Roles of High-Fidelity Acoustic Modeling in Robust ASR

Li Deng

Microsoft Research, Redmond, USA

presented at 2007 IEEE ASRU Workshop, Kyoto, Japan

Outline

- Introduction: Issues in acoustic modeling & robust ASR
- Nature of speech variability & need for high-fidelity models
- A multi-layer model that captures variability
- Variability: acoustic environment
- Variability: speaking behavior
- Conclusions and future directions

(thanks to discussions and collaborative work with H. Ney, C. Lee, A. Acero, D. Yu, J. Li, J. Droppo & other colleagues at MSR)

Introduction

- **Issues in acoustic modeling**
  - **Probabilistic models (& Features)** that embed (imperfect) knowledge (Rabiner/Juang93; Acero93;Ostendorf et.al.96; Bilmes2005; Deng et.al.2006, etc.)
  - Performance Measure (Chou/Juang2003; Povey2004;McDermott et.al.2007)
  - Training's Objective Function & its optimization
    (Ney2006; Schluter et.al 2001; Liao&Gales2007; He&Deng&Chou,2008)
  - Decision Rule & optimization algorithm
    (Goel&Byrne2000; Lee&Huo2000;Ney2006)
- **Models (this talk's focus):**
  - Specify statistical dependency between input (observation) and output (speech class)
  - Can be generative or discriminative
  - Enable all three other ingredients
  - Most difficult ingredient
  - Two case studies: phase sensitive model; articulatory-like constraint
  - Warrant scientific pursuit (nature of speech variability)

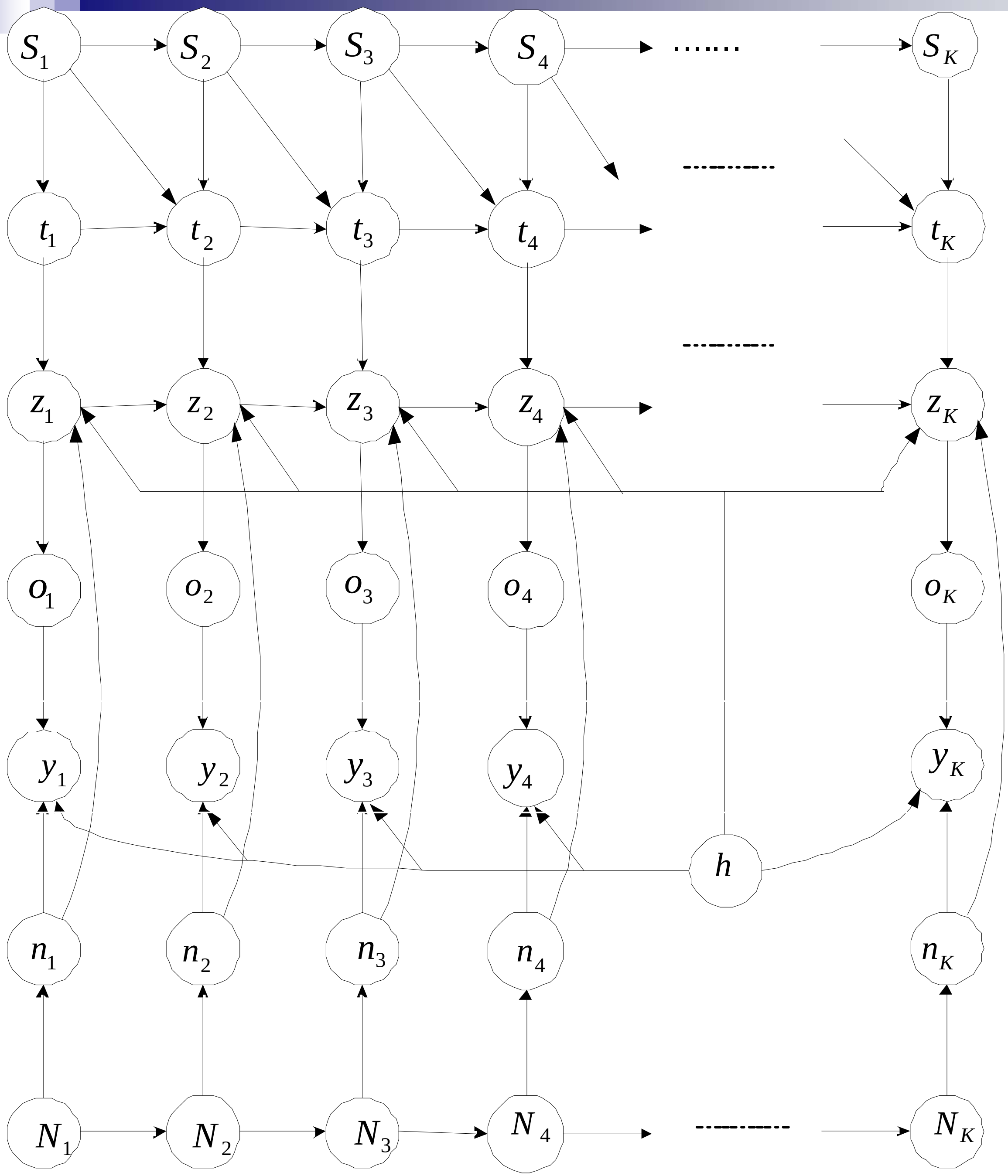
Nature of Speech Variability

- Multiple, interacting sources
  - Pronunciation (phonological & articulatory causes) (Nock&Young,2000)
  - Accent & dialect
  - Prosodic & phonetic contexts
  - **Speaking behavior (rate, style, etc.)** (Pitermann 2000; Deng 2006)
  - **Noisy acoustic environment** (Acero93;Moreno96;Lee98;Zhu&Alwan02;Gong05;Deng&Droppo&Acero04)
  - **Transducer & transmission-channel distortion**
  - Adverse environment that affects articulation

    (Junqua 2000; Hansen 2003)
- To effectively represent these variability sources for robust ASR requires **"high-fidelity"** acoustic models
- →Use of a richer set of knowledge in constructing probabilistic models of the speech process

A General Modeling Framework

- Probabilistic generative model
- Multiple layers, each representing one major cause of speech variability
- Joint distribution among all causes and their relationship
- Multi-layer dynamic Bayesian network



Phonological states (pronunciation)

Phonetic correlates (spatial targets of hidden dynamics)

Articulatory-like hidden dynamics

Dynamics of undistorted speech features (hidden)

Dynamics of environment-distorted speech (observed)

