Trainable Videorealistic Speech Animation

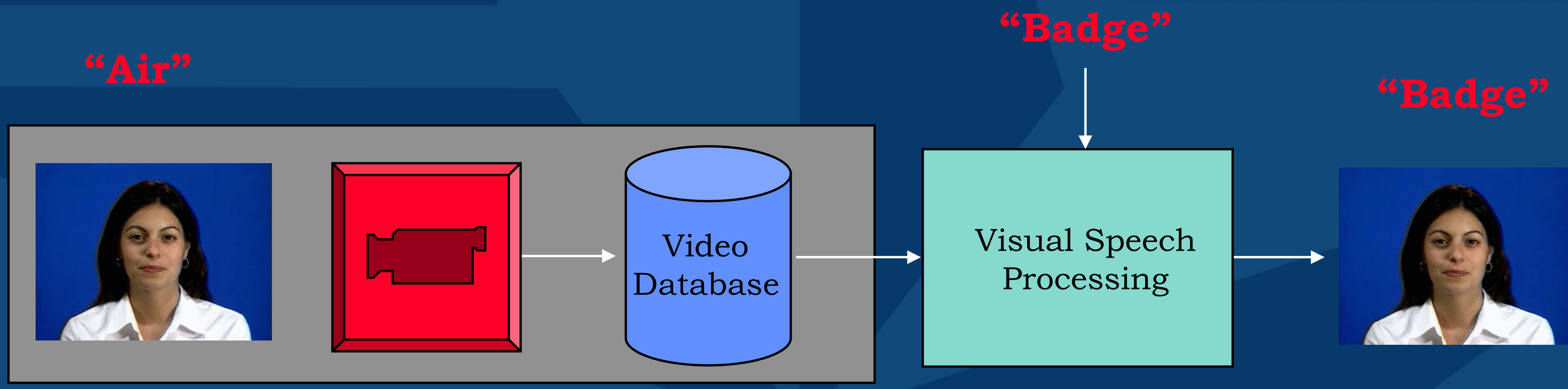
Tony Ezzat
Gadi Geiger
Tomaso Poggio

CBCL/AI Lab
MIT

Outline

- Problem Setting
- Previous Work
- Our Approach
- Results
- Evaluation

Overview



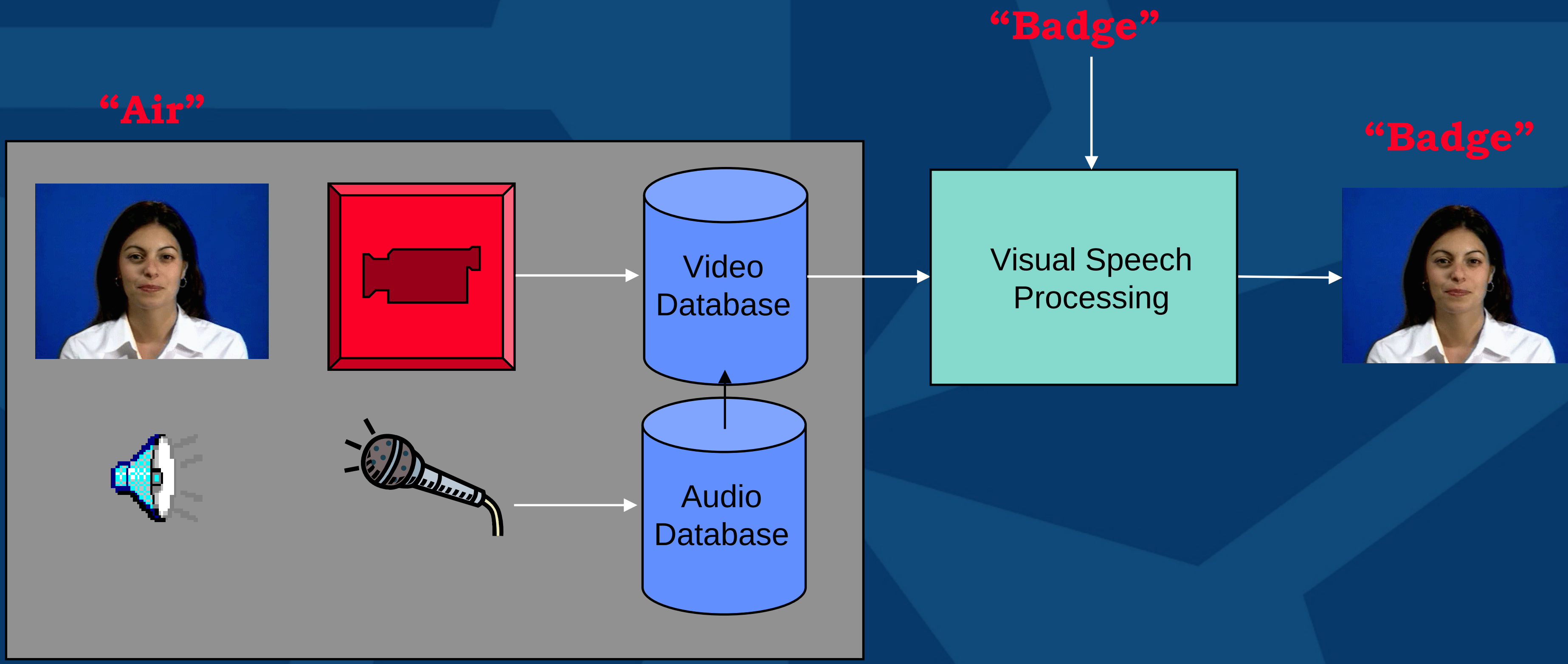
Mary101

2 Themes:

Videorealism

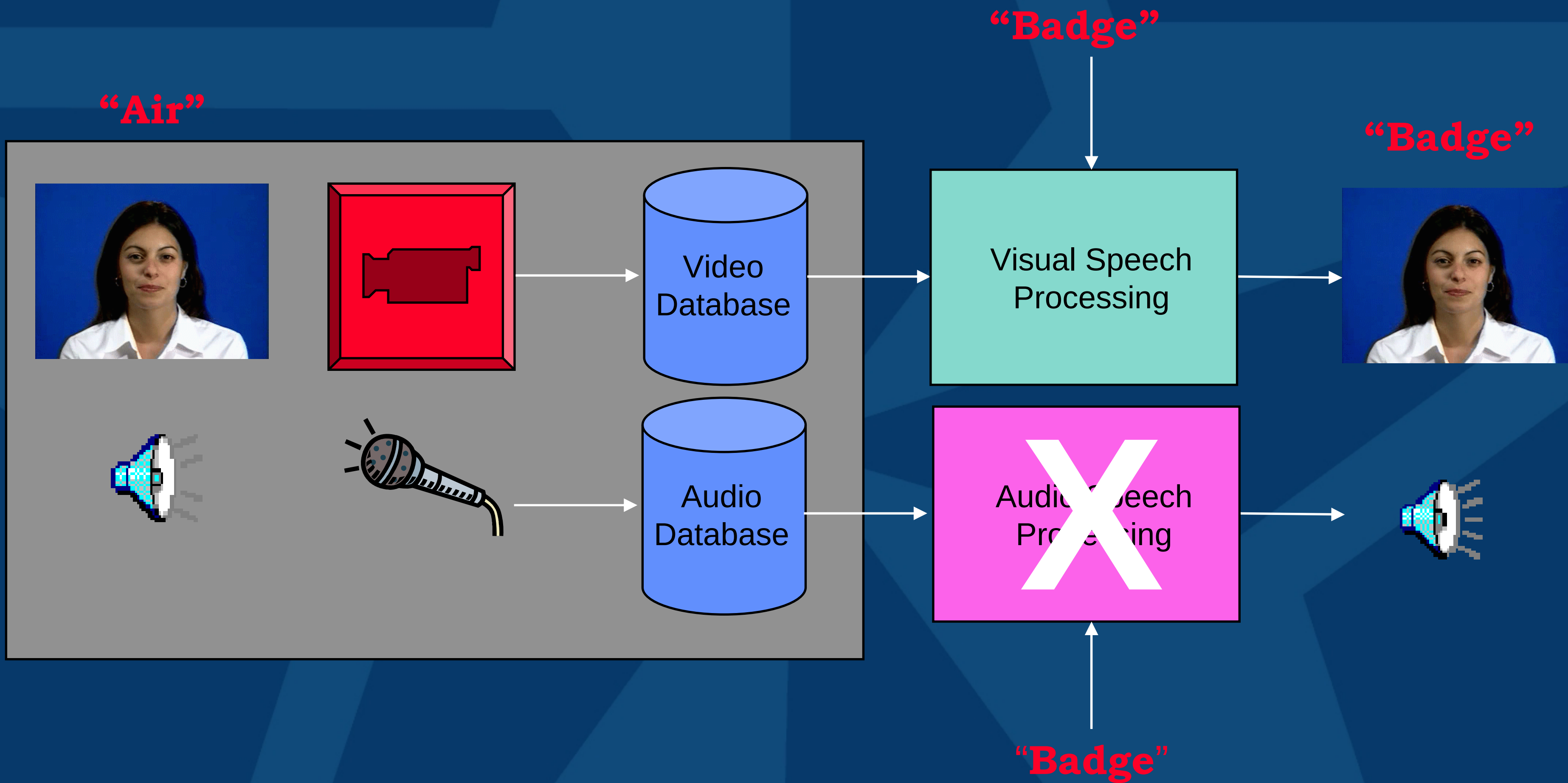
Machine Learning

Audio Analysis



Audio is recorded also to help label video

Audio Synthesis



No Audio Synthesis!

