Computer Speech Recognition: Mimicking the Human System

Li Deng

Microsoft Research, Redmond

July 24, 2005

Banff/BIRS

Fundamental Equations

- Enhancement (denoising):

$$\hat{x} = E[x|y] = \int x p(x|y) dx = \frac{\int x p_{\bar{n}}(y|x) \boxed{p(x)} dx}{p(y)},$$

- Recognition:

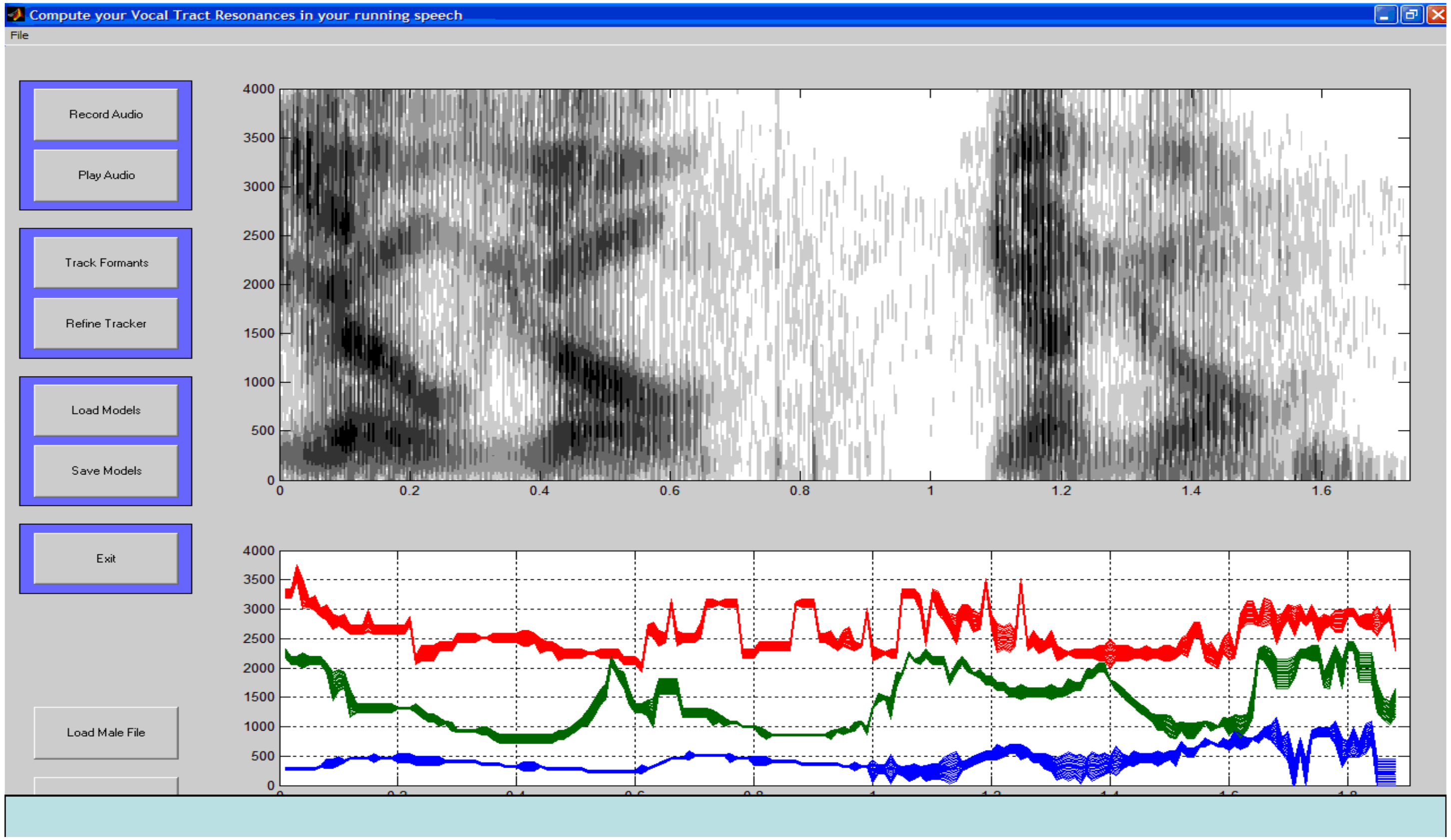
$$\hat{W} = \arg\max_{W} P(W \mid x) = \arg\max_{W} \boxed{P(x \mid W)} P(W)$$

