CVPR LONG BEACH CALIFORNIA June 16-20, 2019

WiCV Women in Computer Vision

Sight and Sound Workshop

Wav2Pix: Speech-conditioned Face Generation using Generative Adversarial Networks

Amanda Duarte, Francisco Roldan, Miquel Tubau, Janna Escur, Santiago Pascual, Amaia Salvador, Eva Mohedano, Kevin McGuinness, Jordi Torres, Xavier Giro-i-Nieto

Barcelona Supercomputing Center, Centro Nacional de Supercomputación · DCU Dublin City University · UPC Universitat Politècnica de Catalunya BarcelonaTech · Insight

http://bit.ly/wav2pix

Abstract

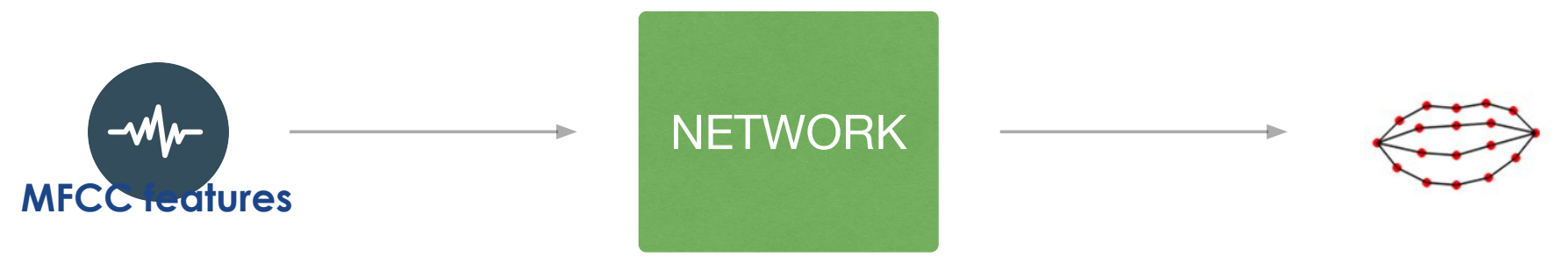
Speech is a rich biometric signal that **contains information** about the **identity**, **gender** and **emotional state** of the speaker. In this work, we explore its potential to **generate face images** of a speaker **by conditioning a Generative Adversarial Network (GAN) with raw speech input**. We propose a deep neural network that is trained from **scratch in an end-to-end fashion**, **generating a face directly from the raw speech waveform** without any additional identity information (e.g reference image or one-hot encoding). Our model is trained in a **self-supervised** fashion by exploiting the **audio** and **visual** signals naturally aligned in videos. 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**.

Contributions:

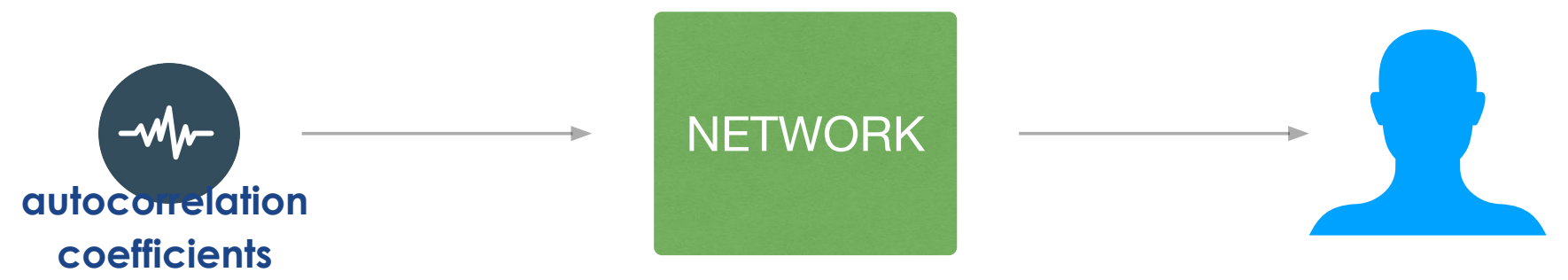
- We present a **conditional GAN** that is able to **generate face** images **directly from the raw speech signal**, which we call **Wav2Pix**.
- We present a **manually curated dataset** of videos from **youtubers**, that contains high-quality data with notable expressiveness in both the speech and face signals.
- We show that our approach is able to **generate realistic** and **diverse faces**.

