Audio-Driven Facial Animation by Joint End-to-End Learning of Pose and Emotion

Authors: Tero Karras, Timo Aila, Samuli Laine, Antti Herva & Jaakko Lehtinen
Presenters: Lucía Conde Moreno & Gurharmeet “Gaffy” Singh
Introduction

Expressive facial animation is essential for CGI.
Introduction

Vision-based performance capture

as part of most production pipelines
Introduction

Vision-based performance capture as part of most production pipelines

... but it has drawbacks

- Production cost of HQ facial animation still very high
- Need of (same) actor to be on location for new shots
- Big variability in voices and physical setups
- Big latency for some applications (e.g. VR avatars)
Introduction

Current trend in the industry:

Vision-based systems only for key animations + Audio-based systems for bulk (in-game) material
Related work

2 types of systems

- Linguistics-based
  1. Audio + transcript for knowledge about phonemes content
  2. Animation based on visemes, through (predefined) rules of articulation

- Machine learning-based
  1. Training based on observed face dynamics (e.g. video)
  2. Learning of coarticulation rules
  3. Testing with audio and/or text

System

- Audio
- Text

2D/3D mesh animation
Captured frames-based video
Related work

2 types of systems

Linguistics-based
- Phonemes ↔ visemes many-to-many mapping (JALI [Edwards et al, 2016], [Taylor et al. 2016])
- FaceFX/Dominance model [Massaro et al. 1993, 2012]

Machine learning-based
- (Most work) Concatenation, blending and/or warping of reused captured video frames [Anderson et al, 2013; Ezzat et al. 2002]
- 3D mesh animation (mostly text as input) [Shabus et al, 2014], separating speech and emotional state [Wampler et al. 2007]
Related work

2 types of systems

- Linguistics-based
  - Phonemes ↔ visemes many-to-many mapping (JALI [Edwards et al, 2016], [Taylor et al, 2016])
  - FaceFX/Dominance model [Massaro et al, 2012]

- Machine learning-based
  - (Most work) Concatenation, blending and/or warping of reused captured video frames [Anderson et al, 2013; Ezzat et al. 2002]
  - 3D mesh animation (mostly text as input) [Shabus et al, 2014], separating speech and emotional state [Wampler et al. 2007]

Example: Ezzat et al. 2002

Related work

**System**
- Audio
- Text

**2D/3D mesh animation**
- Captured frames-based video

**2 types of systems**
- **Linguistics-based**
  - Accumulated **complexity** (very particular and subtle movements of mouth and lips for certain sounds)
  - **Language-specific** rules
  - **Perfect transcript needed**
  - **Lack of rules for full face** animation (aside jaw and lips)
- **Machine learning-based**
  - Limited to **predefined emotions** (or no consideration of emotions)
  - **Audio** generally used just for control parameters, **not as input**
  - (Frames-based) **Fixed viewpoint, rigid animation**
  - [Deep learning] If conflicting outputs, result is averaged (does **not resolve ambiguities**)

**Disadvantages**
What does the paper propose?

Deep learning system for real-time, low latency 3D facial animation based on speech audio input
What does the paper propose?

**Deep learning** system for **real-time, low latency**

**3D facial animation** based on **speech audio** input

<table>
<thead>
<tr>
<th>Data \ Phase</th>
<th>Training</th>
<th>Testing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong></td>
<td>HQ animation data from vision-based performance capture + speech audio</td>
<td>Input waveforms from speech audio (+ emotion vectors)</td>
</tr>
<tr>
<td><strong>Output</strong></td>
<td>Emotion vectors database</td>
<td>3D vertex coordinates of face model</td>
</tr>
</tbody>
</table>
What is the primary goal of the paper?

- **Model** the *speaking style* of an actor...
- ... *to generate* plausible, expressive 3D (complete) facial animation...
- ... *based only on* vocal audio track
What are the possible applications?

