

Computational Modeling Approaches to

Speech Production-Perception as a Closed-Loop Chain

--- Articulation-centric perspective & relevance to NG ASR

Li Deng

Microsoft Research, Redmond

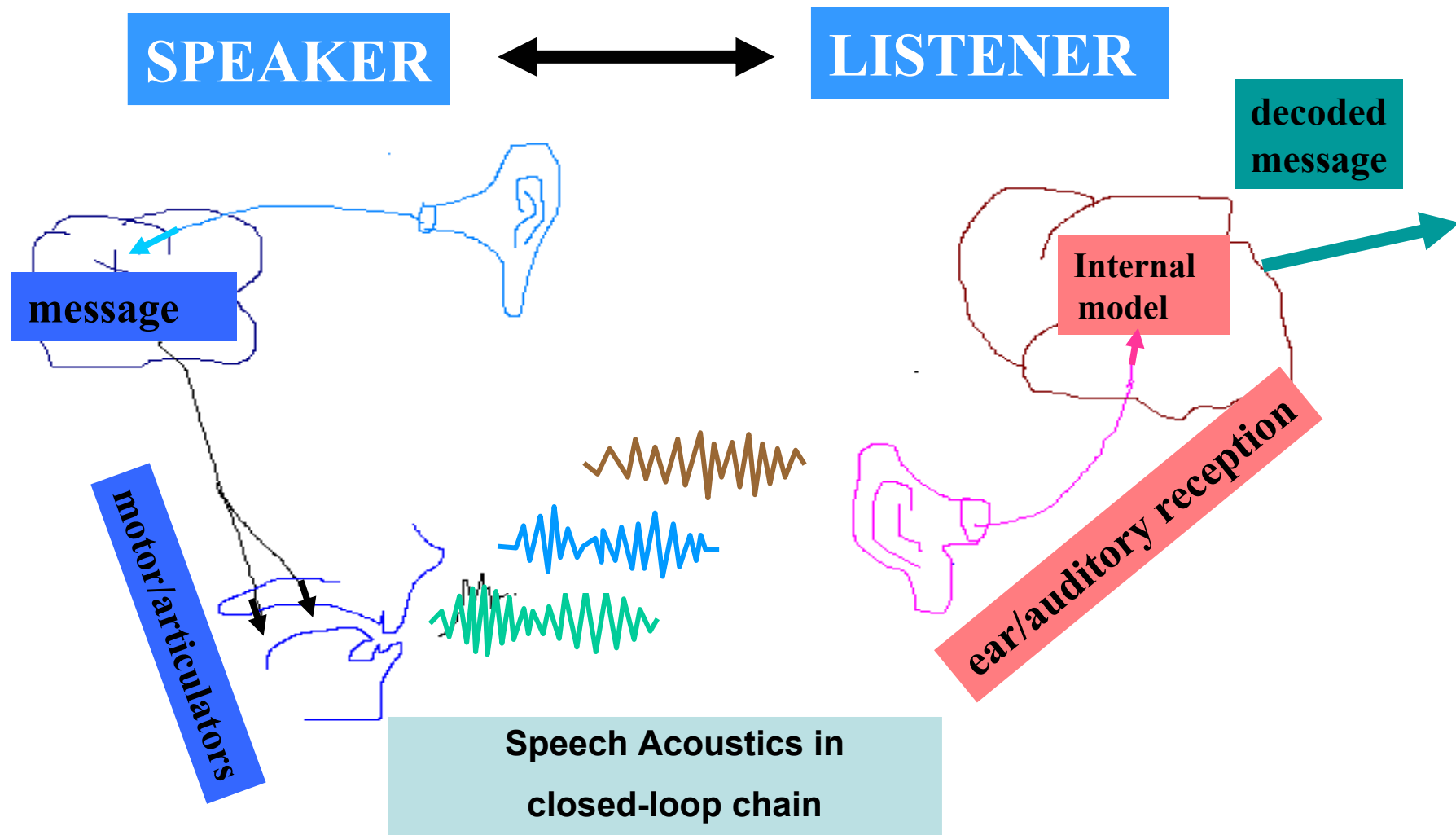
October 7, 2003

at NSF Symposium on Next Generation ASR

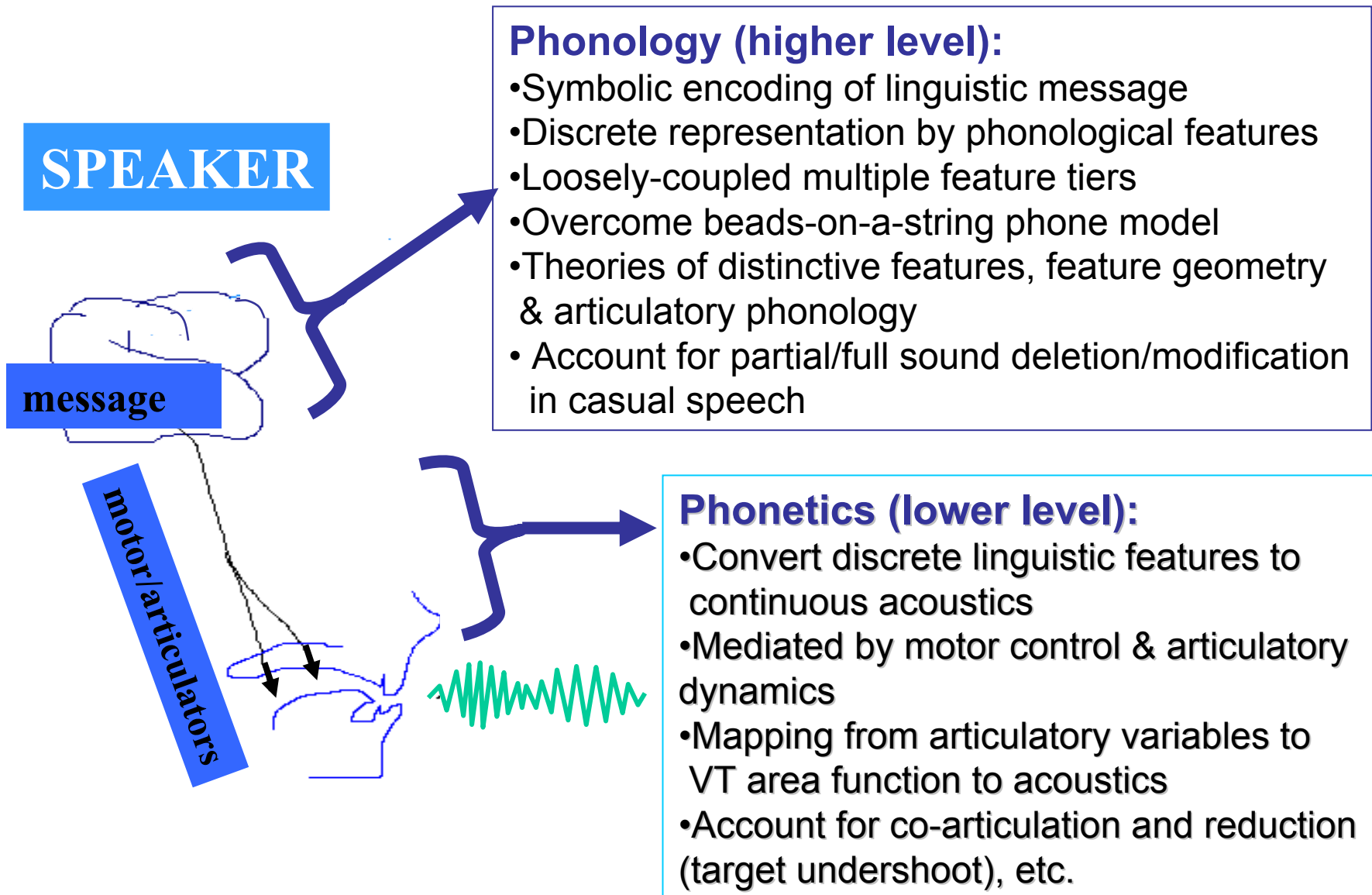
Outline

- Closed-Loop Speech Communication Chain
 - Production & perception --- synergistic view
 - Speaker-listener interactions
- Encoding of linguistic message --- speech production
 - Computation: Dynamic-Bayes-network
 - Representation and adaptive learning
- Decoding of linguistic message --- two stages
 - (auditory) reception --- encoder-independent processing
 - (cognitive) perception --- require structural knowledge of encoder
 - Computation: Dynamic-Bayes-network inference
- Experiments on feasibility of the approach