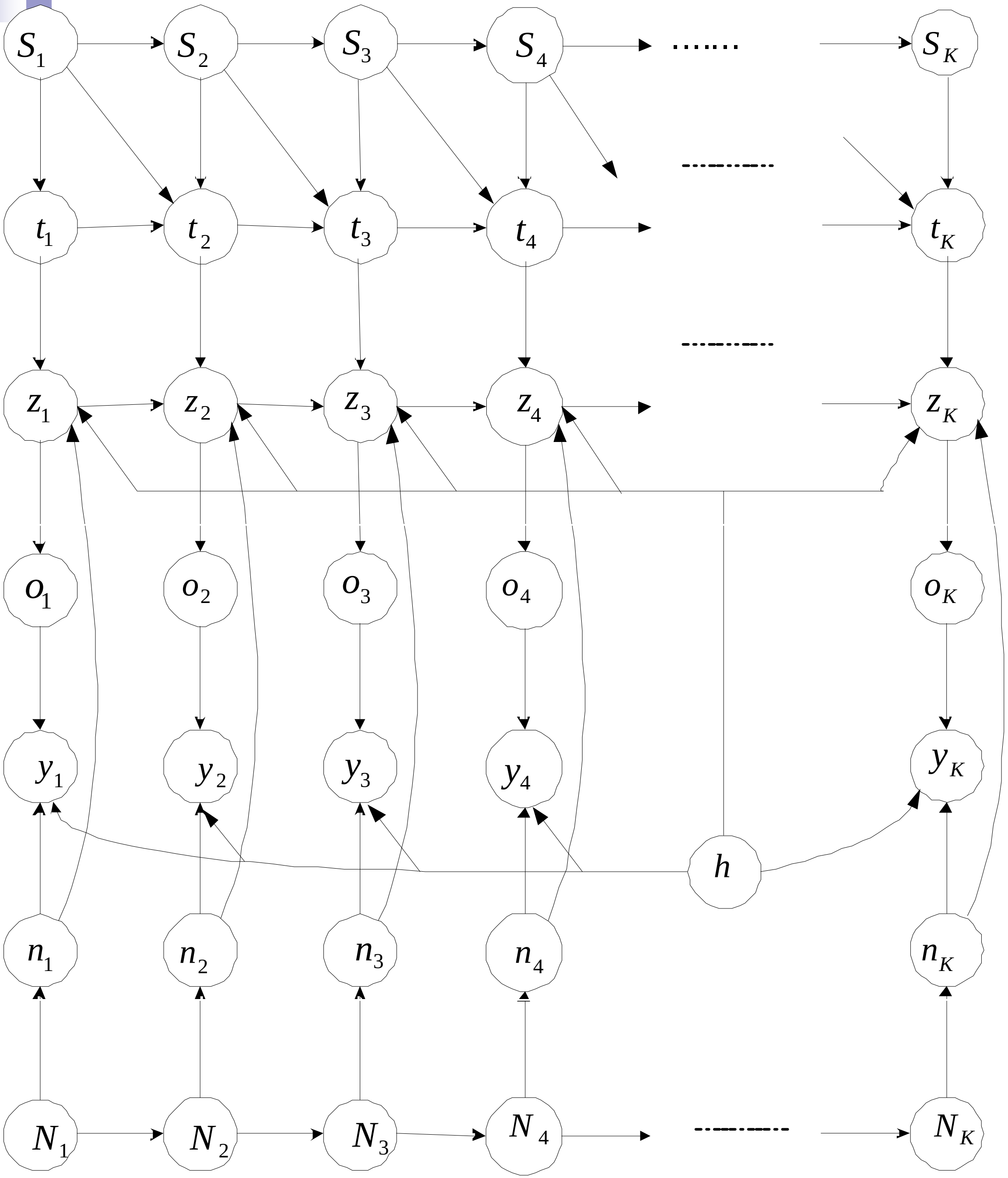
Nonstationary environmental noise

Discrete states of environmental noise

Two Case Studies

- Generative acoustic modeling for robust ASR that accounts for variability due to
  - **Adverse acoustic environment**
    - Sensitivity of cepstra to random phase between speech and mixing noise
  - **Speaking behavior**
    - Interaction between phonetic context and speaking rate/style

Case Study One: Acoustic environment



Dynamics of environment-distorted speech (observed)

Nonstationary environmental noise

Specifying Conditional Dependency in Bayes Net

--- A Phase-Sensitive Model

- Clean-speech=x; noise=n; channel=h; noisy-speech=y
- relationship in waveform-sample and DFT:

$$y[t] = x[t] * h[t] + n[t],$$

$$Y[k] = X[k]H[k] + N[k],$$

Instantaneous mixing phase

Relationship in power-spectrum:

$$|Y[k]|^2 = |X[k]|^2|H[k]|^2 + |N[k]|^2 + 2|X[k]H[k]||N[k]|cos\theta_k,$$

- **The last term was usually assumed zero (phase-insensitive), which is correct only in expected sense**

Phase-Sensitive Model (cont'd)

- relationship in Mel-filter power spectrum:

$$\sum_k W_k^{(l)}|Y[k]|^2 = \sum_k W_k^{(l)}|X[k]|^2|H[k]|^2 + \sum_k W_k^{(l)}|N[k]|^2 + 2\sum_k W_k^{(l)}|X[k]H[k]||N[k]|cos\theta_k,$$

or $$|\tilde{Y}^{(l)}|^2 = |\tilde{X}^{(l)}|^2|\tilde{H}^{(l)}|^2 + |\tilde{N}^{(l)}|^2 + 2\alpha^{(l)}|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}|,$$

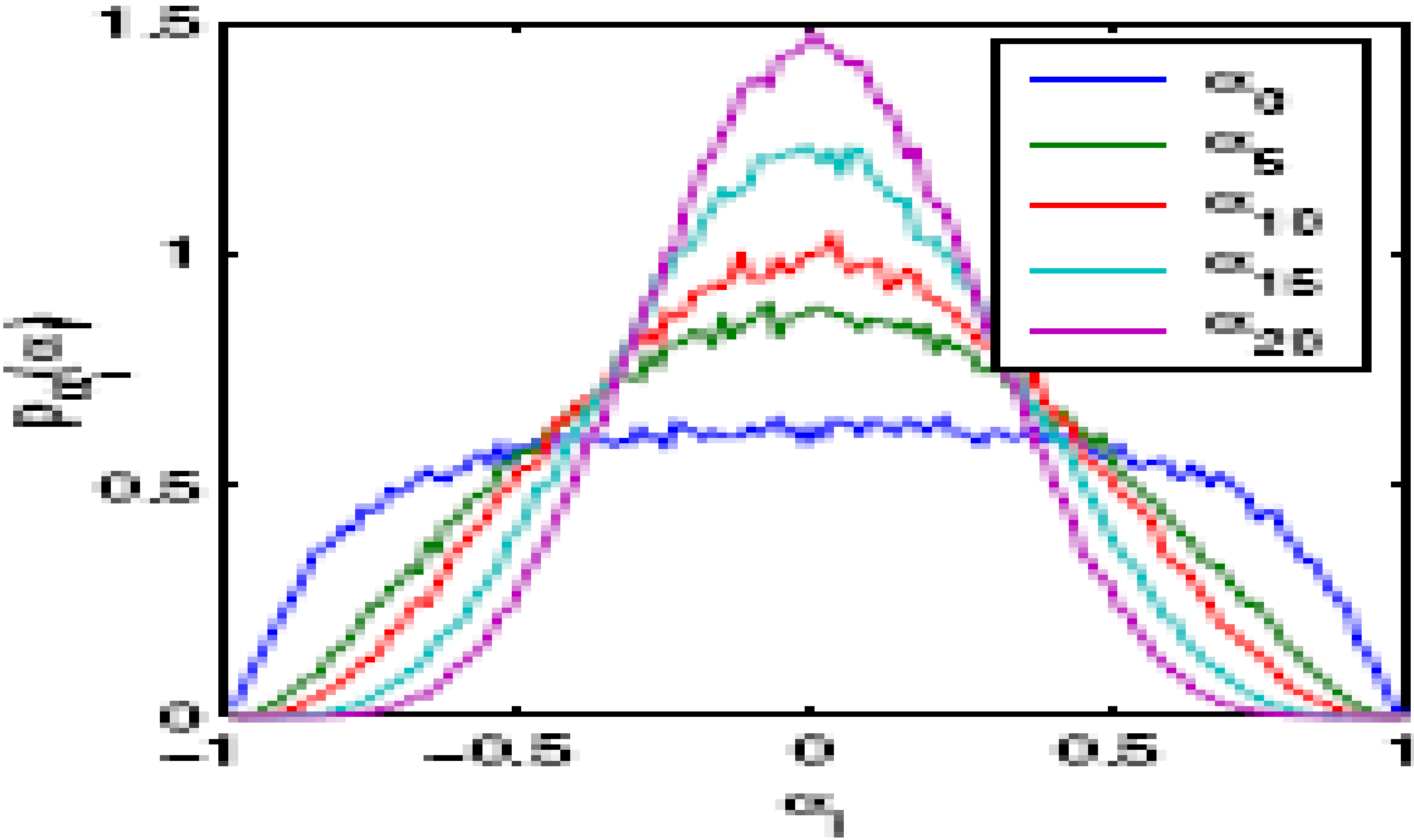
$$\alpha^{(l)} \equiv \frac{\sum_k W_k^{(l)}|X[k]H[k]||N[k]|cos\theta_k}{|\tilde{X}^{(l)}||\tilde{H}^{(l)}||\tilde{N}^{(l)}|}.$$

