SVTS: scalable video-to-speech synthesis

Rodrigo Schoburg Carrillo de Mira, Alexandros Haliassos, Stavros Petridis, Björn W. Schuller, Maja Pantic

Angaben zur Veröffentlichung / Publication details:

Schoburg Carrillo de Mira, Rodrigo, Alexandros Haliassos, Stavros Petridis, Björn W. Schuller, and Maja Pantic. 2022. "SVTS: scalable video-to-speech synthesis." In *Interspeech 2022, Incheon, Korea, 18-22 September 2022*, edited by Hanseok Ko and John H. L. Hansen, 1836–40. Baixas: ISCA. https://doi.org/10.21437/interspeech.2022-10770.

Interspeech 2022
18-22 September 2022, Incheon, Korea

SVTS: Scalable Video-to-Speech Synthesis

Rodrigo Mira[1], *Alexandros Haliassos*[1], *Stavros Petridis*[1], *Björn W. Schuller*[1,2], *Maja Pantic*[1]

[1]Imperial College London, UK
[2]Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg, Germany

{rs2517,ah2214,stavros.petridis04,bjoern.schuller,m.pantic}@imperial.ac.uk

Abstract

Video-to-speech synthesis (also known as lip-to-speech) refers to the translation of silent lip movements into the corresponding audio. This task has received an increasing amount of attention due to its self-supervised nature (i. e., can be trained without manual labelling) combined with the ever-growing collection of audio-visual data available online. Despite these strong motivations, contemporary video-to-speech works focus mainly on small- to medium-sized corpora with substantial constraints in both vocabulary and setting. In this work, we introduce a scalable video-to-speech framework consisting of two components: a video-to-spectrogram predictor and a pre-trained neural vocoder, which converts the mel-frequency spectrograms into waveform audio. We achieve state-of-the art results for GRID and considerably outperform previous approaches on LRW. More importantly, by focusing on spectrogram prediction using a simple feedforward model, we can efficiently and effectively scale our method to very large and unconstrained datasets: To the best of our knowledge, we are the first to show intelligible results on the challenging LRS3 dataset.

Index Terms: video-to-speech, lip-to-speech, speech synthesis, neural vocoder, conformer.

1. Introduction

Lipreading, also known as visual speech recognition (VSR), is defined as the prediction of text transcriptions from silent video of lip movements. The advent of deep learning has enabled practitioners to shift from using only very constrained datasets [3] to training models for lipreading in the wild [22]. The progress in lipreading as well as text-to-speech (TTS) [34] has drawn attention to the idea of predicting speech from silent video directly. This task, known as video-to-speech synthesis, has many impactful applications, such as generating clean speech when videoconferencing under noisy conditions, and helping people suffering from aphonia, who are unable to produce voiced speech. Although video-to-speech can be achieved through a combination of lipreading and text-to-speech, directly predicting speech obviates the need for labels (text transcriptions), meaning that it can be trained on raw video only.

To the best of our knowledge, the first work to train a neural network for video-to-speech synthesis was [8], which predicts the audio clip's spectral envelope from a set of visual features extracted from video. It uses a stack of fully connected layers and feeds this envelope into a vocoder to produce voiced speech. This work was later extended in [7], achieving substantially more intelligible results for a single-speaker subset of GRID [6]. Following this, [10] (an extension of another early video-to-speech approach [11]) was the first to train and evaluate on multiple speakers (in this case, a 4-speaker subset of GRID), achieving a major leap forward in the realism of its outputs. This method set two trends which are widely adopted in following works: predicting speech features directly from raw video, rather than from manually extracted visual features [2, 16, 18, 23, 27, 32, 37–39], and using mel-frequency spectrograms as an intermediate representation [2, 27, 32, 39], which are then converted into raw waveform using the Griffin-Lim algorithm [12]. Notable exceptions include [38], which proposes an end-to-end video-to-waveform generative adversarial network (GAN) capable of producing intelligible speech from raw video without the need for a separate spectrogram-to-waveform system, and [23], which uses a traditional vocoder to synthesize speech, rather than a spectrogram-based approach.