Related work

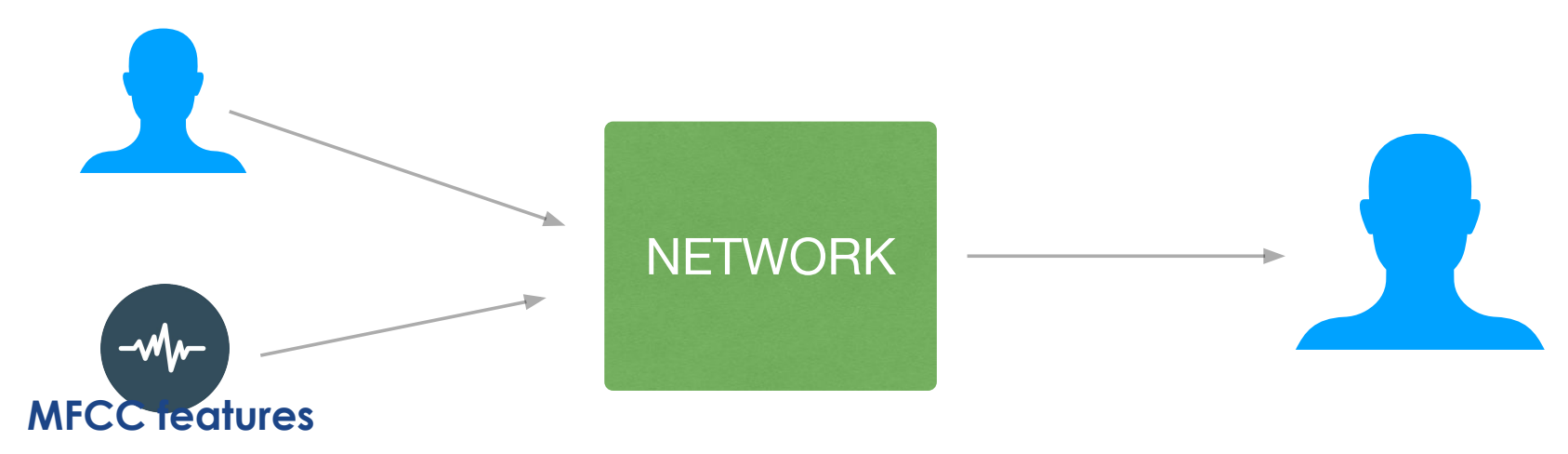
Suwajanakorn et. al [1] focused on animating a point-based lip model to later synthesize high quality videos of President Barack Obama



Karras et. al [2] propose a model for driving 3D facial animation by audio input in real time and with low latency