- Importance of speech modeling

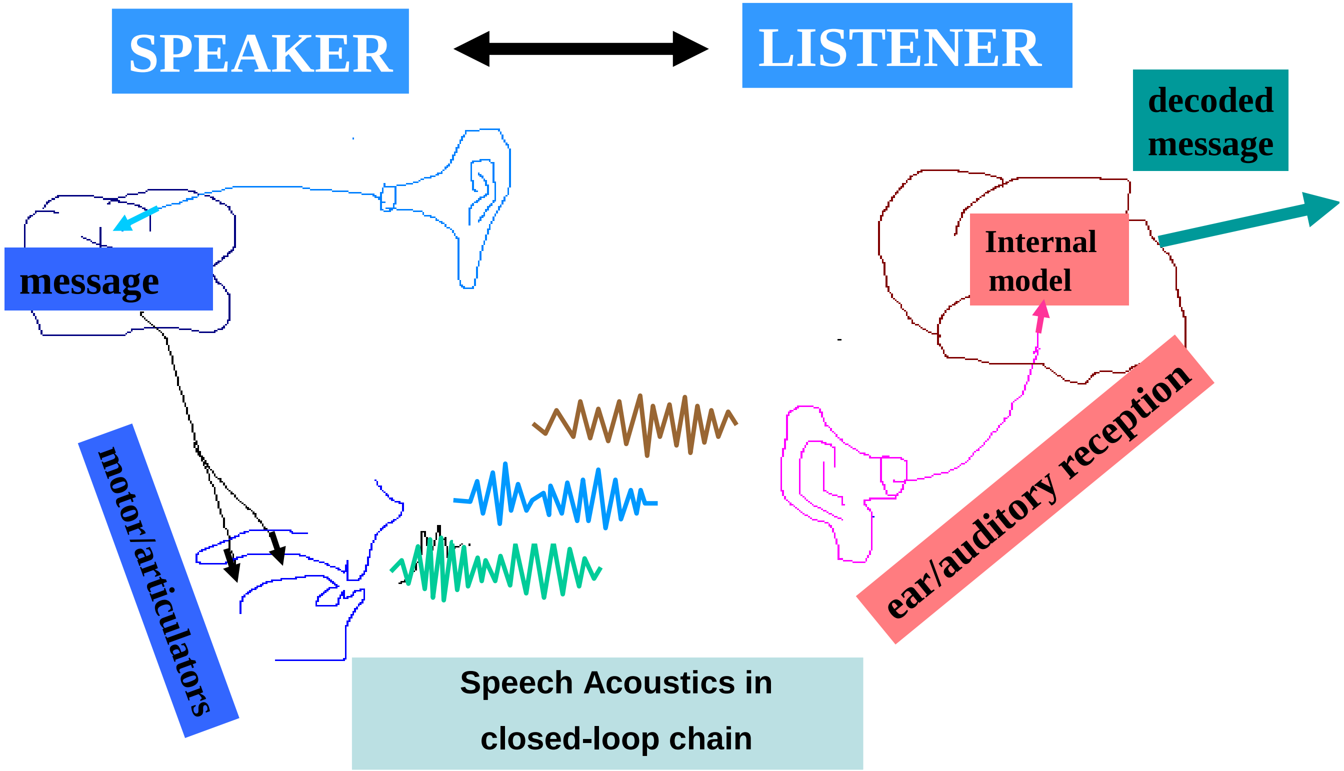
Speech Recognition--- Introduction

- Converting naturally uttered speech into text and meaning
- Conventional technology --- statistical modeling and estimation (HMM)
- Limitations
  - noisy acoustic environments
  - rigid speaking style
  - constrained task
  - unrealistic demand of training data
  - huge model sizes, etc.
  - far below human speech recognition performance
- Trend: Incorporate key aspects of human speech processing mechanisms

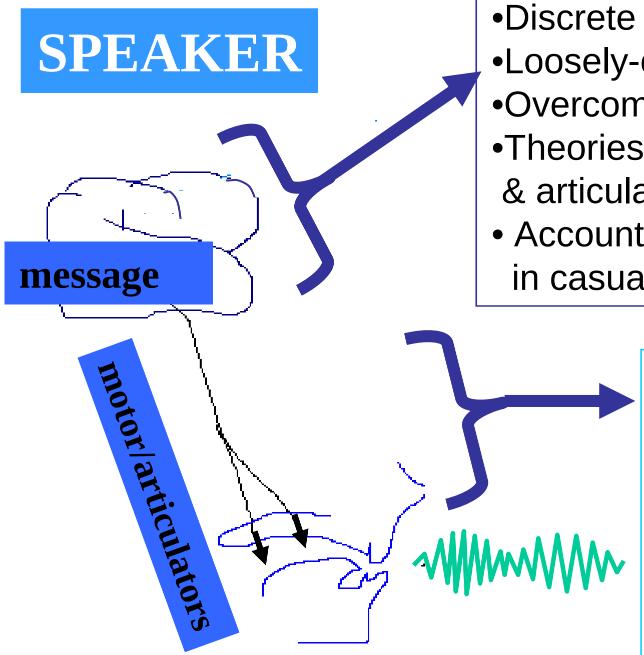
Segment-Level Speech Dynamics



Production & Perception: Closed-Loop Chain



Encoder: Two-Stage Production Mechanisms



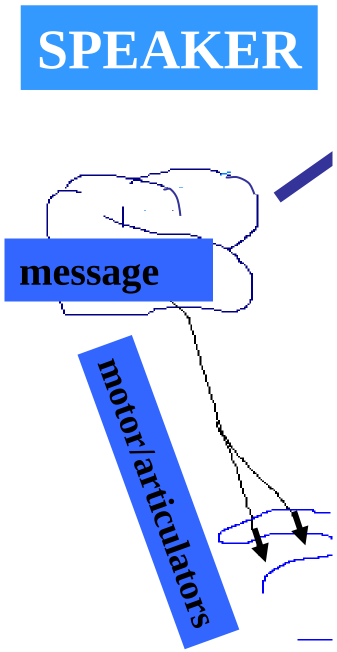
Phonology (higher level):
- Symbolic encoding of linguistic message
- Discrete representation by phonological features
- Loosely-coupled multiple feature tiers
- Overcome beads-on-a-string phone model
- Theories of distinctive features, feature geometry & articulatory phonology
- Account for partial/full sound deletion/modification in casual speech