Remarkably, most recent works focus on corpora with small pools of speakers, constrained vocabularies, and video recorded in studio conditions (e. g., 4-Speaker GRID and 3-Lipspeaker TCD-TIMIT [14]) [2, 16, 23, 27, 37–39], achieving improvements in performance via the use of intricate loss ensembles [18, 24, 37] and complex architectures [16, 32, 37, 39]. While these developments are meaningful within ideal conditions, they fail to leverage the massive amount of audio-visual data available publicly, and propose training procedures which do not easily scale to very large datasets [18, 24]. In this work, we aim to address these issues by proposing a simple video-to-speech system which efficiently scales with more data. It consists of a video-to-spectrogram predictor followed by a spectrogram-to-waveform synthesizer. The former is a ResNet18+conformer network [13, 15], which becomes deeper and wider for larger datasets and is trained using a combination of two established comparative losses. The latter is a pre-trained neural vocoder, which accurately synthesizes the corresponding audio waveform with a low computational overhead.

Our contributions are as follows: **(1)** We present a simple and effective video-to-speech approach that can easily scale to large and complex datasets. **(2)** We conduct a detailed ablation study demonstrating the differences between commonly-used spectrogram inversion methods, as well as validating our choice of loss functions. **(3)** We outperform previous approaches on most metrics on the small but popular GRID dataset and achieve state-of-the-art performance on the larger LRW dataset. **(4)** To the best of our knowledge, we are the first to present intelligible results on the challenging LRS3 [1] dataset, and show that scaling our model even further with a combination of LRS3 and VoxCeleb2 [5] (containing more than 1,500 hours of data) yields significant improvements.

2. Methodology

2.1. Video-to-spectrogram model

Our spectrogram predictor comprises two main components: (1) a visual encoder composed of a 3D convolutional stem followed by a standard 2D ResNet-18 [15], as in [22], and (2) a conformer [13], which receives the features from the visual encoder and aims to model the temporal correlations between

10.21437/Interspeech.2022-10770

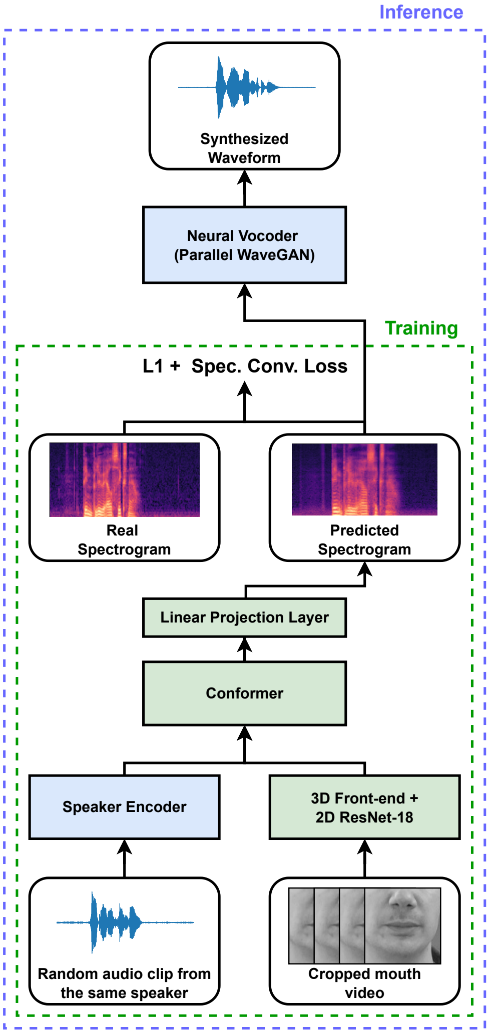

Figure 1: *Summary of our video-to-speech synthesis approach during training and inference. In this figure, the components pictured in blue are pre-trained and kept frozen, while the components pictured in green are trained from scratch.*