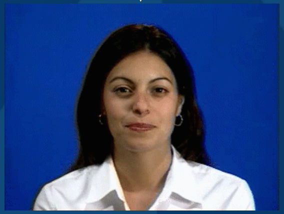
What is the Input **REALLY**?

"Badge"

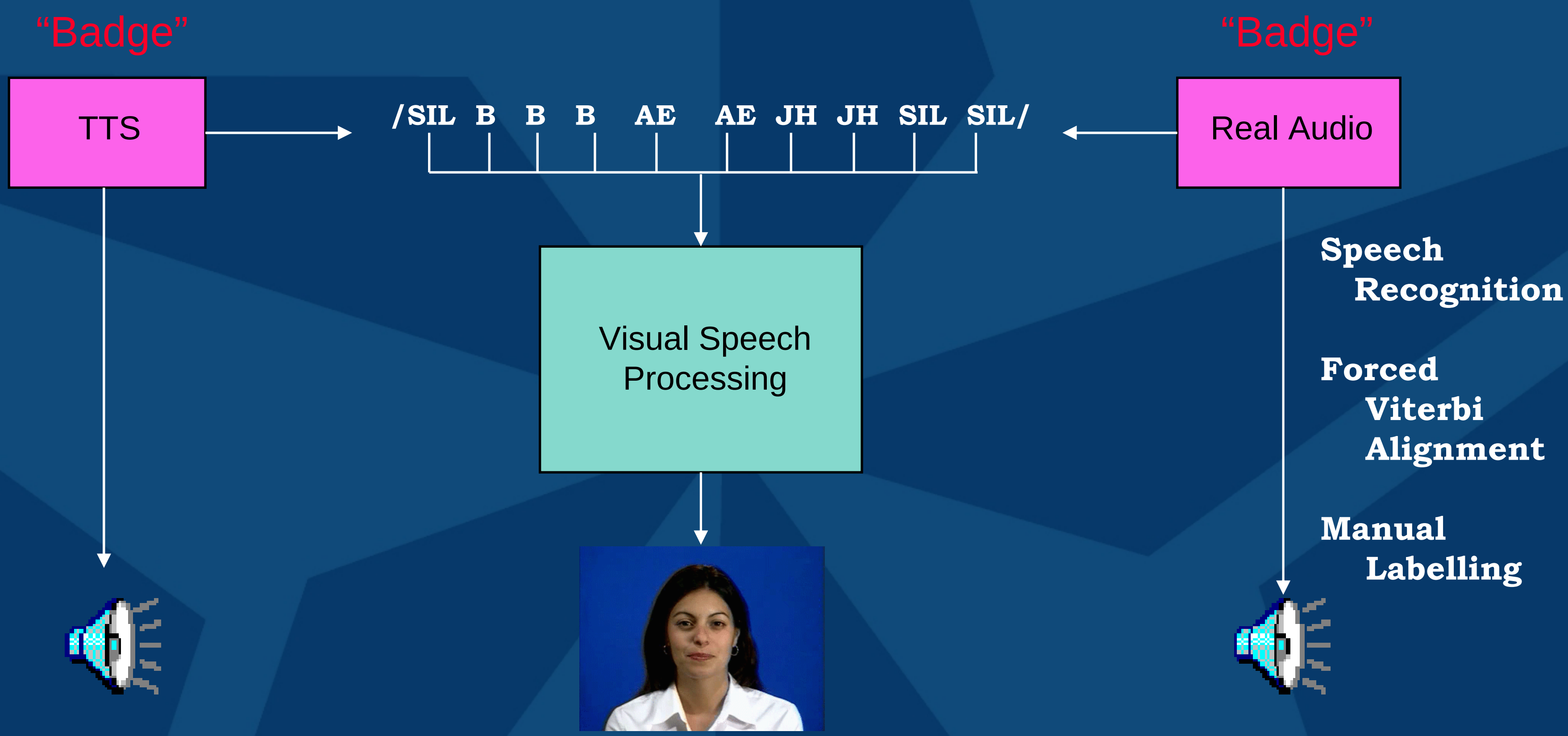
↓

Visual Speech Processing

↓



Input: Phone Stream



Pre- and Post-Processing

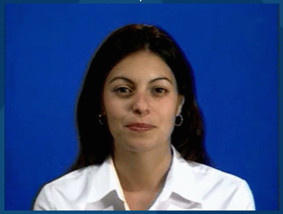
/SIL B B B AE AE JH JH SIL SIL/

Pre-Processing

Visual Speech Processing

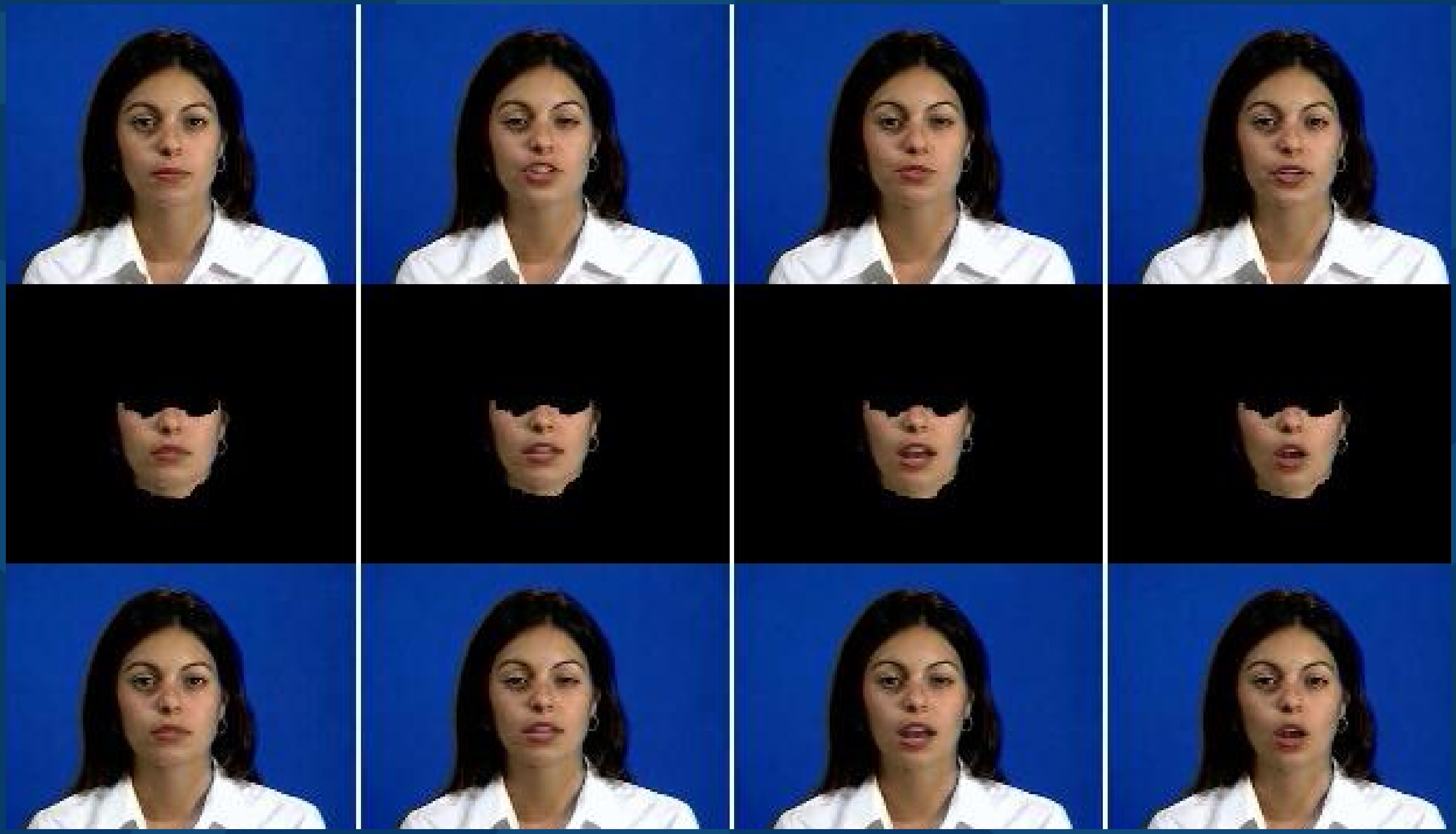
Post-Processing

Remove head movement using planar perspective warping



Mask out mouth
Track & Recomposite
into background sequence

Tracking & Compositing



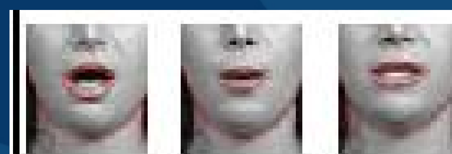
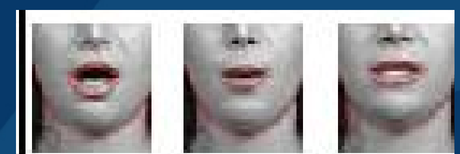
Outline

