SVTS: Scalable Video-to-Speech Synthesis - Extended Abstract

Rodrigo Mira[1,†] Alexandros Haliassos[1] Stavros Petridis[1] Björn W. Schuller[1,2] Maja Pantic[1]
[1]Imperial College London [2]University of Augsburg
{rs2517,alexandros.haliassos14,stavros.petridis04,bjoern.schuller,m.pantic}@imperial.ac.uk

1. Introduction

Video-to-speech synthesis (also known as lip-to-speech) can be described as speech generation from silent video, typically focused on lip movements. Although this task can be achieved through a combination of lipreading and text-to-speech, directly predicting speech sidesteps the need for text transcriptions and thus allows leveraging large amounts of unlabelled audio-visual data. Furthermore, this task has compelling applications, such as audio retrieval from video streams (*e.g.*, videoconferencing) where the speech is either deteriorated or absent altogether, and generating artificial speech for people suffering from aphonia, *i.e.*, who have lost the ability to vocalize.

In recent years, a variety of deep learning-based methods have been proposed for video-to-speech, ranging from simple convolutional architectures [5, 6] to large generative adversarial networks (GANs) with elaborate training procedures and loss ensembles [11, 12]. While these methods have yielded successive improvements on multiple established corpora, they primarily suffer from two recurring limitations: using the Griffin-Lim algorithm [7] to synthesize audio from predicted spectrograms, which introduces noticeable artifacts in the resulting speech, and focusing on datasets recorded under studio conditions with a small pool of speakers and a homogeneous vocabulary (*e.g.*, GRID [4] and TCD-TIMIT [9]).

Aiming to address these shortcoming, we propose a **s**calable **v**ideo-**t**o-**s**peech synthesizer, dubbed SVTS, which combines a video-to-spectrogram predictor with a pre-trained neural vocoder that maps spectrograms to waveforms. Using a powerful off-the-shelf vocoder allows us to focus on spectrogram prediction, which we show can be effectively performed through a scalable ResNet+Conformer architecture and simple comparative losses. We train and evaluate on GRID, outperforming previous works on most metrics, and achieve state-of-the-art performance for LRW [3]. Furthermore, to the best of our knowledge, we are the first to produce intelligible speech for LRS3 [1].

†Corresponding author.

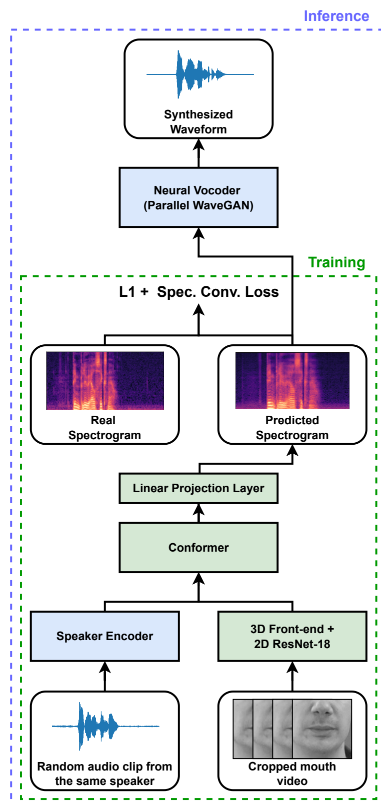

Figure 1. Summary of our video-to-speech synthesis approach during training and inference. In this figure, the components pictured in blue are pre-trained and kept frozen, while the components pictured in green are trained from scratch.

2. Methodology

2.1. Video-to-spectrogram

Our spectrogram prediction model receives video sampled at 20 fps as input and outputs the log-mel spectrogram of the corresponding speech, which contains 80 frames per second. Each video frame is passed through a ResNet18+Conformer architecture [8, 10] and is projected into 4×80 spectrogram frames via a linear projection layer. To capture the speaker's voice profile, we apply a

Method	Corpus	Speaker split (seen/unseen)	Training data (hours)	PESQ	STOI	ESTOI	WER (%)
End-to-end GAN [12]	GRID	seen	24	1.70	0.667	0.466	4.60
VCA-GAN + Griffin-Lim [11]	GRID	seen	20	**1.97**	0.695	0.505	5.13
SVTS-S	GRID	seen	24	**1.97**	**0.705**	**0.523**	**2.36**
Conv. + LSTM + Griffin-Lim [13]	LRW	unseen	157	1.20	0.543	0.344	34.20*
End-to-end GAN [12]	LRW	unseen	157	1.33	0.552	0.330	42.60
VCA-GAN + Griffin-Lim [11]	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	unseen	296	1.25	0.507	0.271	-
SVTS-L	LRS3 + VoxCeleb2	unseen	1556	**1.26**	**0.530**	**0.313**	-

Table 1. Summary of our results. *reported using Google speech-to-text API.

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

Table 2. Summary of our proposed SVTS architectures. *refers to the total number of parameters in the model.

pre-trained speaker encoder[1] on a randomly selected speech clip. We present three versions of our SVTS model in Table 2, ranging from 27.3 to 87.6 million parameters. This model is trained using a combination of the L_1 loss and the spectral convergence loss [17].

