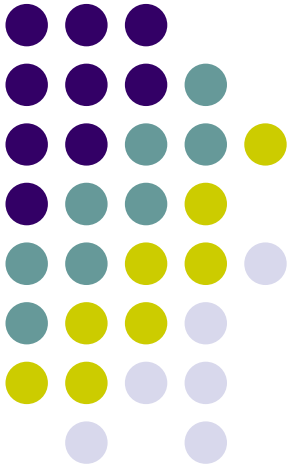
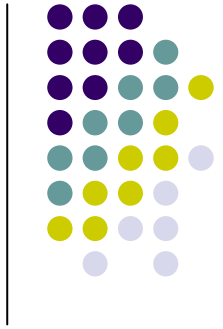
Trainable Videorealistic Speech Animation

Tony Ezzat, Gadi Geiger,
Tomas Poggio@ MIT

Presented by: Yinan Fan
04-26-2007

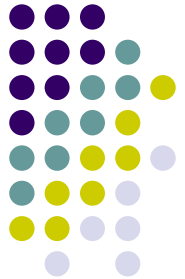


News Coverage



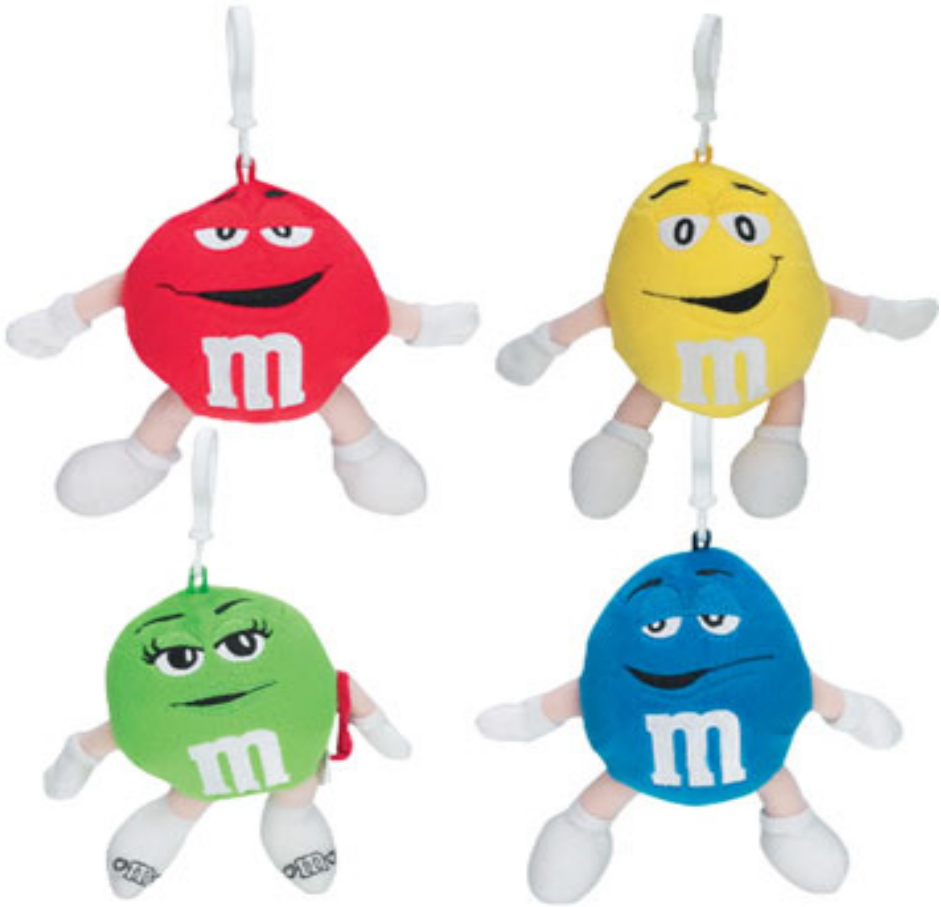
- *(July 23, 2002)* **REPORTS FROM SIGGRAPH-2002** - Wendy Ju: Character Animation
- *(July 2, 2002)* **CNN** : Video Research at MIT Puts Words into Mouths
- *(June 30, 2002)* **ASSOCIATED PRESS** - Theo Emery: Video Research at MIT Puts Words into Mouths, with Startling Results
- *(June 17, 2002)* **THE DISCOVERY CHANNEL** [Toronto, Canada] - Jennifer Scott **Video:*** Science, Lies & Videotape
- *(May 28, 2002)* **DER SPIEGEL** [Germany] - Marco Evers: Videomanipulation: Wie Bilder Luegen Lernen
- *(May 20, 2002)* **NBC TODAY SHOW** - Katie Couric: **Video:*** (100 Kbps) (300 Kbps)
- *(May 20, 2002)* **MIT NEWS OFFICE: TECH TALK** - Deborah Halber: Realistic Animation of Human Face Makes Simulated Talking Look Real
- *(May 16, 2002)* **NPR - "All Things Considered"** - Robert Siegel: **Audio:** MIT Video Lipsync
- *(May 16, 2002)* **TORONTO GLOBE & MAIL** - Graeme Smith: Computers Fake Moving Mouths
- *(May 15, 2002)* **BOSTON GLOBE** - Gareth Cook : At MIT, They Can Put Words in Our Mouths

Background



- Facial Modeling
  - 3D methods
  - Image-based methods: *photorealistic?,videorealistic? Parsimonius?*
    - Video Rewrite
- Speech Animation
  - Keyframe
  - Physics-based
  - Machine learning methods
  - Problem: *Motion, smoothness, dynamics, coarticulation effects…*
- MMM

MMM?

