Trainable Videorealistic Speech Animation

Tony Ezzat
Gadi Geiger
Tomaso Poggio

CBCL/AI Lab
MIT
Outline

• Problem Setting
• Previous Work
• Our Approach
• Results
• Evaluation
Overview

2 Themes:

Videorealism
Machine Learning

“Air”

Video Database

“Badge”

Visual Speech Processing

“Badge”

Mary101
Audio Analysis

“Air”

Audio Database

Video Database

Visual Speech Processing

“Badge”

Audio is recorded also to help label video

“Badge”
Audio Synthesis

“Air”

Audio Database

Video Database

Visual Speech Processing

Audio Speech Processing

“Badge”

“Badge”

“Badge”

No Audio Synthesis!
What is the Input REALLY?

“Badge”

Visual Speech Processing

[Image of a person]
Input: Phone Stream

“Badge”

TTS

/SIL B B B AE AE JH JH SIL SIL/

Visual Speech Processing

Real Audio

“Badge”

Speech Recognition
Forced Viterbi Alignment
Manual Labelling
Pre- and Post-Processing

- **Pre-Processing**: Remove head movement using planar perspective warping
- **Visual Speech Processing**
- **Post-Processing**: Mask out mouth Track & Recomposite into background sequence
Tracking & Compositing
Outline

- Problem Setting
- Previous Work
- Our Approach
- Results
- Evaluation
- More Results
Video Rewrite
(Bregler, Covell, Slaney 1997)

Hello: /H-E-L/ + /E-L-OW/

Triphone basis units
Reorder them to new utterance
Pixel blending at join points

Coarticulation: /utu/ vs /iti/
Video Rewrite Issues

(Bregler, Covell, Slaney 1997)

- **Sampling coarticulation**
  
  20000 triphones ~ 3 hrs!

- **Model of speech is entire video corpus**
  
  No capacity to learn/model/distill
  Not a parsimonious representation

- **Poor capacity for novel image synthesis**
  
  Poor smoothing at join points
  Cannot stretch/shrink to match audio
  Discrete number of paths
  Cannot fill in missing data
Outline

- Problem Setting
- Previous Work
- Our Approach
- Results
- Evaluation
- More Results
Extracting Prototypes

46 prototypes extracted using PCA and K-means clustering
Multidimensional Morphable Model

\((\alpha, \beta)\)
1D Morphing

(Beier & Neely 1992)

\[ \beta_1 \ WARP(I_1, \alpha_1 F_1) \ + \ \beta_2 \ WARP(I_2, \alpha_2 F_2) \]
Optical Flow

(Beymer, Shashua, Poggio 93) (Chen & Williams 93)

\[ C = \{dx(x,y), dy(x,y)\} \]
1D Morphing w/Optical Flow

Forward warping A to B

Forward warping B to A

Blending

Holefilling
MMM Definition

Parameterize using \((\alpha, \beta)\)

46 Image prototypes from Corpus

46 Optical flow between prototypes

Parameterize using \((\alpha, \beta)\)

alpha is 46-dimensional
beta is 46 dimensional
 MMM Synthesis

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

\[ C_i^{synth} = W(C_i - C_1^{synth}, C_i) \]

\[ I_i^{warp} = W(I_i, C_i^{synth}) \]

\[ I_{morph}^{(\alpha, \beta)} = \sum_{i=1}^{N} \beta_i I_i^{warp} \]

Fine, but what about speech?
Mary101 Speech Model

Each phoneme represents a cluster in MMM space.

Speech trajectory passes close to clusters but which is also smooth.
MMM Analysis

\[
I_1 \quad C_2 \quad I_3 \\
I_2 \quad C_3 \quad I_4
\]

\((\alpha, \beta)\)
MMM Analysis (Cntd)

\[ C_{novel} = \sum_{i=1}^{N} \alpha_i C_i \]

Re-orient + Warp

\[ I_{novel} - \sum_{i=1}^{N} \beta_i I_i^{warped} \]

subject to
\[ \beta_i > 0 \quad \forall i \]
\[ \sum \beta_i = 1 \]
MMM Analysis Parameters

badge

Flow

Texture

lavish
Comparison of Real and Synthesized Images

- Tongue is not perfect
- Slight blurring
Analysis of Entire Recorded Corpus

Video Corpus

\[ z_1 = (\alpha_1, \beta_1) \]

\[ z_2 = (\alpha_2, \beta_2) \]

\[ z_{30000} = (\alpha_{30000}, \beta_{30000}) \]

