Caltech

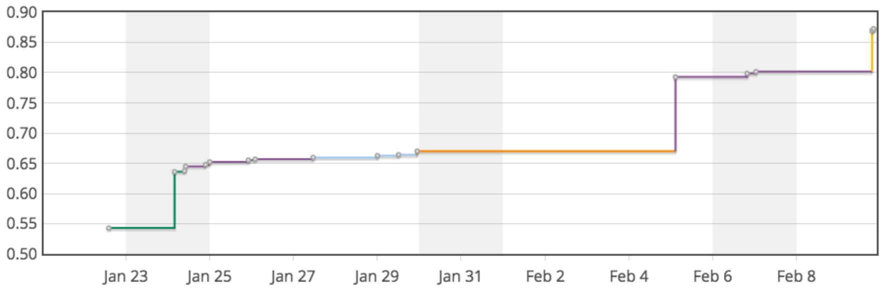
Machine Learning & Data Mining

CS/CNS/EE 155

Lecture 11:

Recent Applications

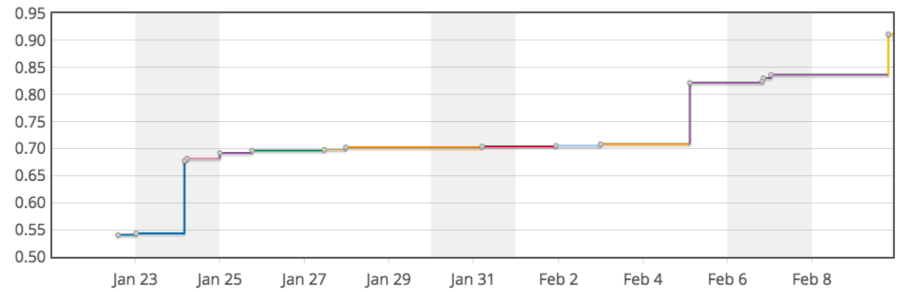
Kaggle Miniproject Closed



This leaderboard is calculated on approximately 50% of the test data.
The final results will be based on the other 50%, so the final standings may be different.

See someone using multiple accounts? Let us know.

#	Δ4d	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	↑42	Black Tornado	0.87278	14	Tue, 09 Feb 2016 21:31:00 (-1.4h)
2	↓1	Yellow Yakuza UnderEducated	0.80178	37	Mon, 08 Feb 2016 01:44:30 (-24.9h)
3	↑9	A.D.D.	0.67160	38	Sat, 06 Feb 2016 00:48:42
4	—	Human Learners	0.67160	52	Tue, 09 Feb 2016 19:51:39 (-0.2h)
5	↓3	Meng Meng Da	0.67012	38	Tue, 09 Feb 2016 10:05:40 (-10.5d)
6	↓3	The Riders of Rohan	0.67012	38	Tue, 09 Feb 2016 19:32:23 (-4.8d)
7	—	adhd	0.67012	26	Sat, 06 Feb 2016 00:25:57 (-0.1h)
8	↑24	monday	0.67012	25	Sat, 06 Feb 2016 00:42:36 (-0.1h)
9	↓4	VoraciousKinkyZebras	0.66716	63	Tue, 09 Feb 2016 05:43:10 (-9.8d)
10	↓4	StickerParty	0.66716	48	Tue, 09 Feb 2016 04:16:34 (-5.1d)
11	↑4	Miss.GreenBean	0.66568	13	Tue, 09 Feb 2016 17:42:49 (-36.3h)
12	↓2	10 Points to Hufflepuff	0.66568	36	Tue, 09 Feb 2016 12:42:22 (-16.6h)
13	new	Victorious Secret	0.66568	13	Tue, 09 Feb 2016 20:04:58
14	↑14	NorthSide StrongSide	0.66420	41	Tue, 09 Feb 2016 19:56:26 (-2.8d)
15	↓2	Prachi	0.66420	55	Tue, 09 Feb 2016 21:56:13 (-2.6d)



This competition has completed. This leaderboard reflects the final standings.

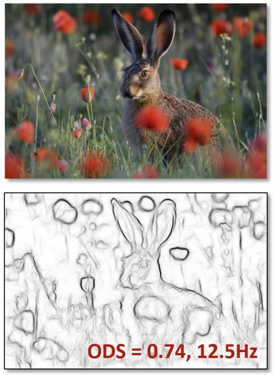
See someone using multiple accounts? Let us know.

#	Δrank	Team Name	Score	Entries	Last Submission UTC (Best – Last Submission)
1	—	Black Tornado	0.90869	14	Tue, 09 Feb 2016 21:31:00 (-1.4h)
2	—	Yellow Yakuza UnderEducated	0.83652	37	Mon, 08 Feb 2016 01:44:30 (-24.9h)
3	↑16	Andrew "Uchiha" Chico	0.71723	5	Tue, 09 Feb 2016 19:13:54
4	—	Human Learners	0.70545	52	Tue, 09 Feb 2016 19:51:39 (-4.8d)
5	↑10	Prachi	0.70250	55	Tue, 09 Feb 2016 21:56:13 (-5d)
6	↓1	Meng Meng Da	0.69809	38	Tue, 09 Feb 2016 10:05:40 (-8.2d)
7	↑2	VoraciousKinkyZebras	0.69809	63	Tue, 09 Feb 2016 05:43:10 (-0.2h)
8	↑18	Do you even train, bro?	0.69661	11	Tue, 09 Feb 2016 12:19:59 (-4.7d)
9	↑3	10 Points to Hufflepuff	0.69514	36	Tue, 09 Feb 2016 12:42:22 (-2.6d)
10	↑10	Nico~Nico~Ni~~☆	0.69219	38	Tue, 09 Feb 2016 20:35:30 (-4.4d)
11	↓1	StickerParty	0.69072	48	Tue, 09 Feb 2016 04:16:34 (-5.1d)
12	↑9	D	0.69072	27	Tue, 09 Feb 2016 16:54:51 (-3.7d)
13	↑3	Walker Mills	0.69072	25	Mon, 08 Feb 2016 23:41:48 (-24.9h)
14	↑34	gg	0.68925	3	Wed, 27 Jan 2016 04:18:49 (-24.1h)
15	↑10	AbysML	0.68925	8	Tue, 09 Feb 2016 07:37:17 (-2h)

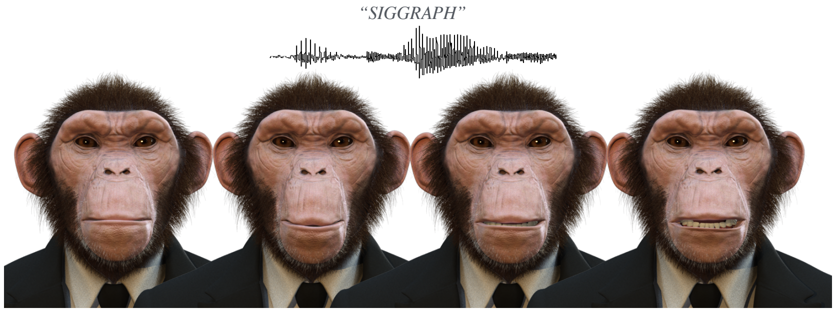
Today

- Recent Applications:

Edge Detection



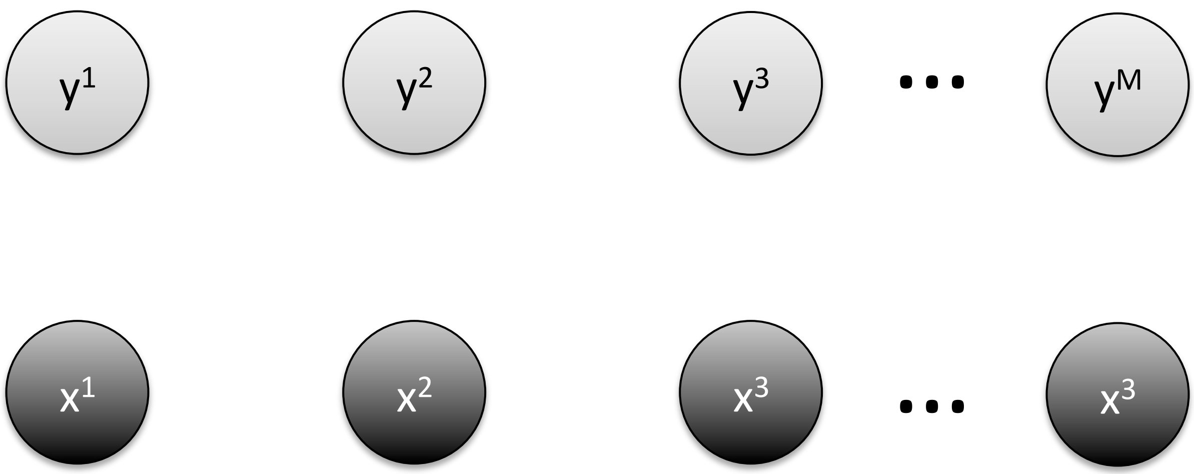
Speech Animation



- Introduction to Learning Reductions

Recall: Sequence Prediction

- X = "The Dog Jumped Over the Fence"
- Y = D N V P D N

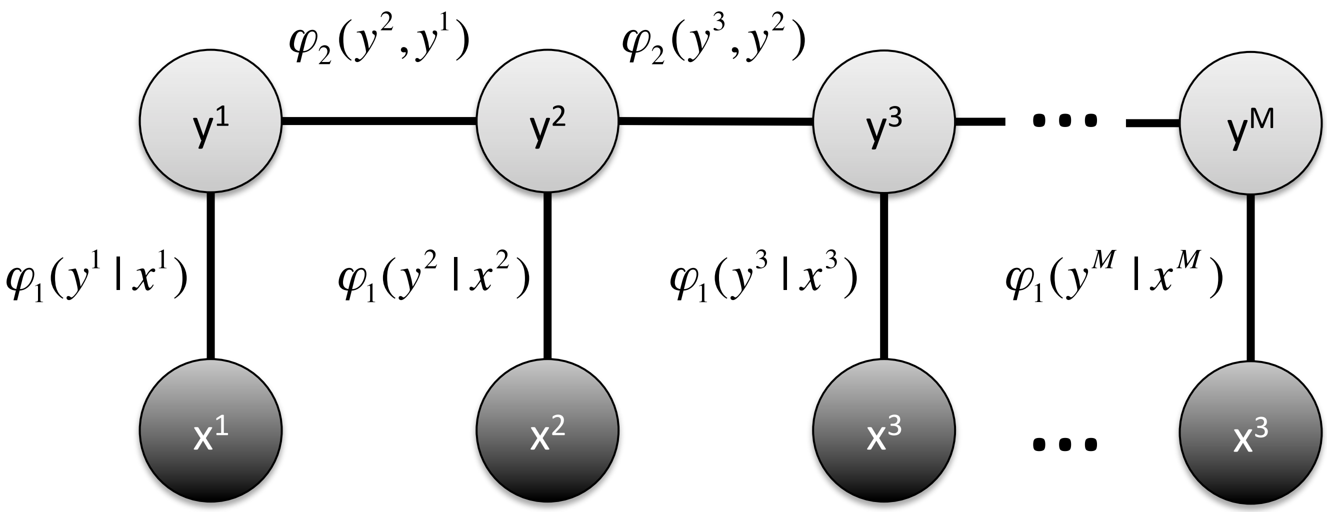


Recall: Conditional Random Field

$$P(y \mid x) = \frac{1}{Z(x)} \exp\left\{ F(x,y) \right\}$$

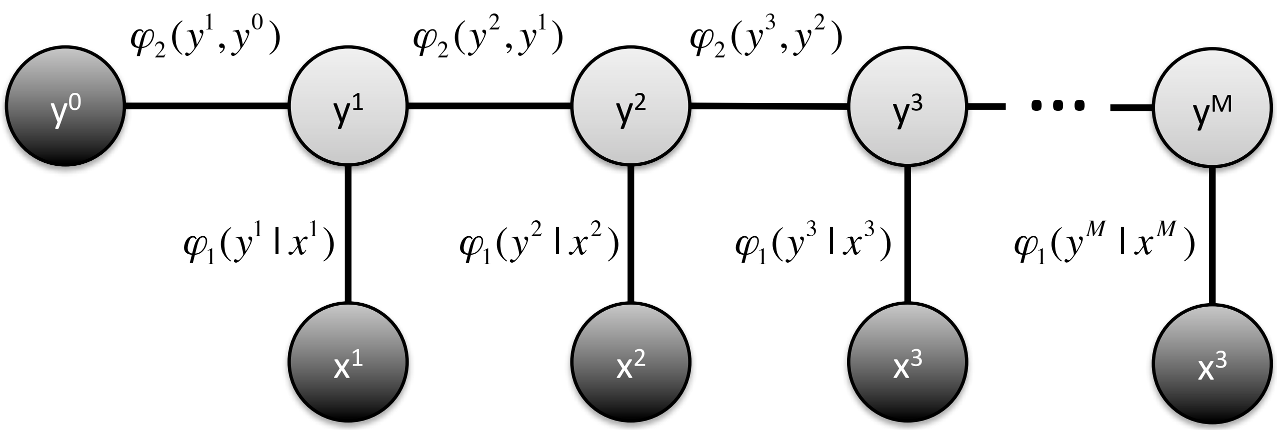
$$F(y,x) \equiv \sum_{j=1}^{M} \left[w^T \varphi^j (y^j, y^{j-1} \mid x) \right]$$

$$\varphi^j (a,b \mid x) = \begin{bmatrix} \varphi_1 (a \mid x^j) \\ \varphi_2 (a,b) \end{bmatrix}$$



Limitations of CRFs

- Linear model
  - Requires good feature representation
- Only first-order effects
  - Cannot model higher-order dependencies



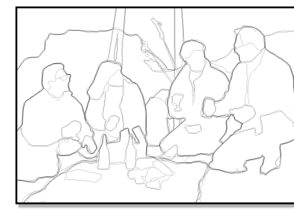
Edge Detection

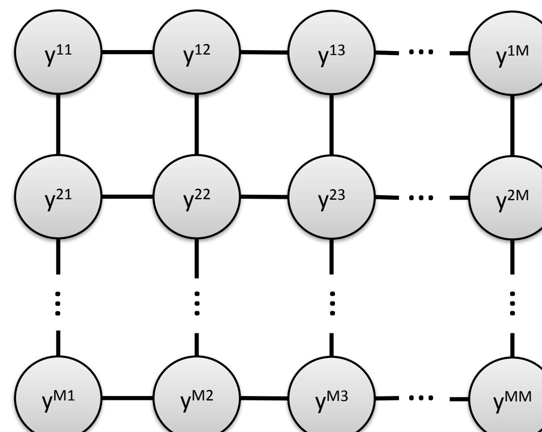
X:

Y:

2D Conditional Random Field

- Each y^{ij} is binary label
  - Edge or Not Edge
- What features?
  - Defined over pixels?