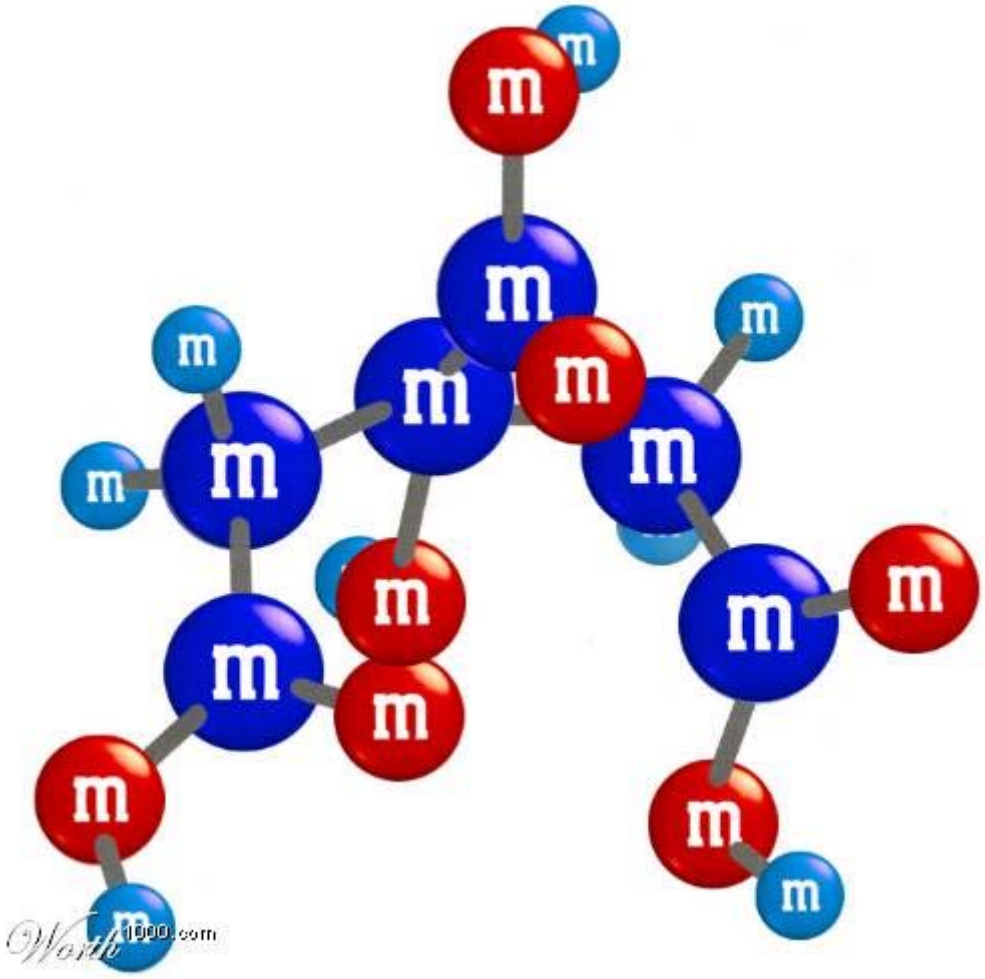


Well in some sense,… yes…

corpus…

...preprocessed
and sorted…

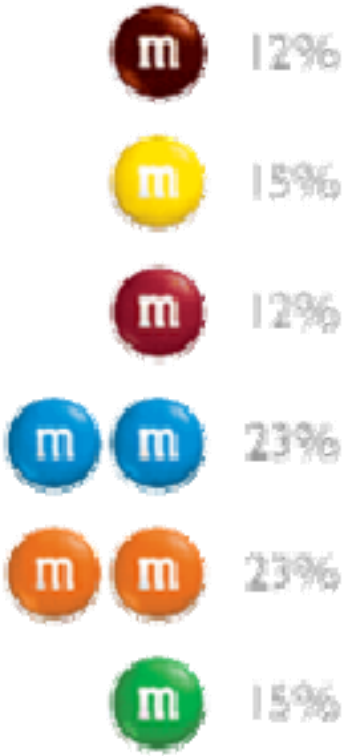
principle component selected…



relationship?

graph?