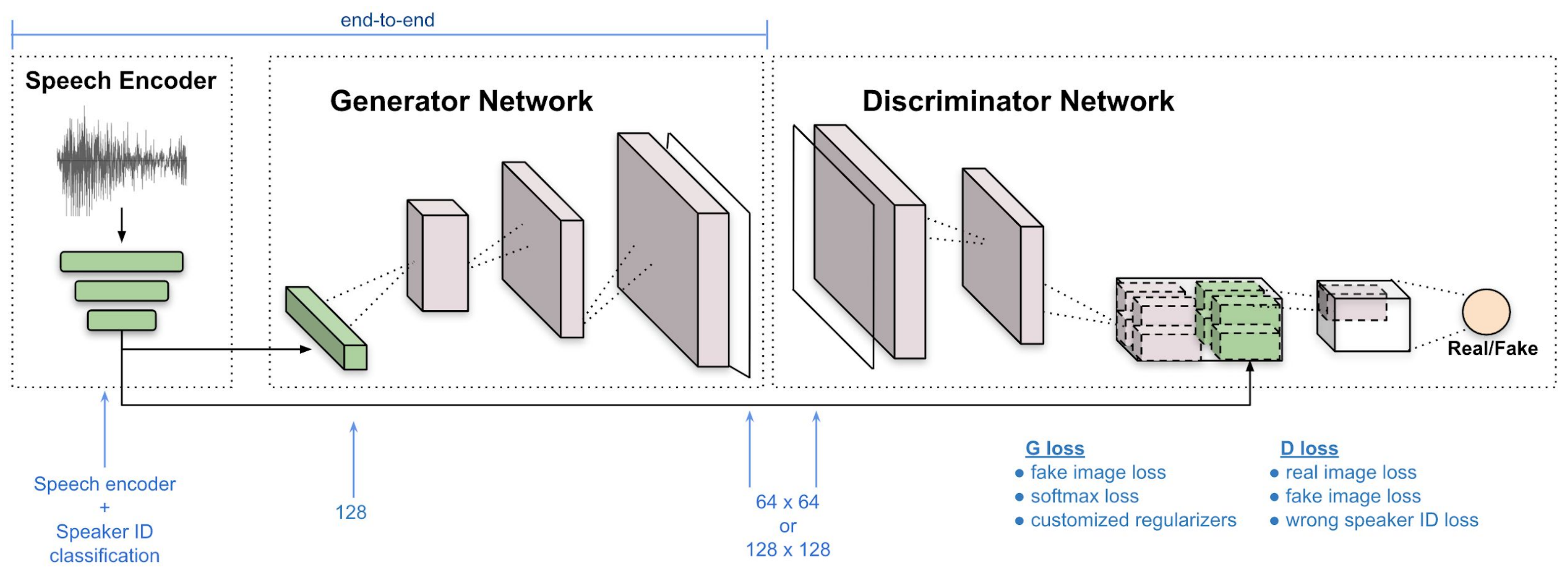
Chung et. al [3] presented a method for generating a video of a talking face starting from audio features and an image of the identity



[1] Supasorn Suwajanakorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman, "Synthesizing obama: learning lip sync from audio," ACM TOG, 2017.
[2] Karras, Tero, et al. "Audio-driven facial animation by joint end-to-end learning of pose and emotion." ACM Transactions on Graphics (TOG).
[3] Chung, Joon Son, Amir Jamaludin, and Andrew Zisserman. "You said that?." BMVC 2017.

Architecture

The **speech encoder** was adopted from the **discriminator** in SEGAN [4], while both the image **generator** and **discriminator** architectures were inspired by [5]. The whole system was trained following a Least Squares GAN [6] scheme.