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

Video Rewrite

(Bregler, Covell, Slaney 1997)

/H-E-L/ /E-L-OW/

Hello:  + 

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

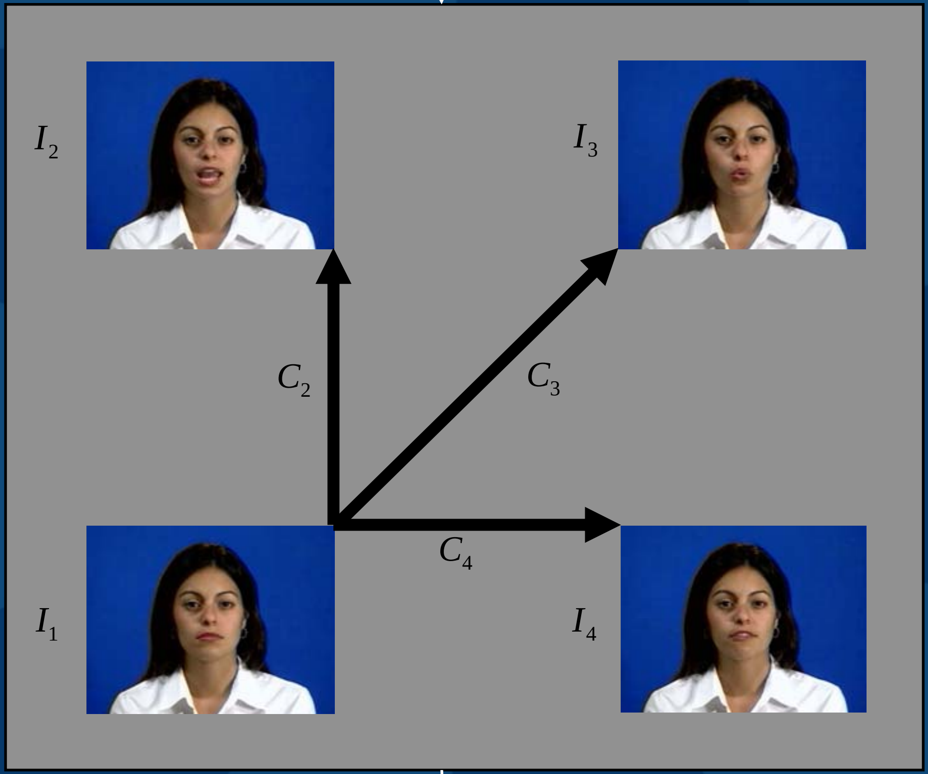
- Problem Setting
- Previous Work
- Our Approach
- Results
- Evaluation
- More Results

Extracting Prototypes

46 prototypes extracted using PCA and K-means clustering

Multidimensional Morphable Model

(α, β)



MMM Background

Tommy Poggio/MIT
David Beymer
Mike Jones
Vinay Kumar

Volker Blanz/
MPI Saabrucken

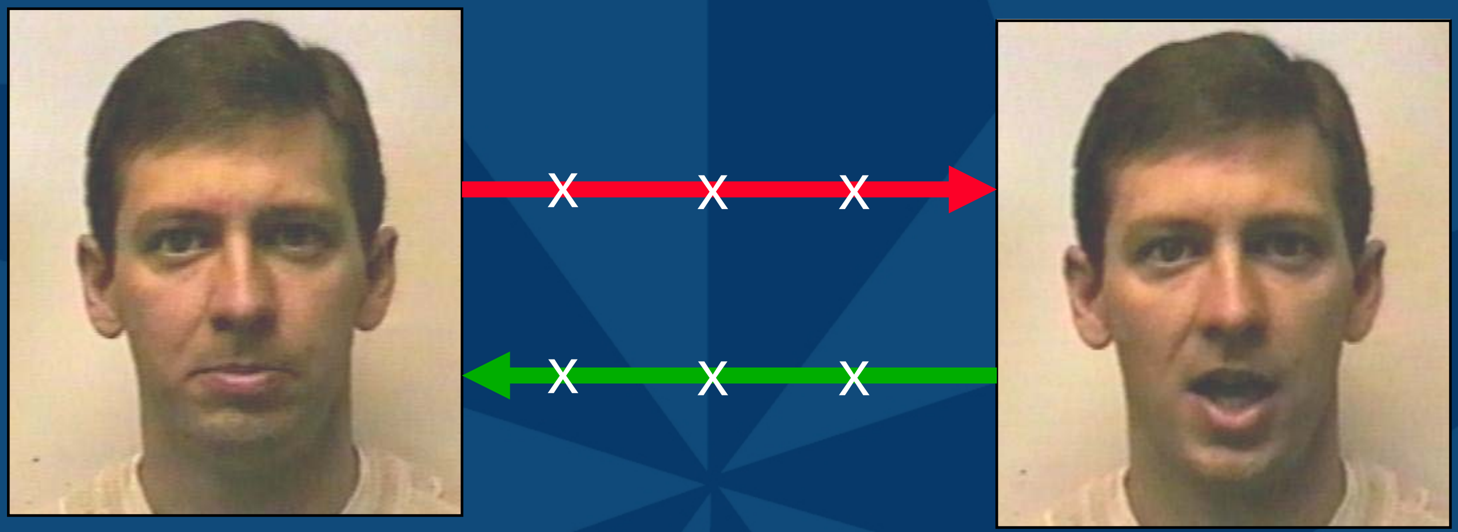
Thomas Vetter/
University of Basel

Tim Cootes/Manchester

Michael Black/Brown

1D Morphing

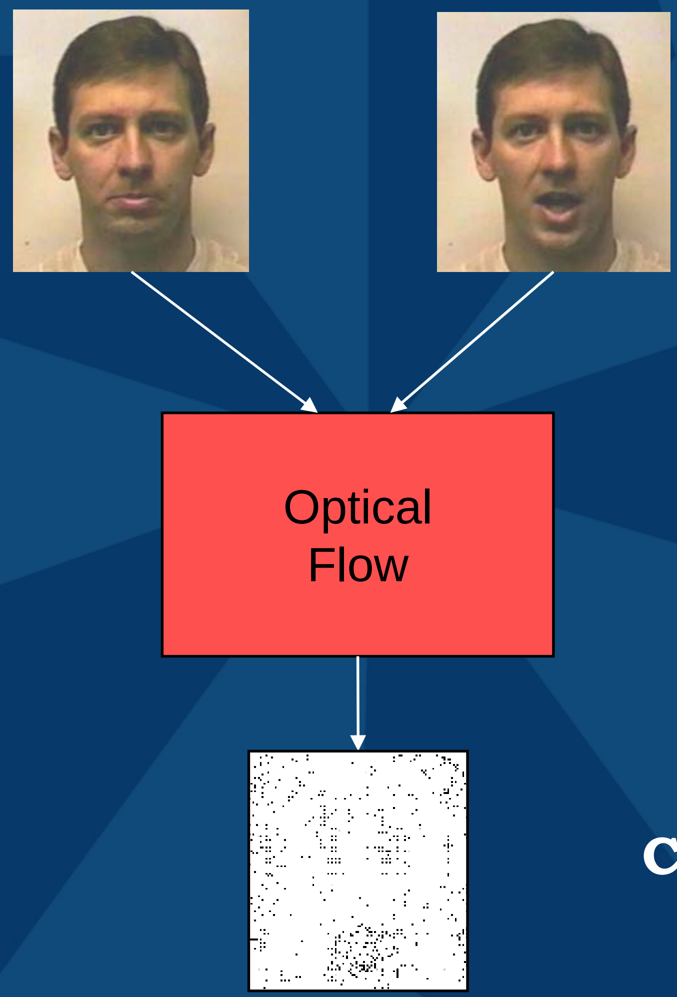
(Beier & Neely 1992)



$$\beta_1 \; WARP(I_1, \alpha_1 F_1) \quad + \quad \beta_2 \; WARP(I_2, \alpha_2 F_2)$$