them. The latter contains an initial linear layer, followed by a set of conformer blocks which vary in depth and width based on the model version, as shown in Table 1. Finally, each feature vector, corresponding to a video frame, is projected into a hidden size of 320 using a linear projection layer, and reshaped into 4 × 80 spectrogram frames. The input video is sampled at 20 fps and the extracted spectrogram contains 80 frames per second. We train our predictor using a combination of the L_1 loss and the spectral convergence loss [40].

As in multi-speaker text-to-speech systems [17], our video-to-speech model requires information about the speaker's voice characteristics, which cannot be derived accurately from silent video only. To this end, we use a pre-trained speaker encoder[1] originally trained for speaker verification on a combination of VoxCeleb [26], VoxCeleb2 [5], and Librispeech [29]. For each video clip, an embedding is extracted from a randomly selected audio clip from the same speaker and concatenated with the visual features extracted by the visual encoder, which are then fed into the conformer. Note that the speaker encoder is kept frozen during training.

Table 1: *Summary of our proposed SVTS architectures.* * *refers to the total number of parameters in the model (ResNet + conformer + projection layer)*

Model	**SVTS-S**	**SVTS-M**	**SVTS-L**
Num. parameters* (M)	27.3	43.1	87.6
Conformer blocks	6	12	12
Attention dim.	256	256	512
Attention heads	4	4	8
Conv. kernel size	31	31	31
Feedforward dim.	2048	2048	2048

2.2. Spectrogram-to-waveform

In order to generate waveform speech from the spectrograms, we opt for the use of a neural vocoder, specifically Parallel WaveGAN [40]. This WaveNet-based [28] model is trained using a combination of comparative and adversarial losses. We employ a version pre-trained on LibriTTS [42] for 1 million iterations. Note that it is used only at inference time, allowing for a substantially simpler training procedure than related video-to-speech works, which train their own vocoder from scratch [16, 37]. An overview of our approach is illustrated in Figure 1.

3. Experimental setup

3.1. Datasets

The first corpus we experiment with is GRID, which has become an established benchmark in video-to-speech literature due to its small vocabulary, predictable structure, and clean recording conditions. GRID is composed of 1,000 unique sentences (with a small vocabulary of 51 words) uttered by 33 speakers; this amounts to roughly 27 hours of audio-visual speech. We experiment with two versions of the dataset: (1) a seen speaker version, originally proposed in [24], where the 33 speakers are present in the training, validation, and testing sets, and (2) an unseen speaker version, introduced in [38], where there is no overlap in the speakers between the sets.

The second corpus is LRW, which features around 150 hours of single-word utterances from hundreds of different speakers recorded 'in the wild.' Although its 500-word vocabulary is not extensive, the filming conditions are significantly less controlled than GRID, with varying lighting, head poses, and background noise. As a result, LRW is considered more challenging than GRID and is substantially closer to a real-world scenario. Due to LRW's lack of speaker labels, it is not possible to select a random audio clip from the same speaker to produce the corresponding speaker embedding. Therefore, for this corpus we generate the speaker embeddings using the audio clip from the corresponding video, which is consistent with previous multi-speaker video-to-speech approaches on LRW [32].

To demonstrate our method's scalability to even larger and less constrained datasets, we run experiments on the 312-hour-long LRS3 dataset. It contains long sentences, a diverse vocabulary of more than 50,000 words, and thousands of speakers. As in GRID, we use two different versions of LRS3: seen speaker, where all speakers' utterances are split into training, validation and testing sets using a 80 – 10 – 10 % ratio, and unseen speaker, following the original split proposed in [1]. Finally, we experiment with combining the LRS3 training dataset with an English-only version [35] of VoxCeleb2 (while keeping the same LRS3 validation and test sets to ease comparison),

[1] https://github.com/CorentinJ/Real-Time-Voice-Cloning.

amounting to around 1,550 hours of footage. For both corpora, utterances exceeding 24 seconds are excluded from training due to hardware limitations.