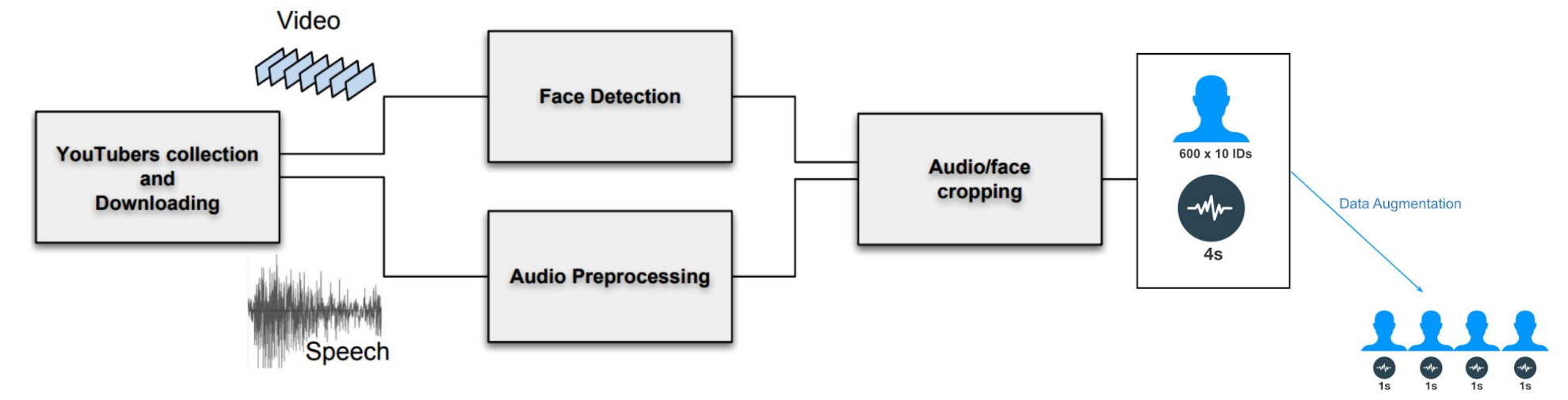


[4] Pascual et al. "SEGAN: Speech enhancement generative adversarial network." INTERSPEECH 2017.
[5] Reed et al, "Generative adversarial text to image synthesis." ICML 2016.
[6] Mao et al, "Least squares generative adversarial networks." ICCV 2017.

YouTubers dataset

We collected videos uploaded to YouTube by 62 well-established users (so-called **youtubers**). Such videos are usually of high quality, with the faces of the subject featured in a prominent way and with notable expressiveness in both the speech and face.

Frontal faces were detected automatically from downloaded videos, but a **manual filtering** of both speech snippets and cropped faces from a subset of 10 identities was done in order to to improve the obtained results.

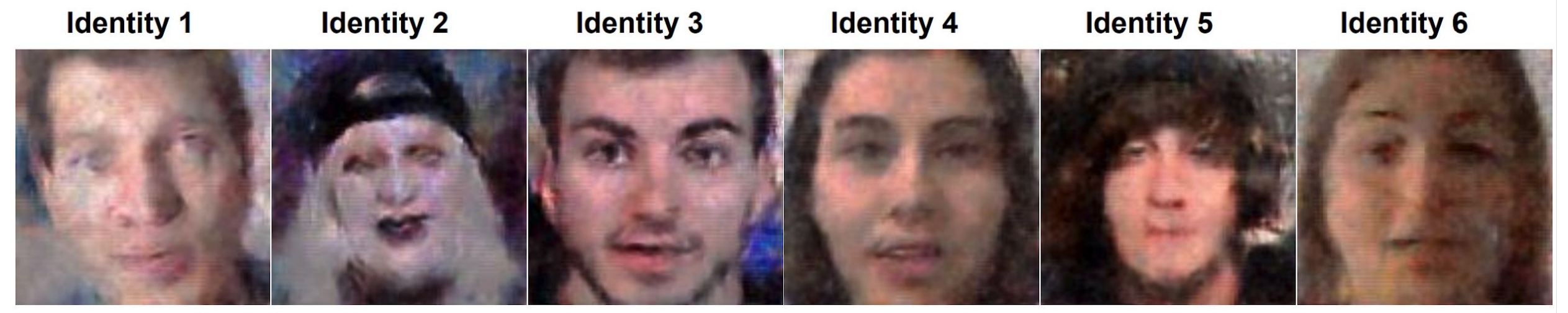


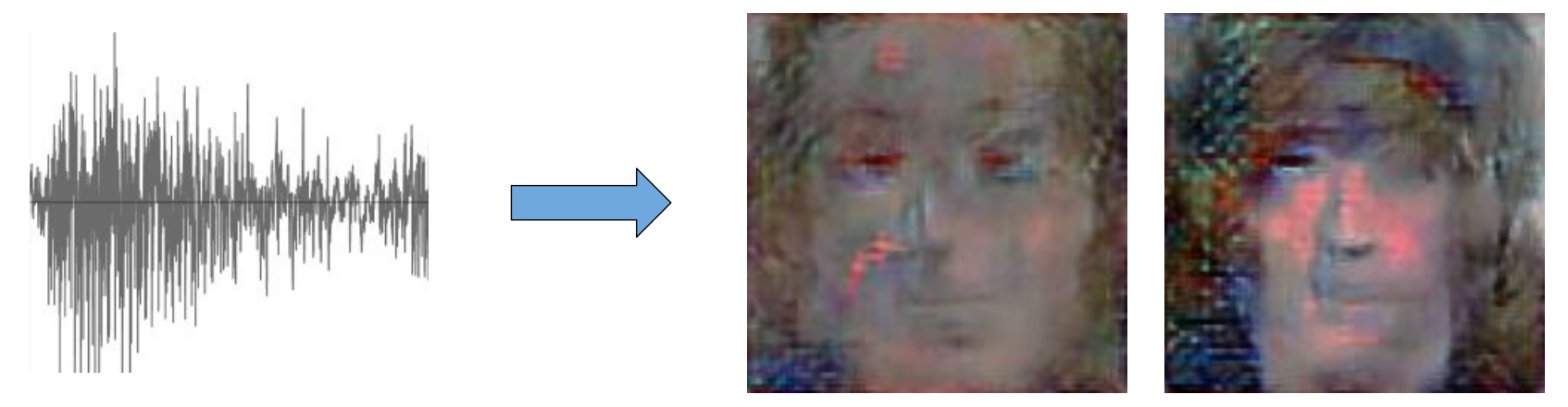
Features	10 youtubers subset
Male	5
Female	5
Speech-face pairs	30k
Speech duration	1s
Manually cleaned	Yes