Optical Flow

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



$$C = \{dx(x,y),\ dy(x,y)\}$$

1D Morphing w/Optical Flow

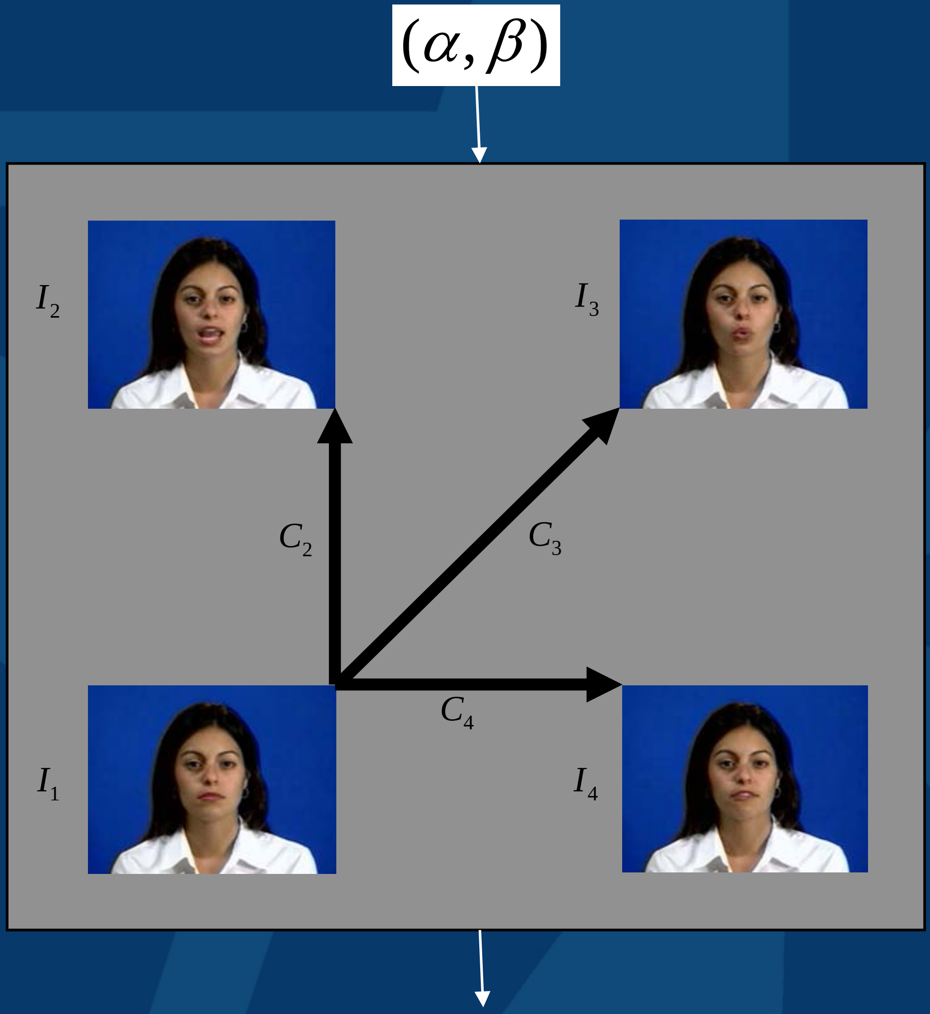
Forward warping A to B

Forward warping B to A

Blending

Holefilling

MMM Definition



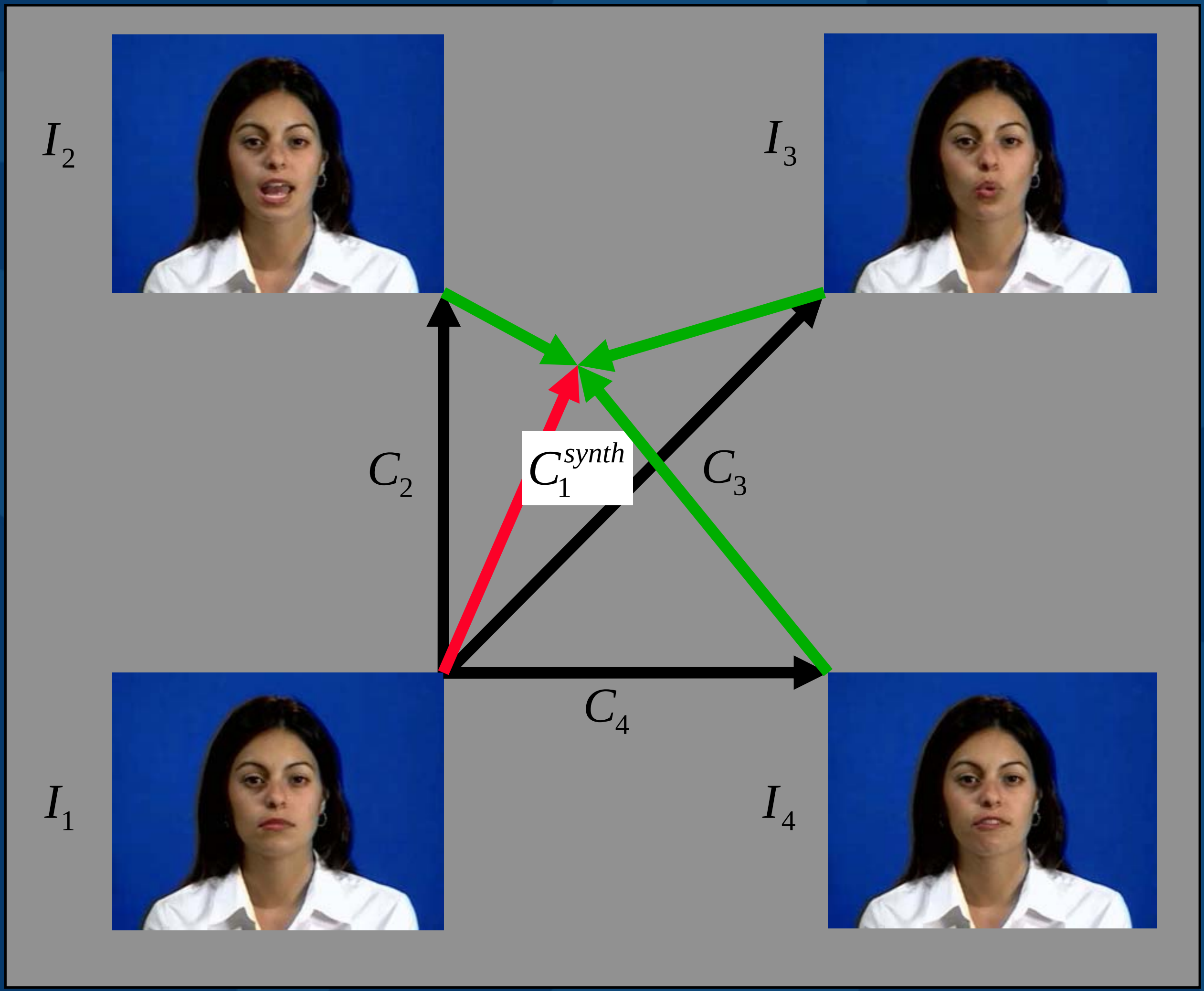
46 Image prototypes from Corpus

46 Optical flow between prototypes

Parameterize using (α, β)

alpha is 46-dimensional
beta is 46 dimensional

MMM Synthesis



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

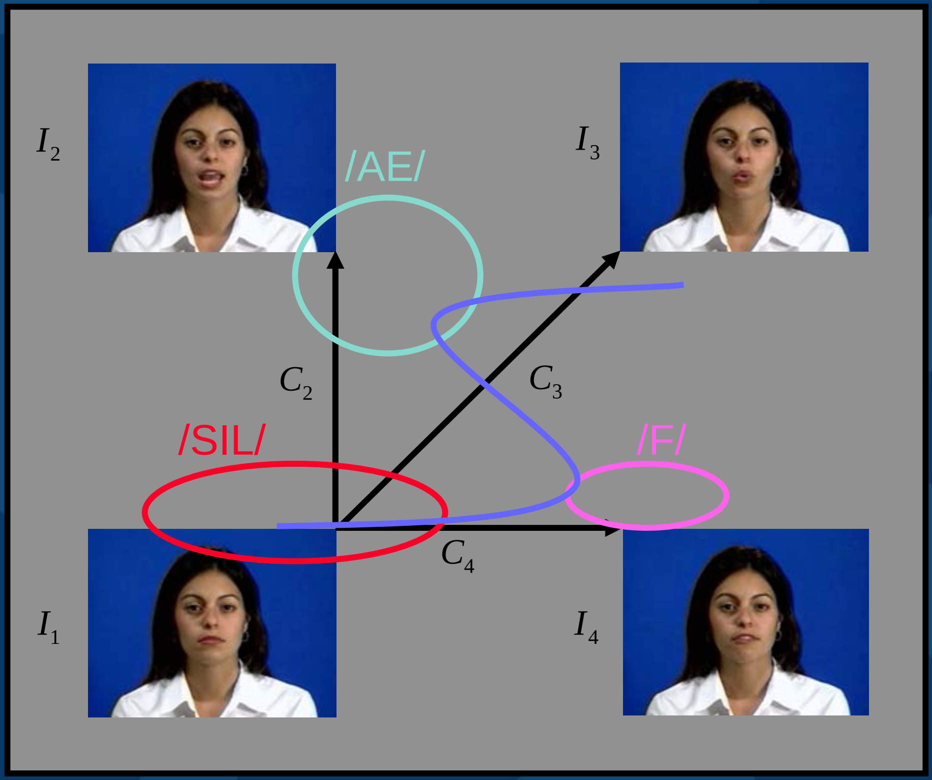
$$C_i^{synth} = W(C_i - C_1^{synth}, C_i)$$

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

$$I^{morph}(\alpha, \beta) = \sum_{i=1}^{N} \beta_i I_i^{warp}$$

Fine, but what about speech?

