End-To-End Video-To-Speech Synthesis using Generative Adversarial Networks with Multiple Critics

Rodrigo Mira[1] Konstantinos Vougioukas[1] Pingchuan Ma[1] Stavros Petridis[1]
Björn W. Schuller[1,2] Maja Pantic[1]

Imperial College London[1]
ZD.B Chair of Embedded Intelligence for Health Care and Wellbeing, University of Augsburg[2]
{rs2517,k.vougioukas,pingchuan.ma16,stavros.petridis04,bjoern.schuller,m.pantic}@imperial.ac.uk

1. Introduction

Cross-modal translation between modalities such as text, images and audio has recently become a prevalent topic in the machine learning community, particularly due to the abundance of publicly available paired data (*e.g.* transcribed speech) and the increasing effectiveness of generative models. The duality between video and audio in speech is especially promising in this regard, since these modalities are almost always recorded simultaneously and are heavily correlated in their content. For these reasons, speech-driven video generation [14] has recently emerged as a relevant research task with encouraging new applications including realistic facial animation for virtual characters. On the other hand, the inverse task of predicting speech from silent video has arguably received less attention, despite also achieving substantial progress.

While there has been a recent attempt to translate between video and speech end-to-end [13, 11], most approaches still rely on intermediate representations such as log-mel filterbanks [1, 5, 12, 15] or acoustic parametric features [4, 10] due to the well-known difficulty of generating realistic waveform audio directly. These features are then passed through vocoders, often resulting in excessively synthetic audio, or phase estimation algorithms such as Griffin-Lim [6], which is remarkably slow and generally incurs a loss in quality.

With our new work, we present a new methodology for video-to-speech synthesis by using an end-to-end encoder-decoder generative adversarial network (GAN), trained using two separate adversarial critics and an ensemble of comparative losses. By applying this methodology, we are able to generate realistic audio from silent video only, which outperforms previous methods on multiple established benchmarks.

2. Methodology

In this section we describe our approach for video-to-speech synthesis, which includes our generative model and the adversarial critics that were used to train it. The full training procedure is portrayed in Figure 1.

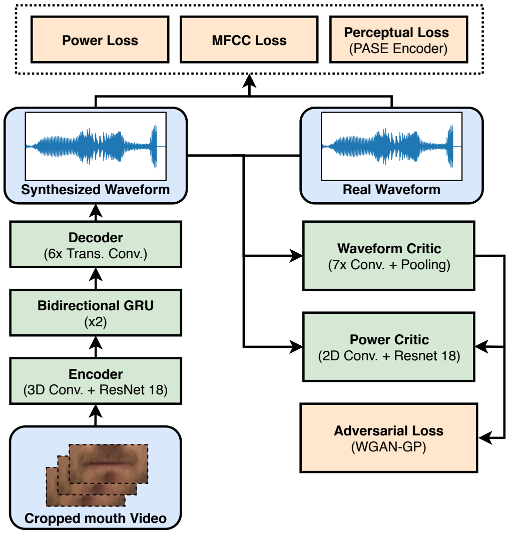

Figure 1. Summarized view of our video-to-speech synthesis approach.

2.1. Generative model

Our generator model receives cropped mouth video sampled at 25 frames per second and predicts raw waveform audio sampled at 16,000 Hz, meaning that each video frame is effectively decoded into 640 waveform samples. The video is fed as input to a 2D ResNet 18 [8] preceded by a 3D front-end layer with a receptive field of 5 video frames (followed by a max pooling layer). The resulting visual features are then fed into a 2-layer bidirectional GRU (Gated Recurrent Unit) recurrent neural network (RNN). Finally, these temporal features are fed into our decoder, which is composed of six stacked transposed convolutions which upsample these features into waveform audio.

2.2. Losses

We combine three comparative losses to train our model: an MFCC Loss, which measures the difference between the MFCC's (Mel-Frequency Cepstral Coefficients) of real and synthesized samples; a Power Loss, which measures the difference between the log-STFT magnitudes extracted from the real and synthesized audio; and a Perceptual Loss, which uses a pre-trained speech encoder to extract features from real and generated audio and compares them. The differences in these three losses are computed using a standard L1 Loss between real and synthesized features. Furthermore, we also apply the improved Wasserstein adversarial loss proposed in [7], through the use of the two critics described below. The model is trained using the Adam optimizer with a learning rate of 0.0001.

2.3. Adversarial Critics

For the waveform critic, we adapt the architecture presented in [9], which is composed of seven stacked convolutional layers followed by a pooling layer, and use it to discriminate real from generated audio, which substantially improves realism. Furthermore, we propose to combine this critic (which is typically found in waveform generation GANs) with a power critic, which aims to discriminate our samples in the spectral domain. This critic is a 2D Resnet 18 which receives the spectrograms extracted from real/generated audio as input and learns to evaluate whether they are real, which further encourages our generator to produce natural sounding speech.

3. Experiments

3.1. Datasets

For the purpose of our experiments, we train our model on a subset of the GRID audio-visual corpus [3], which is referenced in previous works [13, 12]. This dataset features 4 speakers uttering a set of 1000 short sentences made from a constrained vocabulary of six words, amounting to roughly 3 hours of speech. We split the set of sentences from each speaker for training/validation/testing using a 90/5/5 % split, reproducing the split used in previous works [13]. We also train on the full LRW (Lip Reading in the Wild) dataset [2], which features a broad vocabulary of 500 words and more than 160 hours of speech. We use the original train/val/test set for LRW, proposed by the authors.