Experiments

Diversity of identities

Wav2Pix can successfully generate diverse facial images for the 10 known identities ....

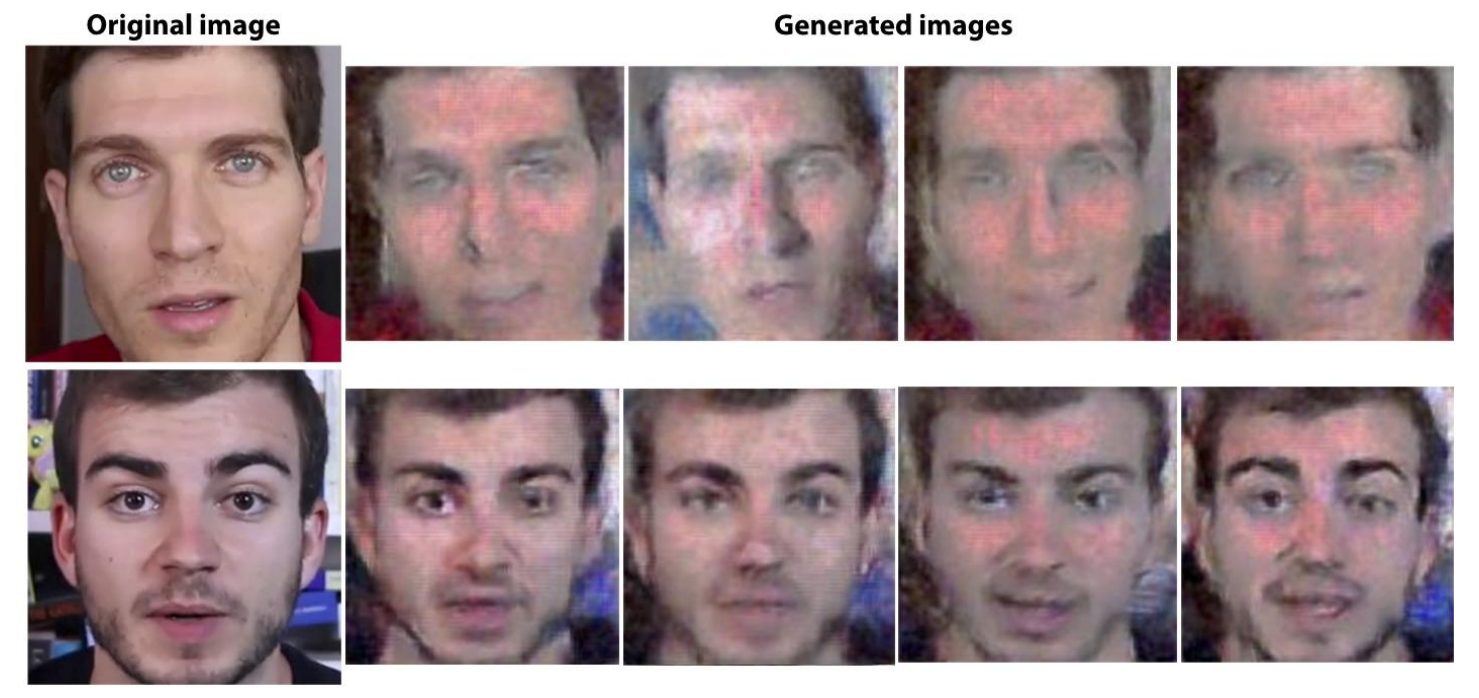




...but fails to generate realistic images when a speech of an unknown speaker is given. We hypothesize that this is due to the Speaker ID classifier included to stabilize training.

