Where do we stand today?

Face to Virtual Face Communication

- What are technologies?
- What is the progress?
- What is still absent?
- What is the future?

What are the technologies?

- Facial Animation
- Speech Synthesis and Recognition
- Face Recognition and Tracking
- NLP and AI

Face to Face Communication

Speech Animation: Hierarchy

<table>
<thead>
<tr>
<th>Step</th>
<th>Technology</th>
<th>Methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temporized phonemes from speech (synthetic or real)</td>
<td>Phoneme recognition</td>
<td>Manual, semi-automatic or automatic</td>
</tr>
<tr>
<td>Phoneme transition</td>
<td>Co-articulation</td>
<td>Rules based, automatic</td>
</tr>
<tr>
<td>Viseme generation and animation</td>
<td>Viseme definition, Synchronization with sound</td>
<td>Automatic</td>
</tr>
</tbody>
</table>
Animated Talking Heads – a typical system

- Temporized phonemes
- Facial animation parameters
- Audio signal
- Synchronization
- Animatable face model

Speech - synthetic/real

Speech Animation from Natural Voice
(for cloned avatars)

(Only) Acoustic Analysis

- Speech signal
- Speech Processing
- Parameter Database for Phonemes
- NN/HMM
- Phonemes

Acoustic and Visual Analysis

- Speech signal
- Speech Processing
- NN/HMM
- Parameters for facial movements
- Video 3D data
- Computer Vision
  - Facial states, lip specking, width, height, protrusion etc.
### Comparison

<table>
<thead>
<tr>
<th>Acoustic Analysis</th>
<th>Acoustic-Visual Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>The output “phonemes”, suitable for any facial animation system</td>
<td>Output parameters (facial states/lip width, height etc.) are tied to a particular animation system</td>
</tr>
<tr>
<td>Resulting facial animation not affected by training database</td>
<td>Resulting facial animation closely affected by the training database</td>
</tr>
<tr>
<td>Ease in training data collection (only speech)</td>
<td>Training data is synchronized speech and video/3D capture</td>
</tr>
<tr>
<td>Only lip/mouth movements can be generated</td>
<td>Technique can be used for synthesis of other facial movements (eyebrows, nod)</td>
</tr>
<tr>
<td>Co-articulation model needs to be applied to resulting phonemes</td>
<td>Co-articulation effect is inherently taken care of in analysis</td>
</tr>
</tbody>
</table>

Greater language dependence | Less language dependence

### Challenges

- Independent of language and speaker
- Independent of face model used for animation
- Minimal training requirements
- Simplicity of tools, algorithm and implementation

### Speech Driven Talking Head: an example

![Diagram of speech driven talking head](image)

**Parameter extraction**

- Choice of LP Analysis
  - LP derived reflection coefficients are directly related to vocal tract shape [Wakita]
  - Phonemes can be characterized by vocal tract shape
- Limitations
  - Works well for vowels so we choose the most common five vowels /a/, /e/, /i/, /o/, /u/.
  - For the consonants?
Use of Neural Network

Typical plots for reflection coefficients for five chosen vowels

Three layer back propagation
12 input nodes, 10 hidden nodes, 5 output nodes
Five vowels used /a/, /e/, /i/, /o/, /u/
12 male and 5 female speakers

Results of NN training

<table>
<thead>
<tr>
<th></th>
<th>/a/</th>
<th>/e/</th>
<th>/i/</th>
<th>/o/</th>
<th>/u/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expected</td>
<td>241</td>
<td>2</td>
<td>15</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>Recognized</td>
<td>0</td>
<td>177</td>
<td>89</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>/i/</td>
<td>0</td>
<td>3</td>
<td>301</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>/o/</td>
<td>10</td>
<td>0</td>
<td>0</td>
<td>224</td>
<td>36</td>
</tr>
<tr>
<td>/u/</td>
<td>4</td>
<td>12</td>
<td>0</td>
<td>88</td>
<td>143</td>
</tr>
</tbody>
</table>

Energy Analysis

Vowel-Vowel Transition
Semi-vowels
Consonants

The parameters corresponding to the vowels are modulated

What more?

