WAV2PIX: Speech-conditioned Face Generation using Generative Adversarial Networks

Amanda Duarte *[1,2], Francisco Roldan[1], Miquel Tubau[1], Janna Escur[1], Santiago Pascual[1], Amaia Salvador[1], Eva Mohedano[3], Kevin McGuinness[3], Jordi Torres[1,2], and Xavier Giro-i-Nieto[1,2]

[1]Universitat Politecnica de Catalunya, Barcelona, Catalonia/Spain
[2]Barcelona Supercomputing Center, Catalonia/Spain
[2]Insight Centre for Data Analytics - DCU, Ireland

1. Introduction

Audio and visual signals are the most common modalities used by humans to identify other humans and sense their emotional state. Features extracted from these two signals are often highly correlated, allowing us to imagine the visual appearance of a person just by listening to their voice, or build some expectations about the tone or pitch of their voice just by looking at a picture of the speaker. When it comes to image generation, however, this multimodal correlation is still under-explored.

In this paper, we focus on cross-modal visual generation, more specifically, the generation of facial images given a speech signal. Unlike recent works, we aim to generate the whole face image at pixel level, conditioning only on the raw speech signal (*i.e.* without the use of any hand-crafted features) and without requiring any previous knowledge (e.g speaker image or face model).

To this end, we propose a conditional generative adversarial model (shown in Figure 1) that is trained using the aligned audio and video channels in a self-supervised way. For learning such a model, high quality, aligned samples are required. This makes the most commonly used datasets such as *Lip Reading in the wild* [6], or *VoxCeleb* [17] unsuitable for our approach, as the position of the speaker, the background, and the quality of the videos and the acoustic signal can vary significantly across different samples. We therefore built a new video dataset from YouTube, composed of videos uploaded to the platform by well-established users (commonly known as *youtubers*), who recorded themselves speaking in front of the camera in their personal home studios. Hence, our main contributions can be summarized as follows: 1) We present a conditional GAN that is able to generate face images directly from the raw speech signal, which we call *Wav2Pix*;

2) We present a manually curated dataset of videos from youtubers, that contains high-quality data with notable *expressiveness* in both the speech and face signals;

3) We show that our approach is able to generate realistic and diverse faces.

The developed model, software and dataset are publicly released[1].

2. Related works

Generative Adversarial Networks: (GANs) [8] are a state of the art deep generative model that consist of two networks, a Generator G and a Discriminator D, playing a min-max game against each other. This means both networks are optimized to fulfill their own objective: G has to generate realistic samples and D has to be good at rejecting G samples and accepting real ones. The way Generator can create novel data mimicking real one is by mapping samples $z \in \mathbb{R}^n$ of arbitrary dimensions coming from some simple prior distribution $\mathcal{Z}$ to samples x from the real data distribution $\mathcal{X}$ (in this case we work with images, so $\mathbf{x} \in \mathbb{R}^{w \times h \times c}$ where $w \times h$ are spatial dimensions width and height and c is the amount of channels). This means each $\mathbf{z}$ forward is like sampling from $\mathcal{X}$. On the other hand the discriminator is typically a binary classifier as it distinguishes *real* samples from *fake* ones generated by G. One can further condition G and D on a variable $e \in \mathbb{R}^k$ of arbitrary dimensions to derive the the conditional GANs [15] formulation, with the conditioning variable being of any type, *e.g.* a class label or text captions [23]. In our work, we generate images conditioned on raw speech waveforms.

*Contact author: amanda.duarte@upc.edu
To appear in *Proc. ICASSP2019, May 12-17, 2019, Brighton, UK*
The developed model, software and a subset of the dataset are publicly released at https://imatge-upc.github.io/wav2pix/.

[1] https://imatge-upc.github.io/wav2pix/

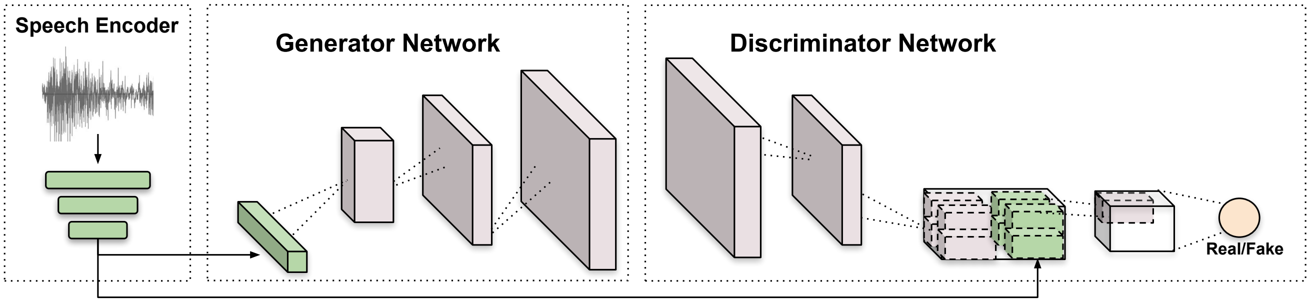

Figure 1. The overall diagram of our speech-conditioned face generation GAN architecture. The network consists of a speech encoder, a generator and a discriminator network. An audio embedding (green) is used by both the generator and discriminator, but its error is just back-propagated at the generator. It is encoded and projected to a lower dimension (vector of size 128). Pink blocks represent convolutional/deconvolutional stages.

