Speech-driven Facial Animation

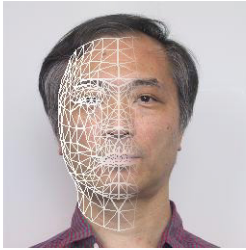
– how to learn a stream-to-stream mapping? –

Hiroshi Shimodaira (ICCS, CSTR)

Junichi Yamagishi, Gregor Hofer, Michael Berger

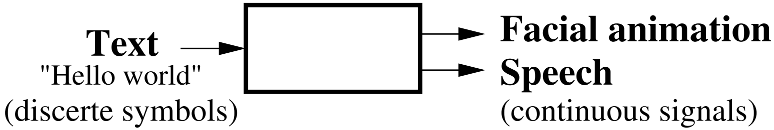
Speech-driven facial animation?

It's a computer animated talking face: "talking head"

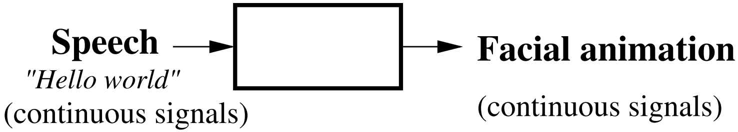


Two types of talking heads:

1. **Text driven**



2. **Speech driven**



Applications

- **Film industries**

  (a good animator can produce 4 $\sim$ 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:

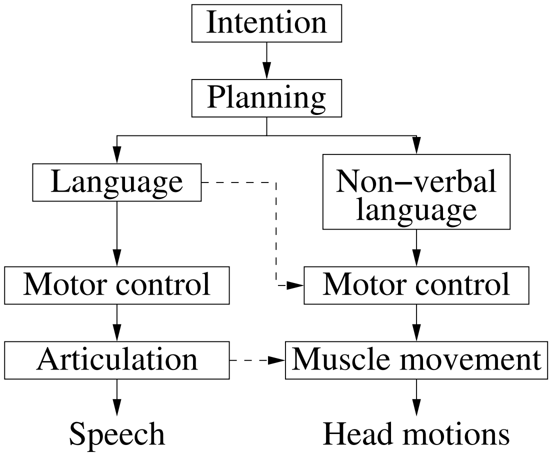
- **Lip motion synthesis synchronised with speech (Lip-sync)**
  - **SyncFace (J. Beskow, KTH, 2004)**
- **Lip-sync + facial expression (rule-based)**
  - **Greta (C. Pelachaud, Université de Paris 8, 2003)**

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



Problem formulation

Define the problem as a probabilistic optimisation problem:

$$\boldsymbol{O}^{M*} = \arg\max_{\boldsymbol{O}^M} P(\boldsymbol{O}^M | \boldsymbol{O}^S)$$

$$\boldsymbol{O}^S = \boldsymbol{o}_1^S, \boldsymbol{o}_2^S, \ldots, \boldsymbol{o}_{L^S}^S$$ **sequence of speech features**

$$\boldsymbol{O}^M = \boldsymbol{o}_1^M, \boldsymbol{o}_2^M, \ldots, \boldsymbol{o}_{L^M}^M$$ **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.**

Input \Output	Discrete	Continuous
Discrete	machine translation	text-to-speech
Continuous	speech recognition	(this problem)

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.**

POI	dependency on speech	literature
mouth & jaw	high	many
head	moderate?	several
eye (gaze, blink)	weak?	very few
eyebrow	weak?	very few

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:

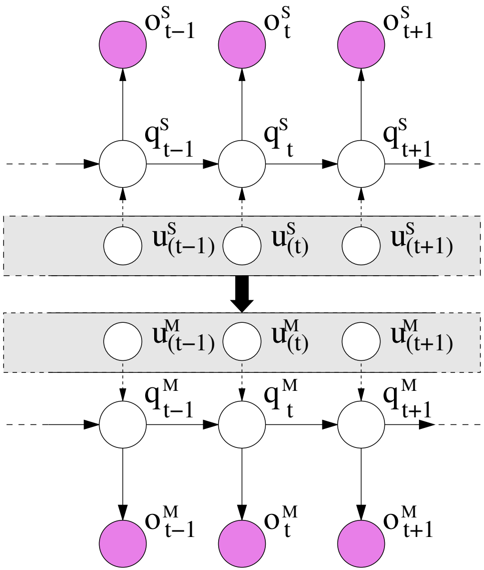
$$\boldsymbol{u}^M = u_1^M, u_2^M, \ldots \quad \textbf{motion label seq.}$$
$$\boldsymbol{u}^S = u_1^S, u_2^S, \ldots \quad \textbf{speech label seq.}$$