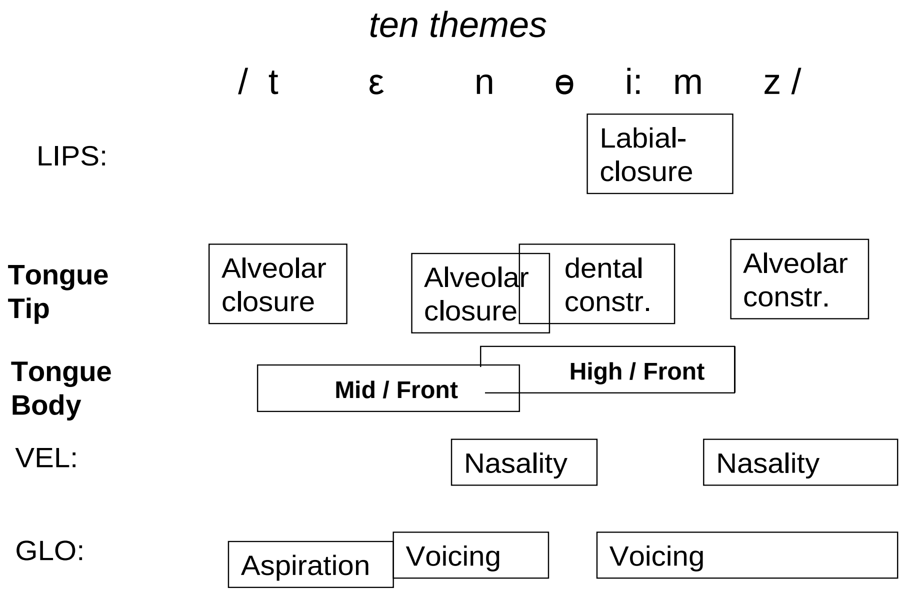
Phonetics (lower level):
- Convert discrete linguistic features to continuous acoustics
- Mediated by motor control & articulatory dynamics
- Mapping from articulatory variables to VT area function to acoustics
- Account for co-articulation and reduction (target undershoot), etc.

Encoder: Phonological Modeling

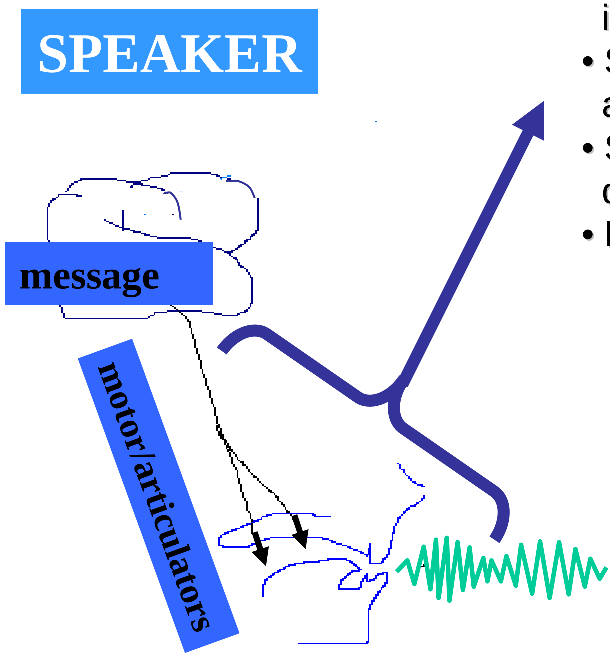
Computational phonology:

- **Represent pronunciation variations as constrained factorial Markov chain**
- **Constraint: from articulatory phonology**
- **Language-universal representation**