Distribution of Phase Factor

(Droppo, Acero, Deng, 2002)

-- Sum of many uniformly distributed random variables (filter banks)

-- Central limit theorem at work



Phase-Sensitive Model (cont'd)

- relationship in log-power-spectrum:

  Define log-power-spectrum vectors:

$$\mathbf{y} = \begin{bmatrix} \log|\tilde{Y}^{(1)}|^2 \\ \log|\tilde{Y}^{(2)}|^2 \\ \cdots \\ \log|\tilde{Y}^{(l)}|^2 \\ \cdots \\ \log|\tilde{Y}^{(L)}|^2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} \log|\tilde{X}^{(1)}|^2 \\ \log|\tilde{X}^{(2)}|^2 \\ \cdots \\ \log|\tilde{X}^{(l)}|^2 \\ \cdots \\ \log|\tilde{X}^{(L)}|^2 \end{bmatrix}, \quad \mathbf{n} = \begin{bmatrix} \log|\tilde{N}^{(1)}|^2 \\ \log|\tilde{N}^{(2)}|^2 \\ \cdots \\ \log|\tilde{N}^{(l)}|^2 \\ \cdots \\ \log|\tilde{N}^{(L)}|^2 \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} \log|\tilde{H}^{(1)}|^2 \\ \log|\tilde{H}^{(2)}|^2 \\ \cdots \\ \log|\tilde{H}^{(l)}|^2 \\ \cdots \\ \log|\tilde{H}^{(L)}|^2 \end{bmatrix},$$

then:

$$e^{\mathbf{y}} = e^{\mathbf{x}} \bullet e^{\mathbf{h}} + e^{\mathbf{n}} + 2\alpha \bullet e^{\mathbf{x}/2} \bullet e^{\mathbf{h}/2} \bullet e^{\mathbf{n}/2} = e^{\mathbf{x}+\mathbf{h}} + e^{\mathbf{n}} + 2\alpha \bullet e^{(\mathbf{x}+\mathbf{h}+\mathbf{n})/2},$$ **or**

$$\mathbf{y} = \log\left[e^{\mathbf{x}+\mathbf{h}} \bullet \left(1 + e^{\mathbf{n}-\mathbf{x}-\mathbf{h}} + 2\alpha \bullet e^{\frac{\mathbf{x}+\mathbf{h}+\mathbf{n}}{2}-\mathbf{x}-\mathbf{h}}\right)\right] = \mathbf{x} + \mathbf{h} + \log\left[1 + e^{\mathbf{n}-\mathbf{x}-\mathbf{h}} + 2\alpha \bullet e^{\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}}\right]$$

Phase-Sensitive Model (cont'd)

- Gaussian assumption for phase factor

$$p(\alpha^{(l)}) = \mathcal{N}(\alpha^{(l)}; 0, \Sigma_{\alpha}^{(l)}),$$

- Computing conditional prob.:

$$p_y(\mathbf{y}|\mathbf{x}, \mathbf{n}, \mathbf{h}) = |J_{\alpha}(\mathbf{y})| \; p_{\alpha}(\boldsymbol{\alpha}|\mathbf{x}, \mathbf{n}, \mathbf{h}),$$

- Jacobian computation:

$$\text{diag}\left(\frac{\partial \mathbf{y}}{\partial \boldsymbol{\alpha}}\right) = \frac{2e^{\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}}}{1 + e^{\mathbf{n}-\mathbf{x}-\mathbf{h}} + 2\boldsymbol{\alpha} \bullet e^{\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}}} = \frac{2e^{\frac{\mathbf{n}+\mathbf{x}+\mathbf{h}}{2}}}{e^{\mathbf{x}+\mathbf{h}} + e^{\mathbf{n}} + 2\boldsymbol{\alpha} \bullet e^{\frac{\mathbf{n}+\mathbf{x}+\mathbf{h}}{2}}} = 2\, e^{\frac{\mathbf{n}+\mathbf{x}+\mathbf{h}}{2} - \mathbf{y}}.$$

- Final result for conditional dependency:

$$p_y(\mathbf{y}|\mathbf{x}, \mathbf{n}, \mathbf{h}) = \frac{1}{2} \; \left| \text{diag}\left(e^{\mathbf{y} - \frac{\mathbf{n}+\mathbf{x}+\mathbf{h}}{2}}\right) \right| \mathcal{N}\left[\frac{1}{2}(e^{\mathbf{y} - \frac{\mathbf{n}+\mathbf{x}+\mathbf{h}}{2}} - e^{\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}} - e^{-\frac{\mathbf{n}-\mathbf{x}-\mathbf{h}}{2}}); \mathbf{0}, \boldsymbol{\Sigma}_{\alpha}\right].$$

Speech Enhancement as Bayes-Net Inference

- After specifying conditional dependency, carry out estimation and inference
- Inference on the clean-speech layer in the Bayes net → speech feature enhancement
- Results (iterative enhancement algorithm):

$$\hat{x} \approx \sum_{m=1}^{M} \gamma_m(x_0, \bar{n}) \left(x_0 - \frac{b_m^{(1)}(x_0, \bar{n})}{b_m^{(2)}(x_0, \bar{n})} \right)$$

(using 2[nd]-order Taylor series expansion)

Noisy Speech Recognition Experiments

(Deng, Droppo, Acero, 2004)

- Aurora 2 noisy speech data
- Using power of true noise (i.e., no est. error)
- Recognition accuracy (%) using enhanced features:

L	1	2	4	7	12
SetA	94.12	96.75	97.96	98.11	98.12
SetB	94.80	97.29	98.10	98.48	98.55
SetC	91.00	94.50	96.50	97.86	98.00
Ave.	93.77	96.52	97.72	98.21	98.27

- Best spectral subtraction (phase insensitive): **95.90%**
- Use of phase model reduces errors by half, if noise “estimate” is accurate

Experiments (cont'd)

Recognition Accuracy	Automatic noise est. algorithm	Assuming no noise est. errors
no phase info (low-fidelity)	**84.80%**	95.90%
phase info (high-fidelity)	**85.74%**	98.27%

--- Much lower relative error reduction when noise estimation errors are introduced

--- Why?

$$|Y[k]|^2 = |X[k]|^2|H[k]|^2 + \underline{|N[k]|^2} + \underline{2|X[k]H[k]||N[k]|cos\theta_k},$$

