Wav2Pix
Speech-conditioned Face Generation using Generative Adversarial Networks

Amanda Duarte  Francisco Roldán  Miquel Tubau  Janna Escur  Xavier Giró-i-Nieto
Santiago Pascual  Amaia Salvador  Eva Mohedano  Kevin McGuinness  Jordi Torres
Wav2Pix
MOTIVATION

Speech Signal → Deep Learning Model → Face
● **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**.

● Roldán et. al. address this correlation proposing a face synthesis method using **exclusively** raw audio representation as inputs.
RELATED WORK
GENERATIVE MODELS

unit gaussian

generative model (neural net)

θ

generated distribution

^p(x)

image space

true data distribution

p(x)

image space

Image taken from https://blog.openai.com/generative-models/
**GENERATIVE ADVERSARIAL NETWORKS**

The diagram illustrates the training process of a Generative Adversarial Network (GAN). It consists of two main components: the **Generator** and the **Discriminator**.

- **Generator**: Takes random noise as input and generates a fake image.
- **Discriminator**: Receives both real images from the training set and fake images from the generator. It tries to distinguish between real and fake images.

The objective is to minimize the loss of the discriminator while simultaneously maximizing the loss of the generator. This is achieved through a min-max game, as described by the following equation:

$$
\min_G \max_D V(D, G) = \mathbb{E}_{x \sim p_{data}(x)} \left[ \log D(x) \right] + \mathbb{E}_{z \sim p_z(z)} \left[ \log (1 - D(G(z))) \right].
$$

---

Imagine we have a counterfeiter (G) trying to make fake money, and the police (D) has to detect whether the money is **true** or **fake**.

FAKE! It’s not even green
Imagine we have a counterfeiter (G) trying to make fake money, and the police (D) has to detect whether the money is true or fake.

FAKE! There's no watermark

Slide credit: Santi Pascual
Imagine we have a counterfeiter (G) trying to make fake money, and the police (D) has to detect whether the money is true or fake.
After enough iterations:
In that case:

- one-hot vector with the corresponding MNIST class

In our case:

- speech embedded in a 128 dimensional vector
SPEECH-CONDITIONED IMAGE SYNTHESIS

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

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

- **Chung** et. al. presented a method for generating a video of a talking face starting from audio features and an image of him/her (identity).

---

Chung et al. presented a method for generating a video of a talking face starting from audio features and an image of him/her (identity).

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

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

Roldán et al. 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).

Chung et al. presented a method for generating a video of a talking face starting from audio features and an image of him/her (identity).

Francisco Roldán Sánchez, “Speech-conditioned face generation with deep adversarial networks”
SPEECH-CONDITIONED FACE GENERATION WITH DEEP GANs

Francisco Roldán Sánchez, “Speech-conditioned face generation with deep adversarial networks”
Santiago Pascual, Antonio Bonafonte, and Joan Serrà, “Segan: Speech enhancement generative adversarial network,” Interspeech, 2017

LSGAN
64x64 resolution
Dropout instead of noise input
Wav2Pix

SPEECH-CONDITIONED FACE GENERATION WITH DEEP GANs

- Roldán et. al. model does **not generalize** for unseen speech

  ![Diagram showing a network with an audio input and a face output with a cross indicating failure to generalize.]

- Inception Score metric used by Roldán et. al. evaluates the images in terms of quality but **not** in terms of **realism**

In this Project

Enhancement

- Search the optimal input audio **length**
- Add an **audio classifier** on top of the embedding
- Augment the capacity to generate **128x128** resolution

Evaluation

- Compute the accuracy of a **fine-tuned** VGG classifier
- Compute a novel approach based on a **face detection** system
- Perform an online **survey** assessed by humans

Francisco Roldán Sánchez, “Speech-conditioned face generation with deep adversarial networks”
DATASET
PREVIOUS DATASET