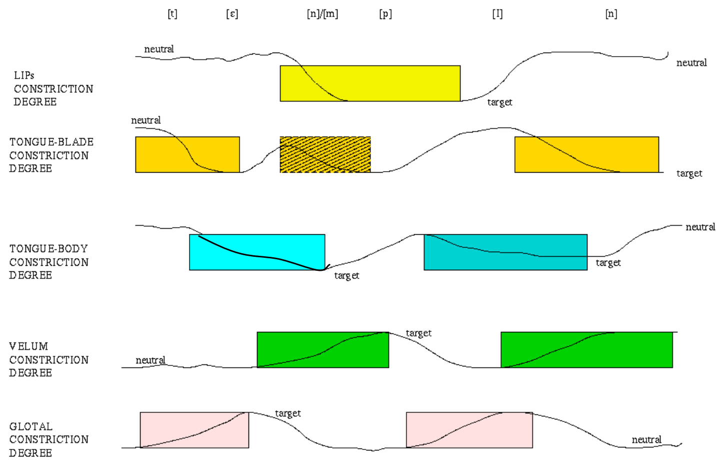


Encoder: Phonetic Modeling

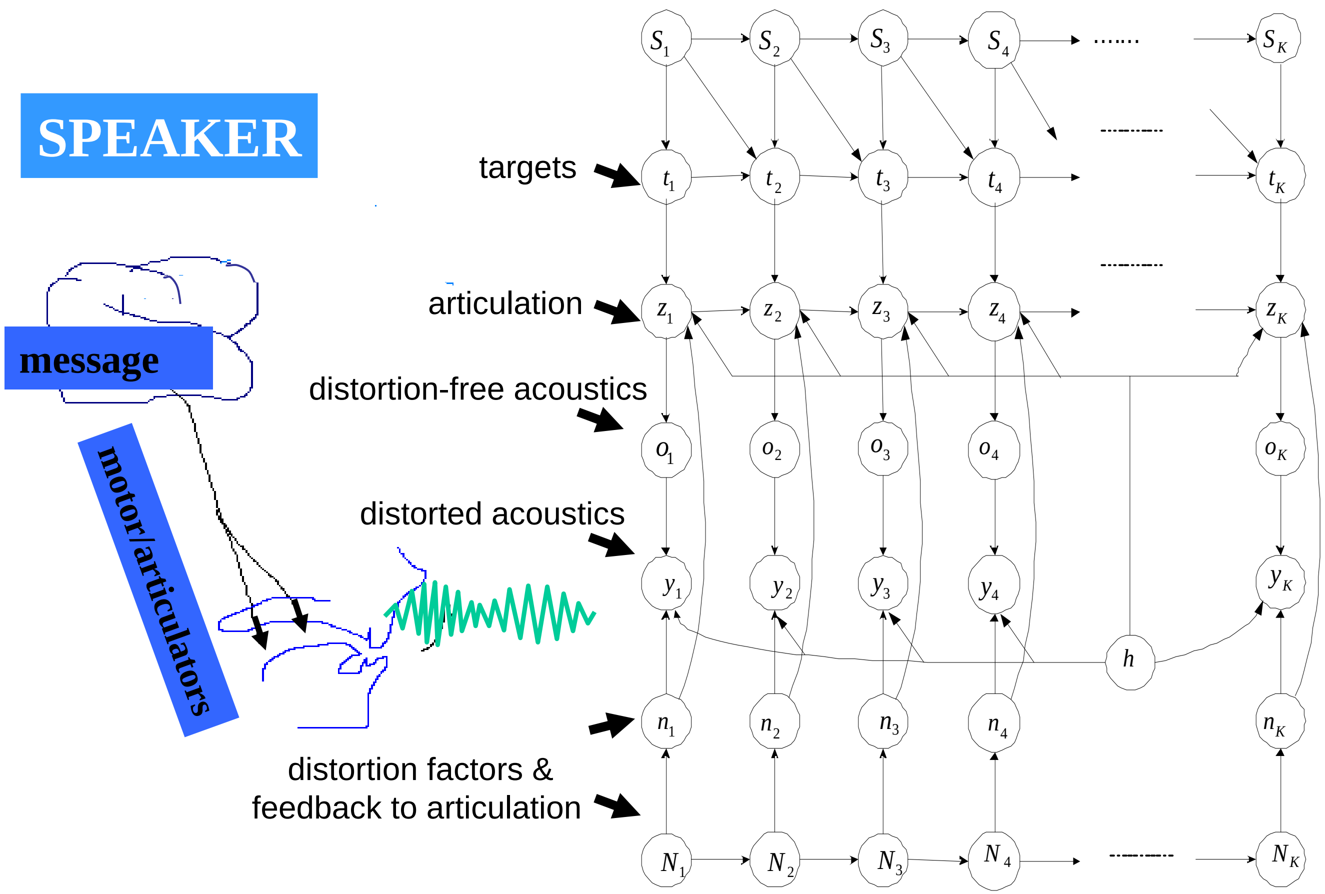


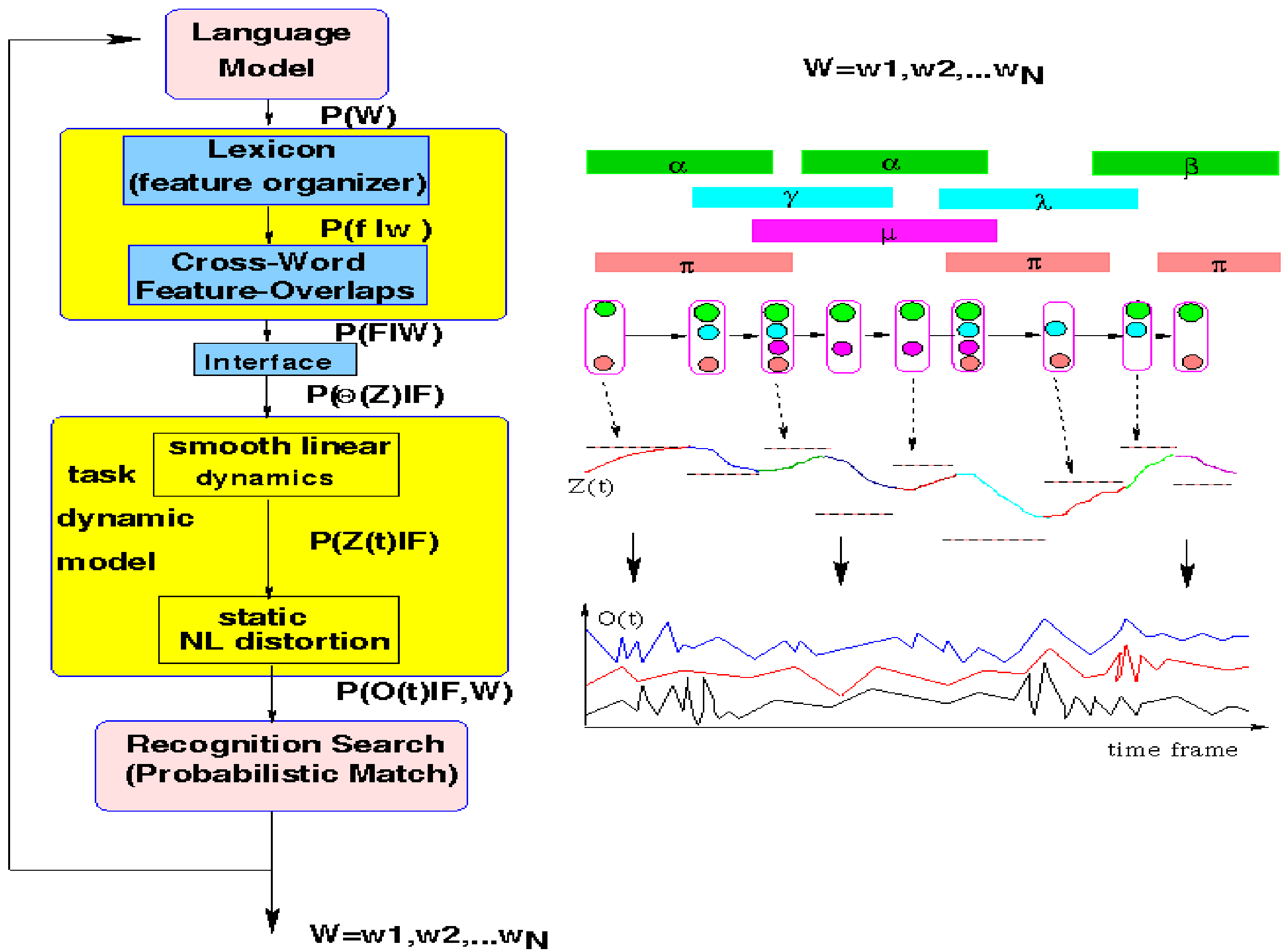
Computational phonetics:

- Segmental factorial HMM for sequential target in articulatory or vocal tract resonance domain
- Switching trajectory model for target-directed articulatory dynamics
- Switching nonlinear state-space model for dynamics in speech acoustics
- Illustration:



Phonetic Encoder: Computation

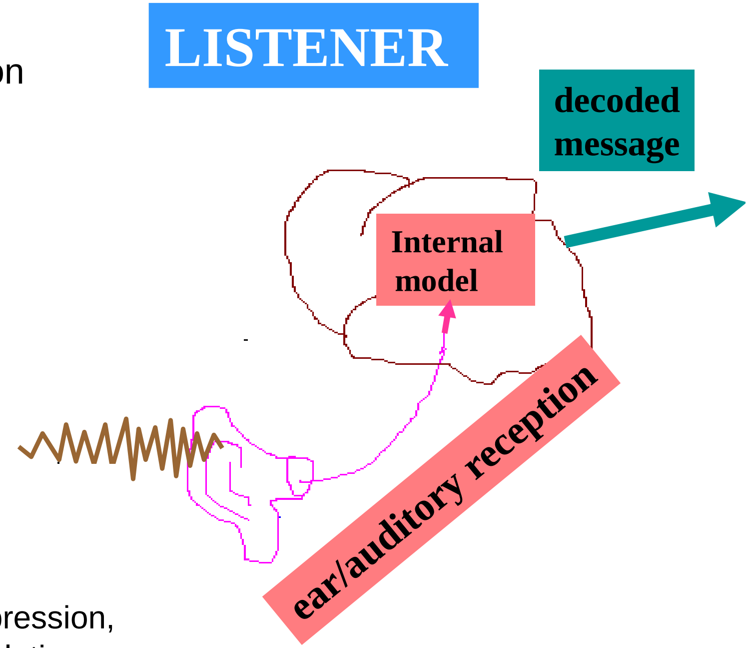




Language Model
P(W)
Lexicon (feature organizer)
P(f lw)
Cross-Word Feature-Overlaps
P(FlW)
Interface
P(Θ(Z)lF)
smooth linear dynamics
task dynamic model
P(Z(t)lF)
static NL distortion
P(O(t)lF,W)
Recognition Search (Probabilistic Match)
W=w1,w2,...wN
W=w1,w2,...wN
α
α
β
γ
λ
μ
π
π
π
Z(t)
O(t)
time frame

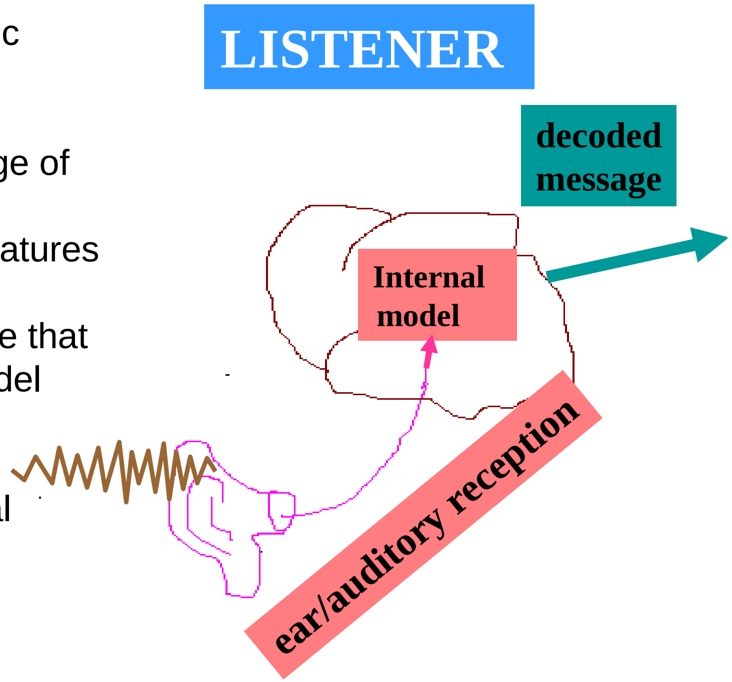
Decoder I: Auditory Reception

- Convert speech acoustic waves into efficient & robust auditory representation
- This processing is largely independent of phonological units
- Involves processing stages in cochlea (ear), cochlear nucleus, SOC, IC,…, all the way to A1 cortex
- Principal roles:
  1) combat environmental acoustic distortion;
  2) detect relevant speech features
  3) provide temporal landmarks to aid decoding
- Key properties:
  1) Critical-band freq scale, logarithmic compression,
  2) adapt freq selectivity, cross-channel correlation,
  3) sharp response to transient sounds
  4) modulation in independent frequency bands,
  5) binaural noise suppression, etc.