3.2. Data pre-processing and augmentation

In order to produce the cropped mouth video, we first extract 68-point landmarks using RetinaFace[2] [9] and a pre-trained 2D-FAN[3] [4]. We average the landmarks across 12 frames through a sliding window to reduce motion jitter, and align each frame to the mean face. We then crop a 96×96 region centred around the mouth and convert the frames to grayscale. The audio is sampled at 24 kHz, and the log-mel spectrograms are extracted using 80 mel bands, frequency bins of size 2048, a hop size of 12.5 ms, a window length of 50 ms, and a Hann window.

During training, we apply random cropping of size 88×88, horizontal flipping with probability of 0.5, and random erasing with a probability of 0.5. The erased area is randomly sampled between 2 and 33 % of the full frame, with an aspect ratio ranging from 0.3 to 3.3. During testing, we perform center cropping of size 88×88. For our LRS3 experiments, we apply time-masking by randomly replacing each frame with the average pixel value in the video, since we find it aids generalization when training on long sentences. We apply one contiguous time-mask for each second of the utterance, and each mask's length is uniformly sampled from 0 to 0.4 seconds.

3.3. Training details

For our GRID and LRW experiments, we train our models using AdamW [19] with a learning rate of 1×10^{-3}, $\beta_1 = 0.9$, $\beta_2 = 0.98$, and a weight decay of 1×10^{-2}. We warm up the learning rate for the first 10 % of iterations, and then decay it with a cosine schedule [20]. For LRS3 seen speakers, we use a maximum learning rate of 7×10^{-3}, while for unseen speakers (including the combination with VoxCeleb2) we use 1×10^{-3}. We train for a total of 200, 150, 500, and 150 epochs for GRID, LRW, LRS3 seen speakers, and LRS3 unseen speakers, respectively. We save a checkpoint at the end of each epoch, and at the end of training select the one with the lowest validation loss.

3.4. Evaluation metrics

We measure the quality and accuracy of our generated samples via 4 objective metrics. The first is Perceptual Evaluation of Speech Quality (PESQ)[4] [33], which aims to measure the clarity and perceptual quality of the generated samples. We also use Short-Time Objective Intelligibility (STOI)[5] [36] and its extended version ESTOI to measure the intelligibility of our samples.

The final metric we apply is word error rate (WER), which has become a benchmark in video-to-speech after its introduction in [38]. It is measured by applying a pre-trained speech recognition model to the generated samples, and comparing the predicted transcription with the ground truth. Hence, WER serves as an easily interpretable intelligibility metric for the generated samples. We propose to forego the use of manual text transcriptions, and use instead the transcription predicted from the corresponding real audio as the ground truth. This increases the interpretability of the reported numbers, as they are a direct measure of the difference in intelligibility between real and generated audio, and it also removes the requirement for labelled datasets in future work. For our GRID experiments, we use a model pre-trained on LRW, LRS2, and LRS3 [21], and fine-tuned on GRID (adopting the split from [3]); it achieves a WER of 0.1 % on the real audio test set. For LRW, we use an ASR model trained only on LRW [31] with a WER of 1.68 %.

Although these metrics are commonly referenced in video-to-speech works and are therefore useful for comparison, it is widely known that no objective speech metric correlates perfectly with human perception of quality and intelligibility [38]. Therefore, we highly encourage readers to listen to the generated samples available on our project website[6] rather than rely solely on the reported metrics.