- In-game dialogue
- Low-cost localization
- VR avatars
- Telepresence
- ...
Keywords & concepts

- Deep learning
- Facial animation
- Phonemes: perceptually distinct sounds in a language
- Visemes: visual counterpart of phonemes
- Coarticulation: characteristic of speech where boundaries between discrete units (e.g. phonemes) are blurred and occur in a smooth transition
**Formants**: resonance frequencies of linear filters, which carry info about phoneme content of speech

**Latent data**: info not inferable from audio alone

**Pangrams**: sentences designed to contain as many phonemes as possible
Task: given short audio window, infer facial expression at center of window
INPUT

- Speech audio signal...
  - 16kHz
  - Mono
  - Normalized to full dynamic range
- ... converted to compact **2D representation**:
  - Linear filter
    - Vocal tract: phoneme content
  - Excitation signal
    - Vocal cords: pitch, timbre
System architecture

INPUT

Emotional state (database)
- Multidimensional vector(s)
- Latent data (e.g. facial expression, speaking style, coarticulation patterns)
- Learned during training
Paper content | System architecture

Output network

OUTPUT

- Per-vertex vectors of difference from neutral pose (fixed topology mesh)

Animation of facial expression:
1. Window-sliding over vocal track
2. Network evaluation on each step
System architecture

Extract raw formant info (through fixed-function autocorrelation analysis)
**Refine raw formant info**
(to obtain short-term speech features e.g. intonation, emphasis, specific phonemes)
Paper content | System architecture

Audio window

Formant analysis network
1 special-purpose layer
5 convolutional layers

Articulation network
5 convolutional layers

Output network
2 fully connected layers

Output time-varying sequence of speech features

Emotional state
Analyze temporal evolution of features
System architecture

- 1 special-purpose layer
- 5 convolutional layers
- Output set of abstract feature vectors (representing facial pose)
- 2 fully connected layers
System architecture

1 special-purpose layer
5 convolutional layers
Output final 3D positions of vertices in mesh

Formant analysis network
Articulation network
Output network

2 fully connected layers
Paper content | System training

Ground truth data:

3D vertex positions of actor face obtained through 9-camera system (DI4D PRO) pointing at head of actor

1. Reconstruction of (unstructured) mesh
2. Projection of template mesh
3. Tracking of template mesh during performance
4. Export of vertex positions of mesh
Ground truth data:

**3D vertex positions** of actor face obtained through **9-camera system** *(DI4D PRO)* pointing at head of actor
Paper content | System training

Training dataset:

2 fragments (3-5 min) per character (2)

- Pangrams (with good coverage of expressions/emotions range)
- In-character material (representative shots, from script)

Lively, various expressions
Similar constant expression
Loss function for CNN - 3 terms (to add stability):

- **Position** (ensure correct location of vertices)
  \[ P(x) = \frac{1}{3V} \sum_{i=1}^{3V} \left( y(i)(x) - \hat{y}(i)(x) \right)^2 \]

- **Motion** (ensure correct movements of vertices)
  \[ M(x) = \frac{2}{3V} \sum_{i=1}^{3V} \left( m \left[ y(i)(x) \right] - m \left[ \hat{y}(i)(x) \right] \right)^2 \]

- **Regularization** (avoid short-term variation)
  \[ R'(x) = \frac{2}{E} \sum_{i=1}^{E} m \left[ e(i)(x) \right]^2 \]

+ normalization scheme for **automatic weight balancing**
+ data augmentation through **random time-shifting**
System testing

Before testing...

Manual **mining** of emotion database **for robust vectors**

Main steps:

1. **Pick few audio windows** (from validation data) containing:
   a. Bilabials (mouth closed)
   b. Vowels (mouth open)
2. **Select vectors matching those windows** by scanning emotion database
3. **Input validation audio** track and inspect inferred facial motion for each selected vector
   a. **Remove vectors** causing **unnatural motion** (short-term effects)
System testing

Before testing...

Manual mining of emotion database for robust vectors

Unnatural motion $\rightarrow$ Removed
System testing

For testing...

Examination of output for novel audio clips with remaining emotion vectors

+ assignment of semantic meaning (e.g. “happy”) to each vector
Blind user study:

- 20 participants unfamiliar with animation
- **Pairwise comparison on perceived naturalness** among:
  - Proposed method
  - Performance capture (ground truth)
  - FaceFX
- Network trained with audio from single speaker (English)

2 user studies:

1. 2 characters, 13 clips, assignment of single emotion vector for each clip
2. 14 clips in 5 different languages
<table>
<thead>
<tr>
<th>Character 1</th>
<th>PC vs. Ours</th>
<th>PC vs. DM</th>
<th>Ours vs. DM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clip 1</td>
<td>14 6</td>
<td>19 1</td>
<td>19 1</td>
</tr>
<tr>
<td>Clip 2</td>
<td>16 4</td>
<td>20 0</td>
<td>20 0</td>
</tr>
<tr>
<td>Clip 3</td>
<td>17 3</td>
<td>20 0</td>
<td>19 1</td>
</tr>
<tr>
<td>Clip 4</td>
<td>16 4</td>
<td>19 1</td>
<td>18 2</td>
</tr>
<tr>
<td>Clip 5</td>
<td>9 11</td>
<td>16 4</td>
<td>17 3</td>
</tr>
<tr>
<td>Clip 6</td>
<td>19 1</td>
<td>19 1</td>
<td>17 3</td>
</tr>
<tr>
<td>Clip 7</td>
<td>15 5</td>
<td>20 0</td>
<td>17 3</td>
</tr>
<tr>
<td>Clip 8</td>
<td>13 7</td>
<td>19 1</td>
<td>15 5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Character 2</th>
<th>Male</th>
<th>Female</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ours</td>
<td>DM</td>
<td>Ours</td>
<td>DM</td>
</tr>
<tr>
<td>English 1</td>
<td>12</td>
<td>8</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>English 2</td>
<td>16</td>
<td>4</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>English 3</td>
<td>16</td>
<td>4</td>
<td>17</td>
<td>3</td>
</tr>
<tr>
<td>French</td>
<td>16</td>
<td>4</td>
<td>18</td>
<td>2</td>
</tr>
<tr>
<td>German</td>
<td>16</td>
<td>4</td>
<td>12</td>
<td>8</td>
</tr>
<tr>
<td>Italian</td>
<td>14</td>
<td>6</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>Spanish</td>
<td>19</td>
<td>1</td>
<td>13</td>
<td>7</td>
</tr>
</tbody>
</table>

| Total Votes | 109  | 31  |
| Total Ratio | 78%  | 22% |
Output of video-based performance capture generally perceived as more natural than the other two [expected]

<table>
<thead>
<tr>
<th>Character 1</th>
<th>PC vs. Ours</th>
<th>PC vs. DM</th>
<th>Ours vs. DM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total</td>
<td>201</td>
<td>246</td>
<td>225</td>
</tr>
<tr>
<td>Votes</td>
<td>59</td>
<td>14</td>
<td>35</td>
</tr>
<tr>
<td>Ratio</td>
<td>77%</td>
<td>95%</td>
<td>87%</td>
</tr>
<tr>
<td></td>
<td>23%</td>
<td>5%</td>
<td>13%</td>
</tr>
</tbody>
</table>
Paper content | Results

- Output of video-based performance capture generally perceived as more natural than the other two [expected]

- Proposed method outperforms FaceFX in every case

<table>
<thead>
<tr>
<th>Character 1</th>
<th></th>
<th>Character 2</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Female</td>
<td>Male</td>
<td>Female</td>
</tr>
<tr>
<td>Ours</td>
<td>DM</td>
<td>Ours</td>
<td>DM</td>
</tr>
<tr>
<td>English 1</td>
<td>12</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>English 2</td>
<td>16</td>
<td>4</td>
<td>18</td>
</tr>
<tr>
<td>English 3</td>
<td>16</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>French</td>
<td>16</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>German</td>
<td>16</td>
<td>4</td>
<td>16</td>
</tr>
<tr>
<td>Italian</td>
<td>14</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td>Spanish</td>
<td>19</td>
<td>1</td>
<td>20</td>
</tr>
</tbody>
</table>

Total votes: 225, 35 votes for Ours, 87%.

Total votes ratio: 78% for Ours, 22% for DM.
Results

- Output of video-based performance capture generally perceived as more natural than the other two [expected]

- **Proposed method outperforms FaceFX in every case**

- Worse results for female speakers (male character, perceived unnatural anyway?)

<table>
<thead>
<tr>
<th>Character 1</th>
<th>Character 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>Male</td>
</tr>
<tr>
<td>Ours</td>
<td>12</td>
</tr>
<tr>
<td>DM</td>
<td>17</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total Votes</th>
<th>109</th>
<th>31</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ratio</td>
<td>78%</td>
<td>22%</td>
</tr>
</tbody>
</table>
Paper content | Results

- Output of video-based performance capture generally perceived as more natural than the other two [expected]

- **Proposed method outperforms FaceFX in every case**

- Worse results for female speakers (male character, perceived unnatural anyway?)