Decoder II: Cognitive Perception

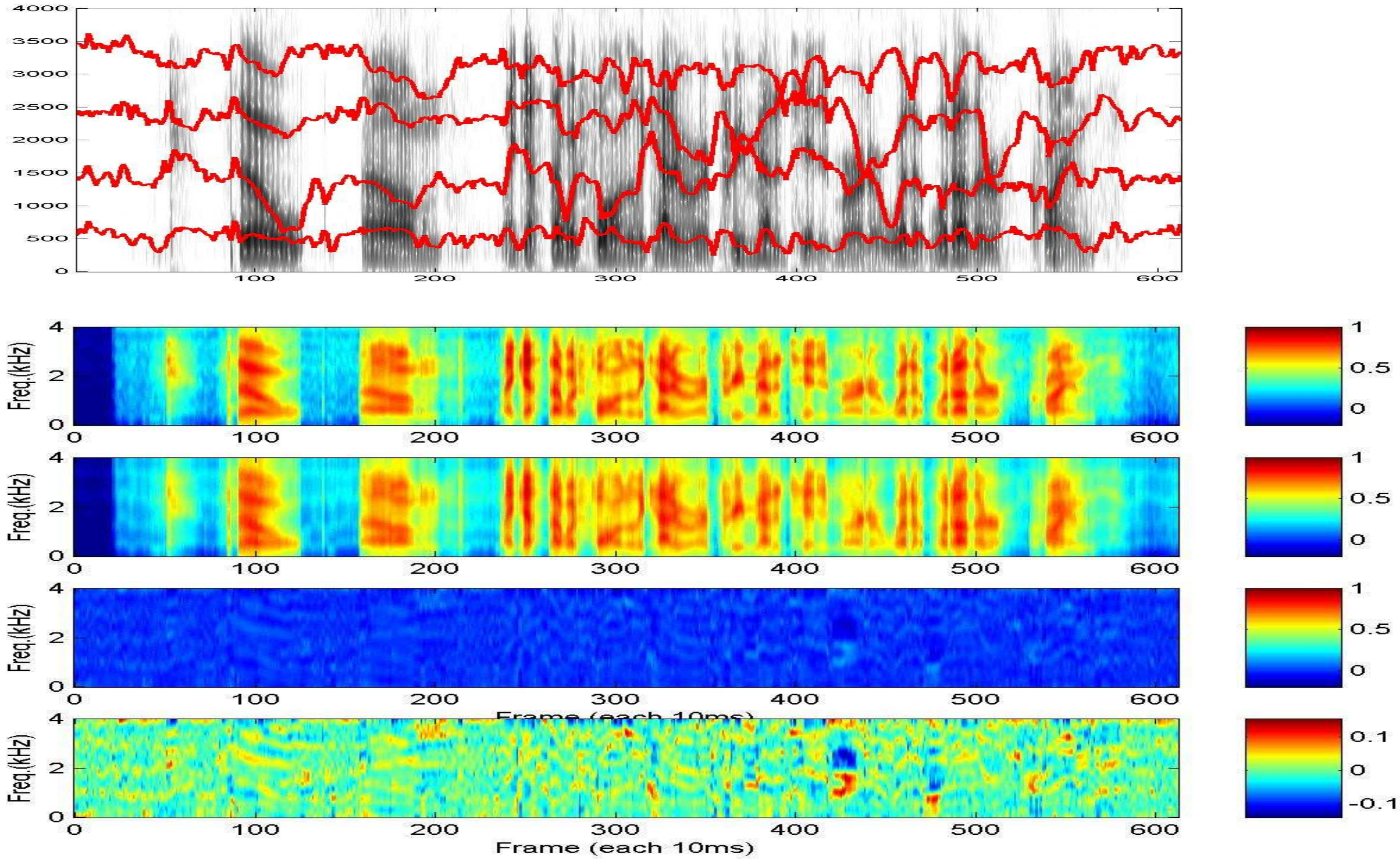
- Cognitive process: recovery of linguistic message
- Relies on
  1) "Internal" model: structural knowledge of the encoder (production system)
  2) Robust auditory representation of features
  3) Temporal landmarks
- Child speech acquisition process is one that gradually establishes the "internal" model
- Strategy: analysis by synthesis
- i.e., Probabilistic inference on (deeply) hidden linguistic units using the internal model
- No motor theory: the above strategy requires no articulatory recovery from speech acoustics