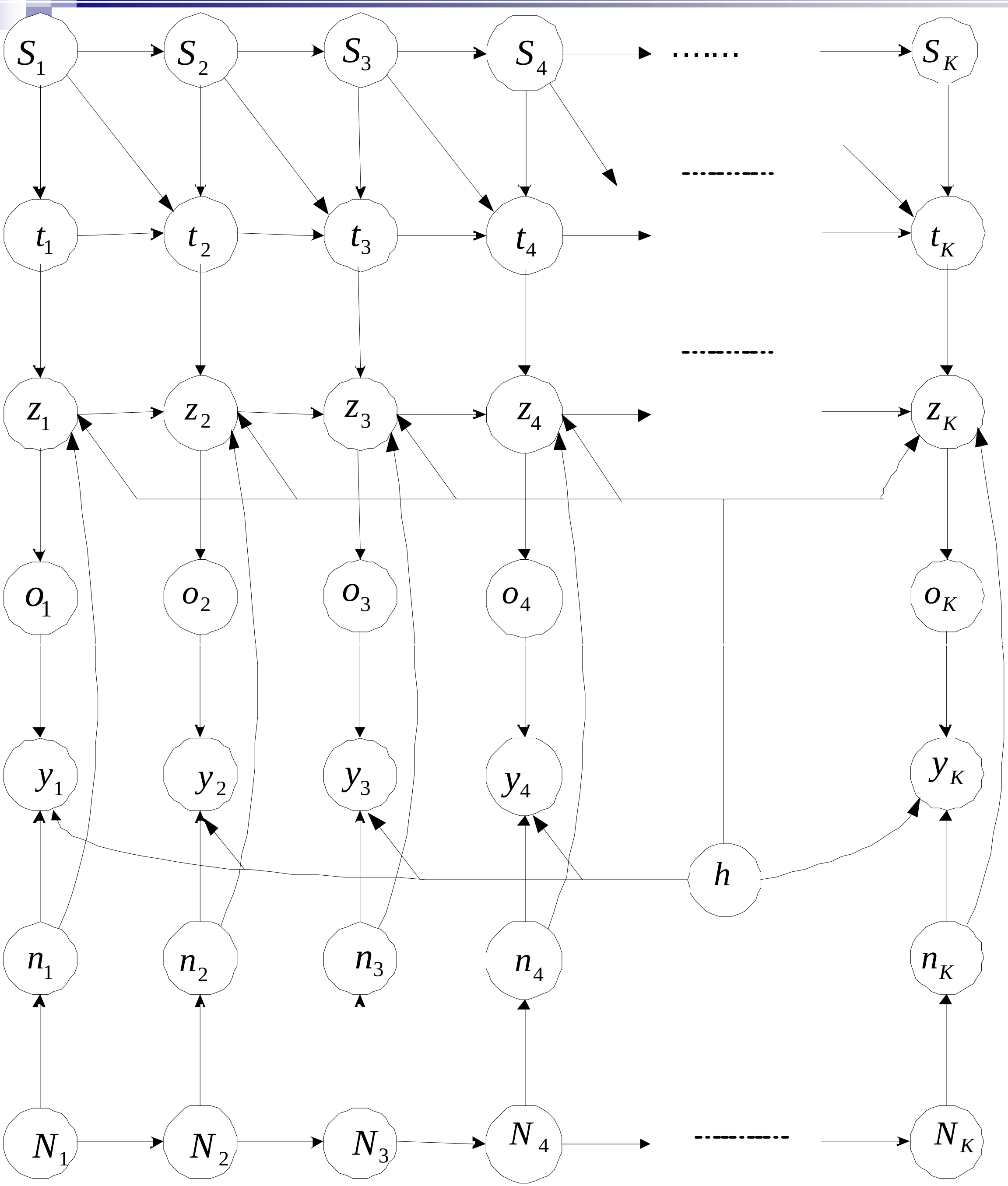
More Recent Experiments

Recognition Accuracy	Automatic noise Est. algorithm	Assuming no noise Est. errors	HMM Adapt (better noise est.)
no phase info (low-fidelity)	84.80%	95.90%	**91.70%** (poster today)
phase info (high-fidelity)	85.74%	98.27%	**93.32%** (ICASSP08 submitted)

$$|Y[k]|^2 = |X[k]|^2|H[k]|^2 + \underline{|N[k]|}^2 + \underline{2|X[k]H[k]||N[k]|cos\theta_k},$$

Case Study Two: speaking behavior

Case Study Two: speaking behavior

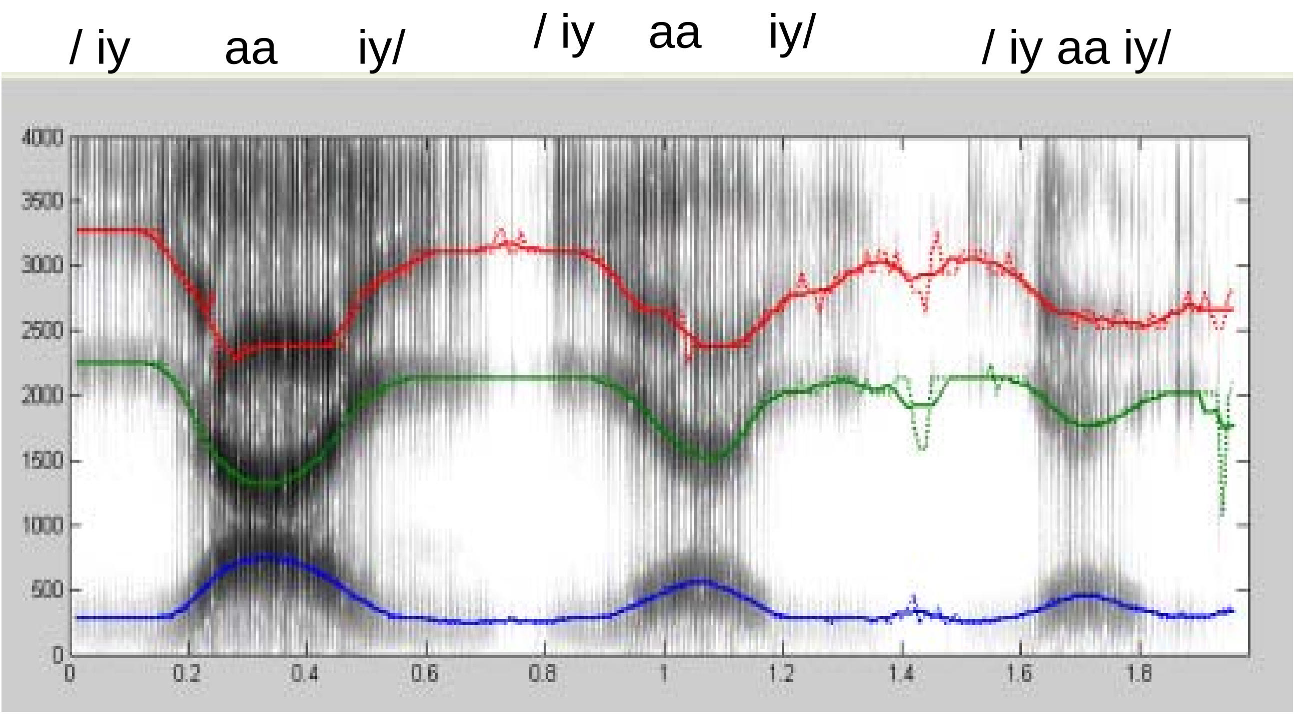


"articulatory" behavior

clean speech acoustic feature sequence (observed)

Temporal Dynamics in Speech: An Illustration

- Fundamental problem: Inherent "static" speech-class overlaps for natural–style speech
- Solution: Dynamic specification of speech