Numerous improvements to the GANs methodology have been presented lately. Many focusing on stabilizing the training process and enhance the quality of the generated samples [27, 3]. Others aim to tackle the vanishing gradients problem due to the sigmoid activation and the log-loss in the end of the classifier [1, 2]. To solve this, the least-squares GAN (LSGAN) approach [14] proposed to use a least-squares function with binary coding (1 for real, 0 for fake). We thus use this conditional GAN variant with the objective function is given by:

$$\begin{aligned}\min_D V_{\text{LSGAN}}(D) = &\frac{1}{2}\,\mathbb{E}_{\mathbf{x},\mathbf{e}\sim p_{\text{data}}(\mathbf{x},\mathbf{e})}[(D(\mathbf{x},\mathbf{e})-1)^2] \\ &+\frac{1}{2}\,\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{z}}(\mathbf{z}),\mathbf{e}\sim p_{\text{data}}(\mathbf{e})}[D(G(\mathbf{z},\mathbf{e}),\mathbf{e})^2].\end{aligned} \tag{1}$$

$$\min_G V_{\text{LSGAN}}(G) = \frac{1}{2}\,\mathbb{E}_{\mathbf{z}\sim p_{\mathbf{e}}(\mathbf{e}),\mathbf{y}\sim p_{\text{data}}(\mathbf{y})}[(D(G(\mathbf{z},\mathbf{e}),\mathbf{e})-1)^2], \tag{2}$$

Multi-modal generation: Data generation across modalities is becoming increasingly popular [22, 23, 18, 26]. Recently, a number of approaches combining audio and vision have appeared, with tasks such as generating speech from a video [7] or generating images from audio/speech [5]. In this paper we will focus on the latter.

Most works on audio conditioned image generation adopt non end-to-end approaches and exploit previous knowledge about the data. Typically, speech has been encoded with handcrafted features which have been very well engineered to represent human speech. At the visual part, point-based models of the face [11] or the lips [24] have been adopted. In contrast to that, our network is trained entirely end-to-end solely from raw speech to generate image pixels.

3. Youtubers Dataset

Our Youtubers dataset is composed of two sets: the complete noisy subset automatically generated, and a clean subset which was manually curated to obtain high quality data.

In total we collected 168,796 seconds of speech with the corresponding video frames, and cropped faces from a list of 62 youtubers active during the past few years. The dataset was gender balanced and manually cleaned keeping 42,199 faces, each with an associated 1-second speech chunk.

Initial experiments indicated a poor performance of our model when trained with noisy data. Thus, a part of the dataset was manually filtered to obtain the high-quality data required for our task. We took a subset of 10 identities, five female and five male, from our original dataset and manually filtered them making sure that all faces were visually clear and all audios contain just speech, resulting in a total of 4,860 pairs of images and audios.

4. Method

Since our goal is to train a GAN conditioned on raw speech waveforms, our model is divided in three modules trained altogether end-to-end: a speech encoder, a generator network and a discriminator network described in the following paragraphs respectively. The speech encoder was adopted from the discriminator in [20], while both the image generator and discriminator architectures were inspired by [23]. The whole system was trained following a Least Squares GAN [14] scheme. Figure 1 depicts the overall architecture.

Speech Encoder: We coupled a modified version of the SEGAN [20] discriminator Φ as input to an image generator G. Our speech encoder was modified to have 6 strided one-dimensional convolutional layers of kernel size 15, each one with stride 4 followed by LeakyReLU activations. Moreover we only require one input channel, so our input signal is $\mathbf{s} \in \mathbf{R}^{T\times 1}$, being $T = 16,384$ the amount of waveform samples we inject into the model (roughly one second of speech at $16\,kHz$). The aforementioned convolutional stack decimates this signal by a factor $4^6 = 4096$ while increasing the feature channels up to 1024. Thus, obtaining a tensor $f(\mathbf{s}) \in \mathbb{R}^{4\times1024}$ in the output of the convolutional stack f. This is flattened and injected into three fully connected layers that reduce the final speech embedding dimensions from $1024 \times 4 = 4096$ to 128, obtaining the vector $\mathbf{e} = \Phi(\mathbf{s}) \in \mathbb{R}^{128}$.