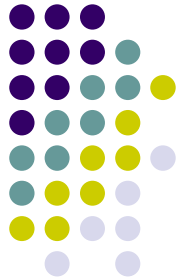
MM space?



data analysis…

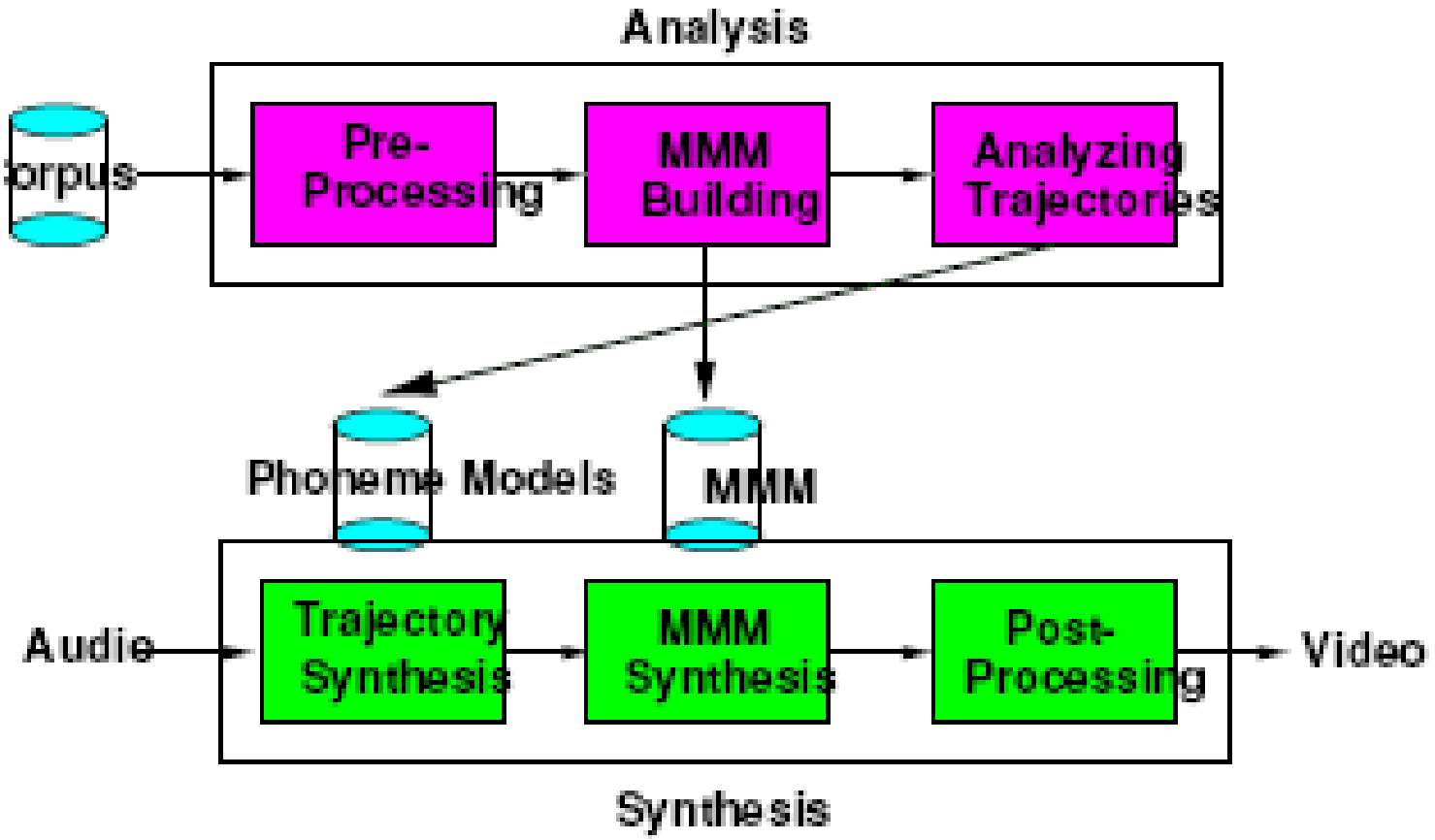
some new stuff!

MMM,...seriously

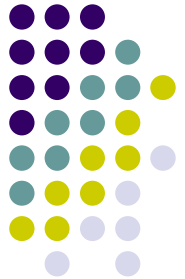


- Morphable Model Representation
- A low-dimensional space --paramaterized by shape parameters α and appearance parameters β
- A ``black box'' capable of performing
  - Synthesis
  - Analysis

System Overview

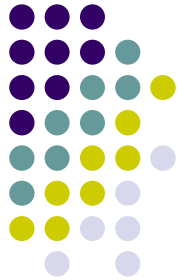


Corpus



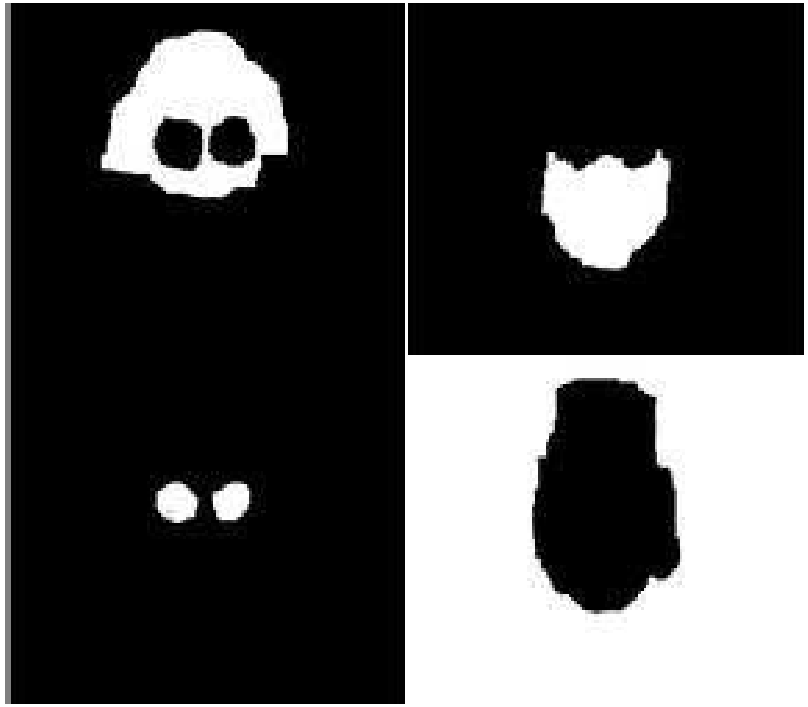
- A human subject uttering various uttaerace, in neutral expression
  - 640*480 of 29.97 fps NTSC, 44.1KHz
  - 15 minutes, 30000 frames
  - 152 one syllable words
  - 156 two syllable words
  - 105 short sentences