3.2. Evaluation

To evaluate our results, we use four objective metrics which are frequently referenced in video-to-speech literature[1]: PESQ (Perceptual Evaluation of Speech Quality), which measures the clarity and perceptual quality of the speech; STOI (Short-Time Objective Intelligibility measure), which measures the intelligibility of the generated samples; MCD (Mel-Cepstral Distance), which measures the distance between real and synthesized speech in the spectral domain; and WER (Word Error Rate), which uses a pre-trained speech recognition model to determine the intelligibility of the generated speech in an easily interpretable manner.

Method	PESQ	STOI	MCD	WER
GAN-based [11]	1.70	0.539	45.37	21.11 %
Lip2Wav [12]	1.77	**0.731**	-	14.08[a] %
Ours	**2.10**	0.595	**26.78**	**7.03 %**

[a] Reported using Google STT API.

Table 1. Comparison between our new model and the previous approaches, using the GRID dataset (4 speaker subset).

Method	PESQ	STOI	MCD	WER
Lip2Wav [12]	1.20	0.543	-	**34.20**[a] %
Ours	**1.45**	**0.556**	**39.32**	42.51 %

[a] Reported using Google STT API.

Table 2. Comparison between our new model and the previous approach, using the full LRW dataset.

3.3. Results

We present our results on GRID and LRW in tables 1 and 2. It is clear that our model outperforms previous approaches on PESQ and WER for GRID, indicating that this method is a step forward in both overall quality, and practical intelligibility. Furthermore, it also outperforms [12] on multiple metrics for LRW, which means our model performs well on more complex corpora. However, it should be noted that the outlined metrics each have their limitations, and that the task of objectively evaluating synthesized speech is notoriously difficult. Therefore, we encourage readers to refer to the generated test samples we have shared in our project website[2].

4. Conclusion

In this work, we outline our new video-to-speech approach, which uses two adversarial critics combined with a complex loss configuration to train an end-to-end video-to-speech model, outperforming previous methods on multiple objective metrics. We believe this is a substantial step forward in video-to-speech modelling and look forward to seeing this methodology applied to more ambitious, less constrained scenarios and datasets.

Acknowledgements

All datasets used in the experiments and all training, testingand ablation studies have been contacted at Imperial College.

[1] Our evaluation procedure is publicly available on https://github.com/miraodasilva/evalaudio, to ease reproducibility

[2] https://sites.google.com/view/video-to-speech

References

[1] H. Akbari, H. Arora, L. Cao, and N. Mesgarani. Lip2audspec: Speech reconstruction from silent lip movements video. In *Proc. of ICASSP*, pages 2516–2520. IEEE, 2018. 1

[2] J. S. Chung and A. Zisserman. Lip reading in the wild. In *Asian Conference on Computer Vision*, 2016. 2

[3] M. Cooke, J. Barker, S. Cunningham, and X. 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. 2

[4] T. L. Cornu and B. Milner. Reconstructing intelligible audio speech from visual speech features. In *Proc. of Interspeech*, pages 3355–3359. ISCA, 2015. 1

[5] A. Ephrat, T. Halperin, and S. Peleg. Improved speech reconstruction from silent video. In *Proc. of ICCV*, pages 455–462. IEEE, 2017. 1

[6] 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, April 1984. 1

[7] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of wasserstein gans. In *Proc. of NeurIPS*, pages 5767–5777, 2017. 2

[8] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proc. of CVPR*, pages 770–778. IEEE, 2016. 1

[9] K. Kumar, R. Kumar, T. de Boissiere, L. Gestin, W. Z. Teoh, J. Sotelo, A. de Brébisson, Y. Bengio, and A. C. Courville. Melgan: Generative adversarial networks for conditional waveform synthesis. In *Proc. of NeurIPS*, pages 14881–14892, 2019. 2

[10] D. Michelsanti, O. Slizovskaia, G. Haro, E. Gómez, Z. Tan, and J. Jensen. Vocoder-based speech synthesis from silent videos. In H. Meng, B. Xu, and T. F. Zheng, editors, *Proc. of Interspeech*, pages 3530–3534. ISCA, 2020. 1

[11] R. Mira, K. Vougioukas, P. Ma, S. Petridis, B. W. Schuller, and M. Pantic. End-to-end video-to-speech synthesis using generative adversarial networks. *CoRR*, abs/2104.13332, 2021. 1, 2

[12] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. V. Jawahar. Learning individual speaking styles for accurate lip to speech synthesis. In *Proc. of CVPR*, pages 13793–13802. IEEE, 2020. 1, 2

[13] K. Vougioukas, P. Ma, S. Petridis, and M. Pantic. Video-driven speech reconstruction using generative adversarial networks. In G. Kubin and Z. Kacic, editors, *Proc. of Interspeech*, pages 4125–4129. ISCA, 2019. 1, 2

[14] K. Vougioukas, S. Petridis, and M. Pantic. End-to-end speech-driven realistic facial animation with temporal gans. In *Proc. of CVPR*, pages 37–40. CVF / IEEE, 2019. 1

[15] R. Yadav, A. Sardana, V. P. Namboodiri, and R. M. Hegde. Speech prediction in silent videos using variational autoencoders. *CoRR*, abs/2011.07340, 2020. 1