2.2. Spectrogram-to-waveform

In order to synthesize waveform audio from log-mel spectrograms, we apply a recently-proposed neural vocoder: Parallel WaveGAN [17]. This WaveNet-based model is trained on a very large speech dataset (LibriTTS [18]) using a combination of adversarial and comparative losses. As highlighted in Figure 1, this module is kept frozen and only used during inference to translate the predicted spectrograms into waveforms, allowing for a simpler training procedure.

3. Experiments

3.1. Datasets

In this work, we experiment with three datasets: **GRID**, which features a small collection of short sentences uttered by 33 different speakers, recorded in studio conditions; **LRW**, which has a wider vocabulary of 500 words and hundreds of different speakers recorded ‘in the wild’ during television broadcasts; and **LRS3**, which contains sentences from thousands of speakers recorded during TED talks, featuring a wide variety of recording conditions, as well as a vocabulary of more than 50,000 words. Furthermore, we augment LRS3’s training set with additional data from the English-only version [15] of VoxCeleb2 [2], containing more than 1,500 hours of video.

3.2. Evaluation metrics

To evaluate the quality of our generated speech samples, we apply four objective metrics: **PESQ** [14], which measures the clarity and overall quality of the speech; **STOI** and **ESTOI** [16], which measure intelligibility; and **WER** (Word Error Rate), which serves as an easily interpretable intelligibility metric. This is measured by using a pre-trained speech recognition model on the real and generated samples and comparing the resulting transcriptions.

3.3. Results

We present our results in Table 1, and encourage readers to listen to the samples presented on our project website[2]. On GRID, our work achieves state-of-the-art performance on all metrics, resulting in a very low WER of 2.36 %. For LRW, which presents a more substantial challenge for video-to-speech, SVTS-M outperforms all previous works on all objective metrics by a wide margin, yielding a particularly impressive improvement on WER.

Finally, we train our largest model SVTS-L on LRS3. To demonstrate our model’s scalability, we compare two experiments with the same validation and testing sets (from LRS3): one trained on 296 hours of LRS3 video, and another trained on a combination of the LRS3 training set and an English-only version of VoxCeleb2, amounting to 1556 hours of data. As shown in Table 1, the addition of the VoxCeleb2 training data yields a noticeable increase on all evaluation metrics, demonstrating our method’s scalability.

[1] https://github.com/corentinj/real-time-voice-cloning

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

4. Conclusion

In this work, we present a new video-to-speech approach which combines a simple spectrogram prediction model with a pre-trained neural vocoder to reproduce speech directly from silent lip movements. This straightforward approach allows us to easily scale to a variety of datasets, ranging from the small and controlled GRID, to a dataset containing >1500 hours of unconstrained speech (LRS3 + Voxceleb2). Through our experiments, we show that our approach is superior to previous works on both GRID and LRW, according to four objective speech metrics.

Acknowledgements

All training, testing, and ablation studies have been conducted at Imperial College London.

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. 1

[2] J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *Interspeech*, 2018. 2

[3] Joon Son Chung and Andrew Zisserman. Lip reading in the wild. In *ACCV*, volume 10112 of *Lecture Notes in Computer Science*, pages 87–103, 2016. 1

[4] Martin Cooke, Jon Barker, Stuart Cunningham, and Xu Shao. An audio-visual corpus for speech perception and automatic speech recognition (l). *The Journal of the Acoustical Society of America*, 120:2421–4, 2006. 1

[5] Ariel Ephrat, Tavi Halperin, and Shmuel Peleg. Improved speech reconstruction from silent video. In *ICCV*, pages 455–462. IEEE Computer Society, 2017. 1

[6] Ariel Ephrat and Shmuel Peleg. Vid2speech: Speech reconstruction from silent video. In *ICASSP*, pages 5095–5099. IEEE, 2017. 1

[7] D. Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 32(2):236–243, 1984. 1

[8] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. Conformer: Convolution-augmented transformer for speech recognition. In Helen Meng, Bo Xu, and Thomas Fang Zheng, editors, *Interspeech*, pages 5036–5040. ISCA, 2020. 1

[9] N. Harte and E. Gillen. Tcd-timit: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia*, 17(5):603–615, 2015. 1

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 1

[11] Minsu Kim, Joanna Hong, and Yong Man Ro. Lip to speech synthesis with visual context attentional GAN. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *NeurIPS*, 2021. 1, 2

[12] Rodrigo Mira, Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, Björn W. Schuller, and Maja Pantic. End-to-end video-to-speech synthesis using generative adversarial networks. *IEEE Transactions on Cybernetics*, pages 1–13, 2022. 1, 2

[13] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *CVPR*, pages 13793–13802. Computer Vision Foundation / IEEE, 2020. 2

[14] Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *ICASSP*, 2001. 2

[15] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *CoRR*, abs/2201.02184, 2022. 2

[16] Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time-frequency weighted noisy speech. *IEEE Trans. Speech Audio Process.*, 19(7):2125–2136, 2011. 2

[17] Ryuichi Yamamoto, Eunwoo Song, and Jae-Min Kim. Parallel wavegan: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In *ICASSP*, pages 6199–6203. IEEE, 2020. 2

[18] Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. Libritts: A corpus derived from librispeech for text-to-speech. In Gernot Kubin and Zdravko Kacic, editors, *Interspeech*, pages 1526–1530. ISCA, 2019. 2