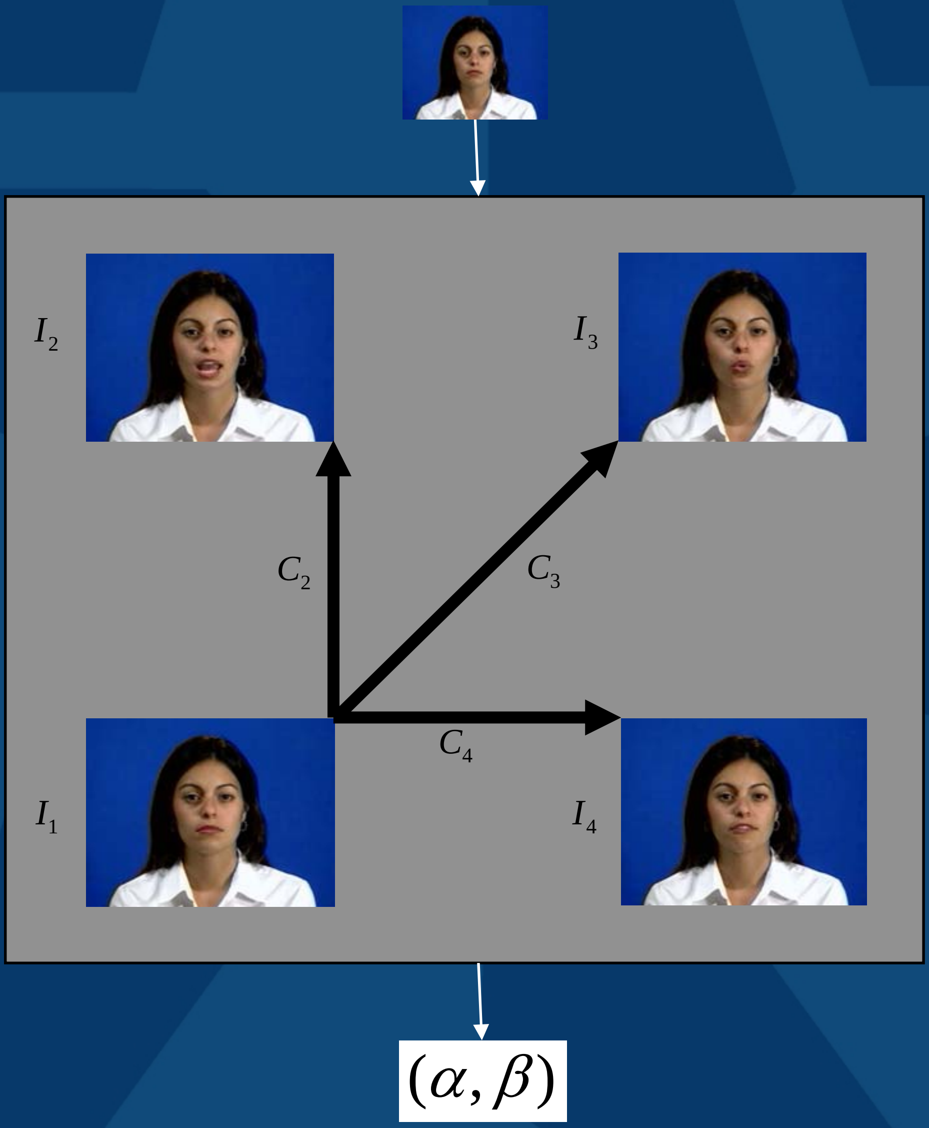
Mary101 Speech Model



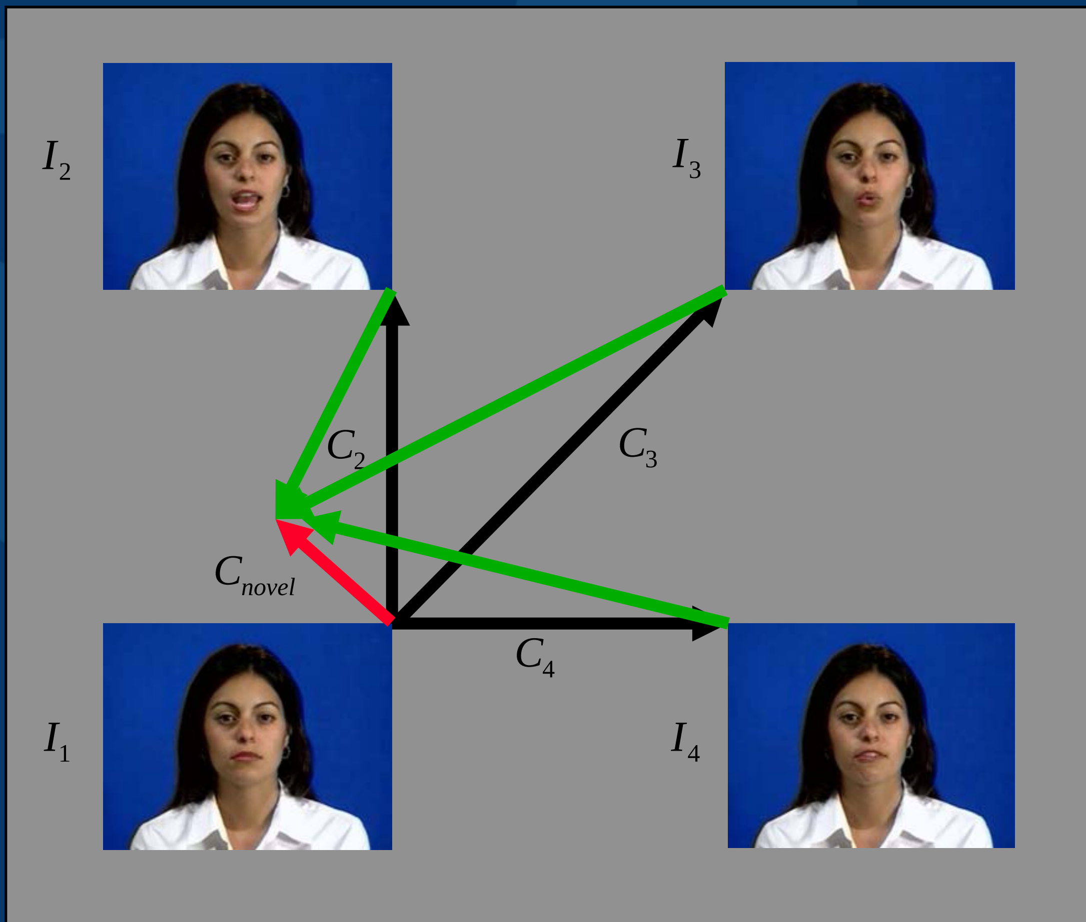
Each phoneme represents a cluster in MMM space

Speech trajectory passes close to clusters but which is also smooth

MMM Analysis



MMM Analysis (Cntd)