Pre-processing



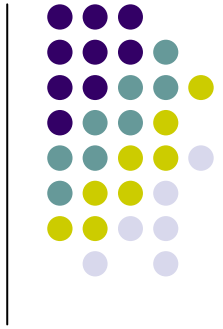
- Audio phonetically aligned
  (using CMU Sphinx system)
- Each image normalized ----head mask
  - Planar perspective deformation
  - Eye mask

Masks



- The only manual work

MMM: Definition



- A set of prototype images

$$\{I_i\}_{i=1}^{N}$$

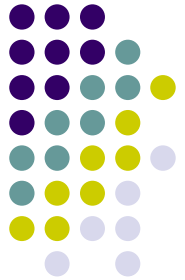
- A set of prototype flows

$$\{C_i\}_{i=1}^{N}$$

$$C_i(p) = \{d_x^i(p), d_y^i(p)\}$$

  - Using coarse-to-fine, gradient based optical flow algorithm

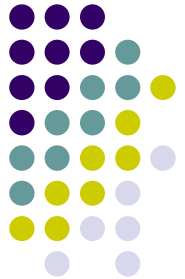
Building MMM



- Task:

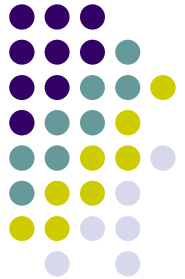
  choose image prototypes and compute correspondence

Building MMM



- EM-PCA
  - 15 PCA dimensions
  - $l_i \rightarrow p_i$
- K-Means Clustering
  - Mahalonobis distance metric: $d(p_m, p_n) = (p_m - p_n)^T \Sigma^{-1} (p_m - p_n)$
  - N=46: No explicit relationship to visems
- Dijkstra
  - Corpus graph
    - K-nearest neighbor frames (k=20), weighted by MD
  - Dijkstra shortest path => 46 correspondences

Synthesis



- Goal:

  Map (α , β) to an image in MMM

  - α : 46-dimensional → mouth shape
  - β : 46-dimensional -> mouth texture

Synthesis

- Steps:
  - Synthesize a new correspondence: $C_1^{synth} = \sum_{i=1}^{N} \alpha_i C_i$.
  - Forward Warp $W(I, C)$

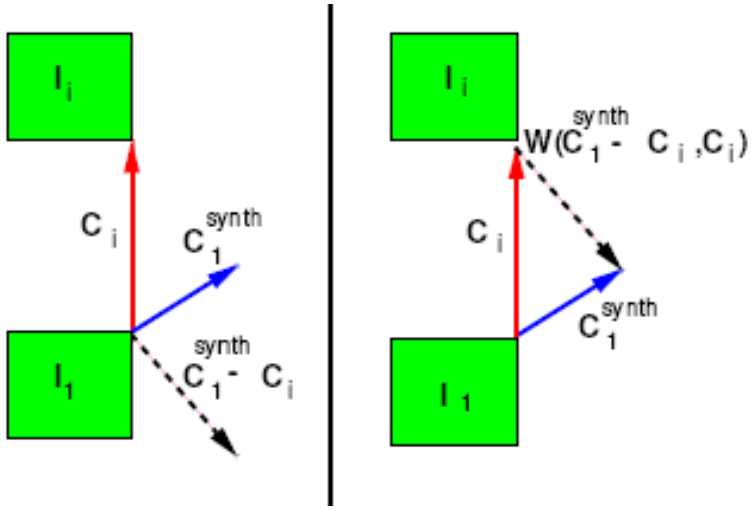
a

```
for j = 0...height,
   for i = 0...width,
      x = ROUND (i + αdx(i,j) );
      y = ROUND (j + αdy(i,j) );
      if (x,y) are within the image
         I^warped(x,y) = I(i,j);
```