Image Generator Network: We take the speech embedding $\mathbf{e}$ as input to generate images such that $\hat{\mathbf{x}} = G(\mathbf{e}) = G(\Phi(\mathbf{s}))$. The inference proceeds with two-dimensional transposed convolutions, where the input is a tensor $\mathbf{e} \in \mathbb{R}^{1\times1\times128}$ (an image of size 1×1 and 128 channels), based on [22]. The final interpolation can either be $64 \times 64 \times 3$ or $128 \times 128 \times 3$ just by playing with the amount of transposed convolutions (4 or 5). It is important to mention that we have no latent variable $\mathbf{z}$ in G inference as it did not give much variance in predictions in preliminary experiments. To enforce the generative capacity of G we followed a dropout strategy at inference time inspired by [10]. Therefore, the G loss, follows the LSGAN loss presented in Equation 2 with the addition of this weighted auxiliary loss for identity classification.

Image Discriminator Network: The Discriminator D is designed to process several layers of stride 2 convolution with a kernel size of 4 followed by a spectral normalization [16] and leakyReLU (apart from the last layer). When the spatial dimension of the discriminator is 4×4, we replicate the speech embedding $\mathbf{e}$ spatially and perform a depth concatenation. The last convolution is performed with stride 1 to obtain a D score as the output.

5. Experiments

Model training: The *Wav2Pix* model was trained on the cleaned dataset described in Section 3 combined with a data augmentation strategy. In particular, we copied each image five times, pairing it with 5 different audio chunks of 1 second randomly sampled from the 4 seconds segment. Thus, we obtained $\approx$ 24k images and paired audio chunks of 1 second used for training our model. Our implementation is based on the PyTorch library [21] and trained on a GeForce Titan X GPU with 12GB memory. We kept the hyper-parameters as suggested in [23], changing the learning rate to 0.0001 in G and 0.0004 in D as suggested in [9]. We use ADAM solver [13] with momentum 0.1.

Evaluation: Figure 4 shows examples of generated images given a raw speech chunk, compared to the original image of the person who the voice belongs to. Different speech waveform produced by the same speaker were fed into the network to produce such images. Although the generated images are blurry, it is possible to observe that the model learns the person's physical characteristics, preserving the identity, and present different face expressions depending on the input speech [2]. Other examples from six different identities are presented in Figure 2.

To quantify the model's accuracy regarding the identity preservation, we fine-tuned a pre-trained VGG-Face Descriptor network [19, 4] with our dataset. We predicted the speaker identity from the generated images of both the speech train and test partitions, obtaining an identification accuracy of 76.81% and 50.08%, respectively.

We also assessed the ability of the model to generate realistic faces, regardless of the true speaker identity. To have a more rigorous test than a simple Viola & Jones face detector [25], we measured the ability of an automatic algorithm [12] to correctly identify facial landmarks on images generated by our model. We define detection accuracy as the percentage of images where the algorithm is able to identify *all* 68 key-points. For the proposed model and all images generated for our test set, the detection accuracy is 90.25%, showing that in most cases the generated images retain the basic visual characteristics of a face. This detection rate is much higher than the identification accuracy of 50.08%, as in some cases the model confuses identities, or mixes some of them in a single face. Examples of detected faces together with their numbered facial landmarks can be seen in Figure 3.

6. Conclusions

In this work we introduced a simple yet effective cross-modal approach for generating images of faces given only a short segment of speech, and proposed a novel generative adversarial network variant that is conditioned on the raw speech signal.

As high-quality training data are required for this task, we further collected and curated a new dataset, the Youtubers dataset, that contains high quality visual and speech signals. Our experimental validation demonstrates that the proposed approach is able to synthesize plausible facial images with an accuracy of 90.25%, while also being able to preserve the identity of the speaker about 50% of the times. Our ablation experiments further showed the sensitivity of the model to the spatial dimensions of the images, the duration of the speech chunks and, more importantly, on the quality of the training data. Further steps may address the generation of a sequence of video frames aligned with the conditioning speech, as well exploring the behaviour of the *Wav2Pix* when conditioned on unseen identities.

Acknowledgements

This research has received funding from la Caixa Foundation funded by the European Unions Horizon 2020 research and innovation programme under the Marie Skodowska-Curie grant agreement No. 713673. This research was partially supported by MINECO and the ERDF under contracts TEC 2015-69266-P and TEC 2016-75976-R, by the MICINN under the contract TIN 2015-65316-P and the Programa Severo Ochoa under the contract SEV-2015-0493 and SFI under grant number SFI/15/SIRG/3283. We gratefully acknowledge the support of NVIDIA Corporation with the

[2] Some examples of images and it correspondent speech as well as more generated images are available at: https://imatge-upc.github.io/wav2pix/