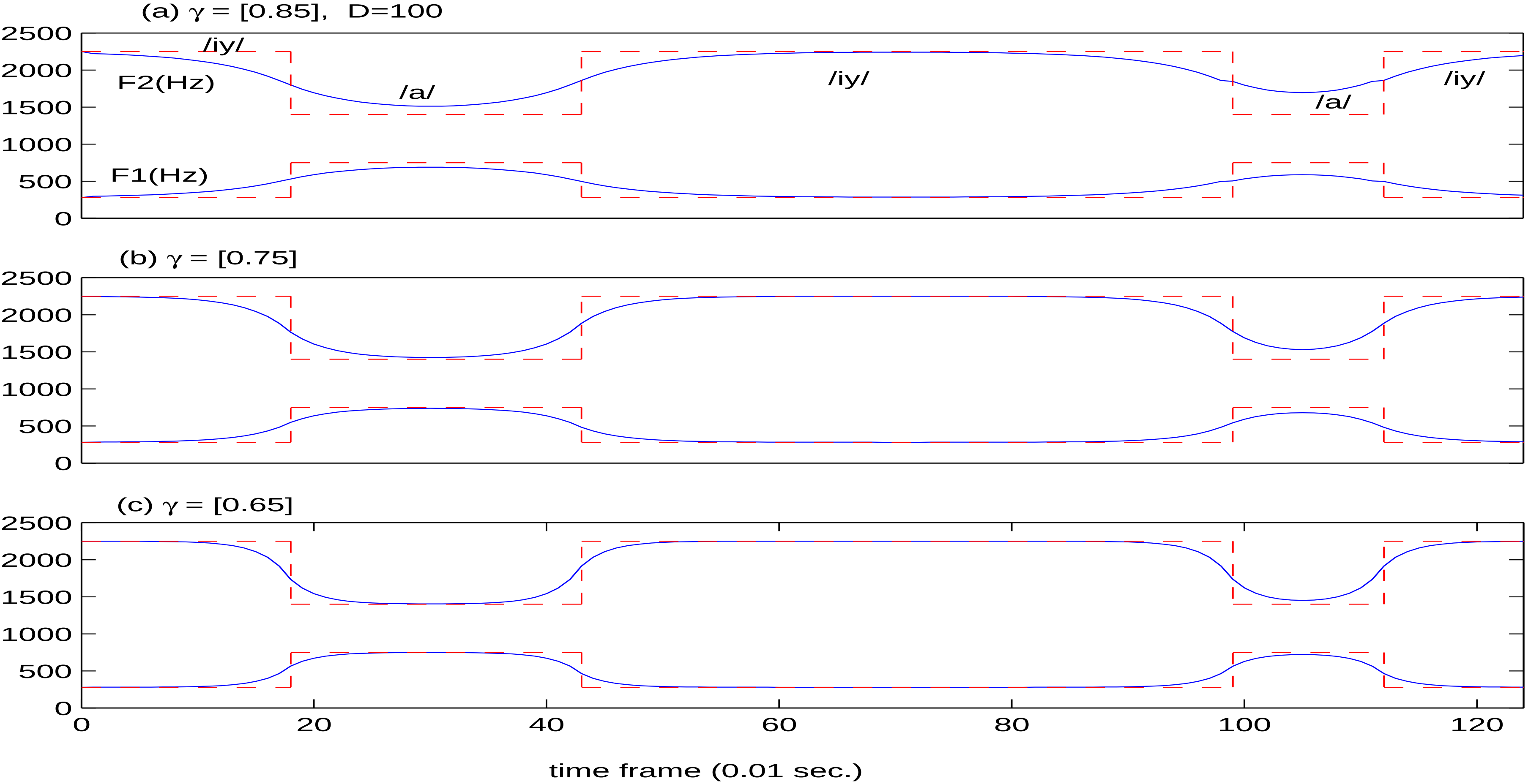
--Same speech content, with drastically different acoustic signatures

--Due partly to articulatory inertia

A Formant Trajectory Model

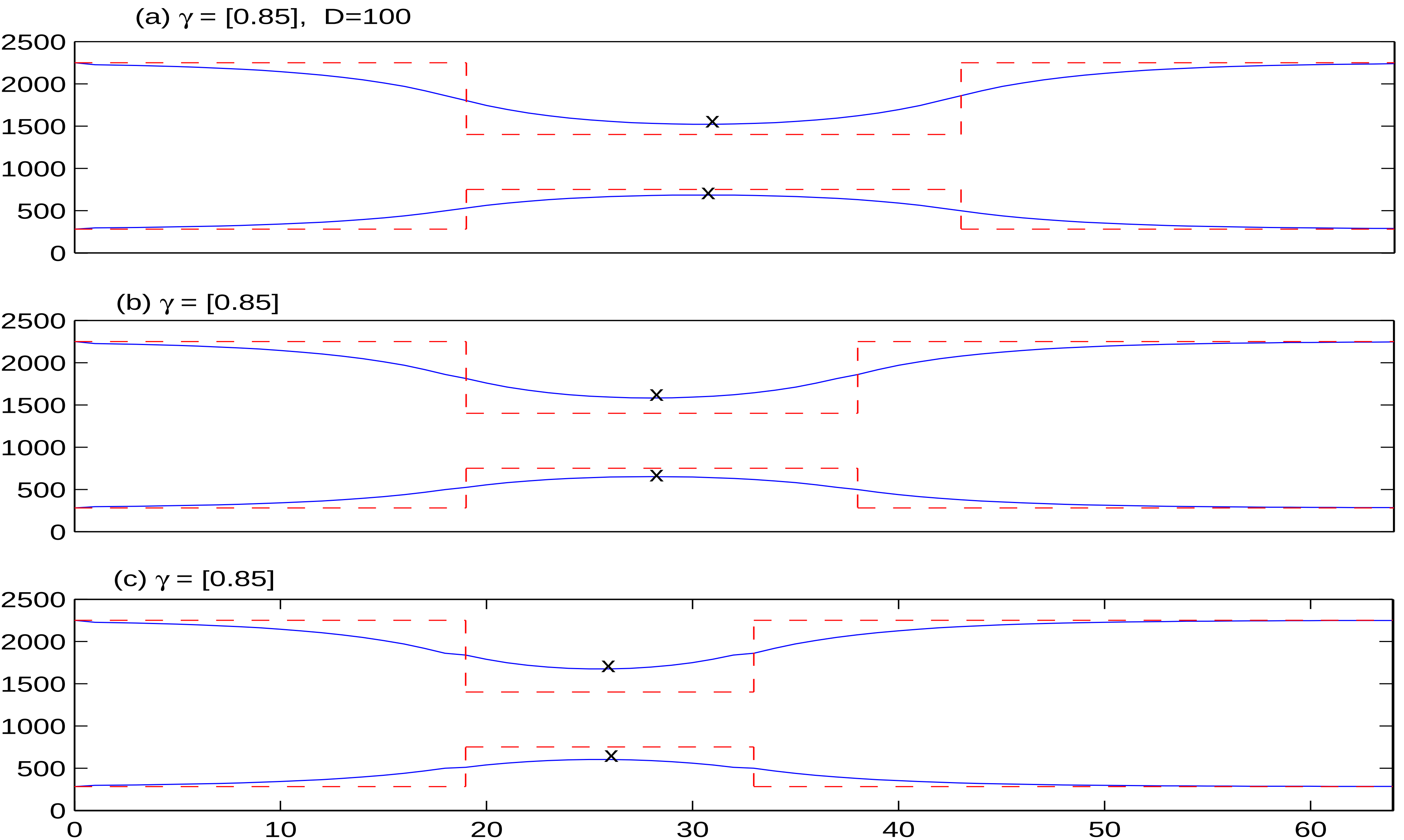
- Conditional dependency in the z-layer of the Bayes Net
- Input to “filter”: target sequence as step functions
- Output of “filter”: formant trajectories
- The output is a convolution between the target sequence and the impulse response of the “filter”

Model Prediction (effects of speaking "efforts")