Synthesis

- Steps:
  - Synthesize a new correspondence:
  - Forward Warp $W(I, C)$

$$C_1^{synth} = \sum_{i=1}^{N} \alpha_i C_i$$

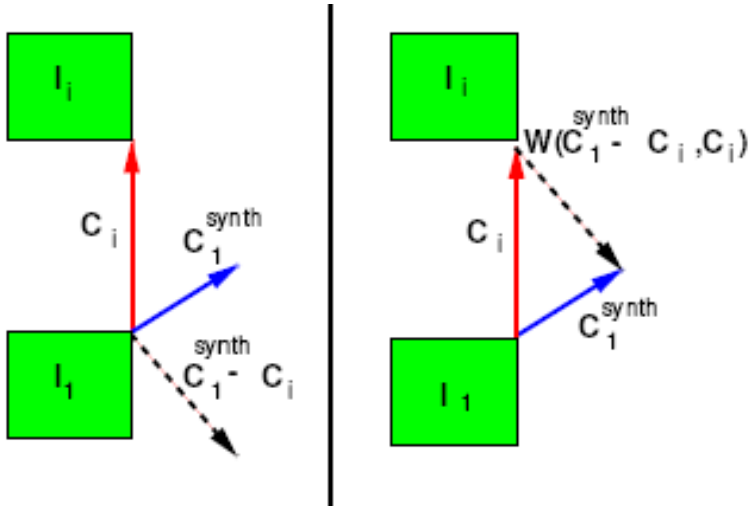


Synthesis

- Steps:
  - Synthesize a new correspondence:
  - Forward Warp $W(I, C)$

$$C_1^{synth} = \sum_{i=1}^{N} \alpha_i C_i.$$

α



$$C_i^{synth} = W(C_1^{synth} - C_i, C_i).$$

$$I_i^{warped} = W(I_i, C_i^{synth}).$$

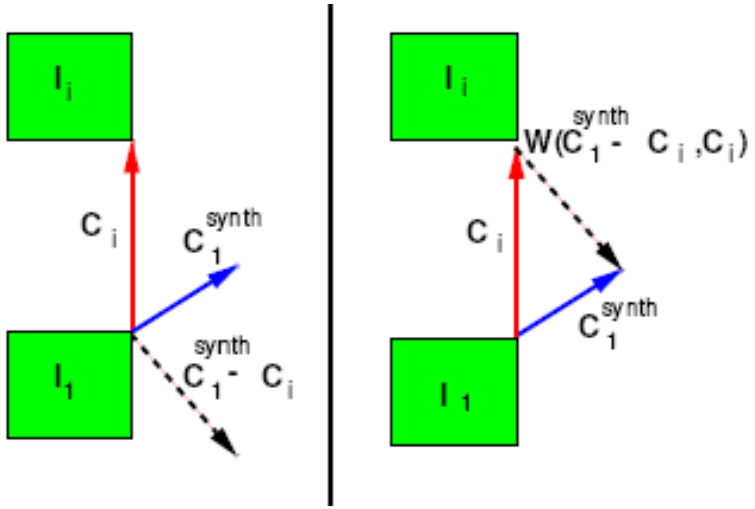
$$I^{morph} = \sum_{i=1}^{N} \beta_i I_i^{warped}.$$

Synthesis

- Steps:
  - Synthesize a new correspondence:
  - Forward Warp $W(I,C)$

$$C_1^{synth} = \sum_{i=1}^{N} \alpha_i C_i.$$

a



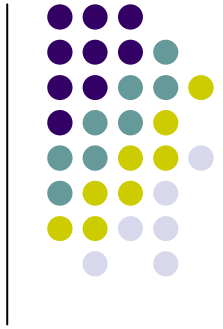
$$C_i^{synth} = W(C_1^{synth} - C_i, C_i).$$

$$I_i^{warped} = W(I_i, C_i^{synth}).$$

$$I^{morph} = \sum_{i=1}^{N} \beta_i I_i^{warped}.$$

$$I^{morph}(\alpha,\beta) = \sum_{i=1}^{N} \beta_i W(I_i, W(\sum_{j=1}^{N} \alpha_j C_j - C_i, C_i)).$$

Analysis



- Goal:

  Project the entire recorded corpus onto the constructed MMM, and produce a time series of parameters (α , β) that represent trajectories of the original mouth motion

- Each utterance analyzed with respect to the 92 dimensional MMM

Analysis

- Estimate parameter α :

$$\|C^{novel} - \sum_{i=1}^{N} \alpha_i C_i\| \quad \Rightarrow \quad \alpha = (C^T C)^{-1} C^T C^{novel}$$