4. Results

4.1. Experiments

Our results are presented in Table 2. We begin by discussing our findings on the small-scale GRID dataset. For the seen speaker split, our SVTS-S model clearly outperforms our previous approach [24], as well as the more recent [18], on STOI and ESTOI. It also achieves a significant improvement in WER. These metrics indicate that our samples are more intelligible than previous works. On unseen speakers, our model achieves a better PESQ, STOI, and ESTOI but is outperformed by our previous GAN-based work [24] in WER. By perceptually evaluating the generated samples, we find that our seen speaker reconstruction is highly realistic and could be mistaken for real audio. On the other hand our unseen speaker samples sound considerably less noisy than previous works and capture the unseen speaker's voice with remarkable accuracy, thanks to our speaker embedding strategy.

On the more challenging and diverse LRW dataset, SVTS-M is superior to previous approaches on all metrics by a wide margin. We achieve a low WER of 13.4 %, indicating that our samples are consistently intelligible. Perceptually, we find that our samples sound substantially more realistic and accurate than previous approaches, including our GAN-based approach [24]. This strong performance is a consequence of our SVTS architecture, which allows us to efficiently scale to this larger dataset.

Finally, we experiment with LRS3, which is undoubtedly the most challenging corpus we approach, as discussed in Section 3.1. On the seen speaker setting, we find that our model achieves reasonable PESQ, STOI and ESTOI performance, comparable to what had been reported by previous works on LRW. The unseen speaker protocol is naturally more challenging, and therefore does not achieve the same level of quality. Interestingly, we find that results are greatly improved with the addition of the VoxCeleb2 data, as shown by the significant boost on all metrics. This empirically demonstrates our model's ability to improve its reconstructions by leveraging additional training data, even if its distribution is different from the testing set (which only contains samples from LRS3). It also suggests that we may have not yet reached a saturation point: There are likely still gains to be made in the future with even more data.

Perceptually, we find that the most intelligible samples are produced by our seen speaker model, closely followed by our model trained on LRS3+VoxCeleb2. Although there is room for improvement, we find that most syllables in the reconstructed speech are discernible, and each speaker's voice profile is re-

[2] https://github.com/biubug6/Pytorch_Retinaface

[3] https://github.com/1adrianb/face-alignment

[4] https://github.com/ludlows/python-pesq

[5] https://github.com/mpariente/pystoi

[6] https://sites.google.com/view/scalable-vts

Table 2: *Summary of our results. Due to LRS3's complex vocabulary and long sentence structure, we are unable to find a speech recognition model that works accurately on our generated samples (e. g., the word "teacher" is often mistaken for "teachers"), and therefore do not report WER for this dataset. *reported using Google speech-to-text API.*

Method	Corpus	Speaker split (seen/unseen)	Training data (hours)	PESQ	STOI	ESTOI	WER (%)
End-to-end GAN [24]	GRID	seen	24	1.70	0.667	0.466	4.60
VCA-GAN + Griffin-Lim [18]	GRID	seen	20	**1.97**	0.695	0.505	5.13
SVTS-S	GRID	seen	24	**1.97**	**0.705**	**0.523**	**2.36**
End-to-end GAN [38]	GRID	unseen	13	1.26	0.494	0.198	32.79
Conv. + GRU + WORLD vocoder [23]	GRID	unseen	13	1.26	0.541	0.227	38.15
End-to-end GAN [24]	GRID	unseen	13	1.37	0.568	0.289	**16.12**
VCA-GAN + Griffin-Lim [18]	GRID	unseen	13	1.39	0.570	0.282	24.57
Conv. + LSTM + WaveNet [16]	GRID	unseen	13	1.33	0.531	0.271	26.17
SVTS-S	GRID	unseen	13	**1.40**	**0.588**	**0.318**	17.85
Conv. + LSTM + Griffin-Lim [32]	LRW	unseen	157	1.20	0.543	0.344	34.20*
End-to-end GAN [24]	LRW	unseen	157	1.33	0.552	0.330	42.60
VCA-GAN + Griffin-Lim [18]	LRW	unseen	157	1.34	0.565	0.364	37.07
SVTS-M	LRW	unseen	157	**1.49**	**0.649**	**0.483**	**13.40**
SVTS-L	LRS3	seen	256	**1.30**	**0.553**	**0.331**	-
SVTS-L	LRS3	unseen	296	1.25	0.507	0.271	-
SVTS-L	LRS3 + VoxCeleb2	unseen	1556	**1.26**	**0.530**	**0.313**	-