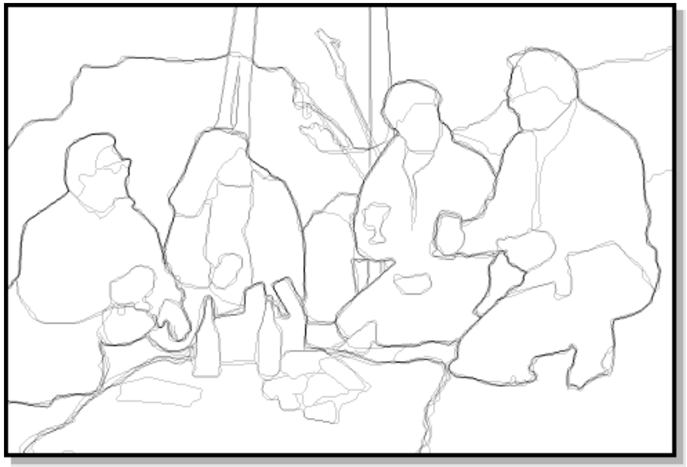
Today: Learning Reductions

- Convert complicated problem into simpler ones
  - Use complex models for simpler problems
  - E.g., decision trees, neural nets

- Recompose predictions for complicated problem

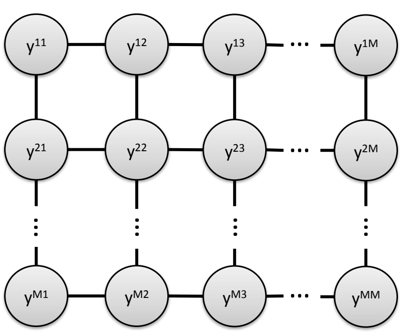
Strong Local Properties

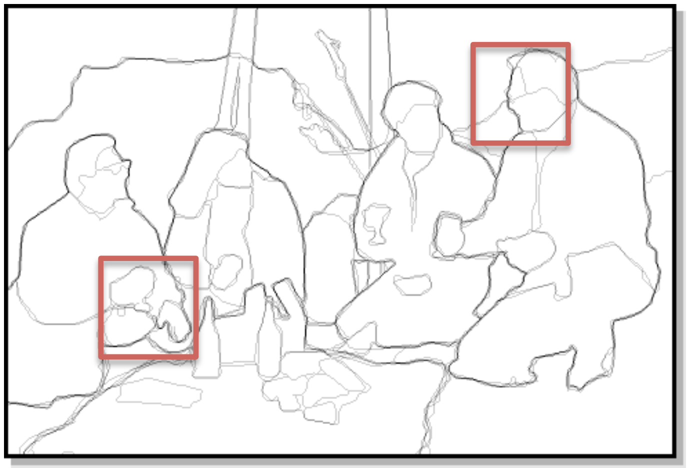
- Local patterns matter
  - E.g., image patches
- Complex relationship
  - Non-linear



Weak Global Properties

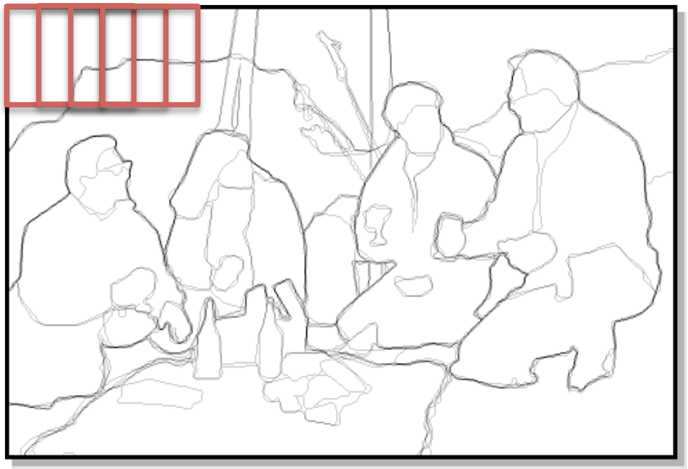
- Edge detections local
- No need to fully connect model



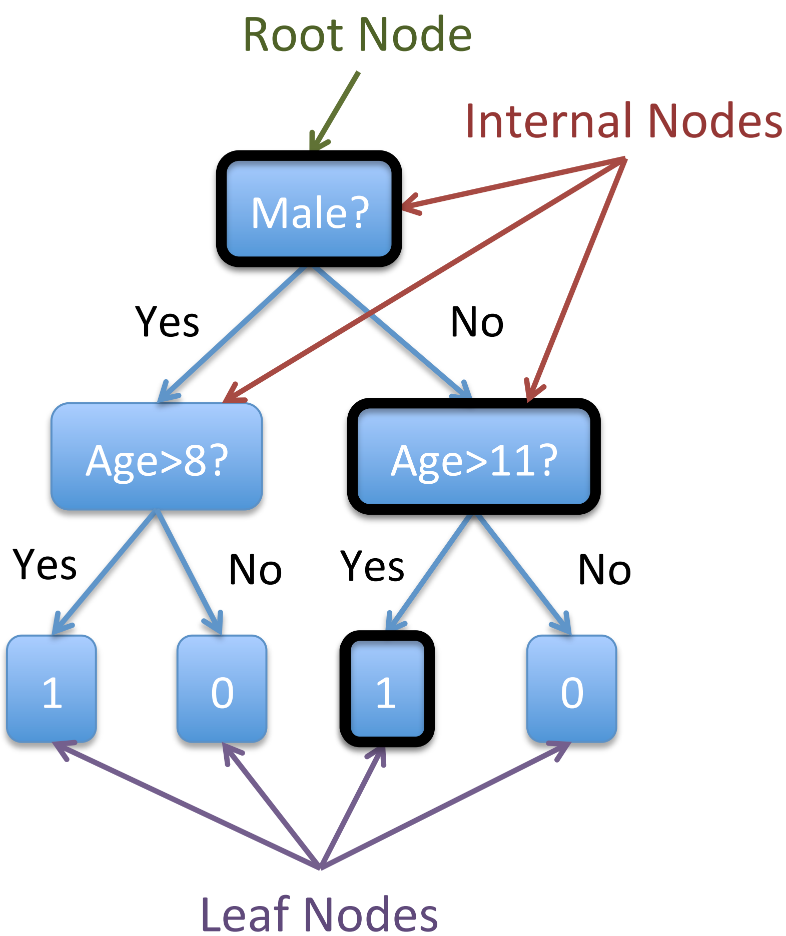


Sliding Window Approach

- Train model to predict patches
  - E.g., 16x16
- Slide across image
- **What model?**



Recall: Binary Decision Tree



Input: 

Alice
Gender: Female
Age: 14

Prediction: Height > 55”

Every **internal node** has a **binary** query function q(x).

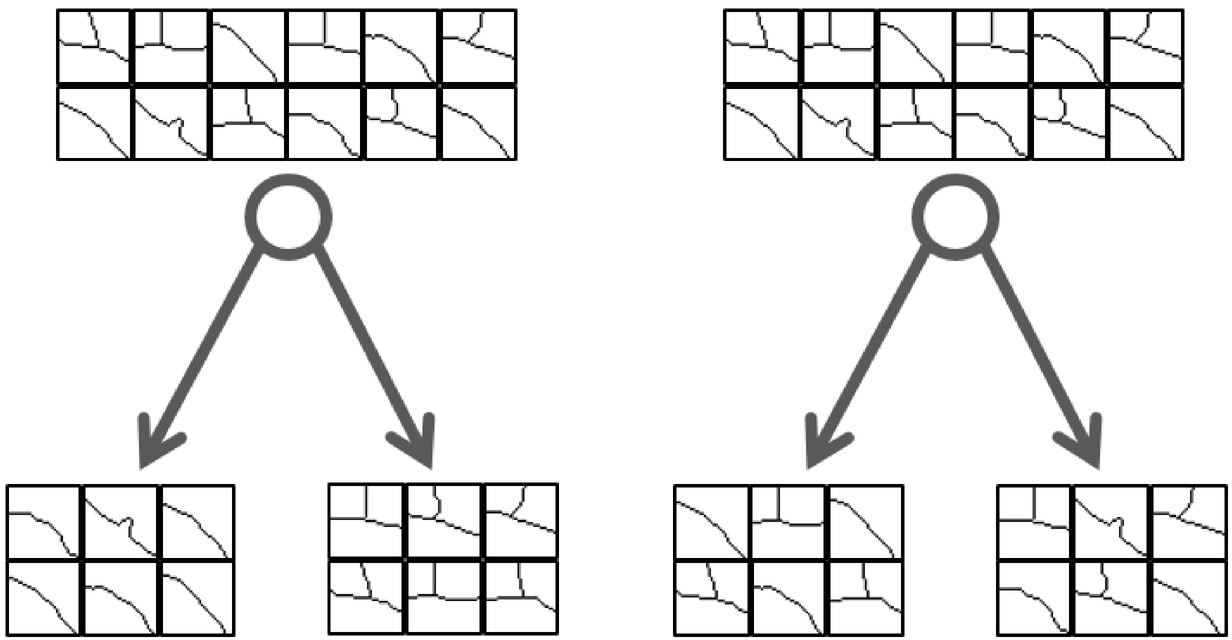
Every **leaf node** has a prediction, e.g., 0 or 1.

Prediction starts at **root node**.
Recursively calls query function.
Positive response → Left Child.
Negative response → Right Child.
Repeat until Leaf Node.

Structured Decision Tree

- Each leaf node predicts a 16x16 edge matrix
  - Average of all training patch labels
- Prediction is very fast!
  - Slide predictor across image, average results
  - No need for Viterbi-type algorithms
- What is splitting criterion?
- What is query set?

Structured Information Gain



Good!

Bad!

“Structured Random Forests for Fast Edge Detection”
Dollár & Zitnick, ICCV 2013

Structured Information Gain

1. First map labels to coordinate system
   A. For each coordinate, choose pair of pixels
   B. Set coordinate to 1 if in same segment, 0 o.w.
      - Coordinate 1 = 0
      - Coordinate 2 = 1
      - Etc…
2. Cluster training labels

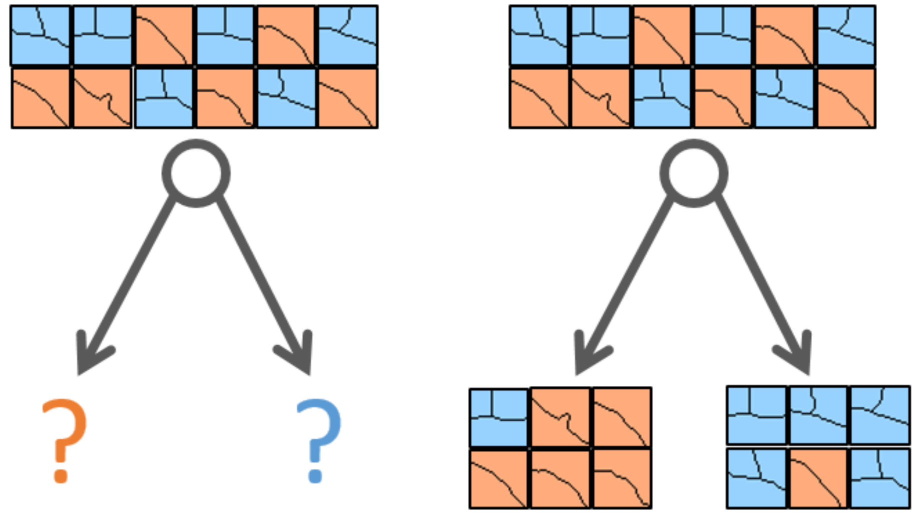
For each training example!

(Actual approach more complicated.)

"Structured Random Forests for Fast Edge Detection"
Dollár & Zitnick, ICCV 2013

Multiclass Entropy

- Reduced training labels to K clusters
  - Can treat as multiclass classification
- Impurity measure = multiclass entropy



Query Set

- Features about color gradients
  - Image gets darker from column 1 to column 5
  - Image gets more blue from row 7 to row 3
  - Etc…
  - 7228 features total

(Actual approach more complicated.)