Diversity of expressions

Examples of **generated** faces compared to the original image of the person who the voice belongs to. In the generated images, we observed that our model is able to **preserve** the **physical characteristics** and produce different **face expressions**.



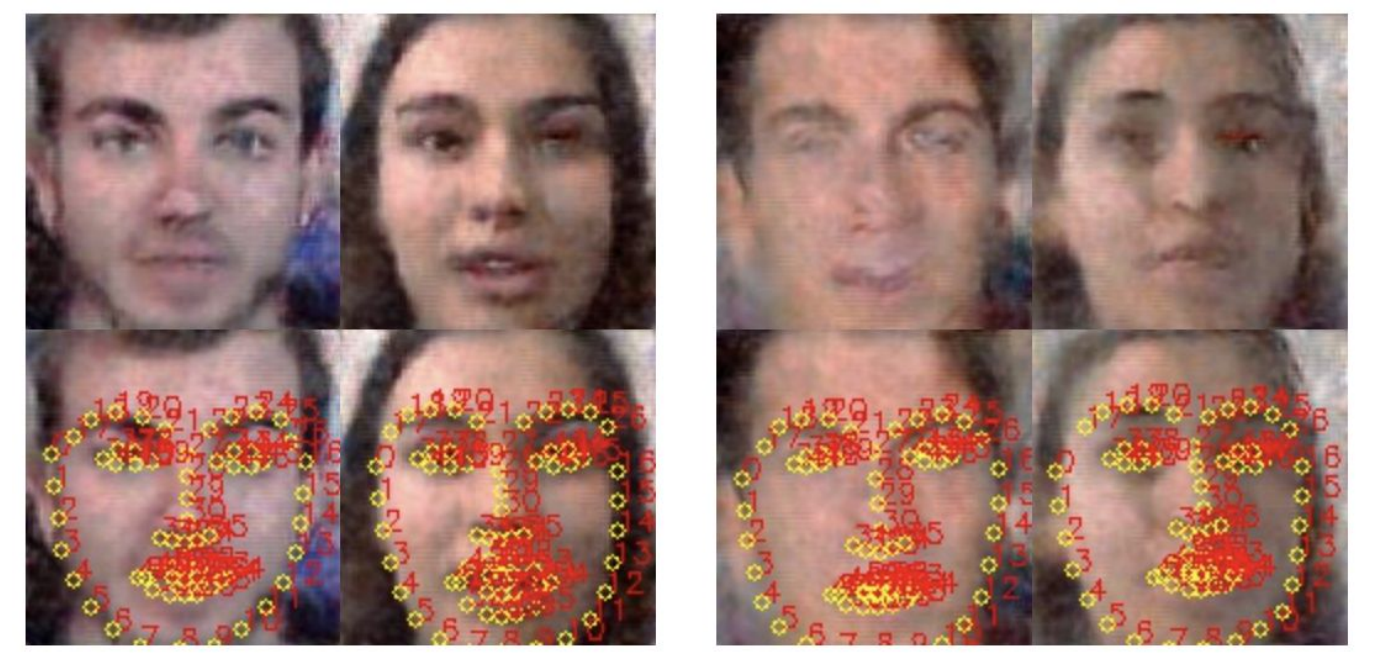
Evaluation

Face detection [7]: **90.25%**

We assessed the ability of our model to generate realistic faces, regardless of the true speaker identity. We define detection accuracy as the percentage of images where the algorithm is able to identify all 68 facial key-points.