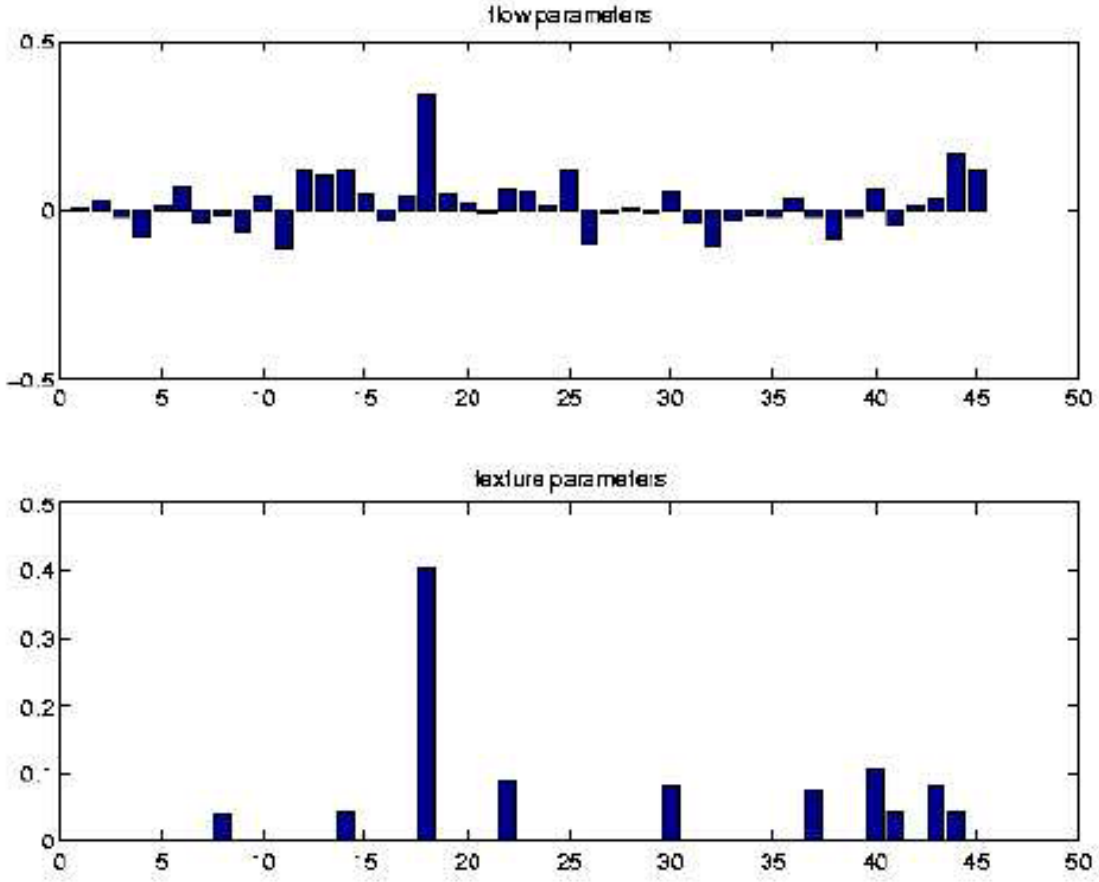
- *N* image warps are synthesized

$$I_i^{warp} = \mathrm{W}(I_i, \mathrm{W}(\sum_{i=1}^{N} \alpha_i C_i - C_i, C_i))$$

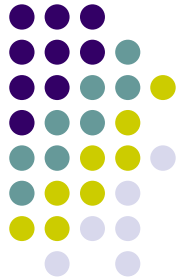
- Estimate β :

$$\|I^{novel} - \Sigma_{i=1}^{N} \beta_i I_i^{warp}\| \quad \text{subject to}$$
$$\beta_i > 0 \,\forall i \text{ and } \Sigma_{i=1}^{N} \beta_i = 1$$

Analysis Result



Trajectory Synthesis



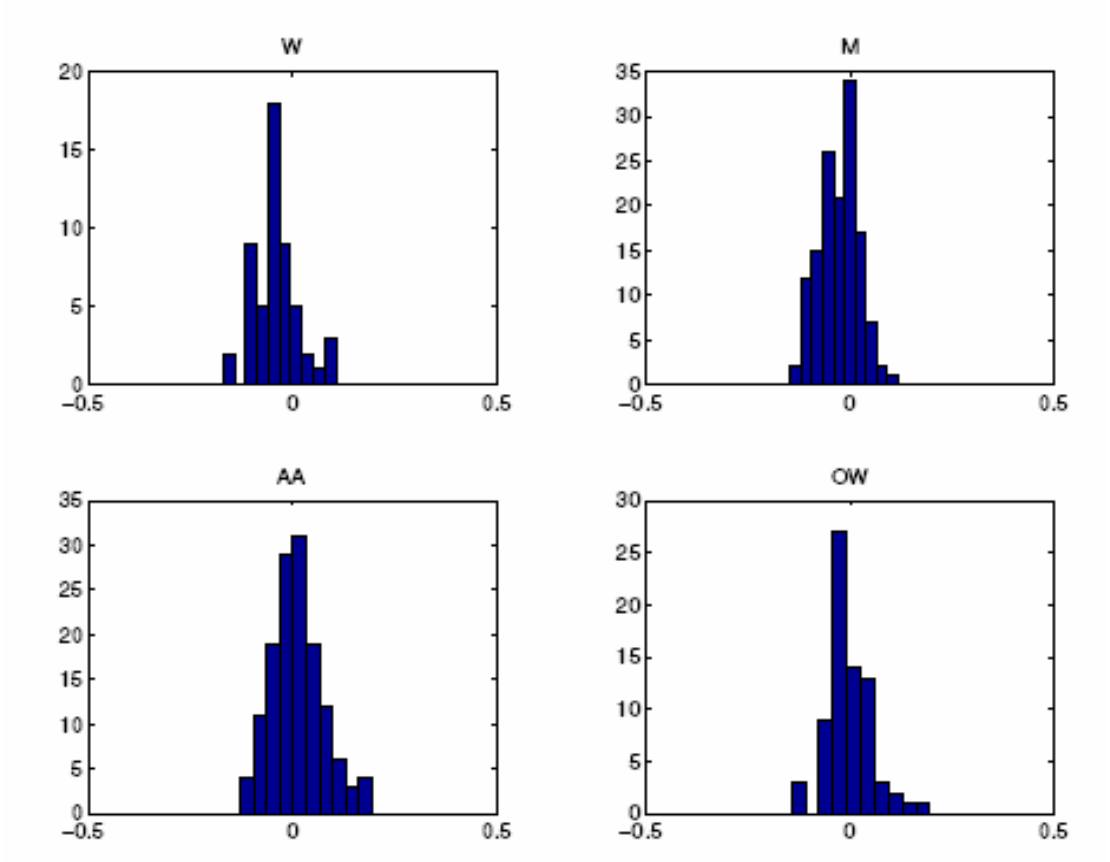
- Goal:

  Map from an input phone stream $\{p_t\}$ to a trajectory of parameters $y_t=(\alpha_t, \beta_t)$ in MMM space.

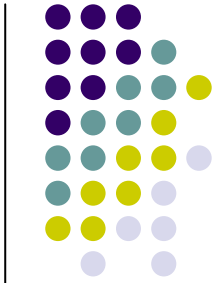
- Phone stream?

  $\{p_t\}_{t=1}^{15}$ ={\w\, \w\, \w\, \w\, \uh\, \uh\, \uh\, \uh\, \uh\, \uh\, \n\, \n\, \n\, \n\, \n\} =>word ‘one’

Histogram



Trajectory Synthesis



- Mathmatically a regularization problem:

$$E = \underbrace{(y-\mu)^T D^T \Sigma^{-1} D(y-\mu)}_{target\ term} + \lambda \underbrace{y^T W^T W y}_{smoothness}.$$

$$y = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix} \quad \mu = \begin{bmatrix} \mu_{P_1} \\ \vdots \\ \mu_{P_T} \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{P_1} & & \\ & \ddots & \\ & & \Sigma_{P_T} \end{bmatrix} \quad D = \begin{bmatrix} \sqrt{I - \frac{D_{P_1}}{T}} & & & \\ & \sqrt{I - \frac{D_{P_2}}{T}} & & \\ & & \ddots & \\ & & & \sqrt{I - \frac{D_{P_T}}{T}} \end{bmatrix} \quad W = \begin{bmatrix} -I & I & & \\ & -I & I & \\ & & \ddots & \\ & & & -I \quad I \end{bmatrix}$$

- Minimization:

$$(D^T \Sigma^{-1} D + \lambda W^T W) y = D^T \Sigma^{-1} D \mu$$

Training

- Adjust the means and variance to better reflect the training data

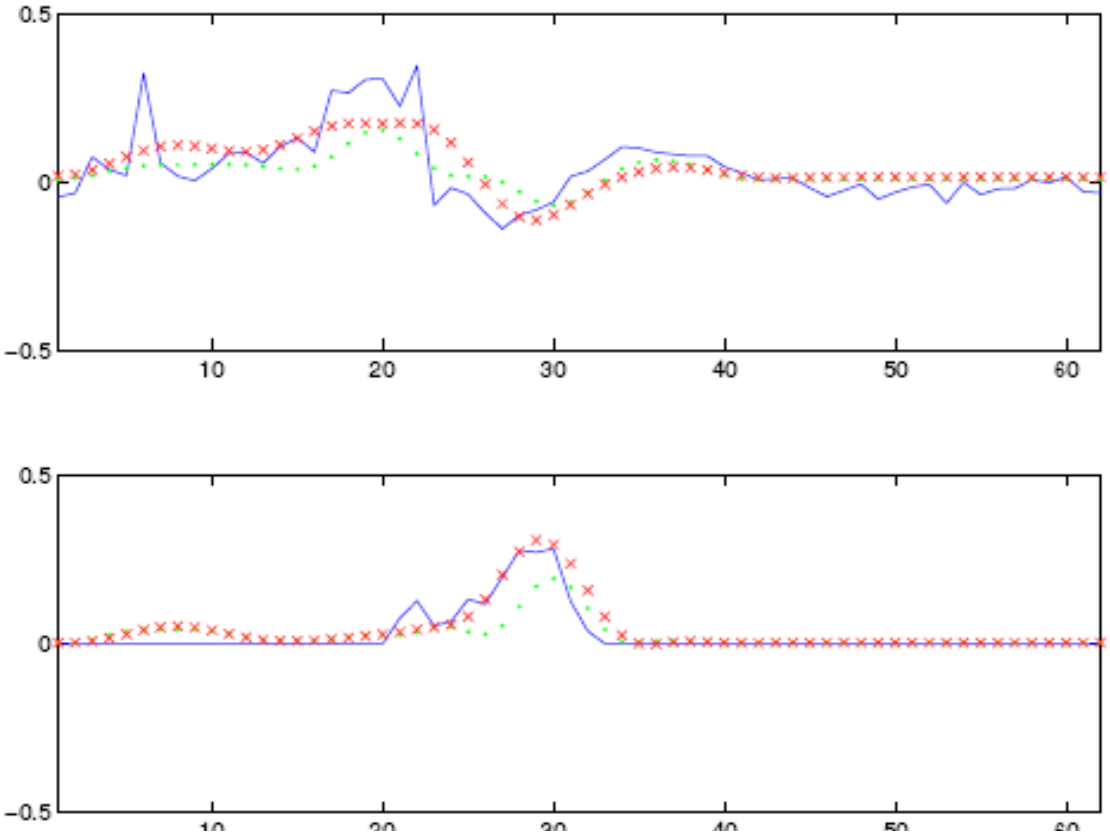
$$E = (z - y)^T (z - y).$$

➔

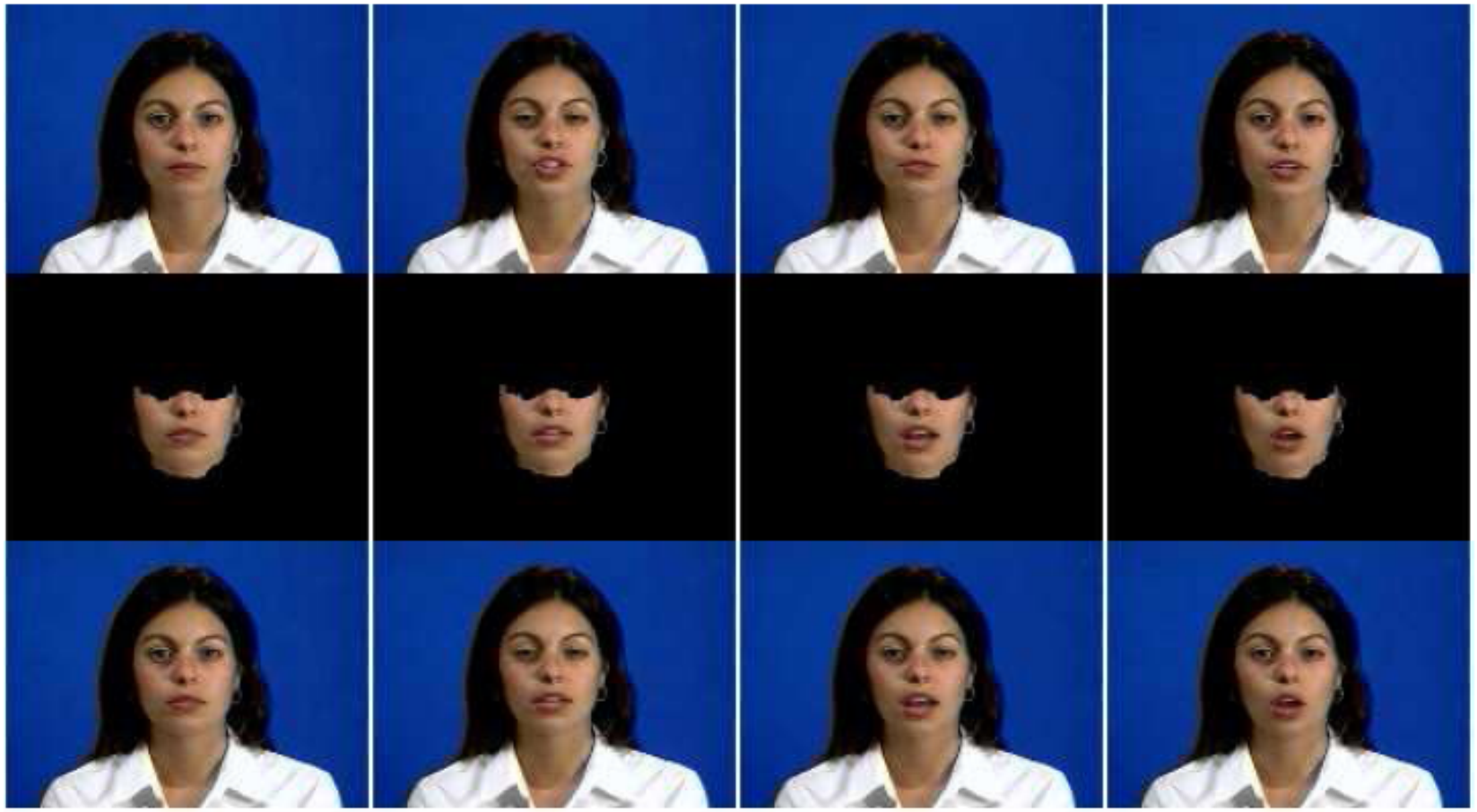
$$\mu^{new} = \mu^{old} - \eta \frac{\partial E}{\partial \mu}$$

$$\Sigma^{new} = \Sigma^{old} - \eta \frac{\partial E}{\partial \Sigma}.$$

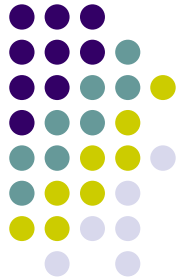
Training Result



Post-Processing



Result

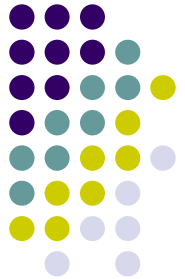


- Demos
- Interviews: Discovery,NBC
- Another Example

Evaluations

Experiment	# subjects	% correct	t	p<
Single pres.	22	54.3%	1.243	0.3
Fast single pres.	21	52.1%	0.619	0.5
Double pres.	22	46.6%	-0.75	0.5

Discussions



- Viewing Conditions?
  - 2D->3D
- Emotion
- Better video-realism
  - Geodesic trajectory synthesis…