"Structured Random Forests for Fast Edge Detection"
Dollár & Zitnick, ICCV 2013

Putting it Together

- Create new training set $\hat{S} = \{(x,\hat{y})\}$
  - x = 16x16 image patch
  - $\hat{y}$ = 16x16 ground truth edges

- Train structured DT on $\hat{S}$

- Predict by sliding DT over input image
  - Average predictions

(Actual approach more complicated.)

"Structured Random Forests for Fast Edge Detection"
Dollár & Zitnick, ICCV 2013


Input
Ground Truth
Four Versions of Method

Comparable accuracy vs state-of-the-art

Much faster!

	ODS	OIS	AP	FPS
Human	.80	.80	-	-
Canny	.60	.64	.58	15
Felz-Hutt [11]	.61	.64	.56	10
Hidayat-Green [16]	.62†	-	-	20
BEL [9]	.66†	-	-	1/10
gPb + GPU [6]	.70†	-	-	1/2‡
gPb [1]	.71	.74	.65	1/240
gPb-owt-ucm [1]	.73	**.76**	.73	1/240
Sketch tokens [21]	.73	.75	**.78**	1
SCG [31]	**.74**	**.76**	.77	1/280
SE-SS, T=1	.72	.74	.77	**60**
SE-SS, T=4	.73	.75	.77	30
SE-MS, T=4	**.74**	**.76**	**.78**	6

Accuracy Measures

Speed

"Structured Random Forests for Fast Edge Detection"
Dollár & Zitnick, ICCV 2013

Speech Animation

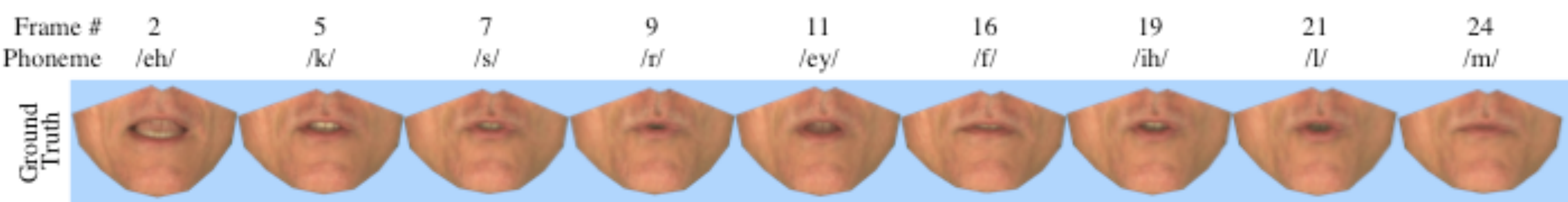
Automatically Animate to Input Audio?

(Given Training Data)

“A Decision Tree Framework for Spatiotemporal Sequence Prediction”
Kim, Yue, Taylor, Matthews, KDD 2015, http://projects.yisongyue.com/visual_speech

Training Data

- ~2500 Sentences
  - Recorded at 30 Hz
  - ~10 hours of recorded speech
- Active Appearance Model
  - Actor’s lower face
  - 30 degrees of freedom (also 100+)



Data from [Taylor et al., 2012]

Prediction Task

Input sequence $X =< x_1, x_2, \ldots, x_{|x|} >$

Output sequence $Y =< y_1, y_2, \ldots, y_{|y|} >$ $, y_t \in R^D$

Goal: learn predictor $h : X \rightarrow Y$

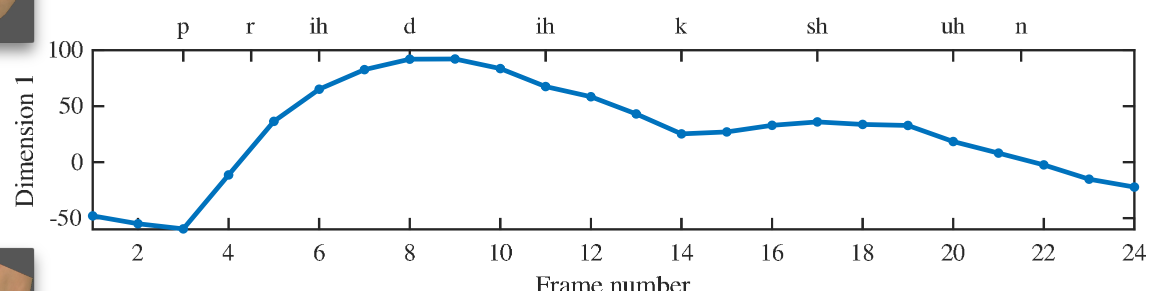
X

Frame	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
Token	-	p	p	r	ih	ih	d	d	ih	ih	ih	ih	k	k	sh	sh	sh	sh	uh	uh	n	-

Phoneme sequence

h

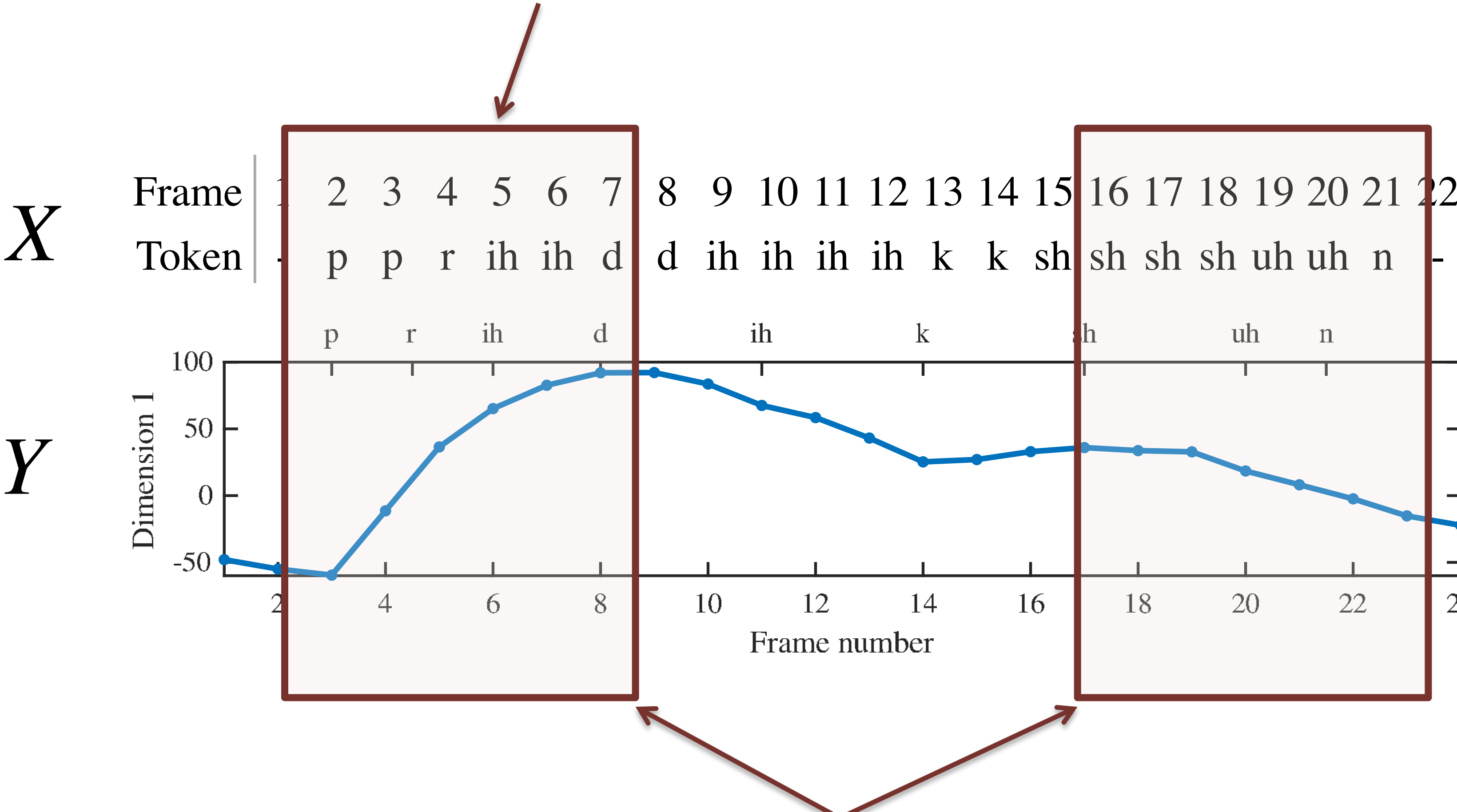
Y



Sequence of face configurations

Temporal curvature can vary smoothly or sharply

(Depends on context – this is the co-articulation problem)

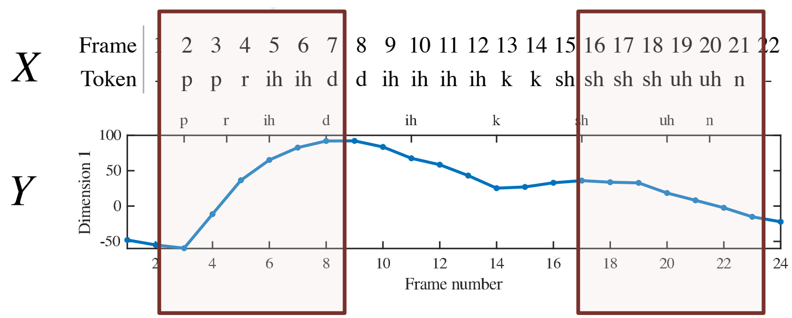


Minimal long-range dependencies

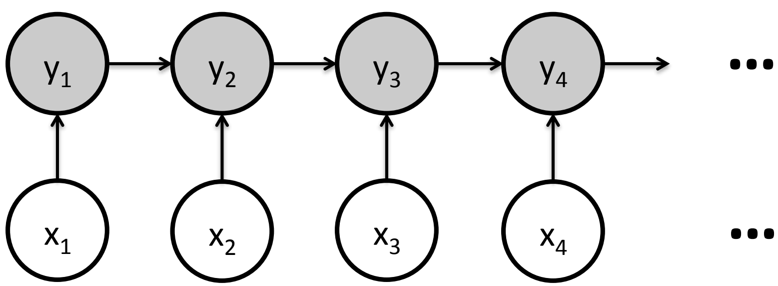
(predi**tion** = constru**tion** = ele**tion**…)

Strong Local Properties

- Need to model arbitrary local curvature

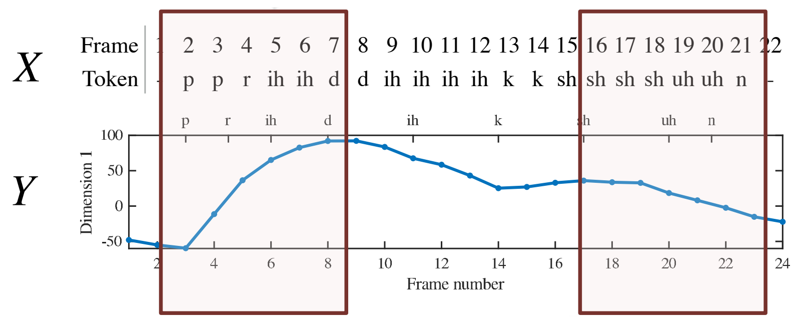


- Not well suited by linear chain models!



Weak Global Properties

- No need to model entire chain directly



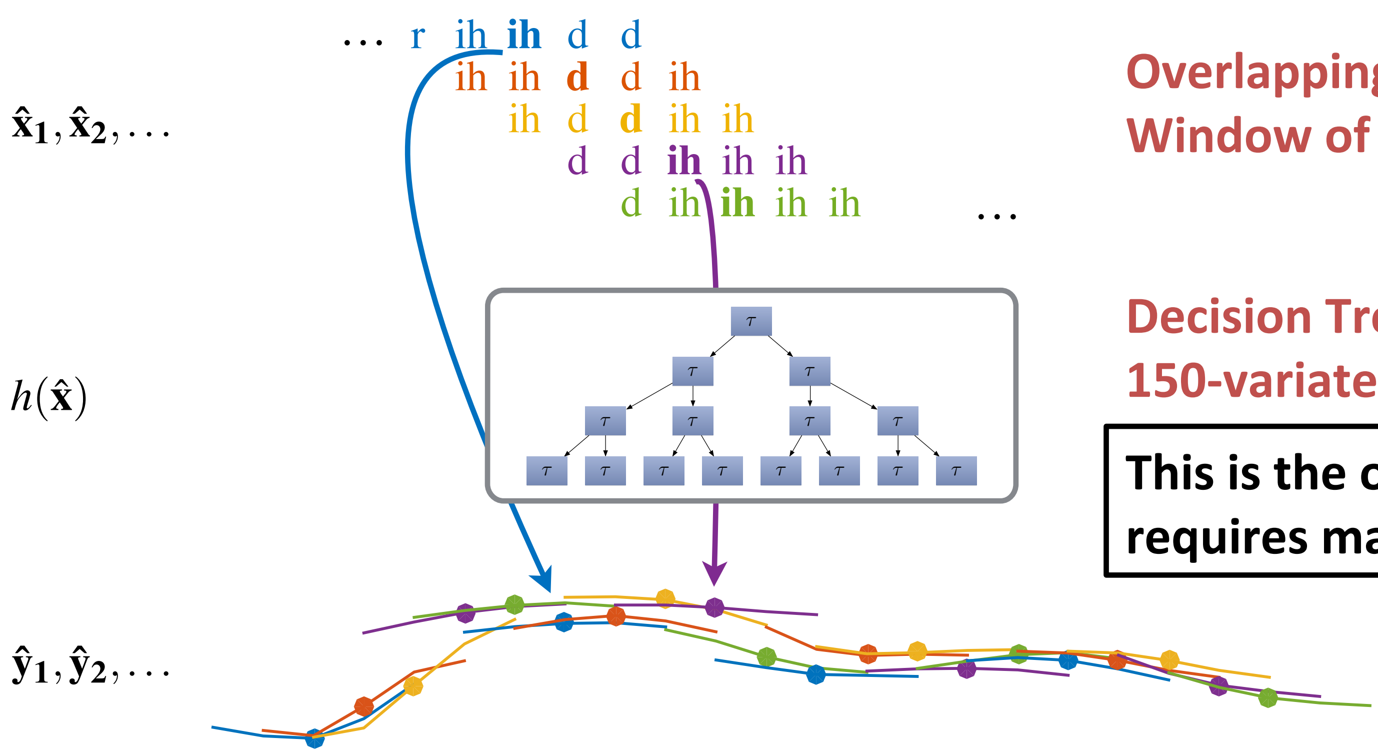
Minimal long-range dependencies

(predi**ction** = constru**ction** = ele**ction**…)

- Motivates sliding window approach!