**Wav2Pix**

**youtubers_v1**

<table>
<thead>
<tr>
<th>Sex</th>
<th>Speakers</th>
<th>Faces</th>
<th>Speech (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>29</td>
<td>26299</td>
<td>105196</td>
</tr>
<tr>
<td>Female</td>
<td>33</td>
<td>15900</td>
<td>63600</td>
</tr>
<tr>
<td>TOTAL</td>
<td>62</td>
<td>42199</td>
<td>168796</td>
</tr>
</tbody>
</table>

- **Good** recording hardware
- High amount of **frontal faces**
- Wide range of **emotions**
drawbacks

- **Imbalanced** dataset. Among the 62 youtubers, the amount of images/audios vary between 2669 and 52 pairs.

- Notable amount of **false positives**

- Most of the speech frames were **noisy**
  - Background music in a post-process edition
  - Voice of a third person
youtubers_v2 - new dataset collection

youtubers_v2_dataaugmented

10 ids

4s

600x10 IDs

1s 1s 1s 1s 1s

30k

DATA AUGMENTATION
<table>
<thead>
<tr>
<th>Features</th>
<th>Roldán</th>
<th>Ours</th>
<th>Ours data_augmented</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>youTube_v1</td>
<td>youTube_v2</td>
<td></td>
</tr>
<tr>
<td>Males</td>
<td>29</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Females</td>
<td>33</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Audio-face pairs</td>
<td>42199</td>
<td>6000</td>
<td>30000</td>
</tr>
<tr>
<td>Average audio-face pairs / ID</td>
<td>694</td>
<td>600</td>
<td>3000</td>
</tr>
<tr>
<td>Std audio-face pairs / ID</td>
<td>616</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Audio duration (s)</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Videos processed / ID</td>
<td>15</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Balanced</td>
<td>False</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>Cleaned</td>
<td>False</td>
<td>True</td>
<td>True</td>
</tr>
<tr>
<td>Size in memory (GB)</td>
<td>7.4</td>
<td>1.8</td>
<td>2.1</td>
</tr>
</tbody>
</table>
ARCHITECTURE
Wav2Pix
ARCHITECTURE

end-to-end

Speech Encoder

Generator Network

Discriminator Network

- segan decoder + classifier
- 64 x 64 | 128 x 128
- 128

D loss
- real image loss
- fake image loss
- wrong image loss

G loss
- fake image loss
- softmax loss
- customized regularizers
Wav2Pix

ARCHITECTURE

contributions

- **Audio segmentation** module

- **Speech classifier**
  - 1-hidden NN with 10 output units

- **Additional** convolutional and deconvolutional layers
  - Kernel size: 4
  - Stride: 2
  - Padding: 1
EVALUATION
EVALUATION

Fréchet Inception Distance

\[
FID(r, g) = \|(\mu_r - \mu_g)\|^2_2 + \text{Tr}(\sum_r + \sum_g - 2(\sum_r \sum_g)^{\frac{1}{2}})
\]

- **Inception-v3** network pre-trained on **ImageNet**

- Results **not consistent** with human judgements
- **Little** amount of data to obtain reliable results
- The measure relies on an ImageNet-pretrained inception network, **far from ideal** for datasets like **faces**

**VGGFace fine-tuned classifier**

- Network proposed by the Visual Geometry Group department of Engineering Science (University of Oxford)

<table>
<thead>
<tr>
<th></th>
<th>Roldán</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real data</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>Generated data for seen speech</td>
<td>56.34</td>
<td>76.81</td>
</tr>
<tr>
<td>Generated data for unseen speech</td>
<td>16.69</td>
<td>50.08</td>
</tr>
</tbody>
</table>

- **Improvement** of our model in preserving the identity
- Bearing in mind the metric is sensible to image quality, and the probability of confusion is 90%, the results are promising.