I_{novel}

C_{novel}

$$\left\| C_{novel} - \sum_{i=1}^{N} \alpha_i C_i \right\|$$

Re-orient + Warp

$$\left\| I_{novel} - \sum_{i=1}^{N} \beta_i I_i^{warped} \right\|$$

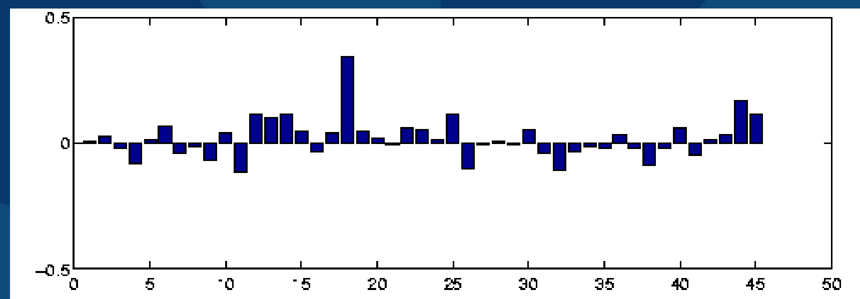
subject to

$$\beta_i > 0 \quad \forall i$$

$$\sum \beta_i = 1$$

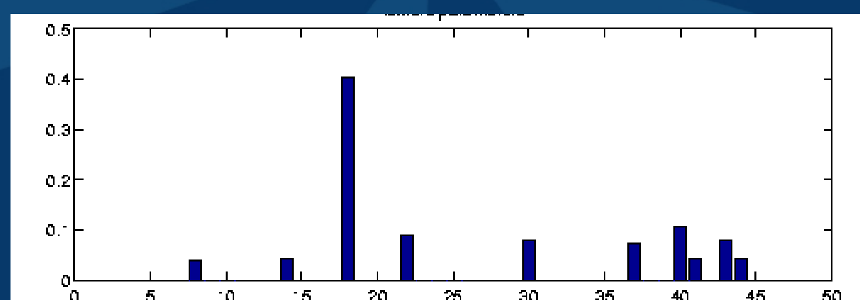
MMM Analysis Parameters

badge



Flow

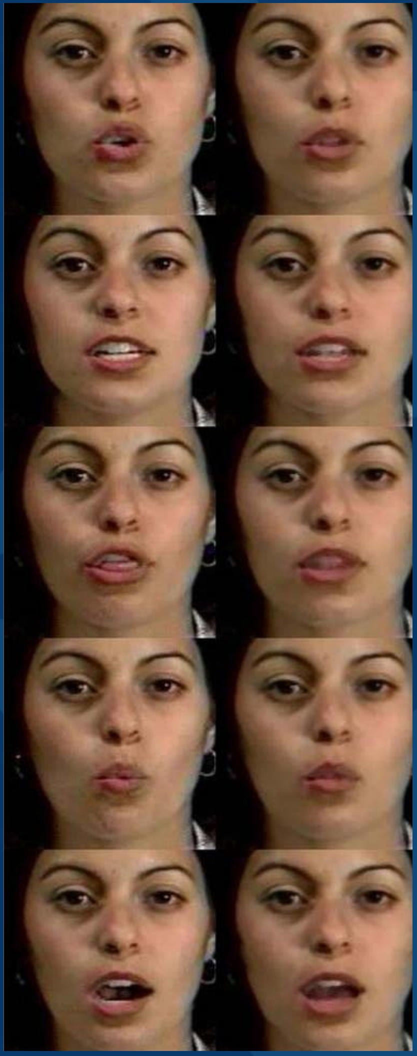
lavish



Texture

Comparison of Real and Synthesized Images

Real Synthetic

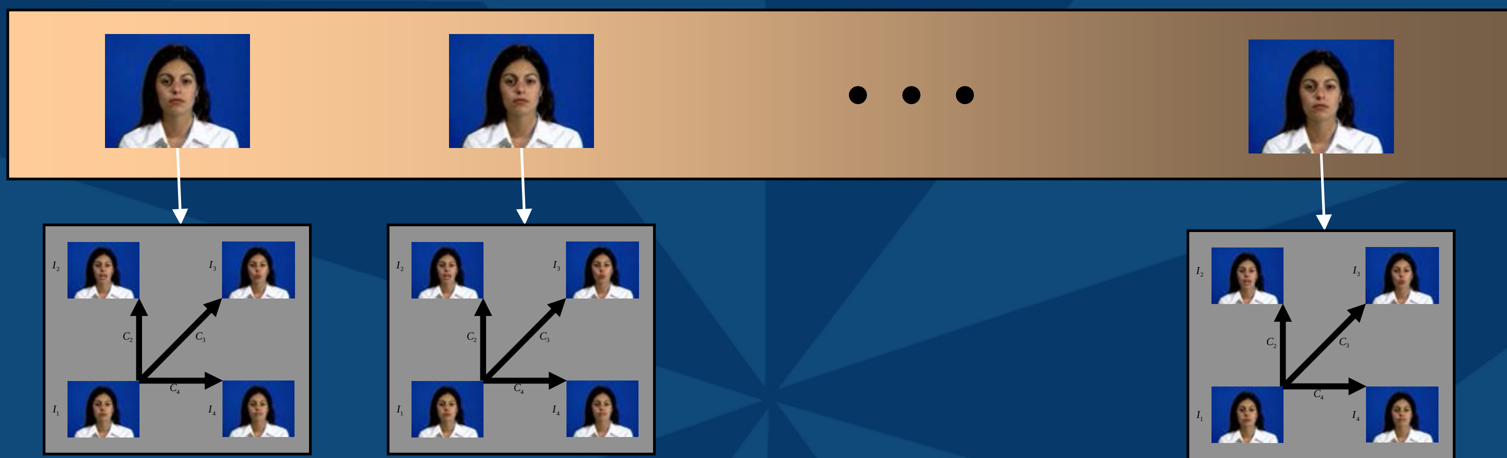


Real Synthetic

Tongue is not perfect
Slight blurring

Analysis of Entire Recorded Corpus

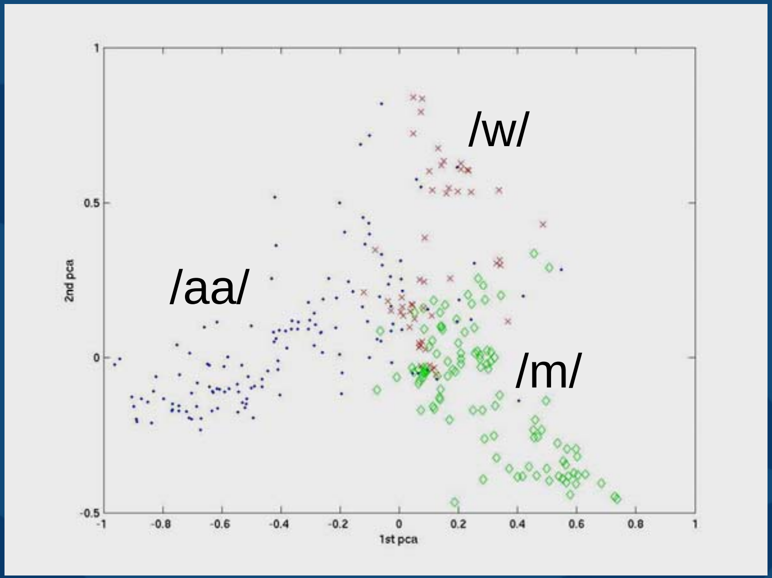
Video Corpus

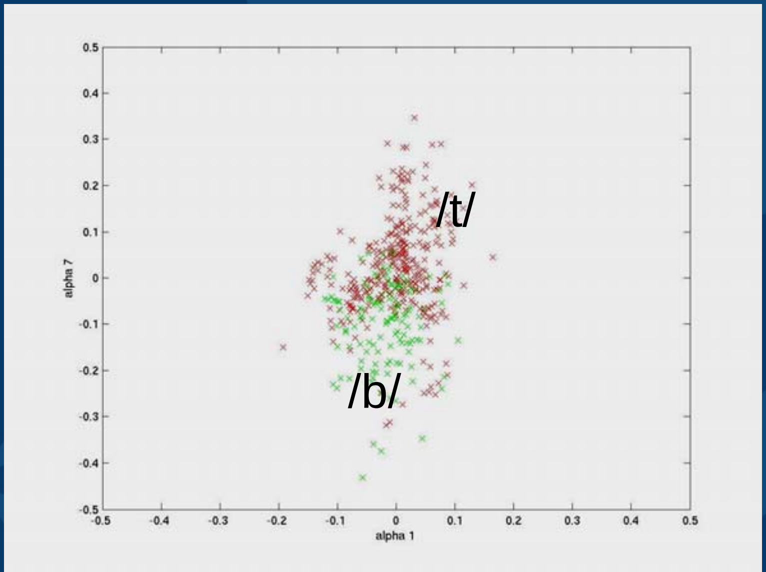


$z_1 = (\alpha_1, \beta_1)$ $z_2 = (\alpha_2, \beta_2)$ $z_{30000} = (\alpha_{30000}, \beta_{30000})$

/b/ **/jh/** **/ae/**

Phonetic Clusters





Represent each phone with μ_p Σ_p

One set for flows, another set for textures

Trajectory Synthesis

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

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix} \qquad \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_T \end{bmatrix} \qquad \Sigma = \begin{bmatrix} \Sigma_{P_1} & & & \\ & \Sigma_{P_2} & & \\ & & \ddots & \\ & & & \Sigma_{P_T} \end{bmatrix}$$

$$\min_{y} \quad (y-\mu)^T \Sigma^{-1} (y-\mu) + \lambda \left\| \Delta y \right\|^2$$

Phonetic Targets

Smoothness

Smoothness

$$\Delta = \begin{bmatrix} -I & I & & & \\ & -I & I & & \\ & & & \ddots & \\ & & & -I & I \end{bmatrix}$$

Higher orders of smoothness: $\Delta\Delta,\ \Delta\Delta\Delta, \ldots$

Order 2, 3, ….