Input speech: " P R E D I C T I O N "

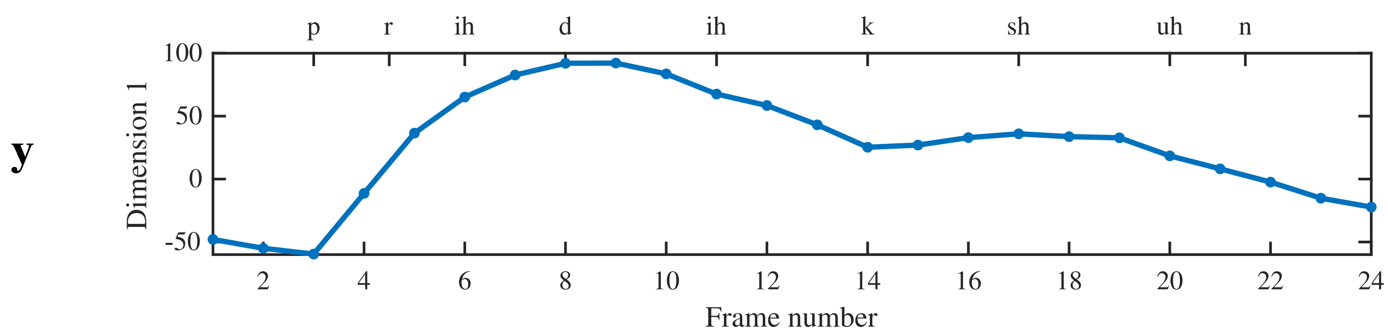
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
	Frame	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
$\mathbf{x}$	Token	-	p	p	r	ih	ih	d	d	ih	ih	ih	ih	k	k	sh	sh	sh	sh	uh	uh	n	-



Overlapping Sliding Window of Inputs

Decision Tree Model 150-variate regression

This is the only thing that requires machine learning!



Aggregate Outputs

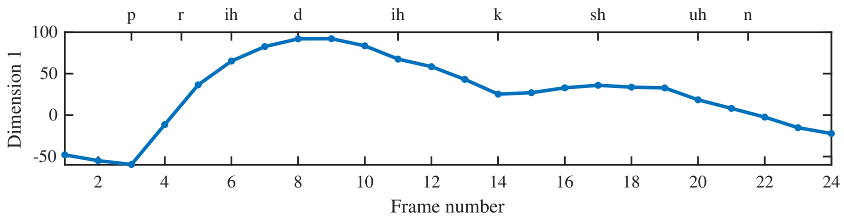
Very fast!

Training

Input speech: **" P R E D I C T I O N "**

	Frame	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22
x	Token	-	p	p	r	ih	ih	d	d	ih	ih	ih	ih	k	k	sh	sh	sh	sh	uh	uh	n	-

y

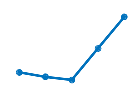
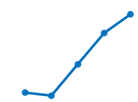


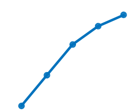
Original Training Data

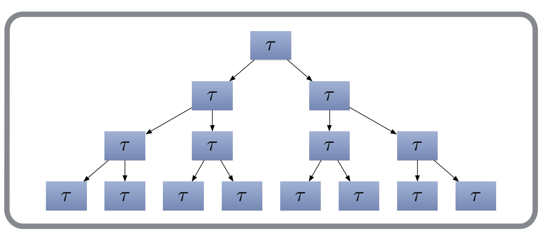
(Variable-Length Trajectory Prediction)

Modified Training Data

(Fixed-Length Multivariate Regression)

$\Big(\langle -,p,p,r,ih\rangle,$  $\Big),\Big(\langle p,p,r,ih,ih\rangle,$  $\Big)$

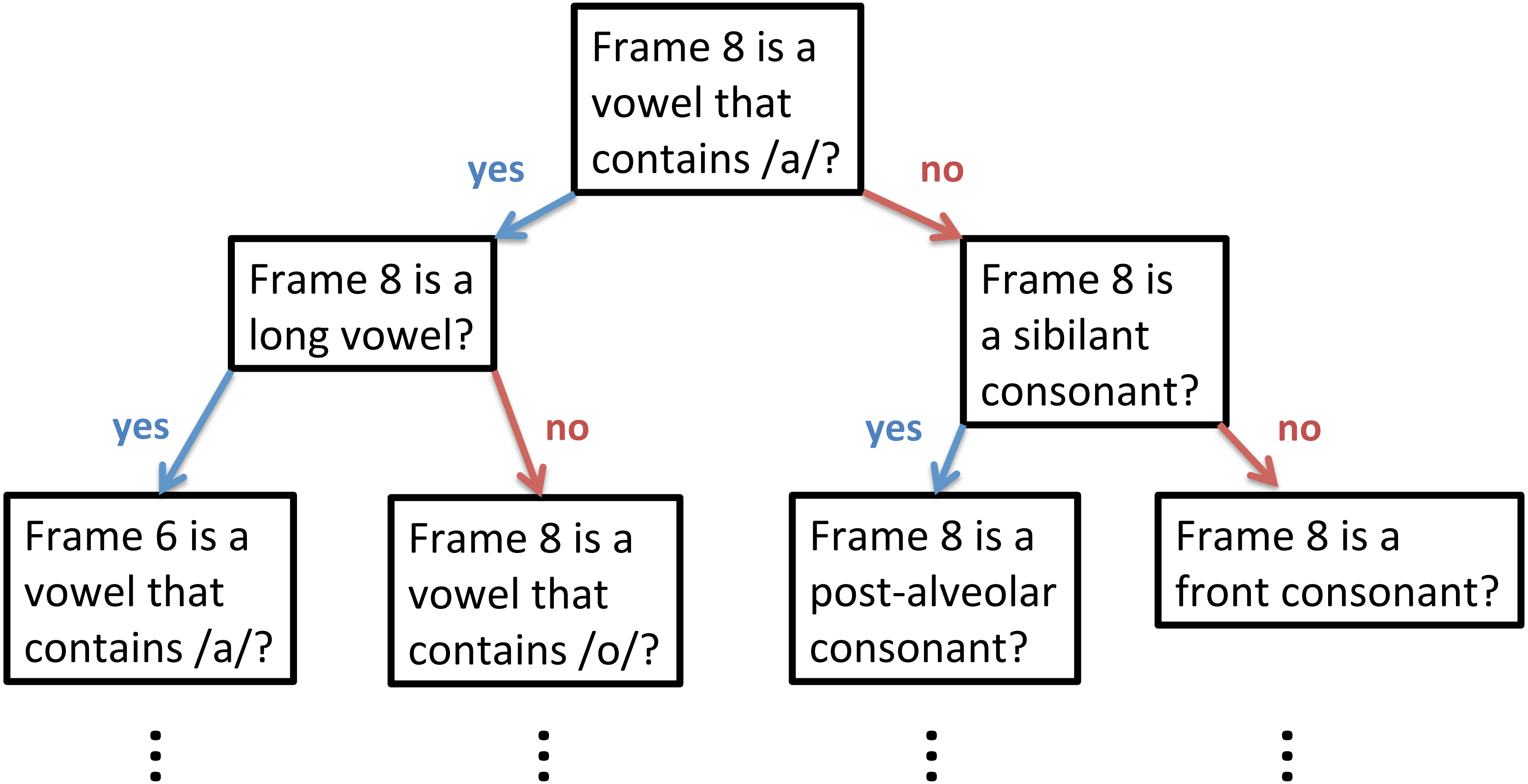
$\Big(\langle p,r,ih,ih,d\rangle,$  $\Big),$ ...



Train Decision Tree

(Or some other regression model)

Query Set for Speech Animation

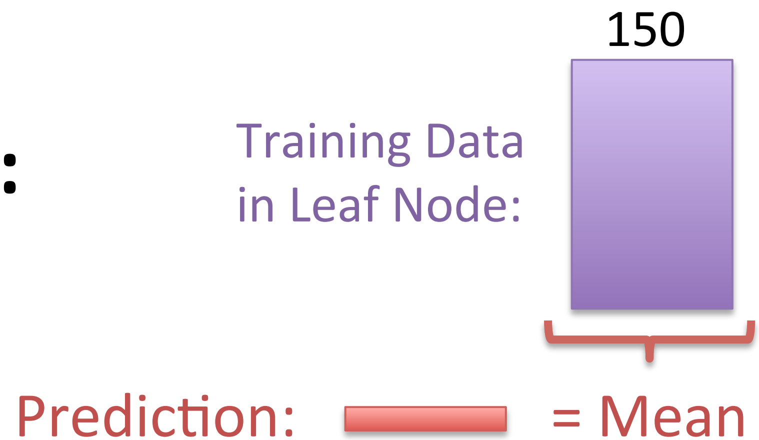


Frames indexed by 1-11 (center is frame 6)

Full tree has 5K+ leaf nodes

Multivariate Regression Tree

- **Prediction:**



- **Training loss:** multivariate squared loss:

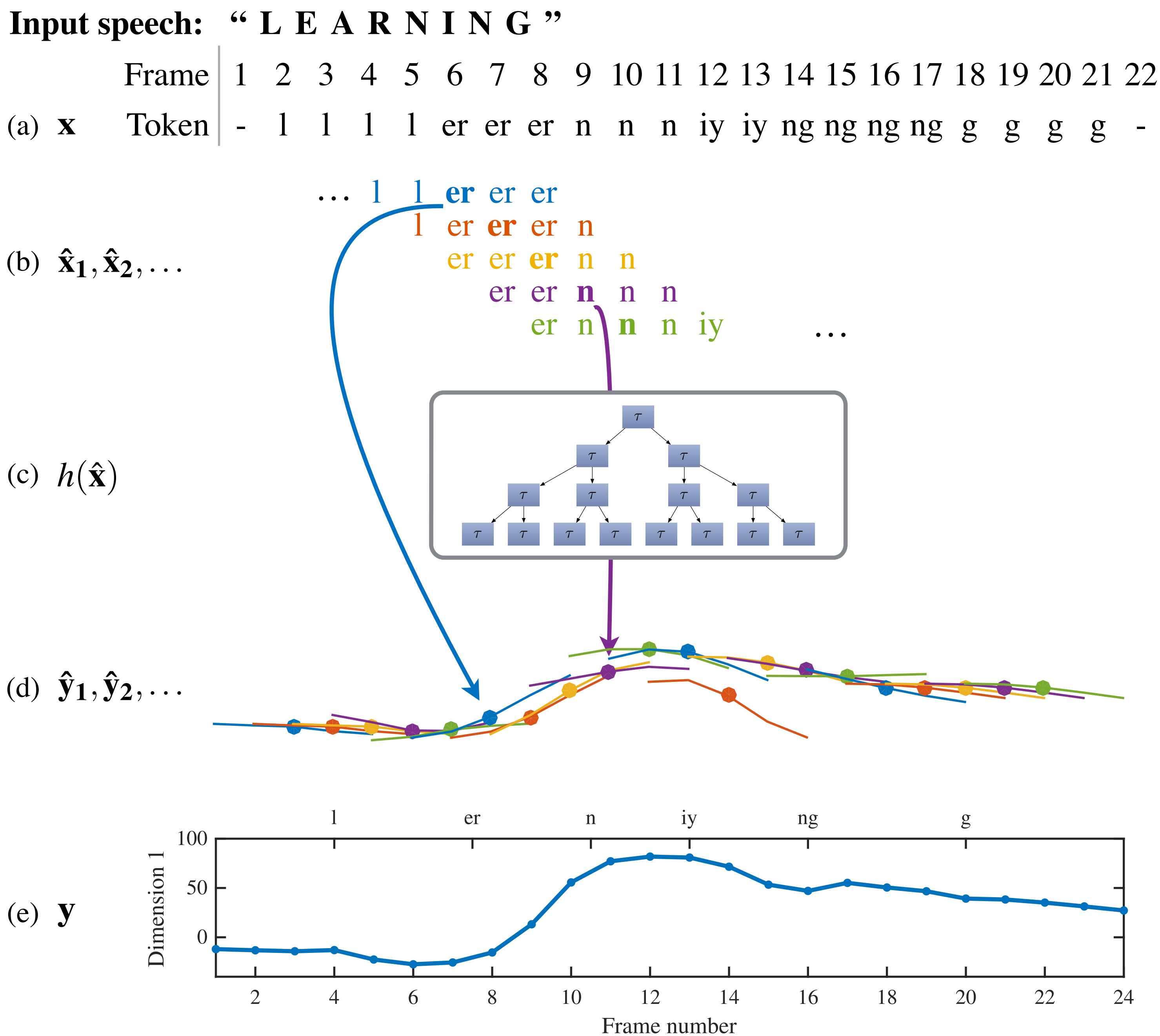
$$\sum_{Leaf} \sum_{\hat{y} \in Leaf} \left\| \hat{y}_{Leaf} - \hat{y} \right\|^2$$

