%) hours of labeled data, respectively. When compared with the state-of-the-art self-supervised method AV-HuBERT [21] that uses an additional 1,759 hours of unlabeled audio-visual data, our method outperforms the AV-HuBERT-BASE model (34.8%) by a large margin. Our method slightly performs better than the AV-HuBERT-LARGE model (28.6%), but with fewer model parameters (our BASE model 250M vs AV-HuBERT-LARGE 390M). Note that we compare with the AV-HuBERT results without self-training as we do not use the pseudo-labeled 933 hours of AVSpeech subset for VSR training.

3.4. High-Resource Labeled Data Setting

We further evaluate the scalability of SynthVSR: when using machine-transcribed AVSpeech [7] and VoxCeleb2 [5] as additional training data. We use the 3,652 hours of synthetic data generated by LAM-LRS3-AVoX-VSR and conduct experiments on BASE and LARGE models. We first train the BASE model with 438 hours of labeled LRS3 and 2,630 hours of pseudo-labeled data, resulting in a strong VSR baseline with the WER 21.2%. By using additional 3,652 hours of synthetic data, the WER of the BASE model improves to 18.7% (19.4% w./o. language model), which outperforms [13] that uses additional labeled dataset LRS2 (223 hours) and LRW (157 hours) for training. Although the VSR model has seen a large amount of labeled data, and the speech-driven lip animation model is trained from part of the VSR training data, synthetic data can still lead to considerable performance gains. Furthermore, increasing the model size from BASE to LARGE results in better VSR performance with the WER of 16.9% (18.2% w./o. language model), which is the current state-of-the-art performance on LRS3, with publicly available data only.

4. Conclusion

We have presented a semi-supervised method for VSR enhanced with synthetic lip movements. The speech-driven lip animation model is proposed to generate synthetic video data from labeled speech datasets and face images for scaling up VSR. Our method achieves state-of-the-art results on LRS3, outperforming prior work trained on more labeled or unlabeled real video data. Our work fosters future research on generating and exploiting synthetic visual data for VSR.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

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

[3] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation*, 2020.

[4] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. LipNet: Sentence-level lipreading. *arXiv:1611.01599*, 2016.

[5] Joon Son Chung, Arsha Nagrani, and Andrew Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

[6] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, pages 87–103. Springer, 2016.

[7] 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. *arXiv:1804.03619*, 2018.

[8] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. In *INTERSPEECH*, 2020.

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

[10] François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *International Conference on Speech and Computer*, pages 198–208, 2018.

[11] Xubo Liu, Egor Lakomkin, Konstantinos Vougioukas, Pingchuan Ma, Honglie Chen, Ruiming Xie, Morrie Doulaty, Niko Moritz, Jachym Kolar, Stavros Petridis, et al. Synthvsr: Scaling up visual speech recognition with synthetic supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18806–18815, 2023.

[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision*, December 2015.

[13] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. Auto-AVSR: Audio-visual speech recognition with automatic labels. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023.

[14] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with Conformers. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7613–7617, 2021.

[15] Pingchuan Ma, Stavros Petridis, and Maja Pantic. Visual speech recognition for multiple languages in the wild. *Nature Machine Intelligence*, 2022.

[16] 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 *IEEE Automatic Speech Recognition and Understanding Workshop*, pages 905–912, 2019.

[17] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5206–5210, 2015.

[18] KR Prajwal, Triantafyllos Afouras, and Andrew Zisserman. Sub-word level lip reading with visual attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5162–5172, 2022.

[19] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. Transformer-based video front-ends for audio-visual speech recognition. In *INTERSPEECH*, 2022.

[20] Dmitriy Serdyuk, Olivier Siohan, and Otavio de Pinho Forin Braga. Audio-visual speech recognition is worth 32x32x8 voxels. In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2021.

[21] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *International Conference on Learning Representations*, 2022.

[22] Brendan Shillingford, Yannis Assael, Matthew W Hoffman, Thomas Paine, Cían Hughes, Utsav Prabhu, Hank Liao, Hasim Sak, Kanishka Rao, Lorrayne Bennett, et al. Large-scale visual speech recognition. In *INTERSPEECH*, 2019.

[23] Themos Stafylakis and Georgios Tzimiropoulos. Combining residual networks with LSTMs for lipreading. *arXiv:1703.04105*, 2017.

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[25] Yuri Viazovetskyi, Vladimir Ivashkin, and Evgeny Kashin. StyleGAN2 distillation for feed-forward image manipulation. In *European Conference on Computer Vision*, pages 170–186. Springer, 2020.

[26] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. End-to-end speech-driven facial animation with temporal GANs. In *British Machine Vision Conference*, 2018.

[27] Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Realistic speech-driven facial animation with GANs. *International Journal of Computer Vision*, 128(5):1398–1413, 2020.

[28] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9299–9306, 2019.

[29] Hang Zhou, Yasheng Sun, Wayne Wu, Chen Change Loy, Xiaogang Wang, and Ziwei Liu. Pose-controllable talking face generation by implicitly modularized audio-visual representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4176–4186, 2021.