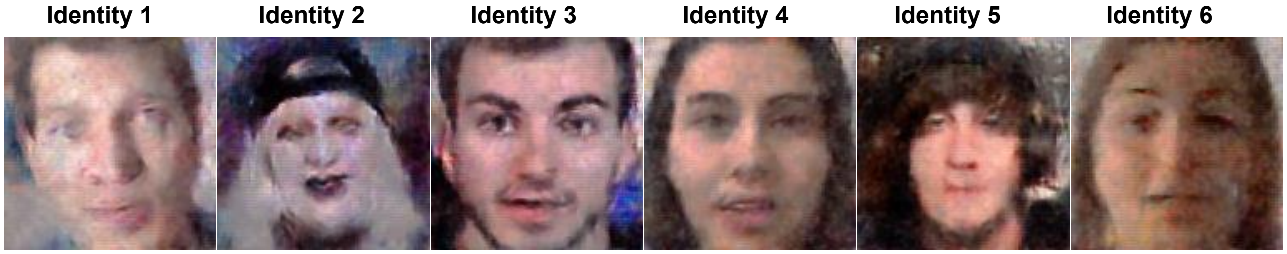

Figure 2. Generated samples conditioned to raw speech produced by our model.

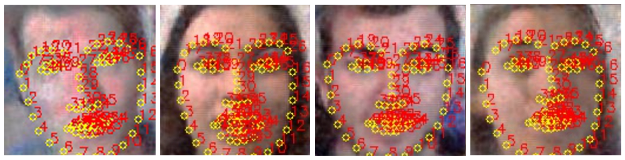

Figure 3. Examples of the 68 key-points detected on images generated by our model. Yellow circles indicate facial landmarks fitted to the generated faces, numbered in red fonts.

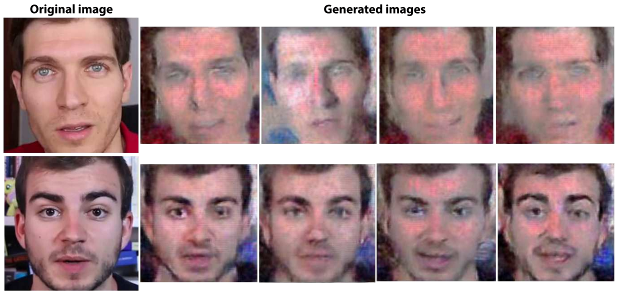

Figure 4. Examples of generated faces compared to the original image of the person who the voice belongs to. In the generated images, we can observe that our model is able to preserve the physical characteristics and produce different face expressions. In the first row we can see examples of the *youtuber Javier Muiz*. In the second row we can see examples of the *youtuber Jaime Altozano*.

donation of GPUs. We would like to thank the youtubers *Javier Muiz and Jaime Altozano* for providing us the rights of publishing their images on this research.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

[2] D. Berthelot, T. Schumm, and L. Metz. Began: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[3] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. In *ICLR*, 2019.

[4] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. Vggface2: A dataset for recognizing faces across pose and age. volume abs/1710.08092, 2017.

[5] L. Chen, S. Srivastava, Z. Duan, and C. Xu. Deep cross-modal audio-visual generation. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pages 349–357. ACM, 2017.

[6] J. S. Chung, A. W. Senior, O. Vinyals, and A. Zisserman. Lip reading sentences in the wild. In *CVPR*, pages 3444–3453, 2017.

[7] A. Ephrat, T. Halperin, and S. Peleg. Improved speech reconstruction from silent video. In *ICCV*, 2017.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, pages 2672–2680, 2014.

[9] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, pages 6626–6637, 2017.

[10] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976. IEEE, 2017.

[11] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen. Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Transactions on Graphics (TOG)*, 36(4):94, 2017.

[12] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

[13] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[14] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. P. Smolley. Least squares generative adversarial networks. In *ICCV*, pages 2813–2821. IEEE, 2017.

[15] M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[16] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *ICLR*, 2018.

[17] A. Nagrani, J. S. Chung, and A. Zisserman. Voxceleb: a large-scale speaker identification dataset. In *Interspeech*, 2017.

[18] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In *ICML*, 2017.

[19] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC*, 2015.

[20] S. Pascual, A. Bonafonte, and J. Serrà. Segan: Speech enhancement generative adversarial network. *Interspeech*, 2017.

[21] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. Automatic differentiation in pytorch. 2017.

[22] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *ICLR*, 2016.

[23] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *ICML*, 2016.

[24] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman. Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (TOG)*, 36(4):95, 2017.

[25] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*. IEEE, 2001.

[26] H. Zhang, T. Xu, H. Li, S. Zhang, X. Huang, X. Wang, and D. Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *ICCV*, 2017.

[27] Z. Zhang, Y. Xie, and L. Yang. Photographic text-to-image synthesis with a hierarchically-nested adversarial network. In *CVPR*, 2018.