Facial Landmark Detection ratio

- **Robustness** to image quality

<table>
<thead>
<tr>
<th></th>
<th>Roldán</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real data</td>
<td>75.02</td>
<td>72.48</td>
</tr>
<tr>
<td>Generated data for seen speech</td>
<td>61.76</td>
<td>84.45</td>
</tr>
<tr>
<td>Generated data for unseen speech</td>
<td>60.81</td>
<td>90.25</td>
</tr>
</tbody>
</table>

- 90% of the generated images of our model for unseen speech can be considered as faces
### Facial Landmark Detection ratio

- **Robustness** to image quality

<table>
<thead>
<tr>
<th></th>
<th>Roldán</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Real data</strong></td>
<td>75.02</td>
<td>72.48</td>
</tr>
<tr>
<td><strong>Generated data for seen speech</strong></td>
<td>61.76</td>
<td>84.45</td>
</tr>
<tr>
<td><strong>Generated data for unseen speech</strong></td>
<td>60.81</td>
<td>90.25</td>
</tr>
</tbody>
</table>

- 90% of the generated images of our model for unseen speech can be considered as faces
Online survey

- **42** people have been asked to answer **2** questions for **32** different pairs of images:

  - Compare the quality of the generated image with respect to the real one (5-identical, 4-good, 3-fair, 2-poor, 1-bad)
  - Could you recognize the real person (appearing in the baseline image) from the generated image?

<table>
<thead>
<tr>
<th>MOS</th>
<th>% NOT SURE</th>
<th>% NO</th>
<th>% YES</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.09</td>
<td>14</td>
<td>52</td>
<td>34</td>
</tr>
</tbody>
</table>

- Not reliable results
- This metric should be further improved
EXPERIMENTS
datasets comparison

Best quality images manually selected

![Images of manually selected best quality images from different datasets](image)

Facial landmark detection ratio (%)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Facial Landmark Detection Ratio (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>youtubers_v1</td>
<td>60.81</td>
</tr>
<tr>
<td>youtubers_v2</td>
<td>71.47</td>
</tr>
<tr>
<td>youtubers_v2 Data Augmented</td>
<td>90.25</td>
</tr>
</tbody>
</table>

The following experiments have been performed with this dataset.
**input audio length**

Best quality images manually selected w.r.t the audio length

![Images](image1.png)  
0.3 s  0.7 s  1s

**Fine-tuned VGG classifier accuracy in % w.r.t the audio length (in seconds)**

<table>
<thead>
<tr>
<th>Length</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.3 s</td>
<td>89.12</td>
</tr>
<tr>
<td>0.7 s</td>
<td>81.16</td>
</tr>
<tr>
<td>1 s</td>
<td>90.25</td>
</tr>
</tbody>
</table>

The following experiments have been performed with this length.
input audio length

The more voice frames in the audio, the easier for the network to learn the identity

true identity: Jaime Altozano

true identity: Mely
The following experiments have been performed with 128x128 image resolution.
Wav2Pix

EXPERIMENTS

**identity classifier**

Fine-tuned VGGFace classifier accuracy in % w.r.t the model

<table>
<thead>
<tr>
<th>Roldán</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.69</td>
<td>50.08</td>
</tr>
</tbody>
</table>

Close to randomness! (10%)
image generation for unseen voice

The network does **not generalize** for unseen IDs!!
audio interpolation

The network does not generate faces for audios which do not contain distinguishable voice. The model has learned to identify speech in audio frames.
The model performs a good generalization for unseen speech of seen IDs.
CONCLUSIONS
CONCLUSIONS

In comparison to Roldán et. al. network, our contributions allows the final model:

- Generate images of **higher quality** due to the network’s capacity increase
- Generate **more face-looking** images for unseen speech
- Preserve the **identity** better for unseen speech
- Obtain better results with a **smaller dataset** (~70% smaller in terms of memory size)
- Obtain results that can be evaluated in terms of quality, face appearance and identity preservation with three different metrics

However,

- No generalization is achieved for **unseen** ID’s
- The dataset needs to be very clean in order to obtain notable results. The building process is very time-consuming