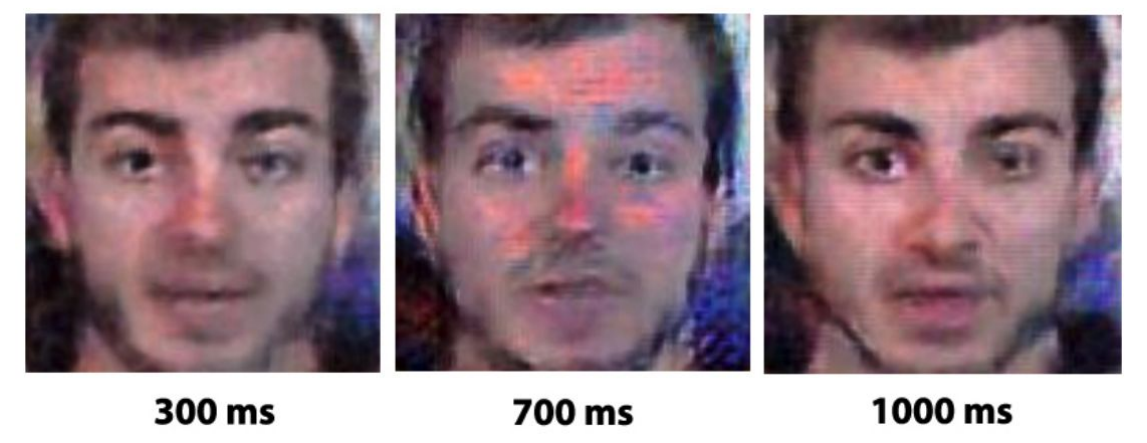
Face recognition [8,9]**: 50.08%** (over 10 IDs)

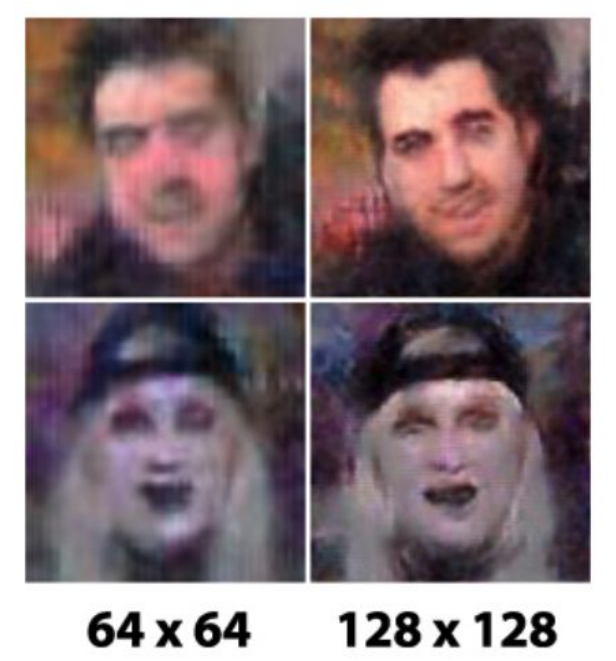
To quantify the model's accuracy regarding the identity preservation, we fine-tuned a pre-trained VGG-Face Descriptor network with our dataset. After that, we predicted the speaker identity from the generated images.



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.

[7] Vahid Kazemi and Josephine Sullivan, "One millisecond face alignment with an ensemble of regression trees," in CVPR2014.
[8] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in BMVC, 2015.
[9] Cao, Qiong, et al. "Vggface2: A dataset for recognising faces across pose and age." 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018.





Additional Experiments

We observed a drop in performance when working with smaller speech chunks and lower image definitions. We observed a visual degradation when using audio chunks of 300 and 700 milliseconds, which was reflected in a decrease of the face detection rate.

Detection accuracy when using 300 and 700 ms chunks was 81.16% and 89.12%, respectively, in both cases worse than the 90.25% accuracy achieved when using 1000 ms chunks.

Powered by: PyTorch

Acknowledgments: Excelencia Severo Ochoa · Obra Social Fundació "la Caixa" · NVIDIA · Gobierno de España · Generalitat de Catalunya