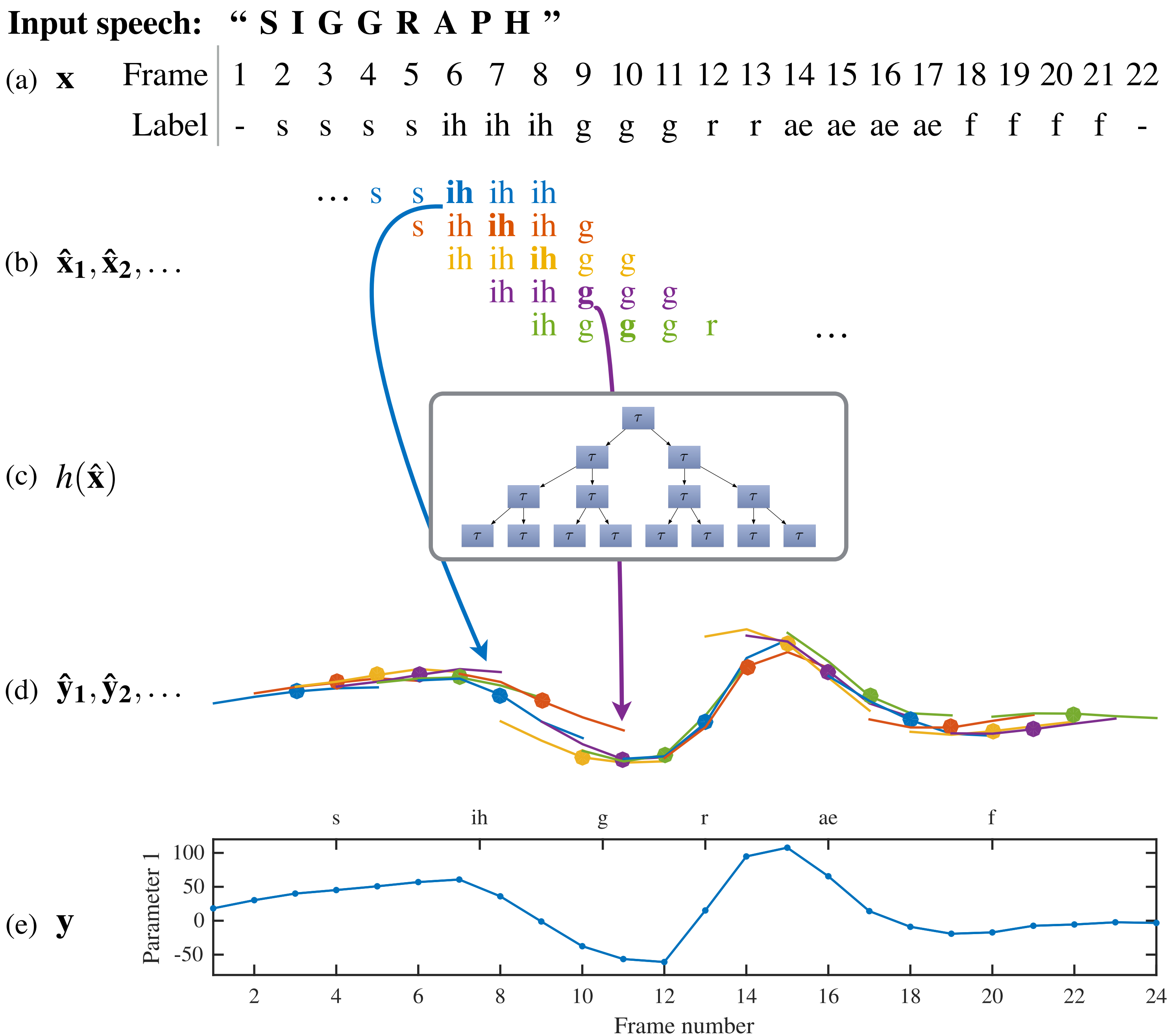
Prediction on New Speaker

"A Decision Tree Framework for Spatiotemporal Sequence Prediction"
Kim, Yue, Taylor, Matthews, KDD 2015, http://projects.yisongyue.com/visual_speech

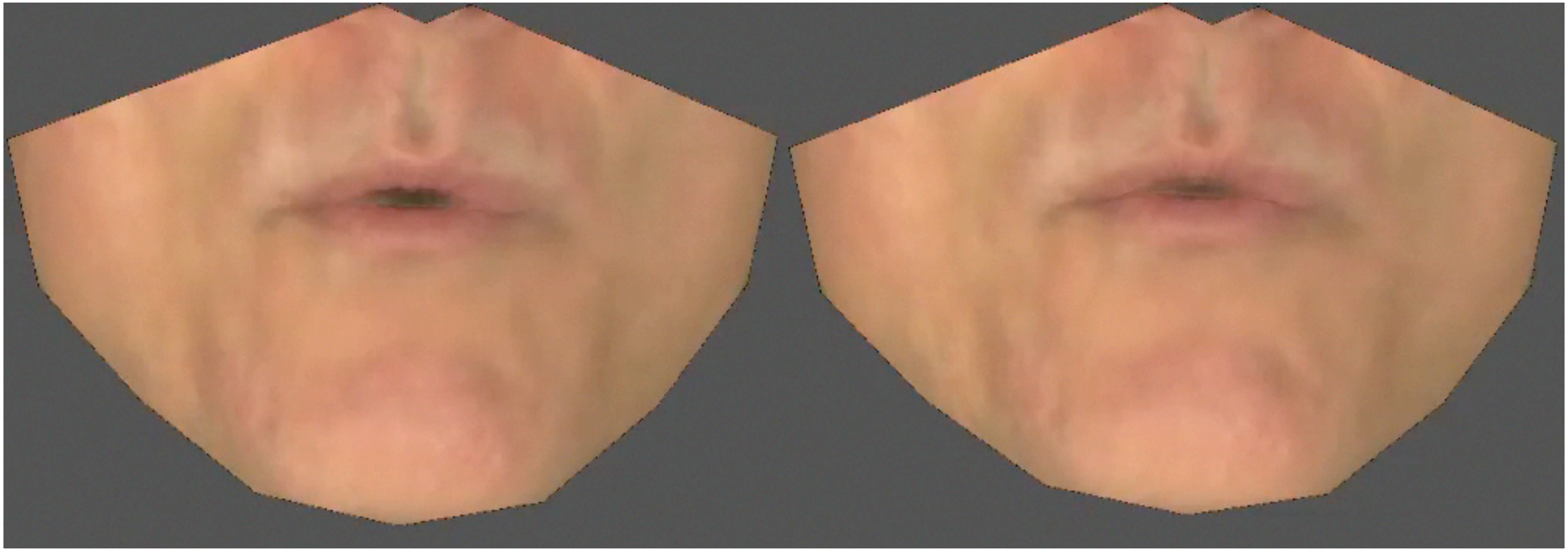
Prediction on New Speaker

"A Decision Tree Framework for Spatiotemporal Sequence Prediction"
Kim, Yue, Taylor, Matthews, KDD 2015, http://projects.yisongyue.com/visual_speech


Input speech: “ L E A R N I N G ”
Frame 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
(a) x Token - l l l l er er er n n n iy iy ng ng ng ng g g g g -
(b) $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots$
… l l er er er
l er er er n
er er er n n
er er n n n
er n n n iy
…
(c) $h(\hat{\mathbf{x}})$
τ
(d) $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots$
(e) y
l
er
n
iy
ng
g
100
50
0
Dimension 1
2
4
6
8
10
12
14
16
18
20
22
24
Frame number


Input speech: “ S I G G R A P H ”
(a) x
Frame 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
Label - s s s s ih ih ih g g g r r ae ae ae ae f f f f -
(b) $\hat{\mathbf{x}}_1, \hat{\mathbf{x}}_2, \ldots$
… s s ih ih ih
s ih ih ih g
ih ih ih g g
ih ih g g g
ih g g g r …
(c) $h(\hat{\mathbf{x}})$
(d) $\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots$
(e) y
s ih g r ae f
Parameter 1
100 50 0 -50
2 4 6 8 10 12 14 16 18 20 22 24
Frame number

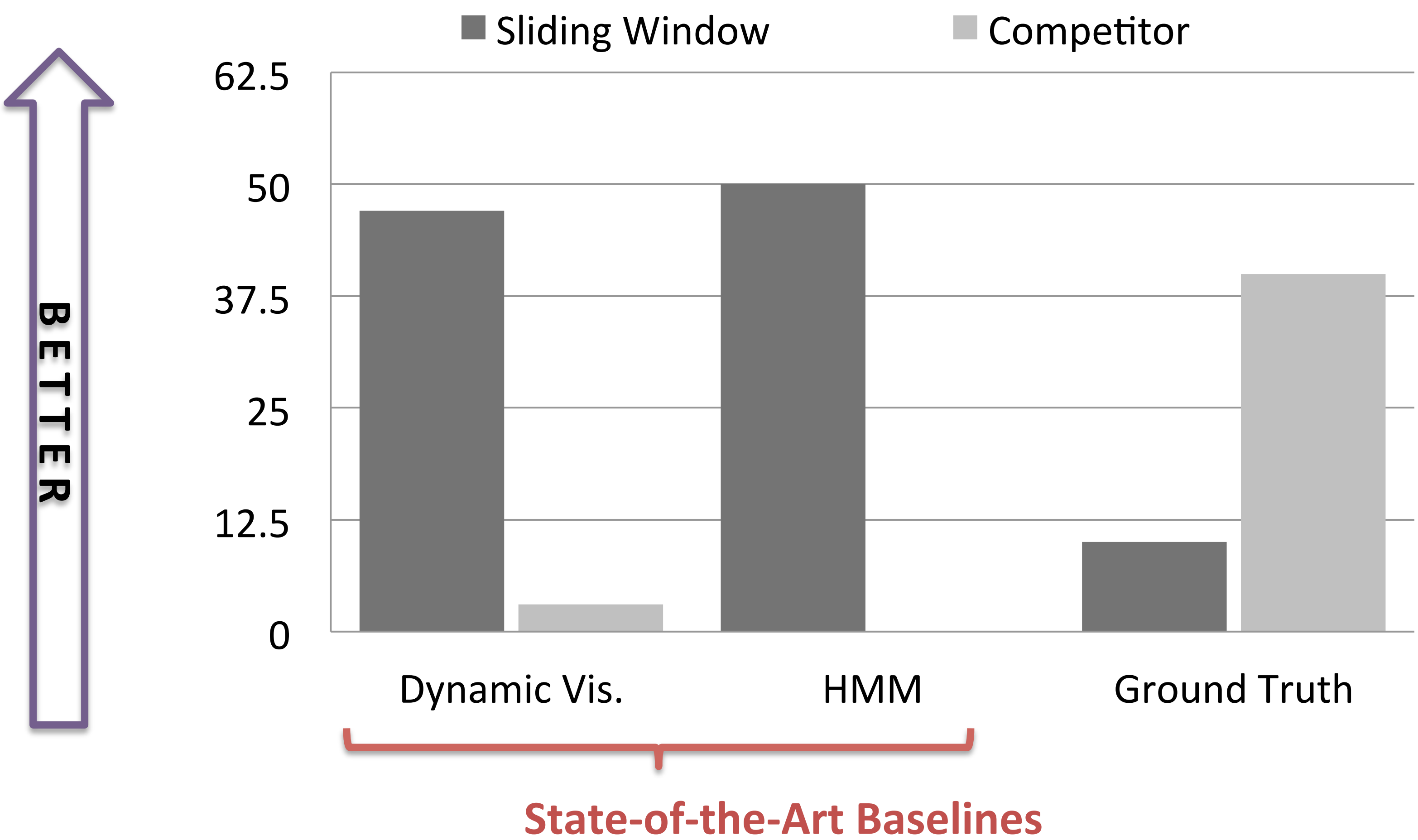
Side-by-Side User Study



Comparing our approach versus competitor on 50 held-out test sentences.