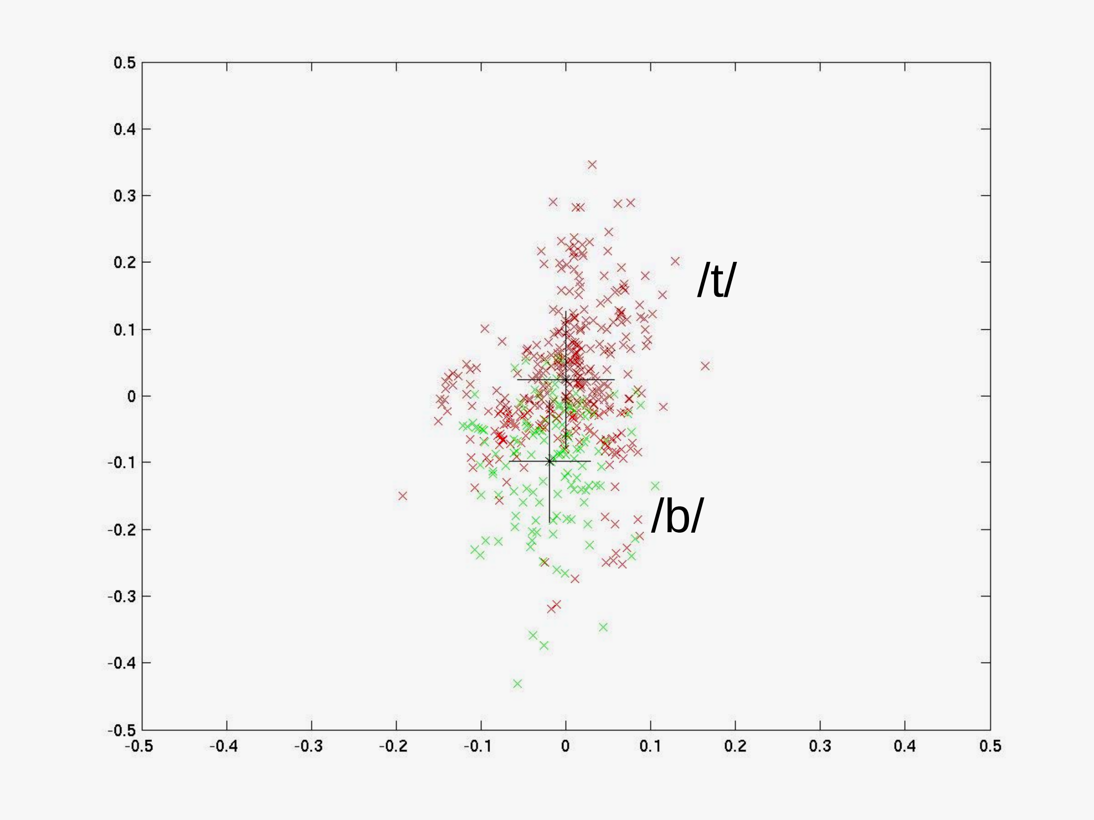
Setting λ, Δ

Cross-validation:

flow:	Δ **order 4,** λ	**= 250:**	**septic splines**
texture:	Δ **order 5,** λ	**= 100:**	**nintic splines**

Setting Phonetic Clusters

Use sample estimates?



Problem: Underarticulation!

Adjusting Phonetic Clusters

Compare synthesized trajectory **with** **original trajectory**

$$y = \{\alpha_t, \beta_t\}$$

$$z = \{\alpha_t, \beta_t\}$$

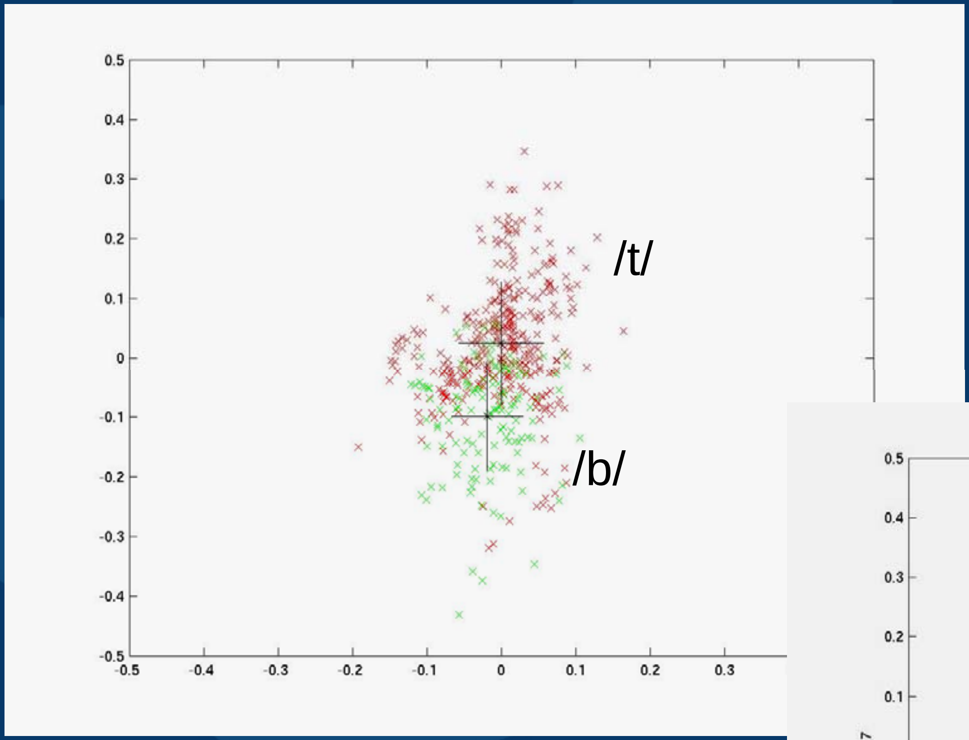
$$E = (z - y)^T (z - y)$$

Use Gradient descent to tweak

$$\frac{\partial E}{\partial \mu_i} = \frac{\partial E}{\partial y} \frac{\partial y}{\partial \mu_i}$$

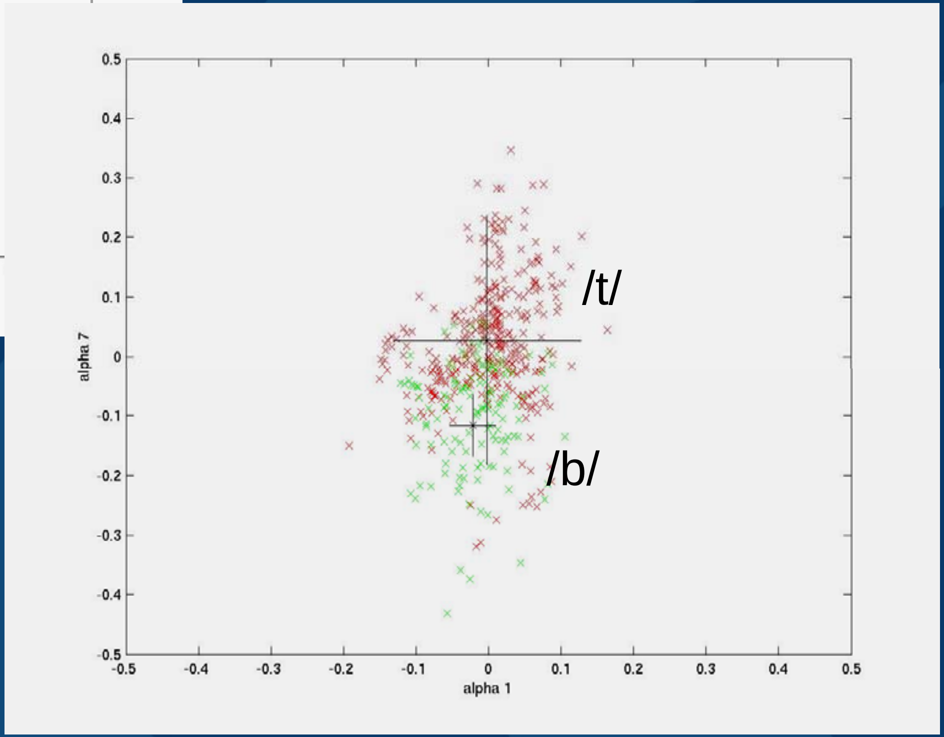
$$\mu^{new} = \mu^{old} - \eta \frac{\partial E}{\partial \mu}$$