Table 3: *Vocoder ablation on GRID (seen speakers). Speed is measured on an Nvidia RTX 2080 Ti. *computed on CPU* .

Metric	PESQ	STOI	ESTOI	WER (%)	Speed (clips/sec.)
Griffin-Lim* [12]	**2.00**	0.696	0.513	2.41	1.2
Multiband MelGAN [41]	1.86	0.683	0.487	2.50	**184.9**
Style MelGAN [25]	1.93	0.702	0.520	2.38	83.7
Parallel WaveGAN [40]	1.97	**0.705**	**0.523**	**2.36**	54.7

Table 4: *Loss ablation on GRID (seen speakers).* .

Metric	PESQ	STOI	ESTOI	WER (%)
w/o Spec. Conv.	**1.97**	**0.705**	**0.523**	2.90
w/o L_1	1.91	0.700	0.514	2.74
L_1+Spec. Conv.	**1.97**	**0.705**	**0.523**	**2.36**

produced with considerable accuracy, which is particularly impressive in the unseen speaker scenario.

4.2. Ablations

In order to motivate our use of Parallel WaveGAN (PWG) as our waveform synthesis model, we compare it in Table 3 with other recently proposed neural vocoders as well as the commonly used Griffin-Lim algorithm. All models, including our version of PWG, are pre-trained on LibriTTS and are publicly available[7]. The Griffin-Lim synthesis is performed using the fast version of the algorithm[8] [30], and runs for 30 iterations. It can be observed that Parallel WaveGAN outperforms its peers Multiband Melgan [41] and Style Melgan [25] on all four evaluation metrics. Furthermore, through perceptual evaluation, we find that PWG produces substantially more realistic audio. Regarding Griffin-Lim, although it achieves a slightly higher PESQ score, we find that it consistently produces noisy speech with frequent artifacts. This highlights the limitations of PESQ as a metric, as it is often not sensitive to artifacts that are immediately noticeable to human listeners. Thanks to efficient GPU implementations, the vocoders are roughly 50× faster than Griffin-Lim, with the fastest vocoder, Multiband Melgan, being able to process almost 200 GRID clips per second.

In Table 4, we experiment with each of our loss functions separately and compare with the combined loss (baseline). We find that the baseline's performance is roughly similar to the individual losses on PESQ, STOI and ESTOI, but is clearly superior on WER. Interestingly, we find that our model achieves comparable performance with only an L_1 loss, which contrasts greatly with previous approaches' reliance on elaborate loss combinations [24, 37].

5. Conclusion

In this paper, we propose SVTS, a scalable approach for video-to-speech synthesis. We present three architectures of varying sizes, which allow us to efficiently adapt our training procedure to datasets ranging from GRID (27 hours) to LRS3+VoxCeleb2 ($>$ 1,500 hours). We show that our method outperforms previous approaches on most metrics for two popular versions of GRID, and establishes a new state-of-the-art for LRW. Finally, we experiment with the large and unconstrained LRS3 corpus, achieving intelligible results, and combine it with VoxCeleb2 to further improve our performance, demonstrating our method's scalability. We hope our work will encourage a shift towards larger corpora, as this aligns with the current ubiquity of unlabelled audio-visual data.

6. Acknowledgements

We would like to thank Pingchuan Ma for sharing his speech recognition models. All training, testing, and ablation studies have been conducted at Imperial College London.

