Speech-driven Facial Animation
– how to learn a stream-to-stream mapping? –

Hiroshi Shimodaira (ICCS, CSTR)
Junichi Yamagishi, Gregor Hofer, Michael Berger
Speech-driven facial animation?

It’s a computer animated talking face: “talking head”

Two types of talking heads:

1. Text driven

   Text
   "Hello world"
   (discrete symbols)

   Facial animation

   Speech
   (continuous signals)

2. Speech driven

   Speech
   "Hello world"
   (continuous signals)

   Facial animation
   (continuous signals)
Applications

- Film industries
  (a good animator can produce 4 ~ 5 frames of high quality speech animation per day)
- Computer games
- Agent-based system (spoken dialogue systems)
- Education (pronunciation training), psychotherapy
- Simulator for scientific research
Examples of facial animation

Current automatic facial animation systems:
- Lip motion synthesis synchronised with speech (Lip-sync)
  - SyncFace (J. Beskow, KTH, 2004)
- Lip-sync + facial expression (rule-based)
  - Greta (C. Pelachaud, Université de Paris 8, 2003)

There is still far to go to achieve something like this:
- High quality motion capture (appearance-based)
  - Meet Emily (Image Metrics Inc., 2008)
What we want to do?

- Synthesise realistic head and facial motions from given speech without using semantics.
  - trainable on real data
  - adaptable to new speakers / styles
  - able to generate stochastic motions

```
Intention
  ↓
Planning
  ↓
Language
  ↓
Motor control
  ↓
Articulation
  ↓
Speech
  ←
Non-verbal language
  ↓
Motor control
  ↓
Muscle movement
  ↓
Head motions
```
Problem formulation

Define the problem as a probabilistic optimisation problem:

\[ O^{M*} = \arg \max_{O^M} P(O^M | O^S) \]

\[ O^S = o_1^S, o_2^S, \ldots, o_{LS}^S \quad \text{sequence of speech features} \]

\[ O^M = o_1^M, o_2^M, \ldots, o_{LM}^M \quad \text{sequence of motion features} \]

- It’s not a point-to-point mapping, but a stream-to-stream mapping of real-valued vectors, in which context should be taken into account.

<table>
<thead>
<tr>
<th>Input \ Output</th>
<th>Discrete</th>
<th>Continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discrete</td>
<td>machine translation</td>
<td>text-to-speech</td>
</tr>
<tr>
<td>Continuous</td>
<td>speech recognition</td>
<td>(this problem)</td>
</tr>
</tbody>
</table>
Problem formulation (cont. 2)

- Difficulty
  - The mapping seems to be complex, non-linear, context dependent.
  - Different POIs have different dependencies and different levels of synchrony on/with speech.
  - It’s not clear what acoustic/language features and model unit should be used to predict motions of POI.

<table>
<thead>
<tr>
<th>POI</th>
<th>dependency on speech</th>
<th>literature</th>
</tr>
</thead>
<tbody>
<tr>
<td>mouth &amp; jaw head</td>
<td>high</td>
<td>many</td>
</tr>
<tr>
<td>eye (gaze, blink)</td>
<td>moderate?</td>
<td>several</td>
</tr>
<tr>
<td>eyebrow</td>
<td>weak?</td>
<td>very few</td>
</tr>
</tbody>
</table>
Our approach

- Employ generative models of reasonably small unit.
- Use human readable model unit
- Use models capable of handling different levels of synchrony between the two streams.

Assuming we give a label sequence to each stream:

\[ u^M = u_1^M, u_2^M, \ldots \text{ motion label seq.} \]
\[ u^S = u_1^S, u_2^S, \ldots \text{ speech label seq.} \]

\[ O^{M*} = \arg \max_{O^M} P(O^M | O^S) \]
\[ = \arg \max_{O^M} \sum_{u^M} \sum_{u^S} P(O^M, u^M, u^S | O^S) \]
\[ = \arg \max_{O^M} \sum_{u^M} \sum_{u^S} P(O^M | u^M, u^S, O^S) P(u^M | u^S, O^S) P(u^S | O^S) \]
Our approach (cont. 2)

Assuming some conditional independencies between variables,

\[
O^M* = \arg \max_{O^M} \sum_{u^M} P( O^M | u^M ) \sum_{u^S} P( u^M | u^S ) P( u^S | O^S )
\]
Our approach (cont. 3)

Using model level synchrony as a constraint, we could assume a common unit \{u\}.

\[ O^{M*} \approx \arg \max_{O^M} \sum_u P(O^M|u)P(u|O^S) \]

\[ \approx \arg \max_{O^M} P(O^M|u^*) \]

\[ u^* = \arg \max_u P(u|O^S) \]
Training & synthesis

Training
Train HMMs with a complete data set (two streams with labels)

Synthesis
1. Decode a given speech into a unit sequence [recognition]
2. Generate a motion sequence from the unit sequence [synthesis]
   (trajectory HMMs)
Model unit for head motion synthesis

- Possible units

<table>
<thead>
<tr>
<th>Domain</th>
<th>Feature</th>
<th>unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>speech</td>
<td>text</td>
<td>phoneme/syllable</td>
</tr>
<tr>
<td></td>
<td>acoustic</td>
<td>word</td>
</tr>
<tr>
<td></td>
<td></td>
<td>phrase</td>
</tr>
<tr>
<td>head motion</td>
<td>direction</td>
<td>manual</td>
</tr>
<tr>
<td></td>
<td>(angles)</td>
<td>clustering</td>
</tr>
</tbody>
</table>

- Selected unit: 4 types of head motions

  postural shift : the head shifts axis of movement
  shake & nod    : lateral movement around one axis
  pause          : no movement / rest position
  default        : non-distinctive movement
                  or slow movement
Speech-driven animation of
- mouth motion (lip-sync)
- head motion
- eyebrow motion
Conclusions

- Record more training data of good quality/resolution
- Investigate more complex models
  \[ O^{M*} = \arg \max_{O^M} \sum_{u^M} P(O^M | u^M) \sum_{u^S} P(u^M | u^S) P(u^S | O^S) \]
  but how to implement this?
  - integrate with physical models
- Synthesise motions of other POIs, e.g. eye blink/gaze
- Evaluate synthesised animation