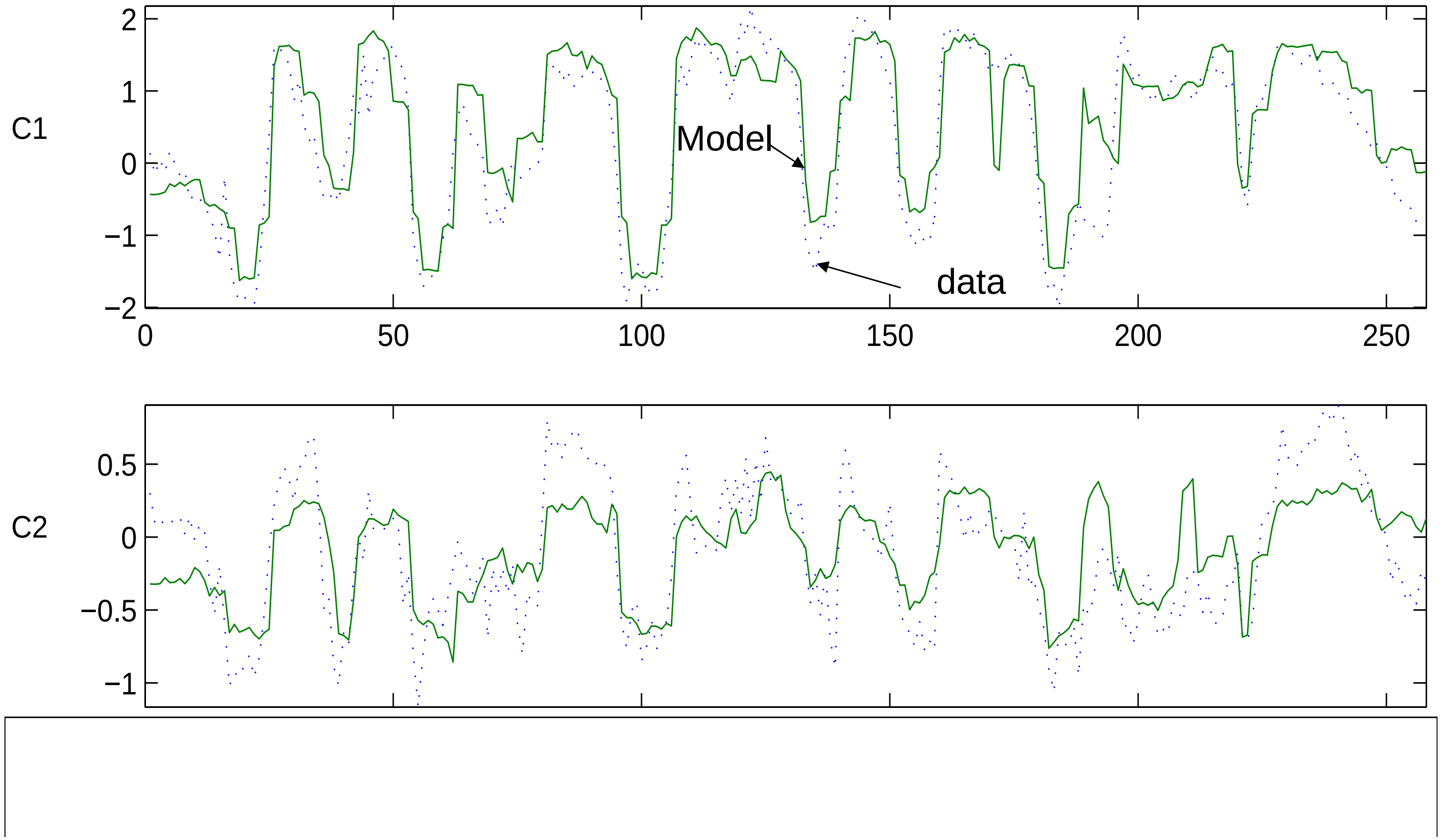
Speaker-Listener Interaction

- On-line modification of speaker's articulatory behavior (speaking effort, rate, clarity, etc.) based on listener's "decoding" performance (i.e. discrimination)
- Especially important for conversational speech recognition and understanding
- On-line adaptation of "encoder" parameters
- Novel criteria:
  - maximize **discrimination** while minimizing articulation **effort**
- In this closed-loop model, the "effort" quantified as "curvature" of temporal sequence of articulatory vector $\mathbf{z}_t$.
- No such concept of "effort" in conventional HMM systems