[7] https://github.com/kan-bayashi/ParallelWaveGAN

[8] https://librosa.org/doc/main/generated/librosa.griffinlim.html

7. References

[1] T. Afouras, J. S. Chung, and A. Zisserman, "Lrs3-ted: A large-scale dataset for visual speech recognition," in *arXiv preprint arXiv:1809.00496*, 2018.

[2] H. Akbari, H. Arora, L. Cao, *et al.*, "Lip2audspec: Speech reconstruction from silent lip movements video," in *ICASSP*, IEEE, 2018, pp. 2516–2520.

[3] Y. M. Assael, B. Shillingford, S. Whiteson, *et al.*, "Lipnet: Sentence-level lipreading," *CoRR*, vol. abs/1611.01599, 2016.

[4] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks)," in *ICCV*, 2017.

[5] J. S. Chung, A. Nagrani, and A. Zisserman, "Voxceleb2: Deep speaker recognition," in *Interspeech*, 2018.

[6] M. Cooke, J. Barker, S. Cunningham, *et al.*, "An audio-visual corpus for speech perception and automatic speech recognition (l)," *The Journal of the Acoustical Society of America*, vol. 120, pp. 2421–4, 2006.

[7] T. L. Cornu and B. Milner, "Generating intelligible audio speech from visual speech," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 25, no. 9, pp. 1751–1761, 2017.

[8] ——, "Reconstructing intelligible audio speech from visual speech features," in *Interspeech*, ISCA, 2015, pp. 3355–3359.

[9] J. Deng, J. Guo, E. Ververas, *et al.*, "Retinaface: Single-shot multi-level face localisation in the wild," in *CVPR*, Computer Vision Foundation / IEEE, 2020, pp. 5202–5211.

[10] A. Ephrat, T. Halperin, and S. Peleg, "Improved speech reconstruction from silent video," in *ICCV*, IEEE Computer Society, 2017, pp. 455–462.

[11] A. Ephrat and S. Peleg, "Vid2speech: Speech reconstruction from silent video," in *ICASSP*, IEEE, 2017, pp. 5095–5099.

[12] D. Griffin and Jae Lim, "Signal estimation from modified short-time fourier transform," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

[13] A. Gulati, J. Qin, C. Chiu, *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech*, H. Meng, B. Xu, and T. F. Zheng, Eds., ISCA, 2020, pp. 5036–5040.

[14] N. Harte and E. Gillen, "Tcd-timit: An audio-visual corpus of continuous speech," *IEEE Transactions on Multimedia*, vol. 17, no. 5, pp. 603–615, 2015.

[15] K. He, X. Zhang, S. Ren, *et al.*, "Deep residual learning for image recognition," in *CVPR*, IEEE Computer Society, 2016, pp. 770–778.

[16] J. Hong, M. Kim, S. J. Park, *et al.*, "Speech reconstruction with reminiscent sound via visual voice memory," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 29, pp. 3654–3667, 2021.

[17] Y. Jia, Y. Zhang, R. J. Weiss, *et al.*, "Transfer learning from speaker verification to multispeaker text-to-speech synthesis," in *NeurIPS*, S. Bengio, H. M. Wallach, H. Larochelle, *et al.*, Eds., 2018, pp. 4485–4495.

[18] M. Kim, J. Hong, and Y. M. Ro, "Lip to speech synthesis with visual context attentional GAN," in *NeurIPS*, A. Beygelzimer, Y. Dauphin, P. Liang, *et al.*, Eds., 2021.

[19] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[20] ——, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[21] P. Ma, S. Petridis, and M. Pantic, "End-to-end audio-visual speech recognition with conformers," in *ICASSP*, 2021, pp. 7613–7617.

[22] ——, "Visual speech recognition for multiple languages in the wild," *CoRR*, vol. abs/2202.13084, 2022.