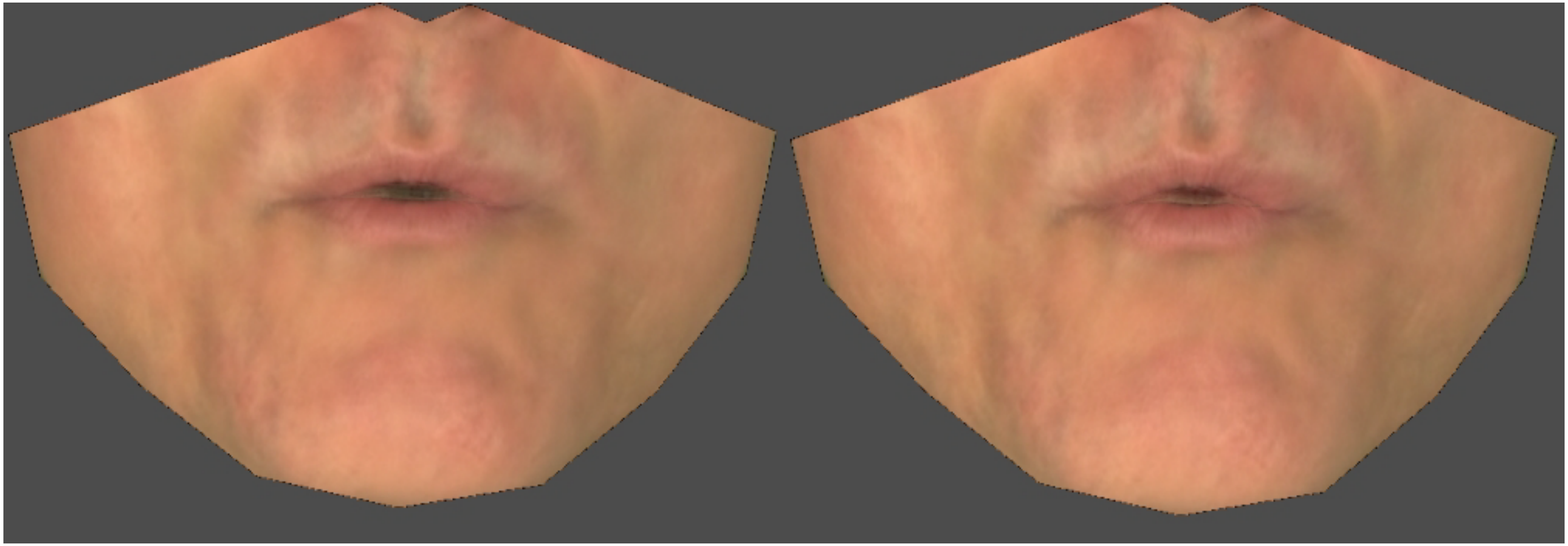
“A Decision Tree Framework for Spatiotemporal Sequence Prediction”
Kim, Yue, Taylor, Matthews, KDD 2015, http://projects.yisongyue.com/visual_speech

Side-by-Side User Study



Comparing our approach versus competitor on 50 held-out test sentences.

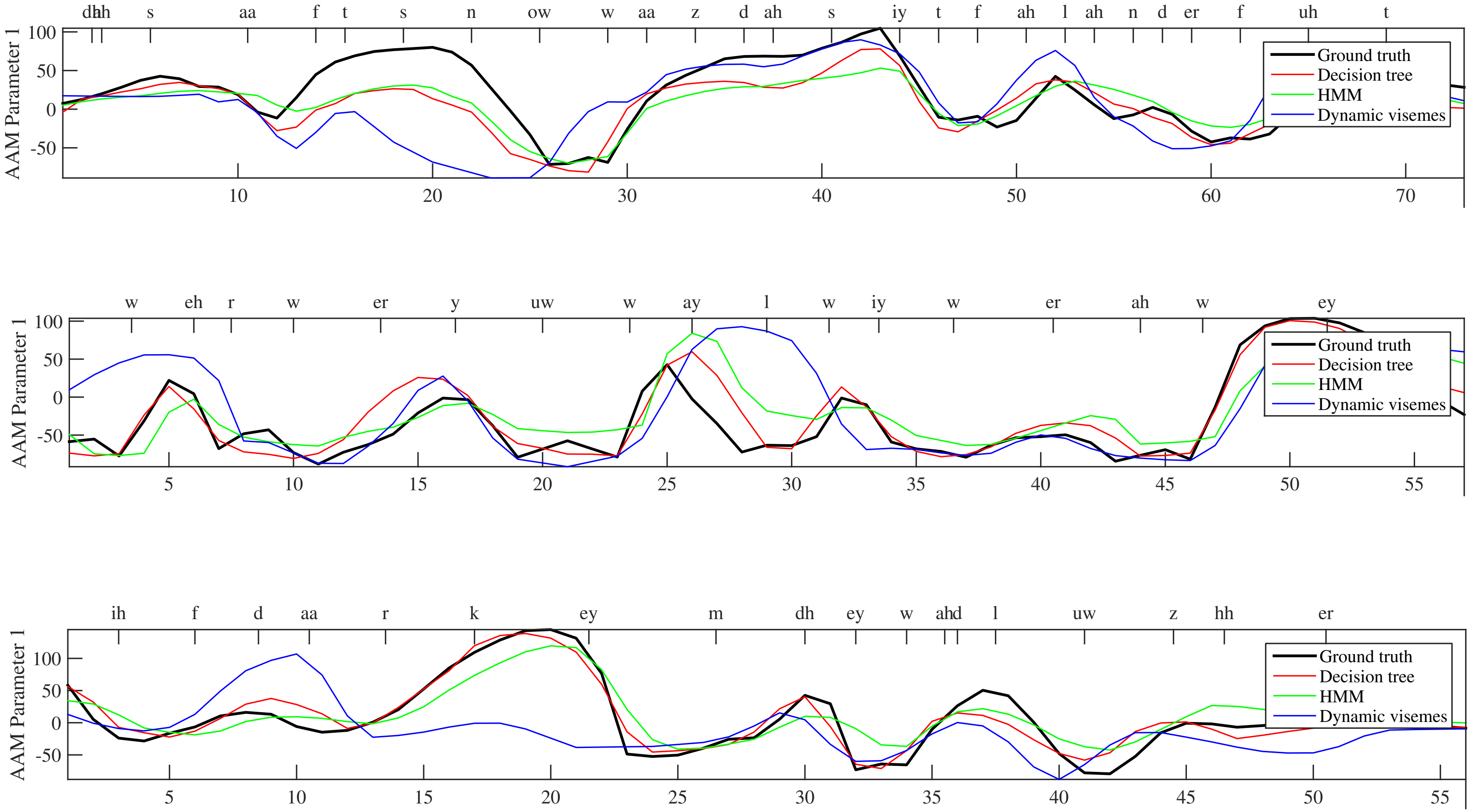
Comparison with Ground Truth

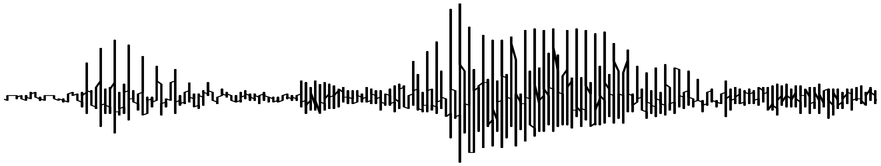


We under-articulate relative to ground truth!
(Could be solved with more training data…)

"A Decision Tree Framework for Spatiotemporal Sequence Prediction"
Kim, Yue, Taylor, Matthews, KDD 2015, http://projects.yisongyue.com/visual_speech

Comparison with Ground Truth

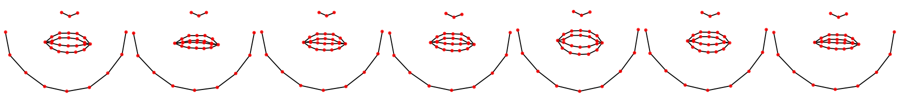




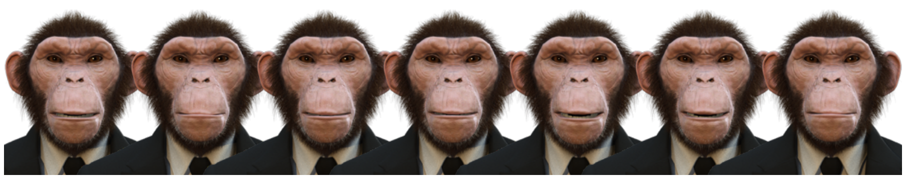
Input Audio

s s s s s ih ih ih g g r r ae ae ae ae f f f

Speech Recognition



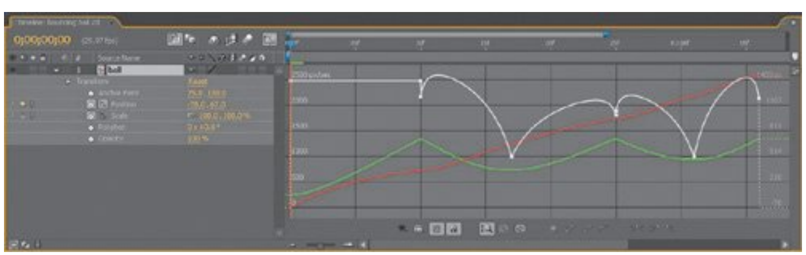
Speech Animation



(chimp rig courtesy of Hao Li)

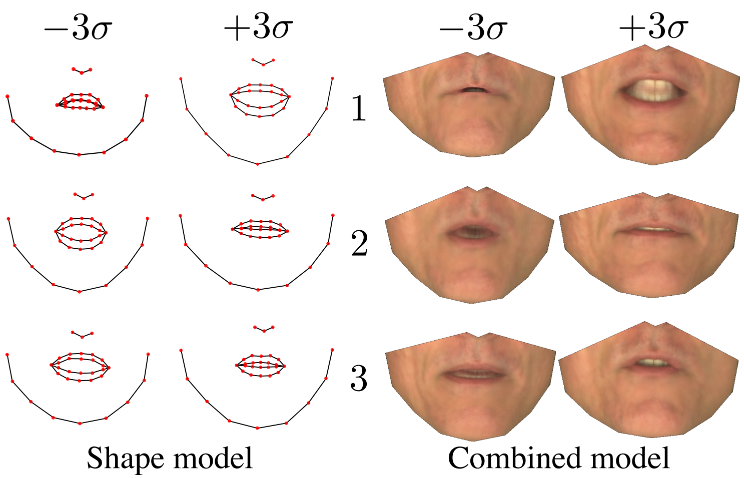
Retargeting

E.g., [Sumner & Popovic 2004]



Editing

Aside: Retargeting



Reference face → target face

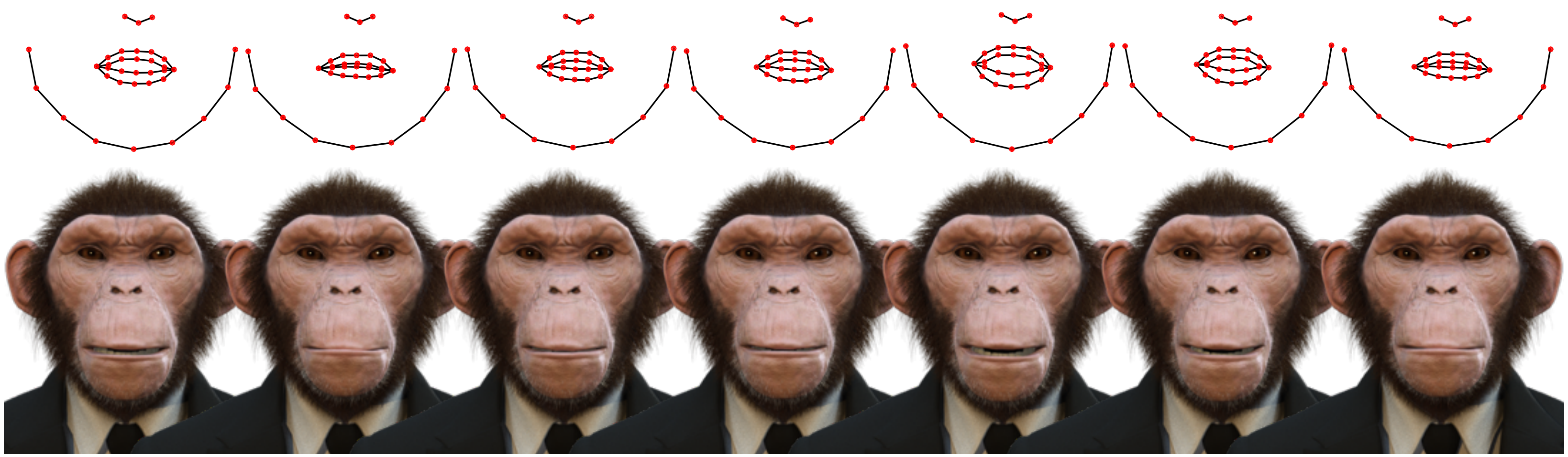
(Semi-)Automatic:

Deformation Transfer [Sumner & Popovic 2004]