Phones Before/After Training

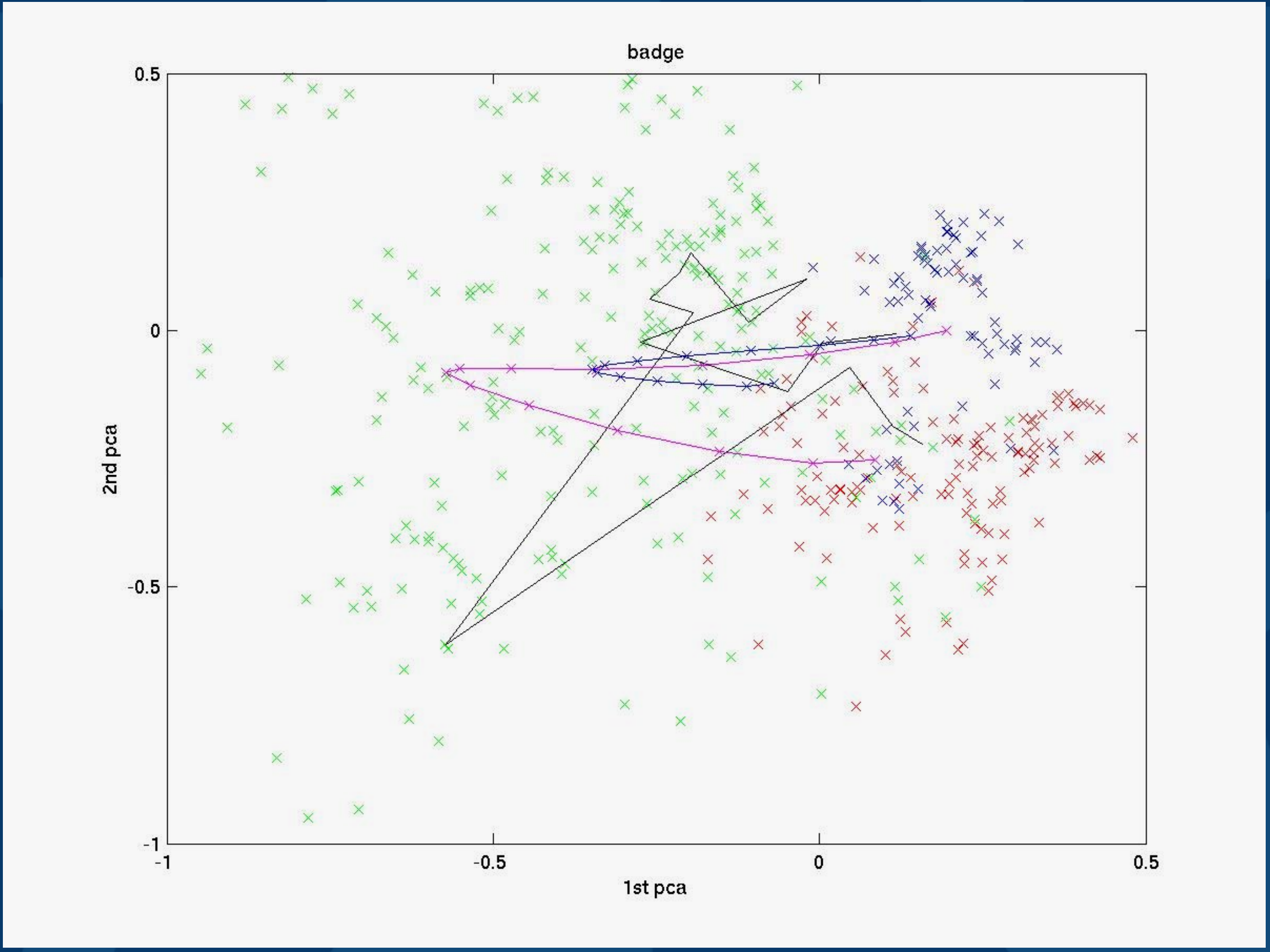


before

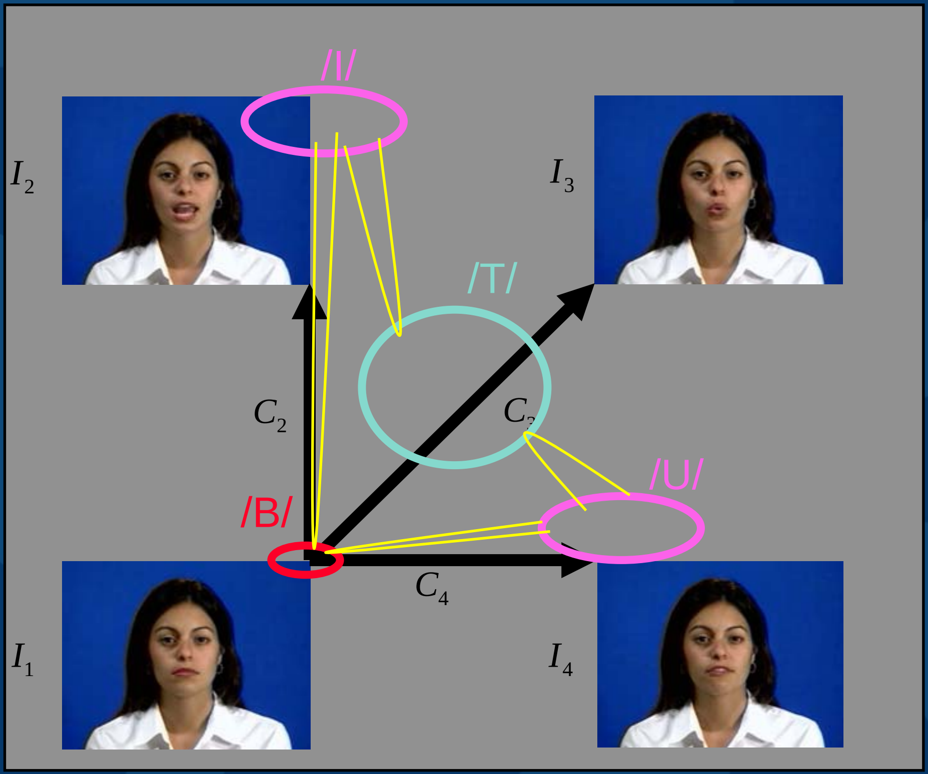
after



Trajectories
Before/After Training

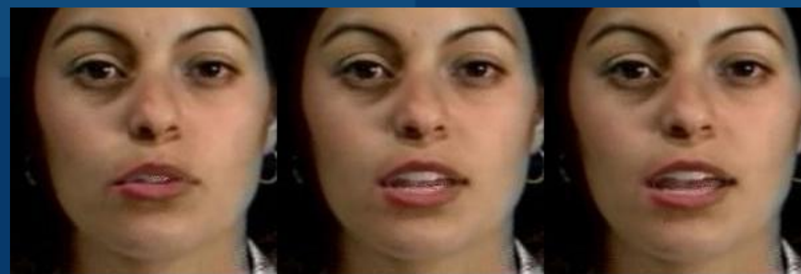


Coarticulation Model

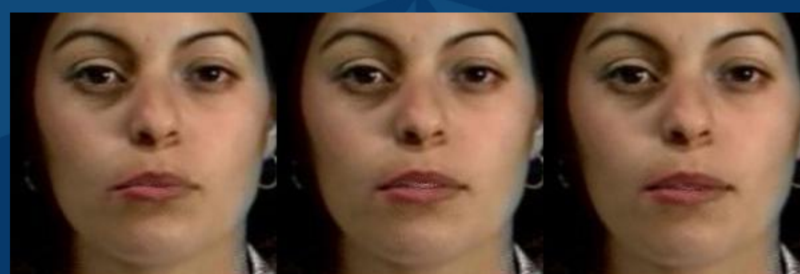


Coarticulation controlled by width of cluster regions

Coarticulation



/utu/ /iti/ /ata/



/ubu/ /ibi/ /aba/

Big Picture

Pre-process

Construct MMM

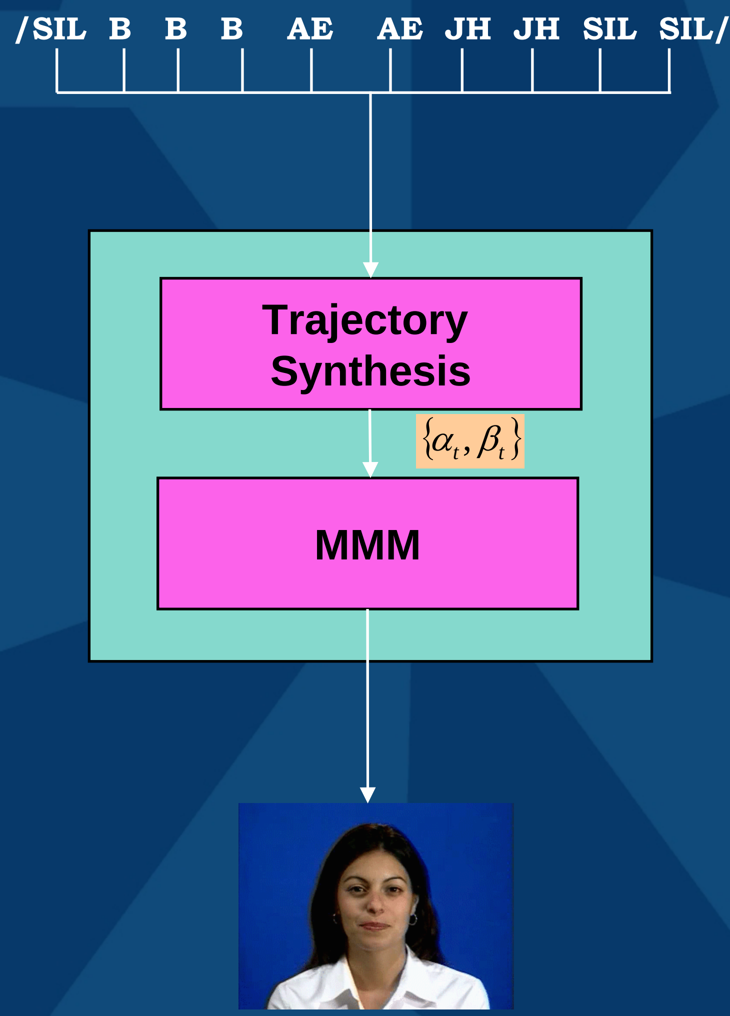
Analyze Corpus

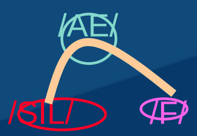
$\{\alpha_t, \beta_t\}$

Train phonetic models

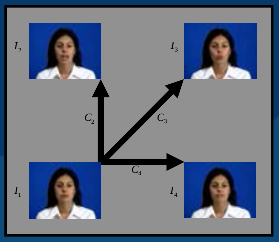
Synthesize!

Post-process





$\{\mu_p, \Sigma_p\}$



$\{I_i, C_i\}$

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

Experiment	% correct	P<
Single presentation	52.1%	0.3
Double presentation	46.6%	0.5

We win!

Visual Intelligibility

Experiment	%correct on N	%correct on S	P<
Words+Sents	30.01%	21.19%	0.001
Words	38.55%	28.07%	0.001
Sents	24.38%	16.52%	0.01

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