[23] D. Michelsanti, O. Slizovskaia, G. Haro, *et al.*, "Vocoder-based speech synthesis from silent videos," in *Interspeech*, H. Meng, B. Xu, and T. F. Zheng, Eds., ISCA, 2020, pp. 3530–3534.

[24] R. Mira, K. Vougioukas, P. Ma, *et al.*, "End-to-end video-to-speech synthesis using generative adversarial networks," *IEEE Transactions on Cybernetics*, pp. 1–13, 2022.

[25] A. Mustafa, N. Pia, and G. Fuchs, "Stylemelgan: An efficient high-fidelity adversarial vocoder with temporal adaptive normalization," in *ICASSP*, IEEE, 2021, pp. 6034–6038.

[26] A. Nagrani, J. S. Chung, and A. Zisserman, "Voxceleb: A large-scale speaker identification dataset," in *Interspeech*, F. Lacerda, Ed., ISCA, 2017, pp. 2616–2620.

[27] D. Oneata, A. Stan, and H. Cucu, "Speaker disentanglement in video-to-speech conversion," in *EUSIPCO*, IEEE, 2021, pp. 46–50.

[28] A. van den Oord, S. Dieleman, H. Zen, *et al.*, "Wavenet: A generative model for raw audio," in *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, ISCA, 2016, p. 125.

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

[30] N. Perraudin, P. Balázs, and P. L. Søndergaard, "A fast griffin-lim algorithm," in *WASPAA*, IEEE, 2013, pp. 1–4.

[31] S. Petridis, T. Stafylakis, P. Ma, *et al.*, "End-to-end audiovisual speech recognition," in *ICASSP*, IEEE, 2018, pp. 6548–6552.

[32] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, *et al.*, "Learning individual speaking styles for accurate lip to speech synthesis," in *CVPR*, Computer Vision Foundation / IEEE, 2020, pp. 13 793–13 802.

[33] A. W. Rix, J. G. Beerends, M. P. Hollier, *et al.*, "Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs," in *ICASSP*, 2001.

[34] J. Shen, R. Pang, R. J. Weiss, *et al.*, "Natural TTS synthesis by conditioning wavenet on MEL spectrogram predictions," in *ICASSP*, IEEE, 2018, pp. 4779–4783.

[35] B. Shi, W. Hsu, K. Lakhotia, *et al.*, "Learning audio-visual speech representation by masked multimodal cluster prediction," *CoRR*, vol. abs/2201.02184, 2022.

[36] C. H. Taal, R. C. Hendriks, R. Heusdens, *et al.*, "An algorithm for intelligibility prediction of time-frequency weighted noisy speech," *IEEE Trans. Speech Audio Process.*, vol. 19, no. 7, pp. 2125–2136, 2011.

[37] S. Um, J. Kim, J. Lee, *et al.*, "Facetron: Multi-speaker face-to-speech model based on cross-modal latent representations," *CoRR*, vol. abs/2107.12003, 2021.

[38] K. Vougioukas, P. Ma, S. Petridis, *et al.*, "Video-driven speech reconstruction using generative adversarial networks," in *Interspeech*, G. Kubin and Z. Kacic, Eds., ISCA, 2019, pp. 4125–4129.

[39] R. Yadav, A. Sardana, V. P. Namboodiri, *et al.*, "Speech prediction in silent videos using variational autoencoders," in *ICASSP*, IEEE, 2021, pp. 7048–7052.

[40] R. Yamamoto, E. Song, and J. Kim, "Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP*, IEEE, 2020, pp. 6199–6203.

[41] G. Yang, S. Yang, K. Liu, *et al.*, "Multi-band melgan: Faster waveform generation for high-quality text-to-speech," in *SLT*, IEEE, 2021, pp. 492–498.

[42] H. Zen, V. Dang, R. Clark, *et al.*, "Libritts: A corpus derived from librispeech for text-to-speech," in *Interspeech*, G. Kubin and Z. Kacic, Eds., ISCA, 2019, pp. 1526–1530.