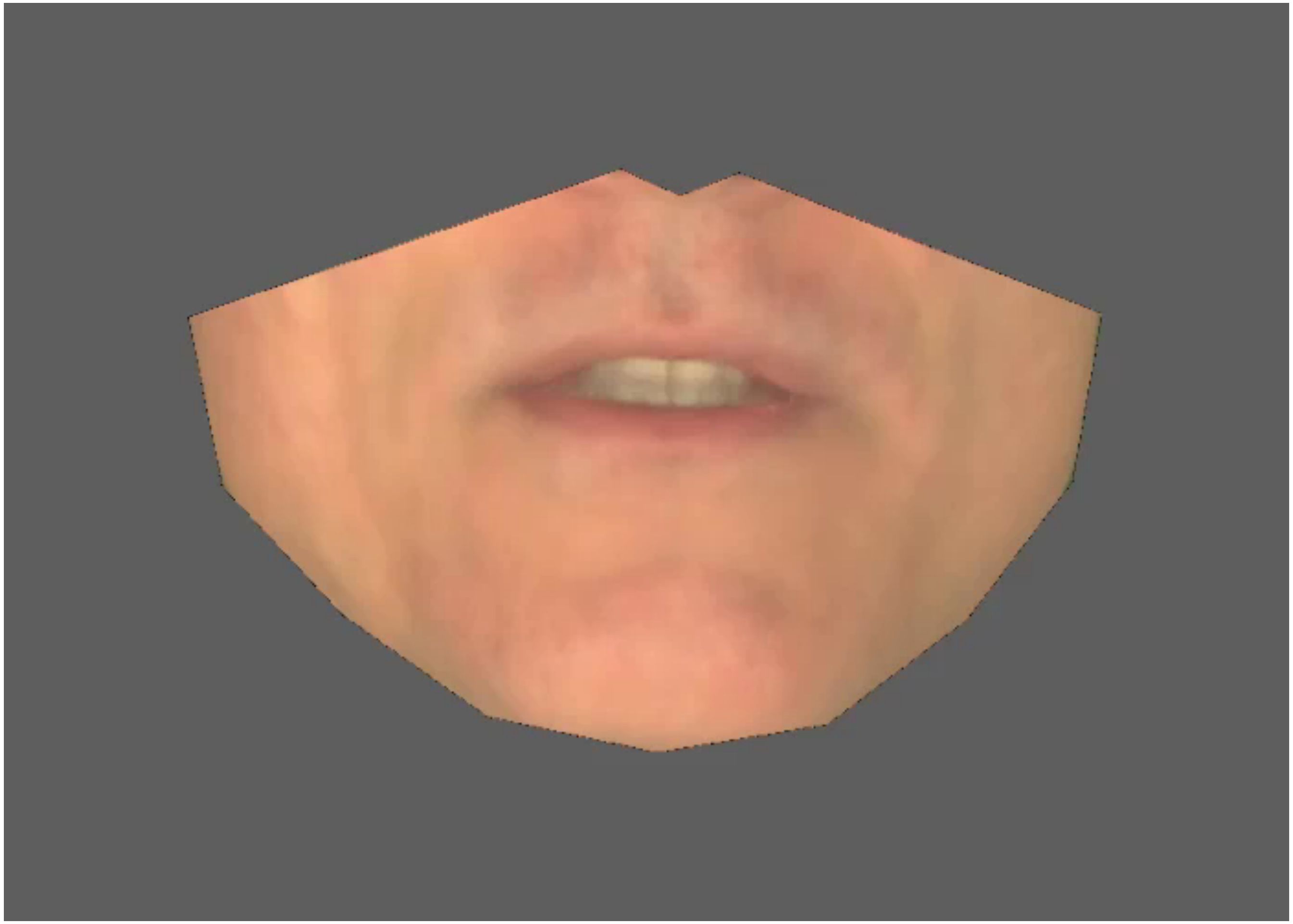
Finds linear transform (requires reference pose)

Manual:

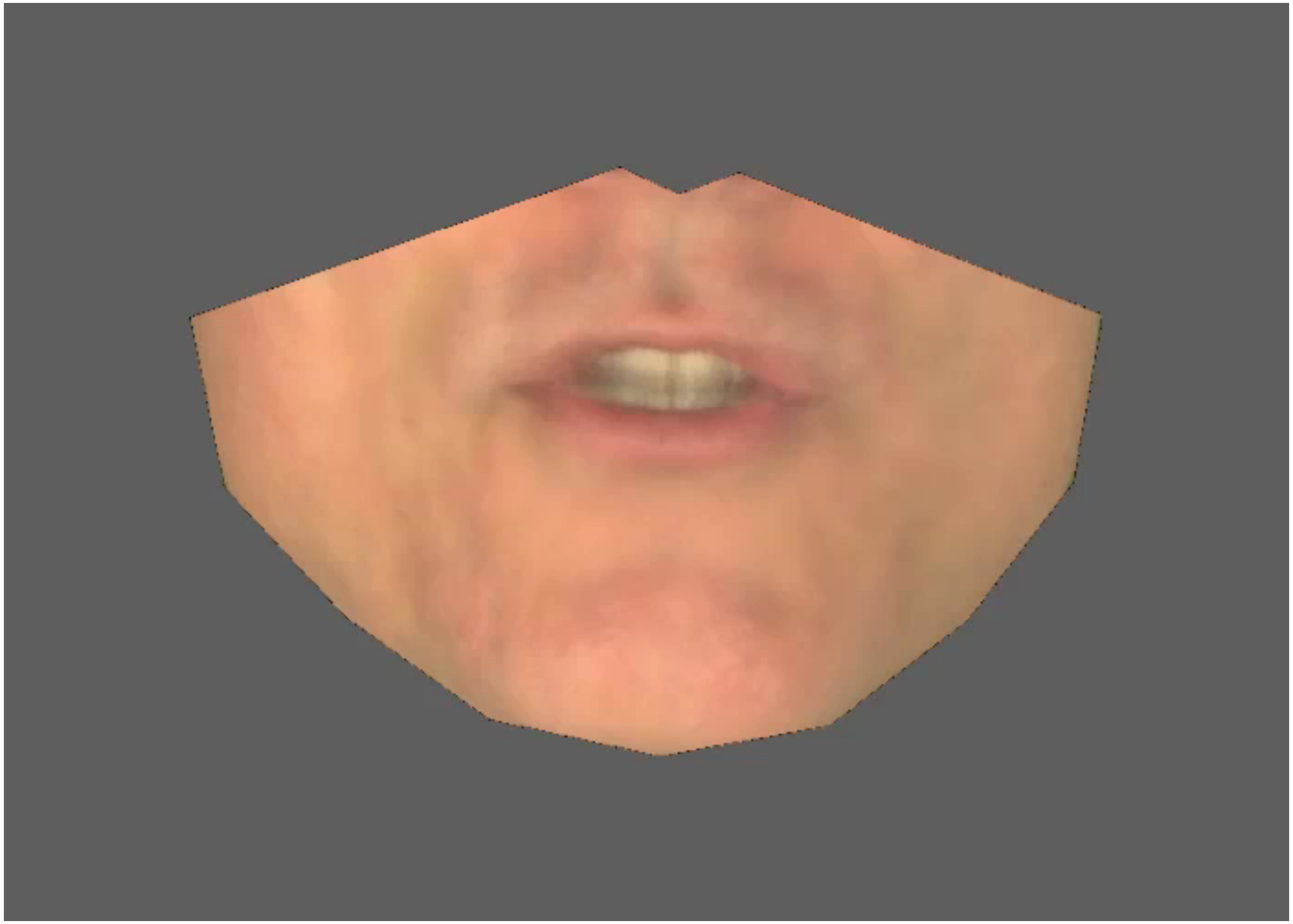
Pose basis shapes & linear blending



Prediction for Very Different Language



Prediction for Very Different Language



Overview of Learning Reductions

Motivation

- Know how to solve "standard" ML problems
  - Classification, regression, etc.
  - SVMs, logistic regression, decision trees, neural nets, etc.

Many toolkits available!

- "Reduce" complex problems to simple ones?
  - Variable-length trajectories ➔ multivariate regression

Still non-trivial!

- Similar to other reduction problems
  - E.g., NP-complete reductions
  - Some learning reductions have provable guarantees

Other Learning Reductions

- **Multiclass → Binary**
- Cost-weighted → Unweighted
- Ranking → Binary
- Sequential → Multiclass
- And many more…

http://hunch.net/~jl/projects/reductions/reductions.html

Why Multiclass → Binary?

- Conventional approach: one-versus-all
  - Scoring function per class
  - Predict class with highest score

- Limitations:
  - Linear in #classes
  - Hard to prove generalization bounds
  - (Binary SVM analyzes generalization via margin)

Learning Reduction Recipe

- Given original training set: $S = \left\{ (x_i, y_i) \right\}_{i=1}^{N}$

Multiclass

- Create modified training set(s):

$$\left\{ \hat{S} = \left\{ (x_i, \hat{y}_i) \right\}_{i=1}^{N} \right\}$$

Binary

  – Train ĥ's on Ŝ's

- Final h = combining predictions ĥ's

Two Flavors of Analysis

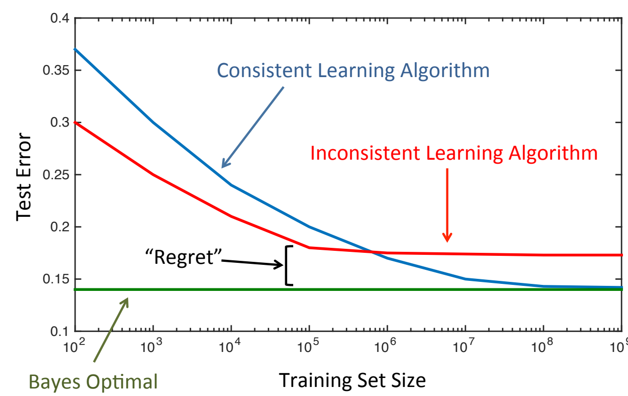
- Error Reduction:
  - Each ĥ achieves 0/1 Loss ε
  - Implication for multiclass 0/1 loss of h?
    - **Answer:** (K-1)ε

$$\varepsilon = L_P(w)$$

Zero 0/1 Test Error typically not possible

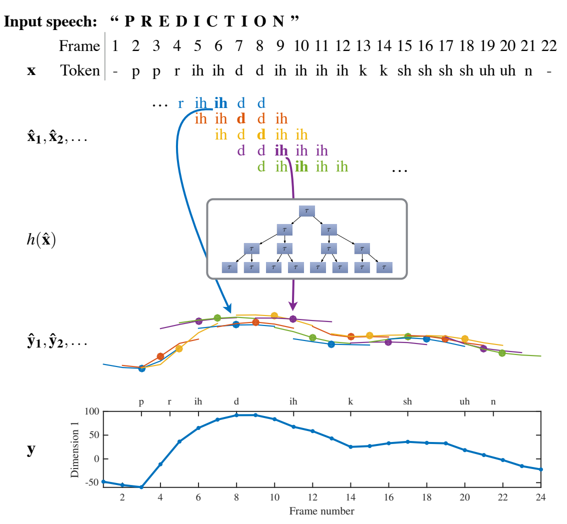
- Regret Reduction:
  - Each ĥ achieves 0/1 regret r
  - Implication of multiclass regret?
    - E.g., Kr?
  - More powerful result

$$r = L_P(w) - L_P(w^*)$$

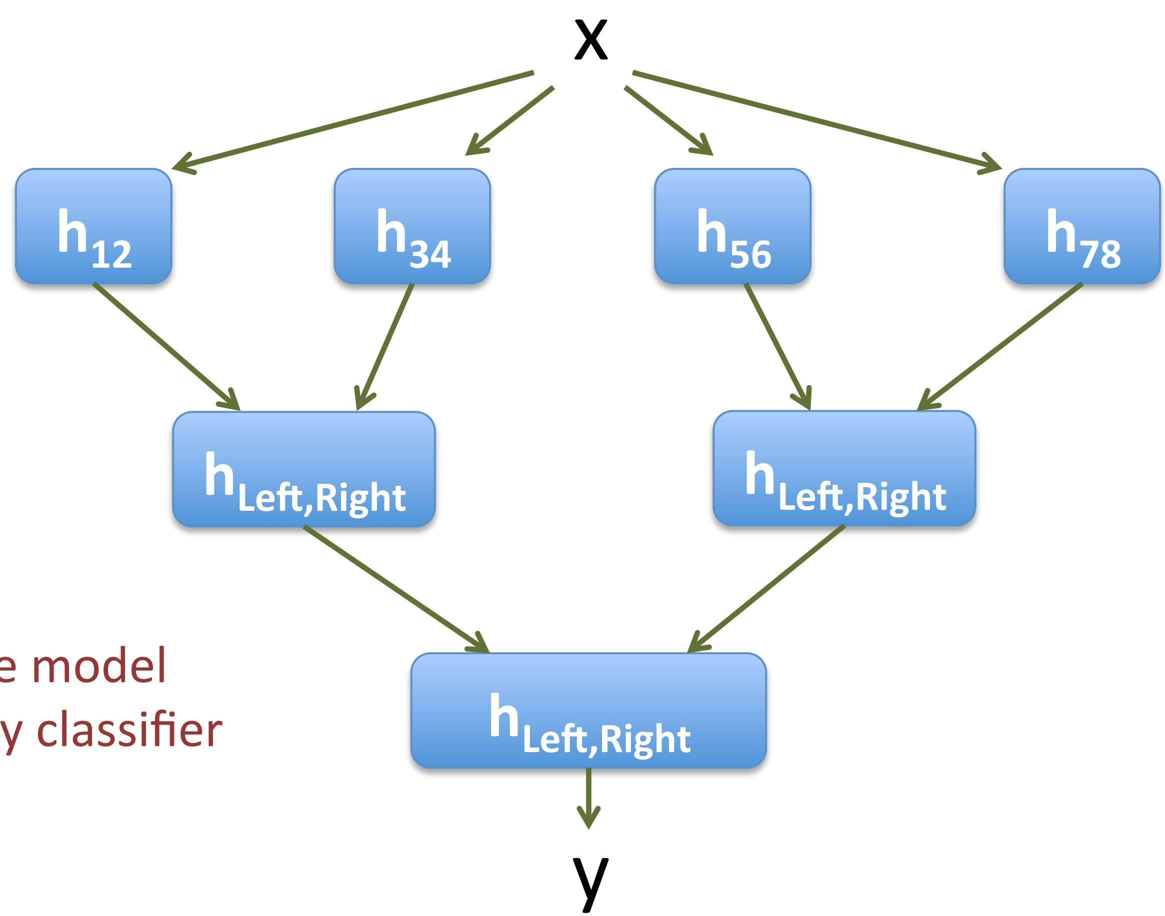


Aside: Sliding Window Regression

- If base model ĥ has 0 error
  - Then sliding window prediction has 0 error
- What about when ĥ has >0 error?
  - As regret of ĥ decreases…
  - … decrease in regret of h?
  - **Open question!**
    - Need to formalize lack of global dependencies