$$\begin{aligned}
\boldsymbol{O}^{M*} &= \arg\max_{\boldsymbol{O}^M} P(\boldsymbol{O}^M | \boldsymbol{O}^S) \\
&= \arg\max_{\boldsymbol{O}^M} \sum_{\boldsymbol{u}^M} \sum_{\boldsymbol{u}^S} P(\boldsymbol{O}^M, \boldsymbol{u}^M, \boldsymbol{u}^S | \boldsymbol{O}^S) \\
&= \arg\max_{\boldsymbol{O}^M} \sum_{\boldsymbol{u}^M} \sum_{\boldsymbol{u}^S} P(\boldsymbol{O}^M | \boldsymbol{u}^M, \boldsymbol{u}^S, \boldsymbol{O}^S) P(\boldsymbol{u}^M | \boldsymbol{u}^S, \boldsymbol{O}^S) P(\boldsymbol{u}^S | \boldsymbol{O}^S)
\end{aligned}$$

Our approach *(cont. 2)*

Assuming some conditional independencies between variables,

$$\boldsymbol{O}^{M*} = \arg\max_{\boldsymbol{O}^M} \sum_{\boldsymbol{u}^M} P(\boldsymbol{O}^M|\boldsymbol{u}^M) \sum_{\boldsymbol{u}^S} P(\boldsymbol{u}^M|\boldsymbol{u}^S) P(\boldsymbol{u}^S|\boldsymbol{O}^S)$$



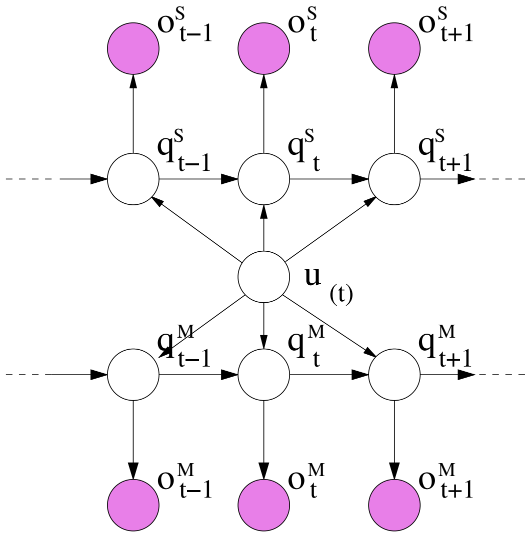
Our approach(cont. 3)

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

$$\boldsymbol{O}^{M*} \approx \arg\max_{\boldsymbol{O}^M} \sum_{\boldsymbol{u}} P(\boldsymbol{O}^M|\boldsymbol{u})P(\boldsymbol{u}|\boldsymbol{O}^S)$$

$$\approx \arg\max_{\boldsymbol{O}^M} P(\boldsymbol{O}^M|\boldsymbol{u}^*)$$

$$\boldsymbol{u}^* = \arg\max_{\boldsymbol{u}} P(\boldsymbol{u}|\boldsymbol{O}^S)$$



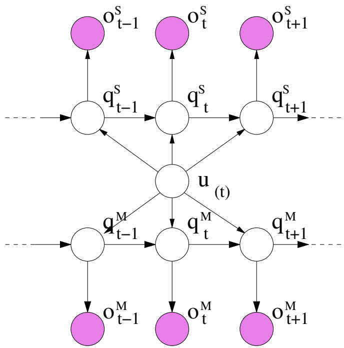
Training & synthesis

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

Domain	Feature	unit
speech	text acoustic	phoneme/syllable word phrase
head motion	direction (angles)	manual clustering

- 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

Video clip of a current system

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**