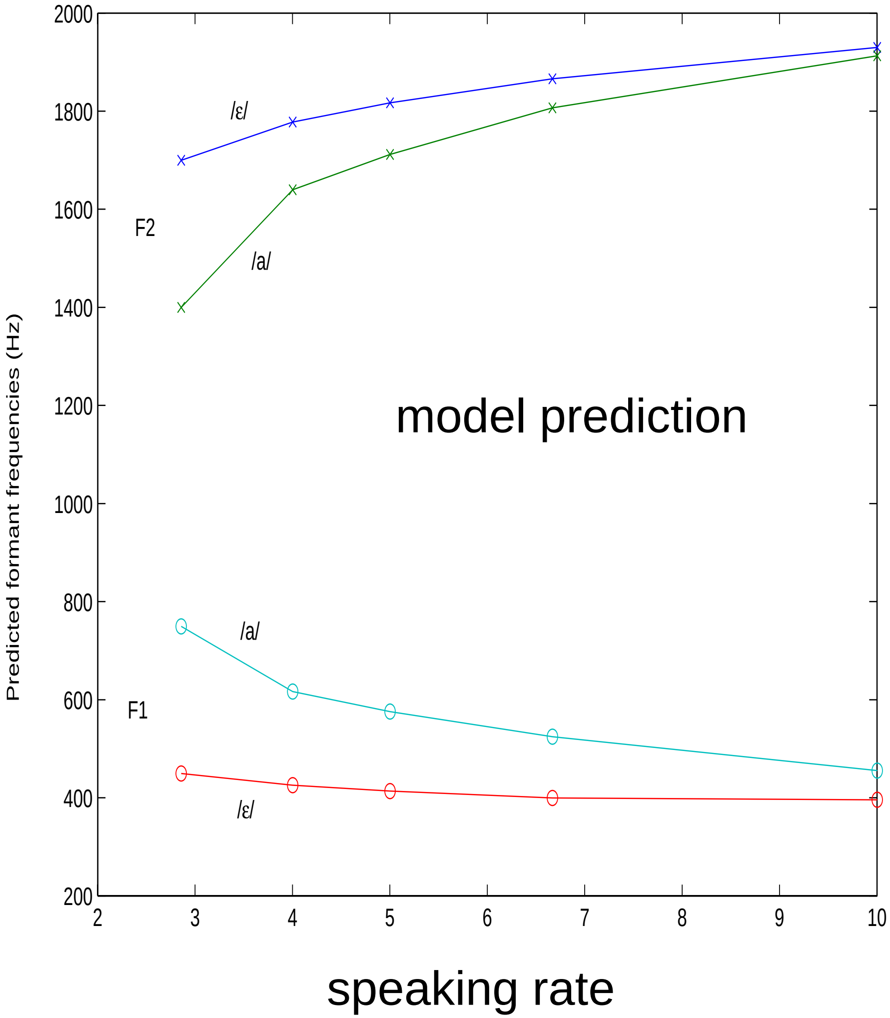


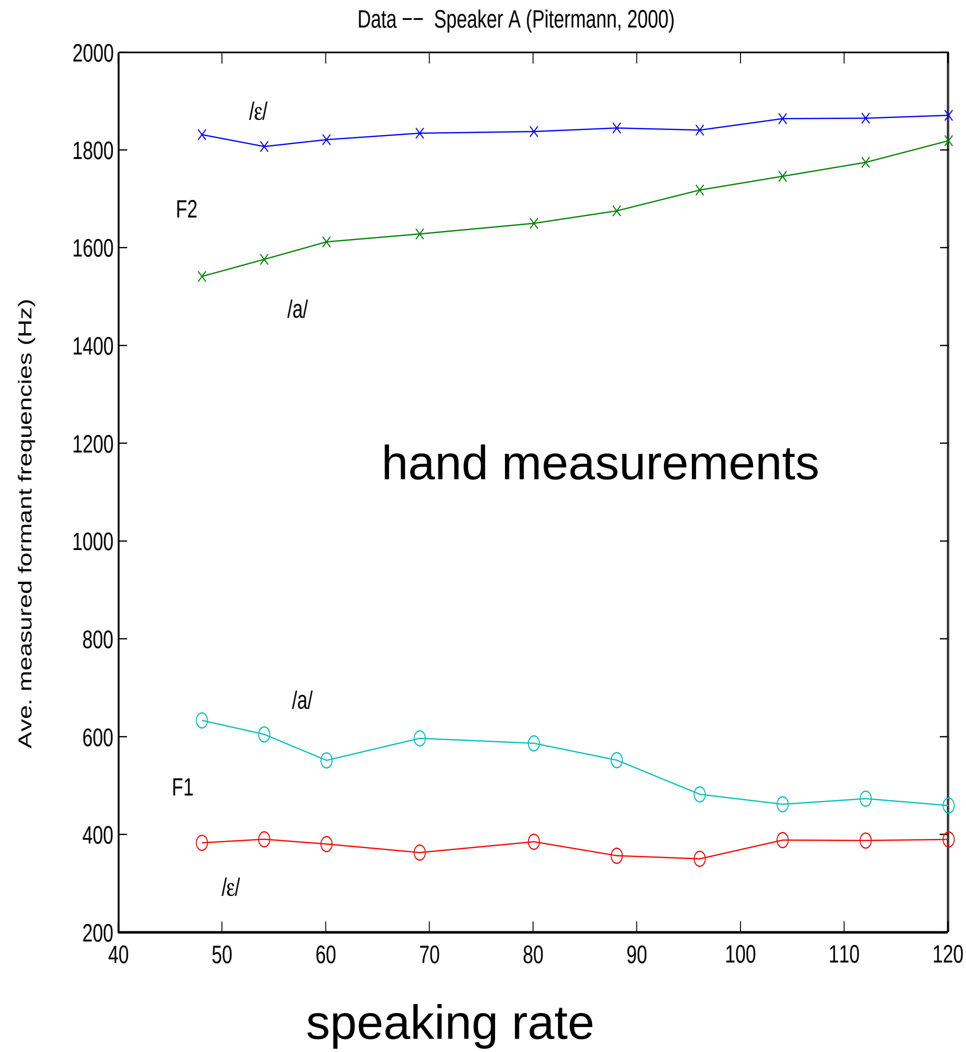
- The same speech content (/iai/) has different formant values
- Speaking effort/rate/style is a big factor
- The model predicts exactly the same kind of effects

Model Prediction (effects of speaking rate)



Sound Confusion for Casual Speech (model vs. data)



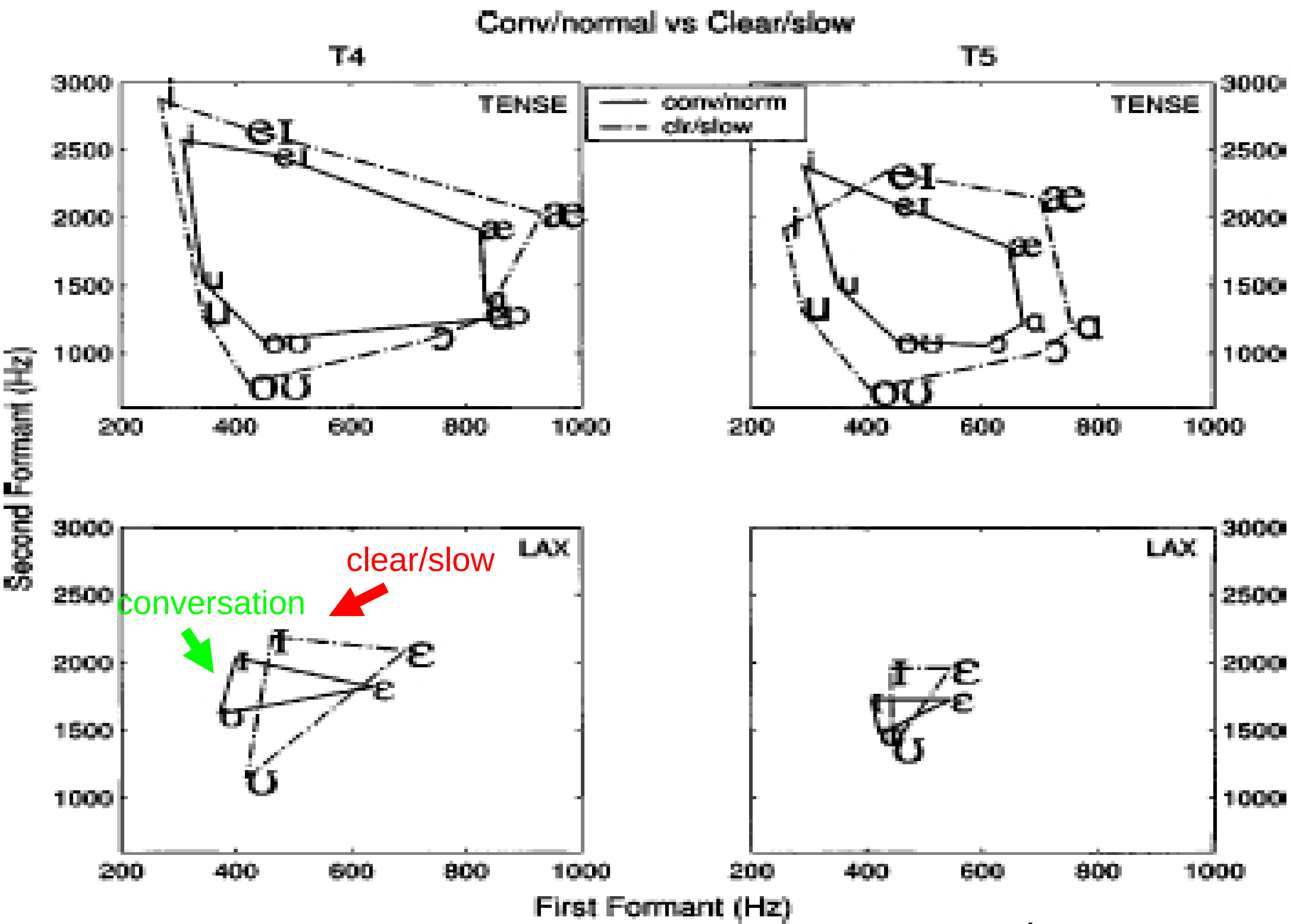


- Two sounds merge when they become “sloppy”
- Human perception does “extrapolation”; so does our model