Zero crossing for affricates and unvoiced fricatives (/sh/, /dzh/) and /h/
Zero crossing rate is 49 per 10 msec for unvoiced, and 14 per 10 msec for voiced speech
Speech Animation from Text/Synthetic Speech (for autonomous virtual humans)

Synthetic Speech Driven Talking Head

Text to Speech

Temporized phonemes
Facial animation parameters

Text

Co-articulation

Co-articulation is a phenomenon observed during fluent speech, in which facial movements corresponding to one phonetic or visemic segments are influenced by those corresponding to the neighboring segments.

Example: a V1-C-V2 sequence where V1 is un-protruded (eg. ‘a’) and V2 is protruded (eg. ‘u’)

Articulatory Gesture Model

- Each speech segment (typically a viseme) has dominance that increases and decreases over time
- Adjacent visemes have overlapping dominance functions that will blend over time
- Each viseme may have a different dominance function for each articulator


Co-articulation Models for Talking Head

Pelachaud (1991):
“Look ahead” model based on deformability of phonemes
Also considered muscle contraction times

Cohen & Massaro (1992):
Non-linear dominance and blending functions designed for each phoneme

In Summary
Define weight (dominance), and overlap according to phoneme group.

Performance Driven Facial Animation
Optical tracking with several cameras
Parameterized (FAP) synthetic face
Enhances realism to a great degree
Enables design of the building blocks
Limitations: complex equipment, availability of skilled performer

Realism in Talking Heads
Can we combine flexibility of facial animation design and realism of performance driven facial animation? How?
Optical tracking data
Realistic building blocks: expressions and visemes
Rules for expression mixing and blending
Expression-Viseme space
Realistic Facial Animation
Statistical analysis
**What is PCA**

PCA is a well-known multivariate statistical analysis technique aimed at:

- reducing the dimensionality of a dataset, which consists of a large number of interrelated variables
- retaining as much as possible of the variation present in the dataset
- transforming the existing dataset into a new set of variables called the principal components (PC)

The PCs are uncorrelated and are ordered so that the first few PCs retain the most of the variation present in all of the original dataset.

---

**Why PCA**

For facial capture:
- High correlation between facial feature points
- Large amount of capture data for speech
- Capturing individual as well as collective movement dynamics important during expressive speech

Use of MPEG4 Feature points and FAP

---

**Data Capture**

- Optical Tracking system: Vicon
- 27 optical markers, 6 Cameras
- Extraction of 3D positions of markers
- 100 phoneme rich sentences from TIMIT database
- 3D position data of 14 markers around lips and cheeks used for PCA

---

**Data Analysis**

- 3D position data for every frame
- Principal Component Analysis
- Principal Components (basis vectors of the PC space)
- Transformation matrix ‘T’
- 'expression/viseme' space

Analysis results into a transformation between 3D position space and the newly constructed expression/viseme space
What are the Principal Components

The facial movements are controlled by single parameters, as opposed to several MPEG4 parameters needed to control the same facial movement.

Eg. ‘Open Mouth ’ affects not only lips, but jaw and cheek region also.

Thus the Principal Components take care of global facial movements using minimum number of parameters and provide higher level parameterization for facial animation design.

(a) Open mouth
(b) Lip protrusion
(c) Lip sucking
(d) Raise corner lips

Expression and Viseme Space

- The ‘Principal Components’ form the basis or the ‘principal axes’ of the abstract Expression and Viseme space.
- Each point in the Expression and Viseme space is a facial expression, a viseme, or a combination.
- Transition in this space from one point (expression) to another, results in smooth and realistic transition in the 3D position space giving a new way of achieving keyframe animations.
- A combination of points in this space results in realistic blending and combination of visemes and expressions in 3D position space, and hence a realistic expressive speech animation.

Application to Speech Animation

Each expression and viseme is a vector in the Expression and Viseme space.

Mixing between Viseme and Expression is a simple vector addition in that space.

Transforming back to 3D position space results into “Expressive Speech.”

For happy expression, PC2 and PC3 are most effective, as it controls lip protrusion.

For sad expression, PC4 and PC6 is found to be most effective, that controls corner lip movements.
Blending Speech with Expressions

Expression vector

Viseme vector

Expression and Viseme space

Transformation \( T \)

Expressive viseme

3D position space (convertible to FAP)

Further Reading...