Production & Perception: Closed-Loop Chain



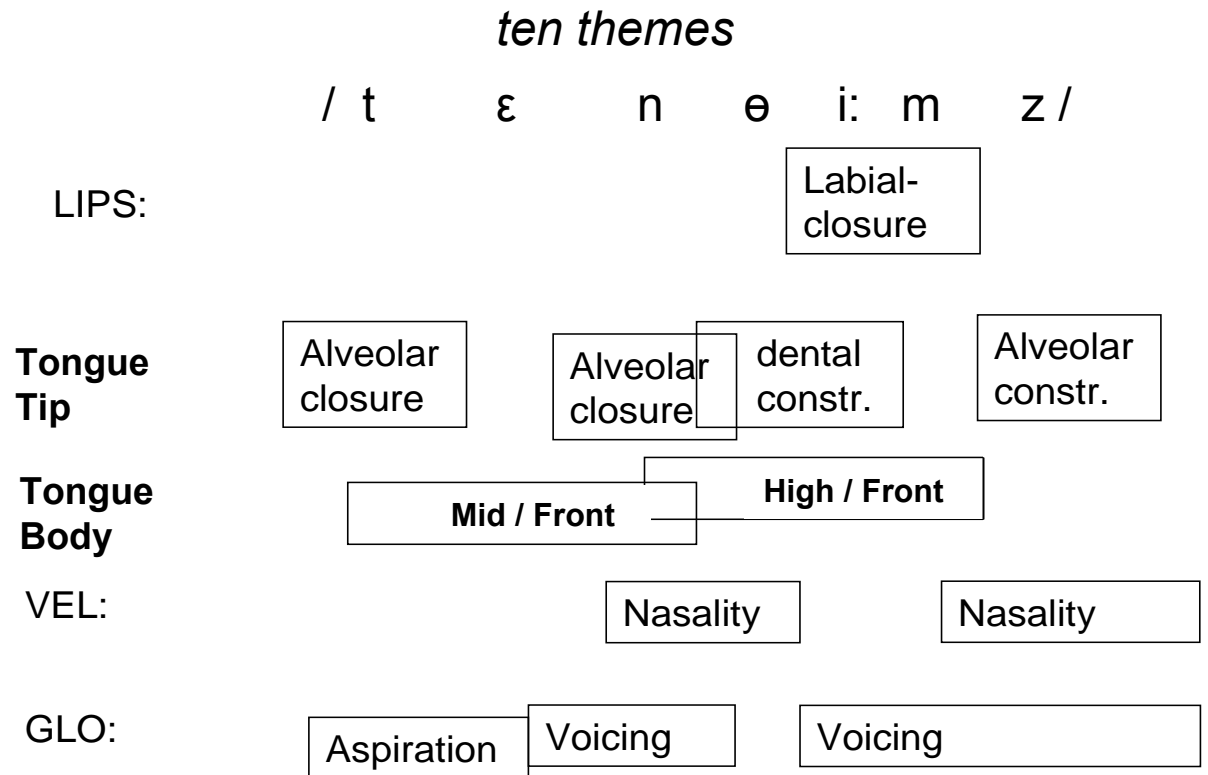
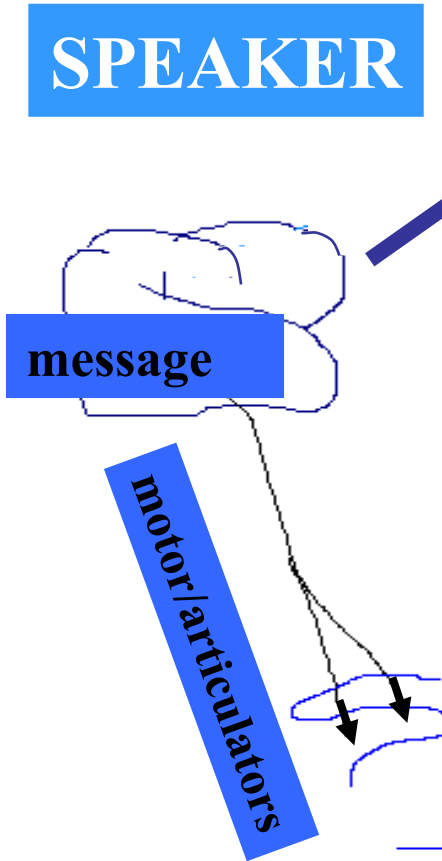
Encoder: Two-Stage Production Mechanisms



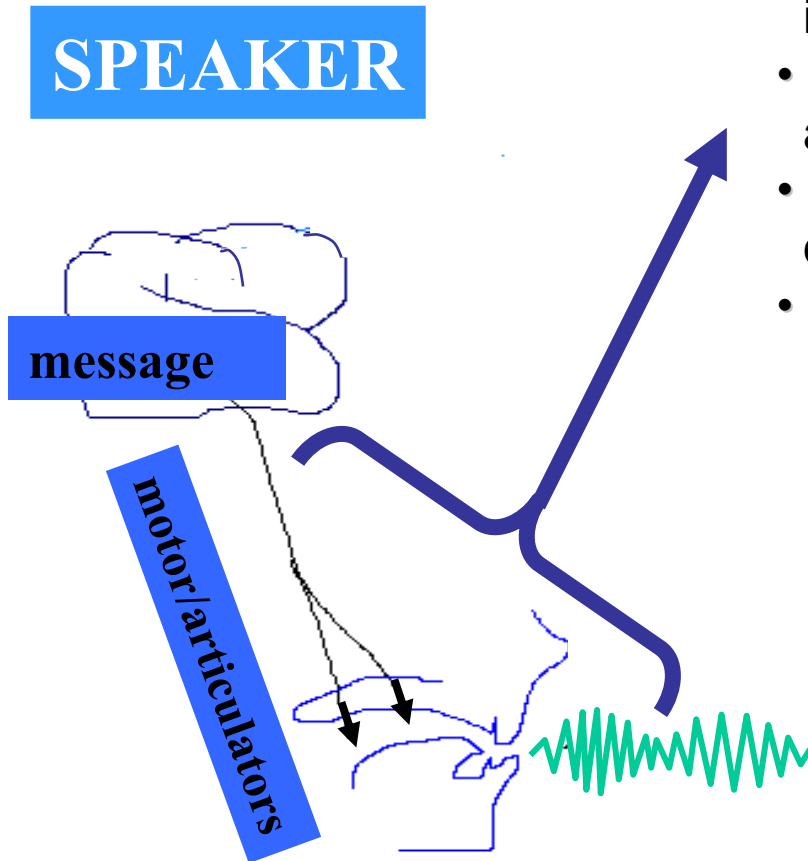
Encoder: Phonological Modeling

Computational phonology:

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

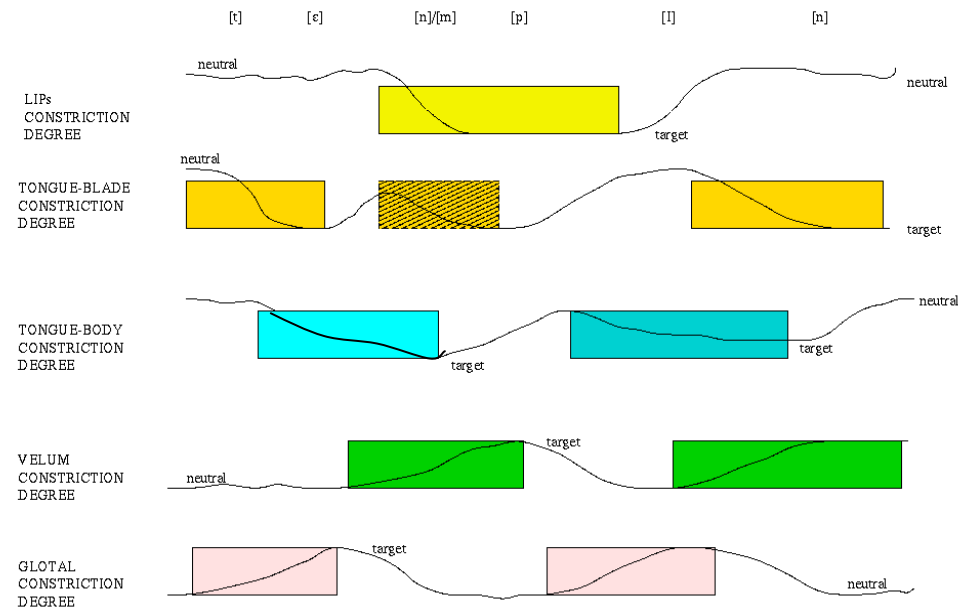


Encoder: Phonetic Modeling

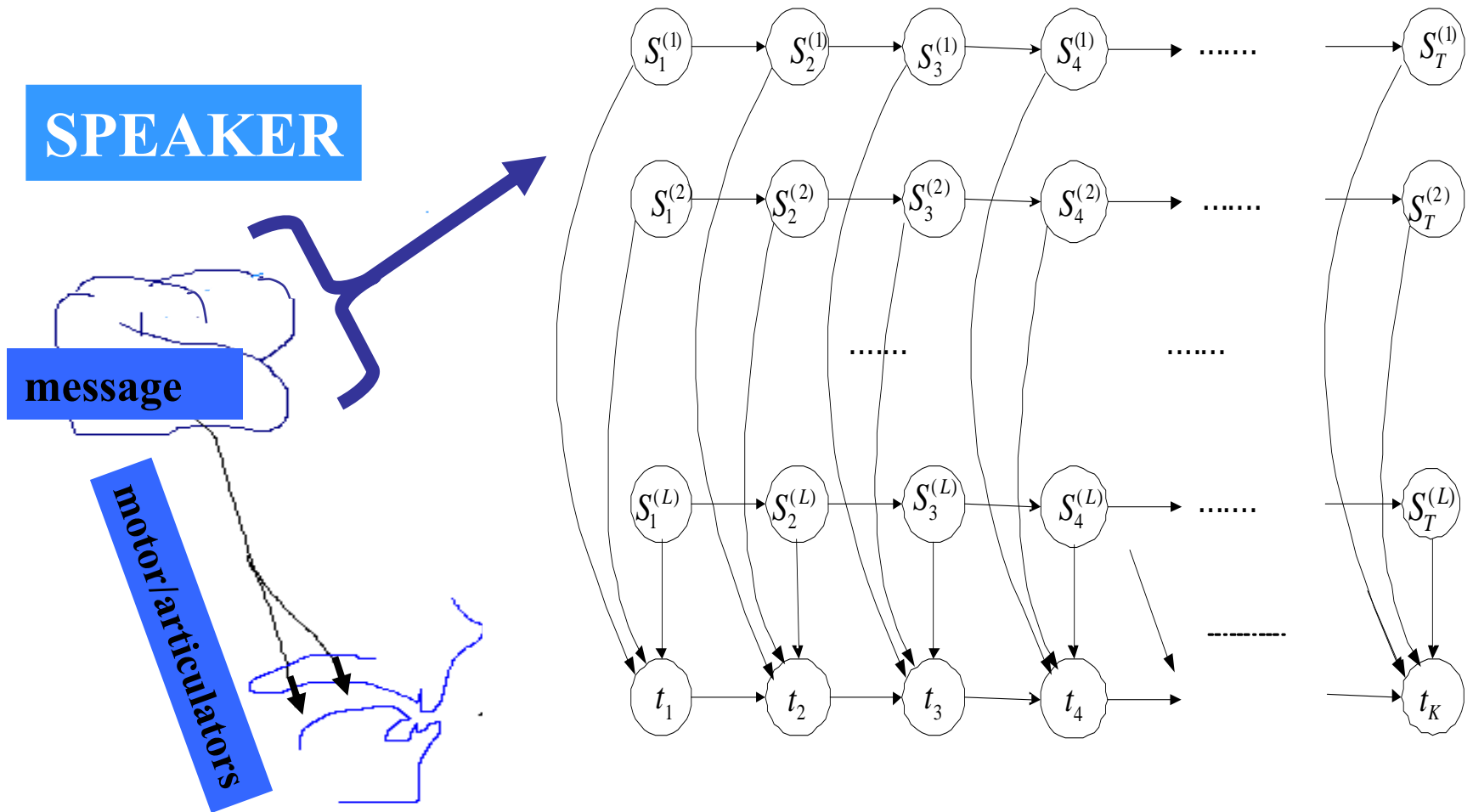


Computational phonetics:

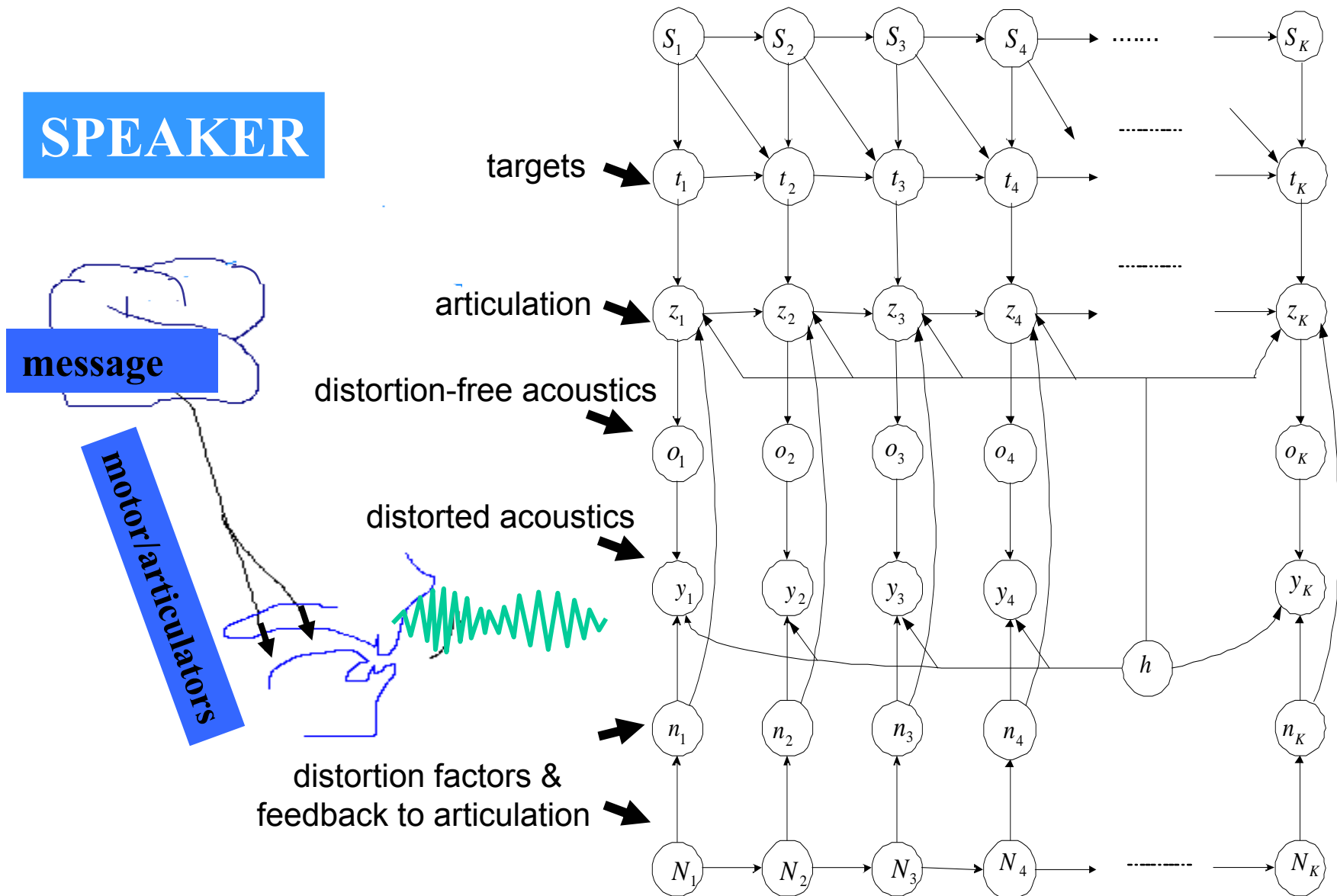
- Segmental factorial HMM for sequential target in articulatory or vocal tract resonance domain
- Switching trajectory model for target-directed articulatory dynamics
- Switching nonlinear state-space model for dynamics in speech acoustics
- Illustration:



Phonological Encoder: Computation



Phonetic Encoder: Computation

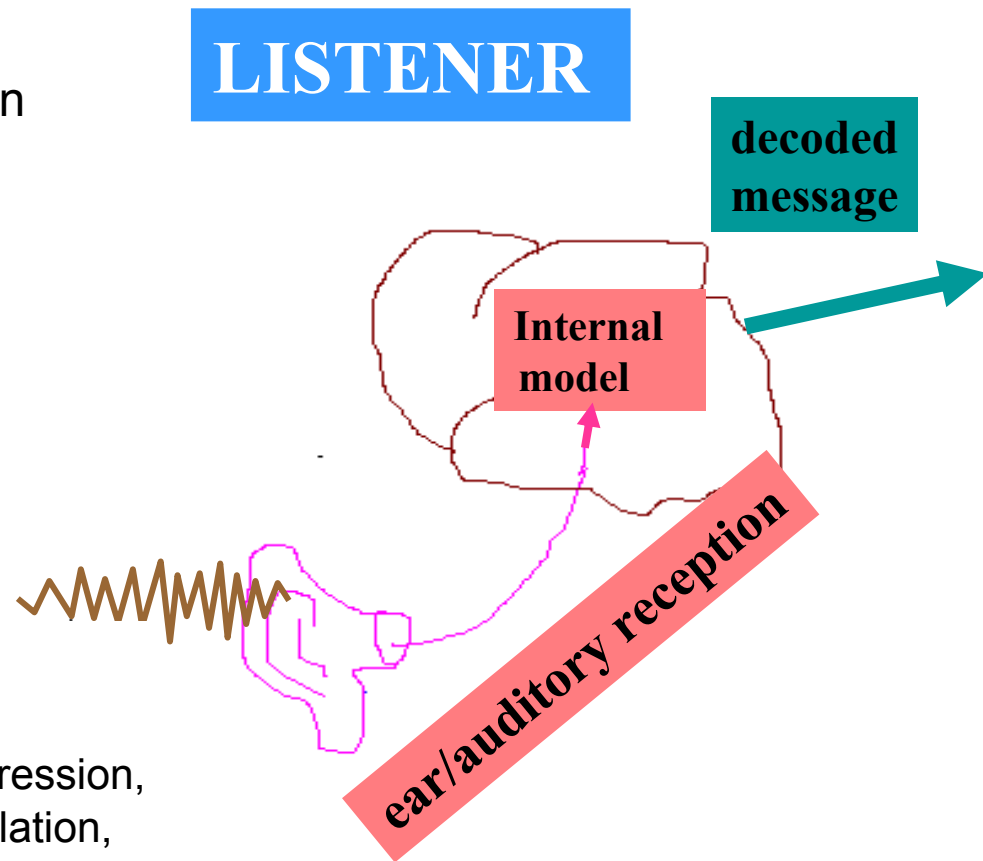


Outline

- Closed-Loop Speech Communication Chain
 - Production & perception --- synergistic view
 - Speaker-listener interactions
- Encoding of linguistic message --- speech production
 - Computation: Dynamic-Bayes-network
 - Representation and learning
- Decoding of linguistic message --- two stages
 - (auditory) reception
 - (cognitive) perception --- require structural knowledge of encoder
 - Computation: Dynamic-Bayes-network inference
- Experiments on feasibility of the approach