- Proposed method, **sensitive to voice/speaking style/tempo**
  - Variation between languages << variation between different speakers of same language

<table>
<thead>
<tr>
<th></th>
<th>Character 1</th>
<th>Character 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Male Ours</td>
<td>Male DM</td>
</tr>
<tr>
<td>French</td>
<td>16</td>
<td>15</td>
</tr>
<tr>
<td>German</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Female Ours</td>
<td>Female DM</td>
</tr>
<tr>
<td>French</td>
<td>18</td>
<td>15</td>
</tr>
<tr>
<td>German</td>
<td>12</td>
<td>16</td>
</tr>
</tbody>
</table>
Paper content | Results

1st user study

2nd user study
Contributions

Research problem (disadvantages of previous work):

- Audio not containing enough info to resolve ambiguities (e.g. same sound, different facial expressions)
- Tendency of neural networks towards mean when resolving ambiguities

3 main contributions to tackle research problem:

- CNN tailored for human speech processing and different speakers generalization
- Novel way of CNN to discover variations (latent info) in data
- 3-way loss function to ensure temporal stability
Contributions | Advantages of the method

- **No predefined emotions** (≠ most machine learning methods, e.g. Cao et al, 2005)

- **Full face animation** (≠ linguistics-based methods)

- Applicable when **free-viewpoint, flexible 3D** models rendering/animation is needed (≠ methods based on concatenation/blending of video frames e.g. Ezzat et al, 2012)

- **Applicable** to **different face meshes** (retargeting) and **synthetic audio**

- **Good temporal stability** (thanks to random time-shifting and motion term)

- **Good interpolation** of emotional states (thanks to manual vector removal)
Drawbacks of the method

- **Incapacity** to cover residual motion (e.g. eye-saccades, blinks, head, tongue)
  - Only motion related to articulation
Drawbacks of the method

- **Incapacity** to cover residual motion (e.g. eye-saccades, blinks, head, tongue)
  - Only motion related to articulation

Fixable through external animation procedures?
Drawbacks of the method

- **Incapacity** to cover *residual motion* (e.g. eye-saccades, blinks-, head, tongue)
  - Only motion related to articulation

- **Emotion vectors** have **no semantic meaning** (e.g. “happy”, “sad”) and potentially not generalizable
  - Not useful for inference if training data not covering all phonemes + coarticulation rules of every emotional state
Drawbacks of the method

- **Incapacity** to cover residual motion (e.g. eye-saccades, blinks-, head, tongue)
  - Only motion related to articulation

- **Emotion vectors** have no semantic meaning (e.g. “happy”, “sad”) and potentially not generalizable
  - Not useful for inference if training data not covering all phonemes + coarticulation rules of every emotional state

- **Bad performance** if testing data hugely differs from training data, mainly:
  - Different (mostly faster) **tempo**
  - Input **volume** level of testing data not **normalized**

→ Solvable by increasing size of training data? (longer recordings, various languages?)
Drawbacks of the method
Drawbacks of the method

- Possibility of incorrect inference of articulation rules
  - e.g. confusion of B and G
Thanks for your attention!
Questions?
Discussion

In your opinion...

What should be the main approach in the industry/for future research?

- Focusing on audio-based systems? (improving their performance)
- Focusing on visual-based systems? (decreasing their computational cost)
Discussion

In your opinion...

What should be the main approach in the industry/for future research?

- Focusing on **audio-based** systems? (improving their performance)
- Focusing on **visual-based** systems? (decreasing their computational cost)

Some ideas:
- Audio-based systems can be perceptively similar to visual-based ones in lower-quality settings, which could be enough for some applications (plus considering their most probable improvement in some years)
- Computational cost is mostly tied to improvements in hardware, so the major developments to influence visual-based systems would probably take place within that field (instead of within Computer Animation research)
- If production (monetary) cost is not an issue, there is always the chance of using visual-based systems anyway
Discussion

Open debate:

Ethical implications of the system (and similar alternatives)?

Should they be taken into account (in every case, in specific cases)?

Should they never limit research?

E.g.:

- Possibility of creating fake video of real person saying something specific (legal implications?)
- Substitution of human workers (assistants, actors)
Discussion

Open debate:

Ethical implications of the system (and similar alternatives)?

Should they be taken into account (in every case, in specific cases)?

Should they never limit research?

Some ideas:

• The results of almost every kind of research are in a way ambivalent (could be applied for either ethically good or bad purposes) but that falls out of the scope of that research

• If the creation of fake videos were indeed possible, legislation would probably be adjusted not to consider these files as valid legal proof (e.g. WhatsApp messages), which avoids its risk but could also affect lawsuits negatively

• Research should not be limited as long as it follows an established ethical policy and its guidelines, rules and considerations (e.g. set by the university where it takes place)