- 5000 hand-labeled speech tokens
- Source: J. Acoustical Society of America, 2000

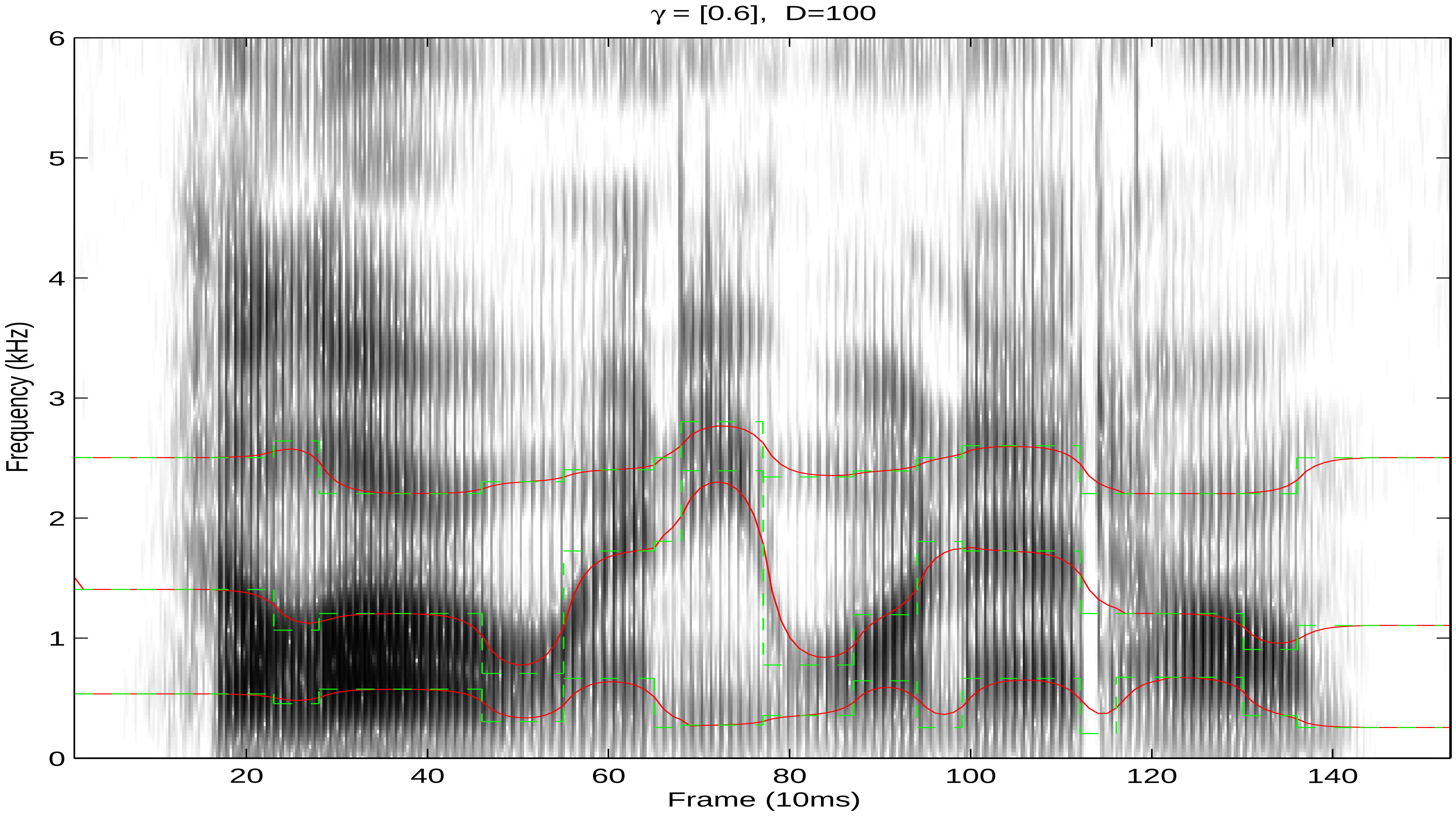
Discriminative-Space Reduction explained

--- consequence of speaking-behavior variability

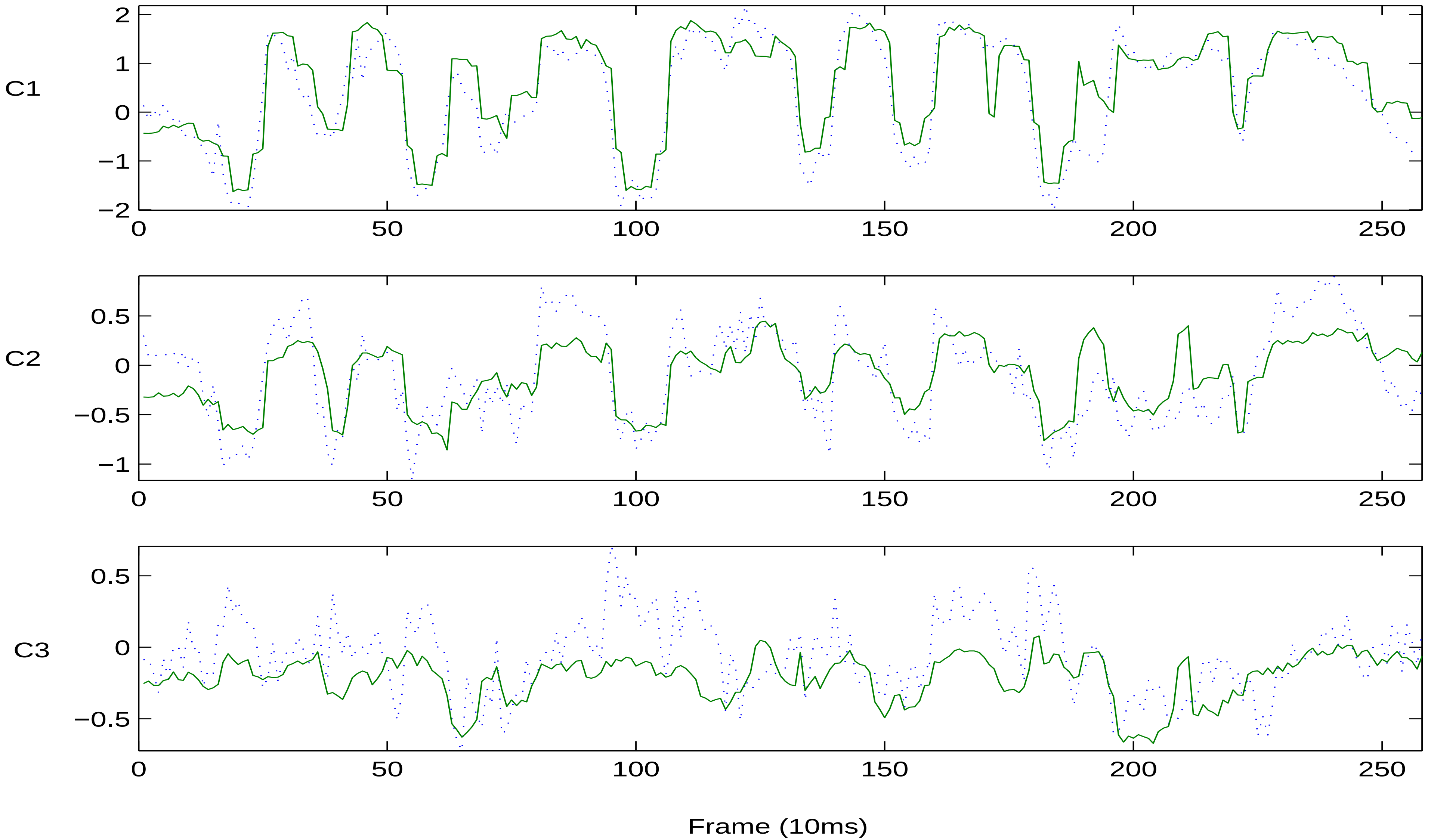


(Krause & Braida, JASA 2004)