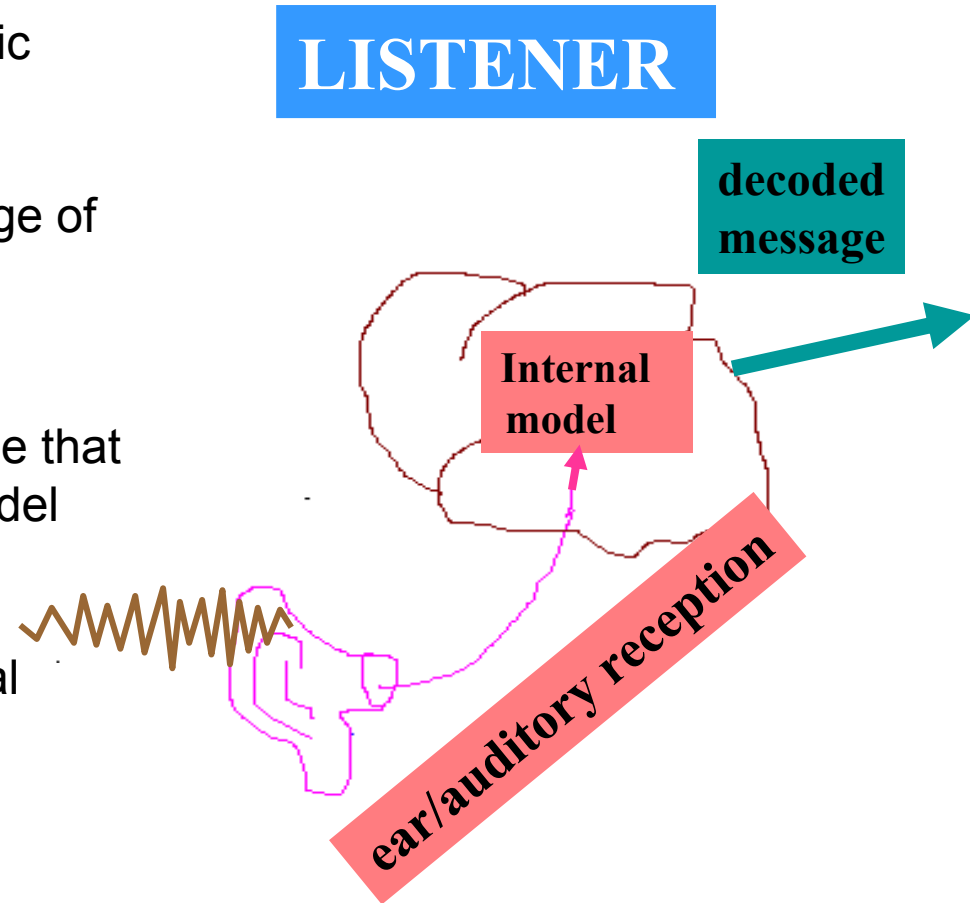
Decoder I: Auditory Reception

- 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
- Two principal roles:
 - 1) combat environmental acoustic distortion;
 - 2) 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 (CN),
 - 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
 - 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

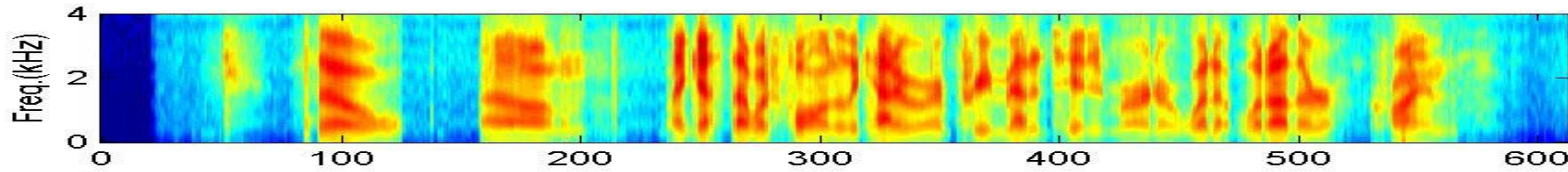
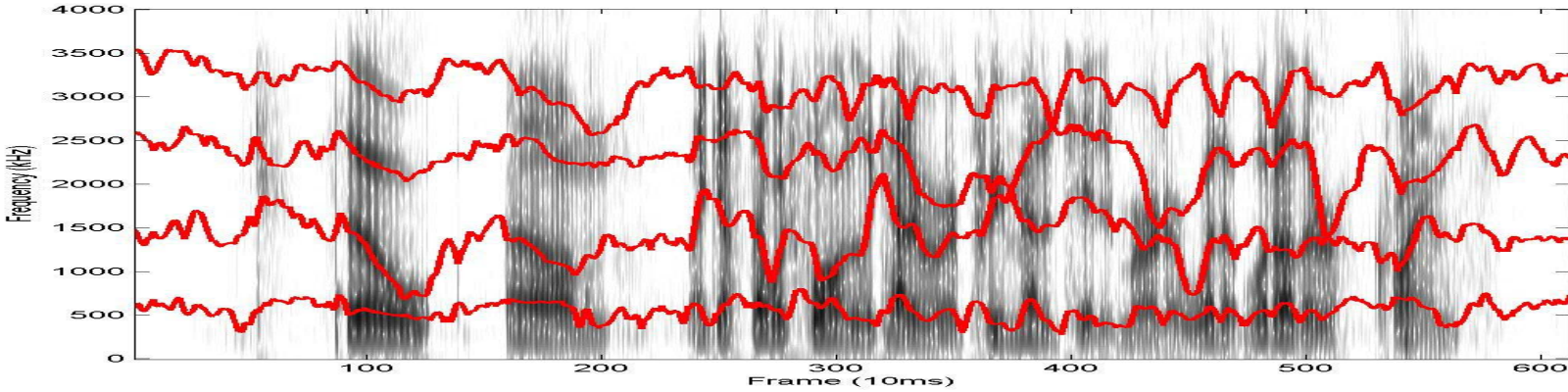
- On-line modification of speaker's articulatory behavior (speaking effort, rate, clarity, etc.) based on listener's "decoding" performance (i.e. discrimination)
- Especially important for conversational speech recognition and understanding
- On-line adaptation of "encoder" parameters
- Novel criteria: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