/b/

/jh/

/ae/
Phonetic Clusters

Represent each phone with $\mu_p$, $\Sigma_p$

One set for flows, another set for textures
Trajectory Synthesis

\[
\begin{align*}
\min_y & \quad (y - \mu)^T \Sigma^{-1} (y - \mu) + \lambda \|\Delta y\|^2 \\
\end{align*}
\]

Phonetic Targets

Smoothness
Smoothness

Higher orders of smoothness:

\[ \Delta = \begin{bmatrix} -I & I \\ -I & I \\ \cdot & \cdot \\ -I & I \end{bmatrix} \]

\[ \Delta \Delta, \Delta \Delta \Delta ,... \]

Order 2, 3, ....
Cross-validation:

flow: \[ \Delta \] order 4, \[ \lambda \] = 250: septic splines

texture: \[ \Delta \] order 5, \[ \lambda \] = 100: nintic splines
Setting Phonetic Clusters

Use sample estimates?

Problem: Underarticulation!
Adjusting Phonetic Clusters

Compare synthesized trajectory \( y = \{\alpha_t, \beta_t\} \) with original trajectory \( z = \{\alpha_t, \beta_t\} \)

\[
E = (z - y)^T (z - y)
\]

Use Gradient descent to tweak

\[
\frac{\partial E}{\partial \mu_i} = \frac{\partial E}{\partial y} \frac{\partial y}{\partial \mu_i}
\]

\[
\mu^{new} = \mu^{old} - \eta \frac{\partial E}{\partial \mu}
\]
Phones Before/After Training

before

after
Trajectories
Before/After Training
Coarticulation controlled by width of cluster regions
Coarticulation

/utu/  /iti/  /ata/

/ubu/  /ibi/  /aba/
Big Picture

Pre-process

Construct MMM

Analyze Corpus
\{\alpha_i, \beta_i\}

Train phonetic models

Synthesize!

Post-process

Trajectory Synthesis

\{\alpha_i, \beta_i\}

MMM

\{\mu_p, \Sigma_p\}

\{I_i, C_i\}
Results

Mary101:

8 minutes of training data

1-syllable words: 132 training/20 test
2-syllable words: 136 training/20 test

46-prototype MMM

Sentences not even included in training.
Comments So Far

“She looks like she’s been Botox’ed”
-- Nobel Laureate

“Has she had a frontal lobotomy?”
-- ATT executive

Send me your comments to tonebone@ai.mit.edu
### Visual Turing Tests

<table>
<thead>
<tr>
<th>Experiment</th>
<th>% correct</th>
<th>P&lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Single presentation</strong></td>
<td>52.1%</td>
<td>0.3</td>
</tr>
<tr>
<td><strong>Double presentation</strong></td>
<td>46.6%</td>
<td>0.5</td>
</tr>
</tbody>
</table>

We win!
# Visual Intelligibility

<table>
<thead>
<tr>
<th>Experiment</th>
<th>%correct on N</th>
<th>%correct on S</th>
<th>P&lt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Words+Sents</td>
<td>30.01%</td>
<td>21.19%</td>
<td>0.001</td>
</tr>
<tr>
<td>Words</td>
<td>38.55%</td>
<td>28.07%</td>
<td>0.001</td>
</tr>
<tr>
<td>Sents</td>
<td>24.38%</td>
<td>16.52%</td>
<td>0.01</td>
</tr>
</tbody>
</table>

**Correct Phoneme ID**

Still some work to do........
Stay Tuned!

Acknowledgments:

Association Christian Benoit
NSF
NTT
ITRI

Mary101
Craig Milanesi
Joanne Flood
Marypat Fitzgerald
Vinay Kumar
Chao Wang
Danielle Suh
Volker Blanz
Demetri Terzopoulos
Rehema Ellis/NBC

Dynasty Models
Dave Konstine
Jay Benoit
Casey Johnson
Sayan Mukherjee
Adlar Kim
Osamu Yoshimi
Thomas Vetter
Jenny Shapiro/BMG
Kevin Chang