Model Prediction of Formants (red)



Model Prediction of Cepstra (vs. data)



Experimental Results

(phonetic recognition in TIMIT core testset)

DENG *et al.*: STRUCTURED SPEECH MODELING (2006)

TABLE I

TIMIT PHONETIC RECOGNITION PERFORMANCE COMPARISONS BETWEEN AN HMM SYSTEM AND THREE VERSIONS OF THE HTM SYSTEM. HTM-1: N-BEST RESCORING WITH HTM SCORES ONLY; HTM-2: N-BEST RESCORING WITH WEIGHTED HTM, HMM, AND LM SCORES; HTM-3: LATTICE-CONSTRAINED A^* SEARCH WITH WEIGHTED HTM, HMM, AND LM SCORES. IDENTICAL ACOUSTIC FEATURES (FREQUENCY-WARPED LPCCs) ARE USED

		Corr %	Sub %	Del %	Ins %
HMM		73.64	17.14	9.22	2.21
HTM-1		77.76	16.23	6.01	3.45
HTM-2		77.73	15.61	6.65	3.14
HTM-3		78.28	15.94	5.78	3.20

Experimental Results
(phonetic recognition in TIMIT core testset)

TABLE II
COMPARISONS OF HMM AND HTM PERFORMANCES (PERCENT CORRECT) WITHIN EACH OF FOUR BROAD PHONE CLASSES

		Fricatives	Closures
Occurrences		1252	1578
HMM		75.64	88.72
HTM		75.74	90.94

Generative vs. Discriminative Models

- Modeling joint vs. conditional distributions
- For high-complexity tasks w/ many sources of variability (speech), generative approach more straightforward in conceptualization
- Longer history of research

  (e.g., HMM: Jelinek75; Baker75; CRF: Pereira 05)
- Easier to systematically embed knowledge
- Easier to diagnose recognizer errors
- Tend to be more complex
- Rely more on “physical modeling” instead of “feature engineering”
- Both approaches have merits

Summary

- Complex, multiple, interacting sources of speech variability →robustness in ASR
- → Need for "high-fidelity" acoustic modeling
- Rich sets of useful, albeit incomplete, knowledge
- What kind of knowledge?
  - Capture essence of speech variability
  - Be amenable to computation and automatic learning
- Example 1: phase-sensitive model of acoustic distortion
- Example 2: hidden dynamic model for variability in speaking behavior
- Both models specify conditional dependency in two separate layers in a Bayesian network

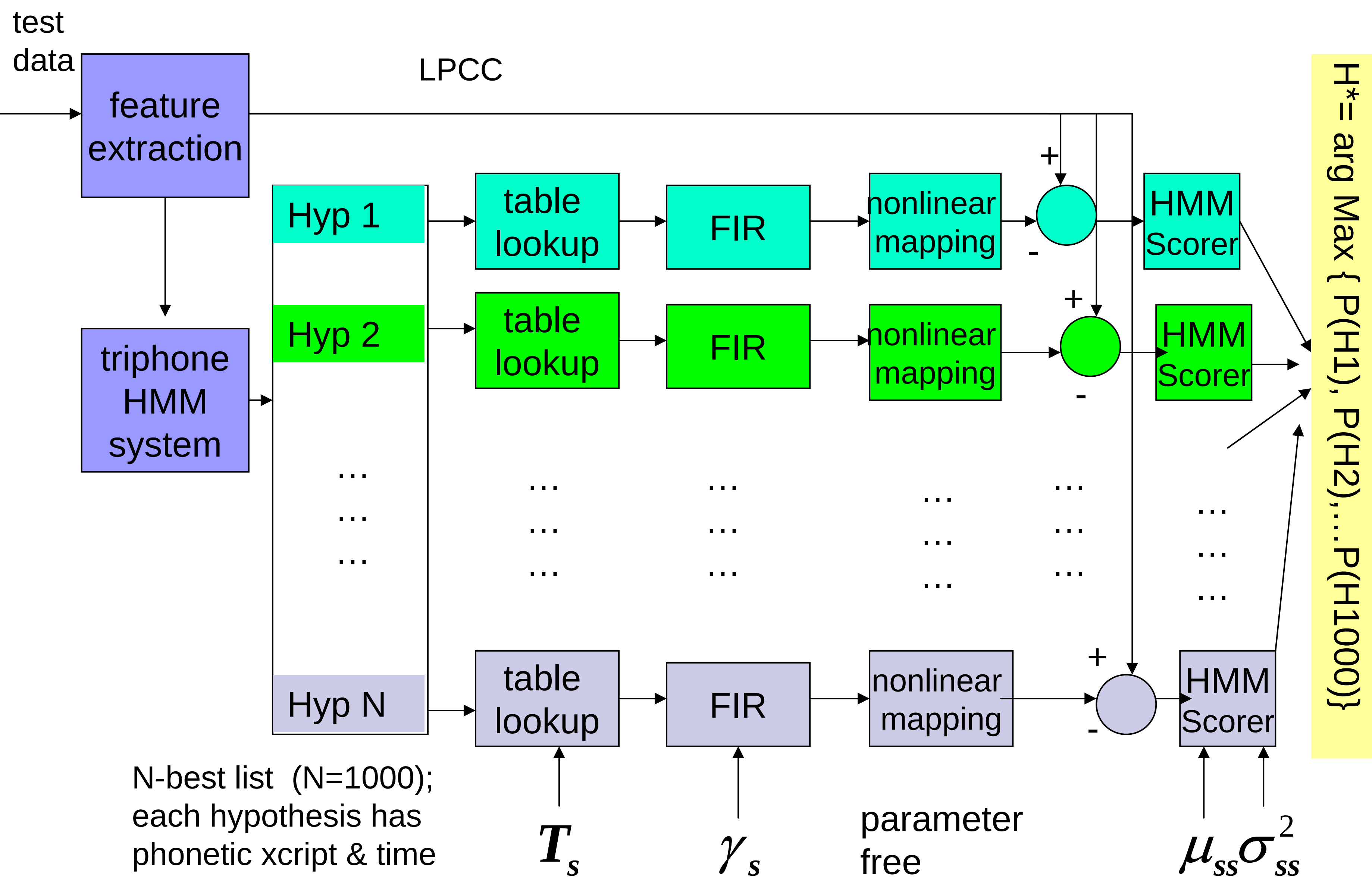
Future Directions

- **Recent NIST MINDS Report** (Baker, Deng, Khudanpur, Lee, Glass, Morgan, 2007)
- **Advanced acoustic models for "everyday audio"**
- **Adaptation and self learning**
- **Cognition-derived speech models**
- **Better use of human speech production & perception knowledge (e.g., masking & attention; discriminative features & learning, etc.)**
- **Require much higher "fidelity" in acoustic models than presented in this talk**

Thank you

Q/A

Procedure



test data
feature extraction
LPCC
triphone HMM system
Hyp 1
Hyp 2
Hyp N
table lookup
FIR
nonlinear mapping
HMM Scorer
+
-
H*= arg Max { P(H1), P(H2),…P(H1000)}
N-best list (N=1000); each hypothesis has phonetic xcript & time
T_s
γ_s
parameter free
$\mu_{ss}\sigma_{ss}^2$