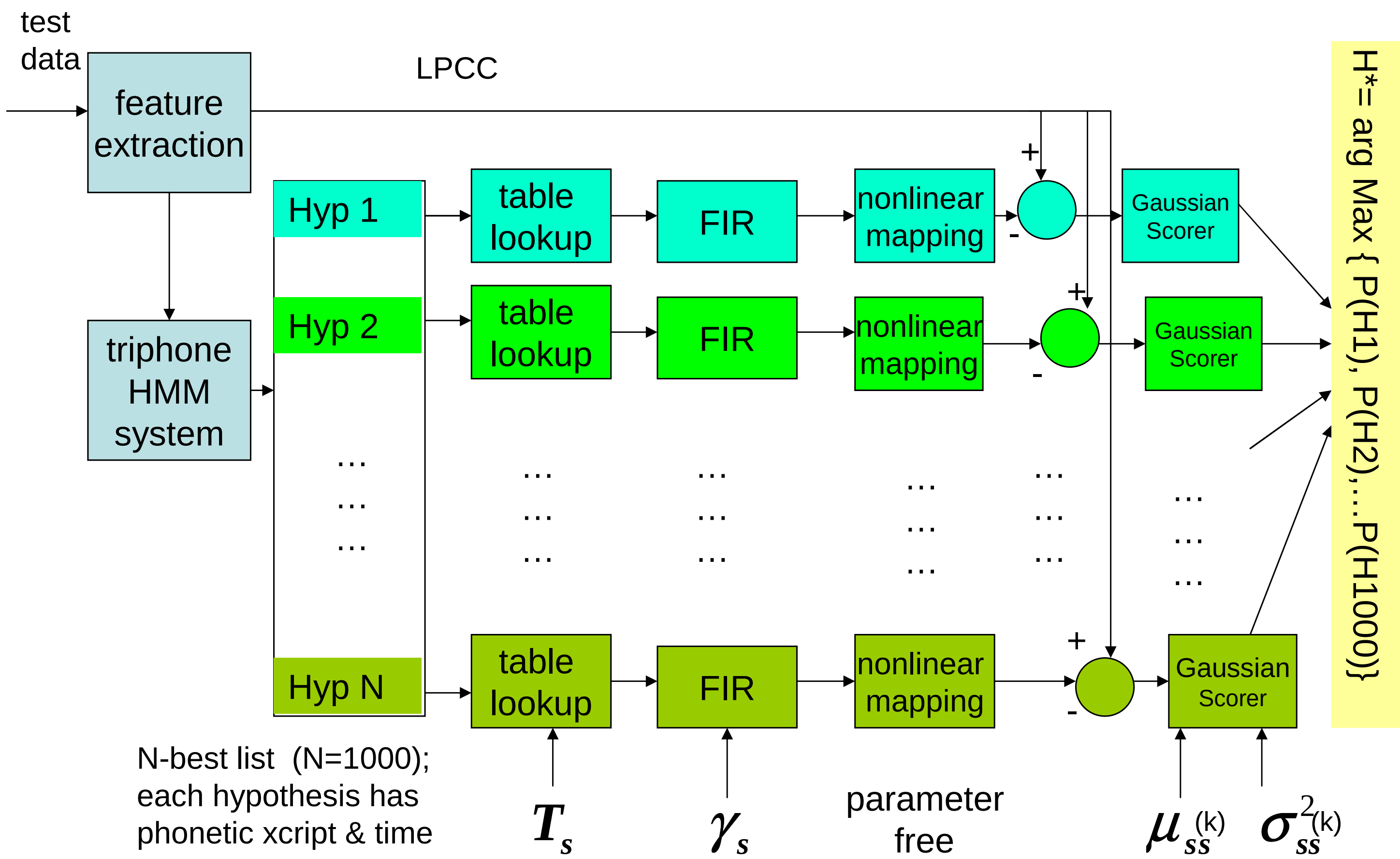
Model synthesis in FT



Model synthesis in cepstra

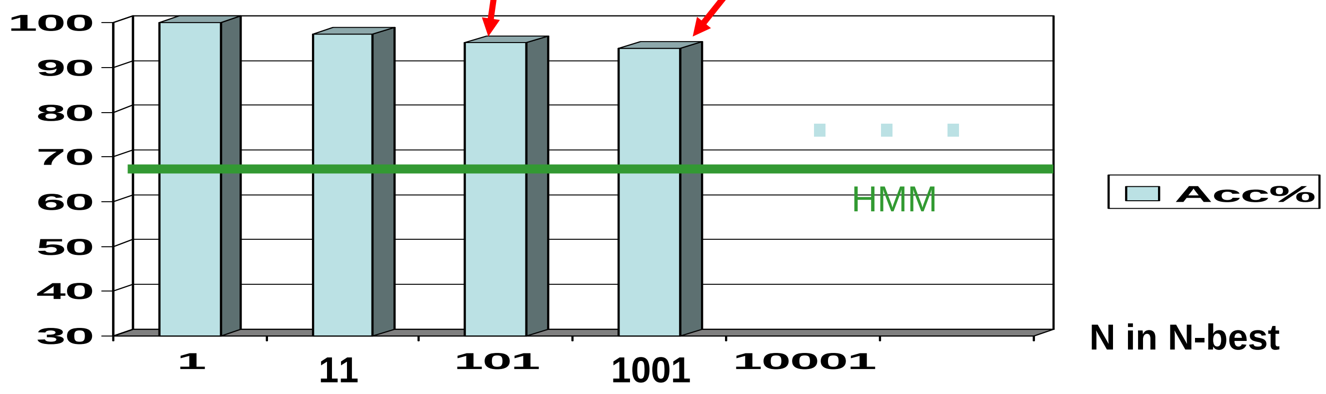


Procedure --- N-best Evaluation



Results (recognition accuracy %)
(work with Dong Yu)

Models	101-Best (with ref.)		1001-Best (with ref.)		Lattice Decode (no ref.)	
	sent	phn	sent	phn	sent	phn
New model	83.3	95.6	78.1	**94.3**	0.5	75.1
HMM system	0.0	72.5	0.0	72.5	0.0	72.5



Summary & Conclusion

- Human speech production/perception viewed as synergistic elements in a closed-looped communication chain
- They function as encoding & decoding of linguistic messages, respectively.
- In human, speech “encoder” (production system) consists of phonological (symbolic) and phonetic (numeric) levels.
- Current HMM approach approximates these two levels in a crude way:
  - phone-based phonological model (“beads-on-a-string”)
  - multiple Gaussians as phonetic model for acoustics directly
  - very weak hidden structure

Summary & Conclusion (cont'd)

- "Linguistic message recovery" (decoding) formulated as:
  - auditory reception for efficient & robust speech representation & for providing temporal landmarks for phonological features
  - cognition perception using "encoder" knowledge or "internal model" to perform probabilistic analysis by synthesis or pattern matching
- Dynamic Bayes network developed as a computational tool for constructing encoder and decoder
- Speaker-listener interaction (in addition to poor acoustic environment) cause substantial changes of articulation behavior and acoustic patterns

Issues for discussion

- Differences and similarities in processing/analysis techniques for audio/speech and image/video processing
- Integrated processing vs. modular processing

$$\hat{W} = \arg\max_{W} P(W \mid x) = \arg\max_{W} P(x \mid W) P(W)$$

- Feature extraction vs. classification
- Use of semantics (class) information for feature extraction (dim reduction, discriminative features, etc.)
- Arbitrary signal vs. structured signal (e.g. face image, human body motion, speech, music)