Filter Tree for Multiclass → Binary



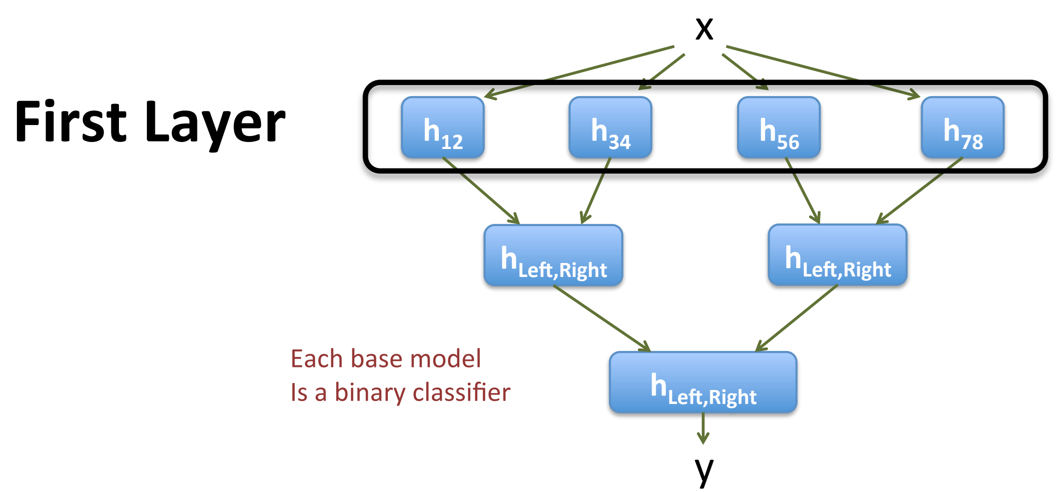
Each base model
Is a binary classifier

http://mi.eng.cam.ac.uk/~mjfg/local/Projects/filter_tree.pdf

The Learning Reduction

- First Layer
  - Train each h_{ij} using

$$S_{ij} = \left\{ (x, 1_{[y=i]}) \middle| \forall (x,y) \in S : y \in \{i,j\} \right\}$$

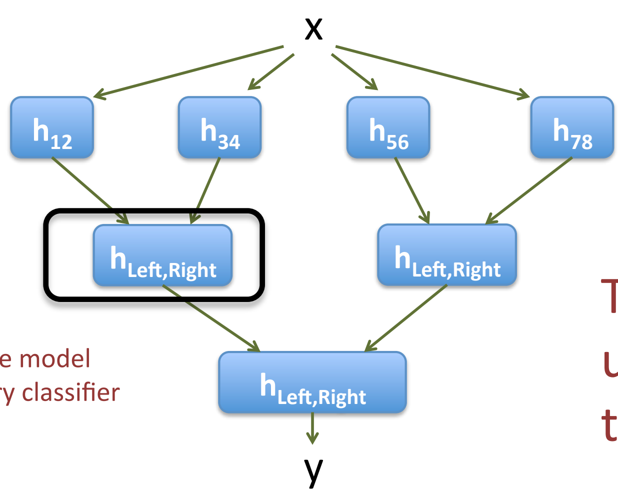


The Learning Reduction

- Second Layer
  - Train $h_{Left,Right}$ using

$$S_{Left,Right} = \left\{ (x, 1_{[y \in \{L,R\}]}) \middle| \forall (x,y) \in S : y \in \{1,\ldots,4\} \wedge \left(\text{no mistake by } h_{12}, h_{34}\right) \right\}$$

Second Layer

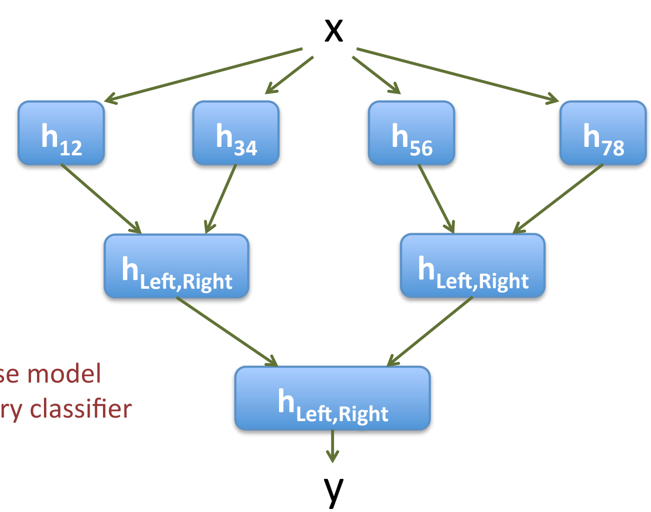


Each base model
Is a binary classifier

Train Lower Layers only using mistake-free training data.

The Learning Reduction

- Classification problem dependent on classifiers learned in previous layers
- Reduction happens iteratively
  - I.e., adaptively



Each base model
Is a binary classifier

Recall: Two Flavors of Analysis

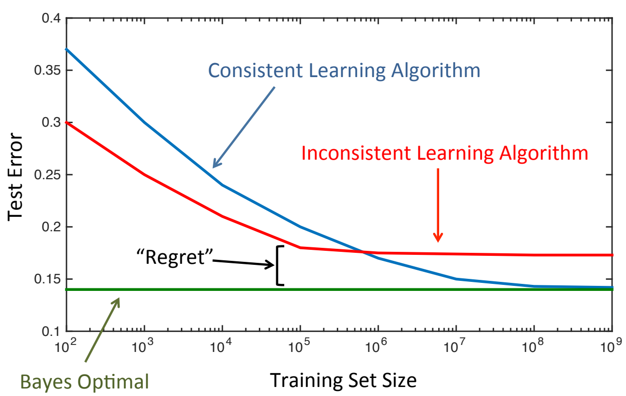
- Error Reduction:
  - Each ĥ achieves 0/1 Loss ε
  - Implication for multiclass 0/1 loss of h?
    - **Answer:** (K-1)ε

$$\varepsilon = L_P(w)$$

Zero 0/1 Test Error typically not possible

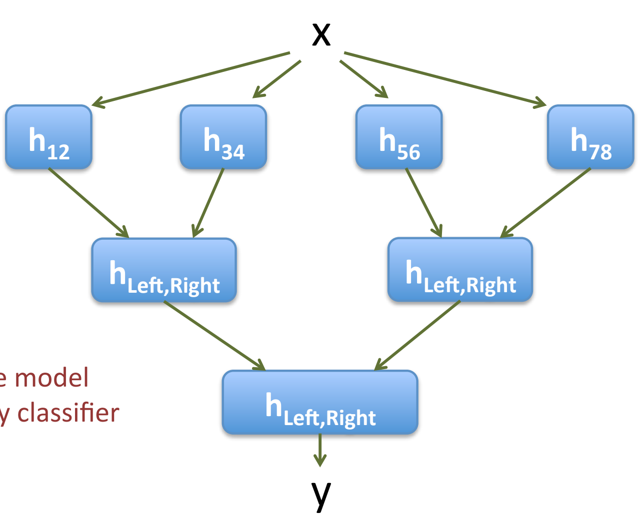
- Regret Reduction:
  - Each ĥ achieves 0/1 regret r
  - Implication of multiclass regret?
    - E.g., Kr?
  - More powerful result

$$r = L_P(w) - L_P(w^*)$$



Filter Tree Regret Guarantee

- If each classifier has regret r
- Filter Tree has multiclass regret ≤ $(\log_2 K)r$
  - Good dependence on K
- Inductive proof
- See details in paper

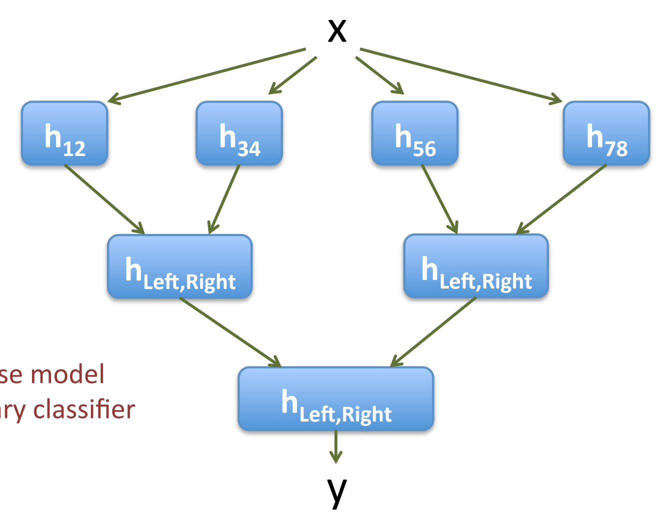


Each base model
Is a binary classifier

http://mi.eng.cam.ac.uk/~mjfg/local/Projects/filter_tree.pdf

Runtime Computational Benefits

- Logarithmic test time
  - With respect to #classes

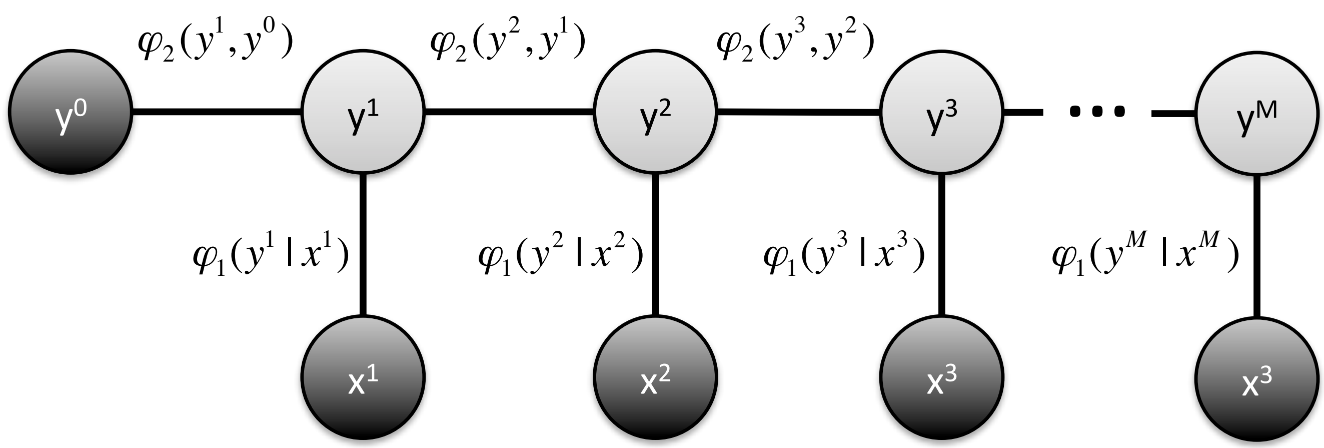


Each base model
Is a binary classifier

See also: **Logarithmic Time Online Multiclass Prediction**
http://arxiv.org/abs/1406.1822

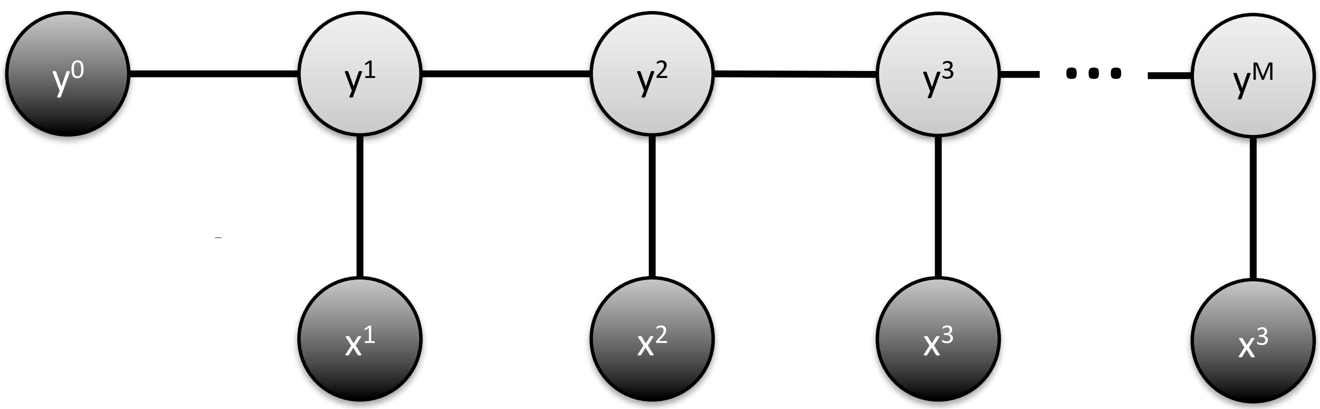
Very Briefly: Sequential → Multiclass

- Suppose we want to use decision trees for first-order sequence prediction



Recurrent Multiclass Classifier

- $h(x,y_{prev})$
  - Takes in current x, previous y
  - Predicts next y



http://www.umiacs.umd.edu/~hal/searn/
http://arxiv.org/abs/1011.0686

Next Week

- No Lecture Thursday
  - Student Faculty Conference
- Recitation Thursday
  - Conditional Random Fields Review
- Kaggle Miniproject Writeup due Thursday
  - Via Moodle
- Next Week:
  - Unsupervised, Clustering, Dim. Reduction