Experiments on Speech Production

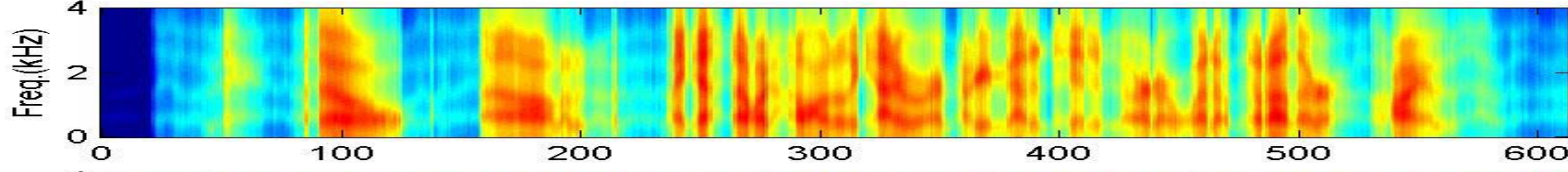
- A simplified version of the model for “encoder”:
 - Use pseudo-articulatory variables (vocal tract resonances)
 - Use analytical nonlinear mapping from resonances to LPC-cepstra
 - Construct piecewise approximation to this mapping
 - Approximate the variational decoding rule by Kalman filter
- Lesson: Importance of employing compact trainable parameters to compensate for encoder-model’s inaccuracy

Model-Driven Production

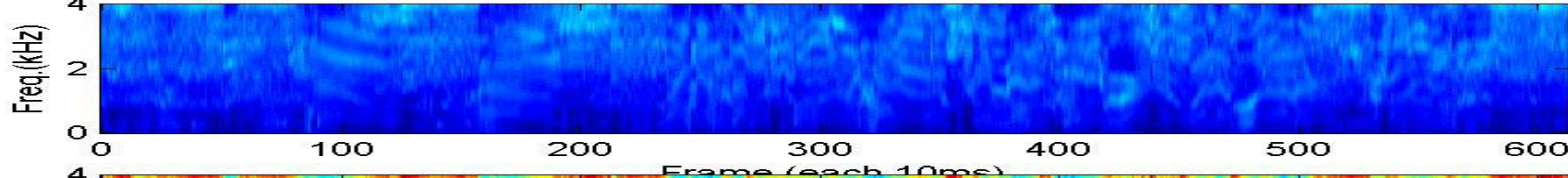
(Ex: SWBD conversational utterance)



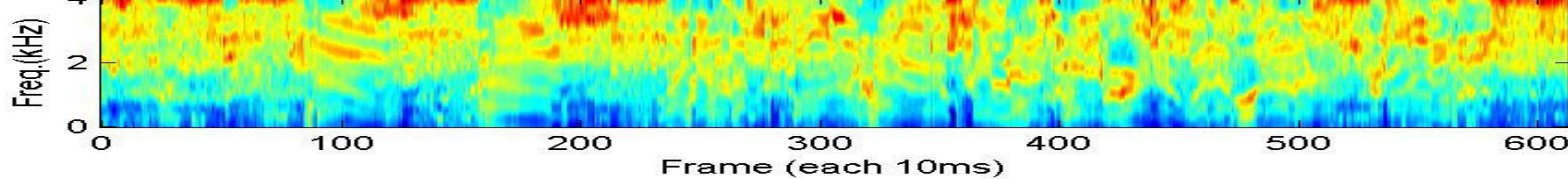
Spectrogram
(cep-smoothed)



Output of
“production”
system

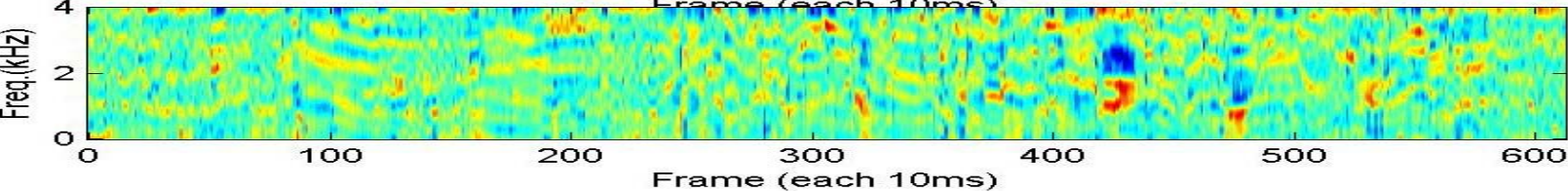
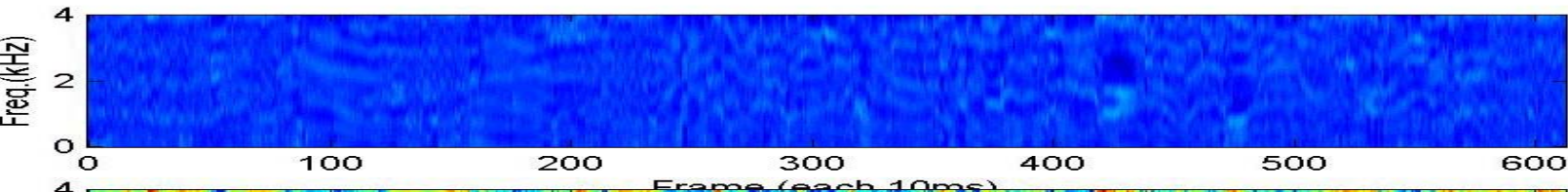
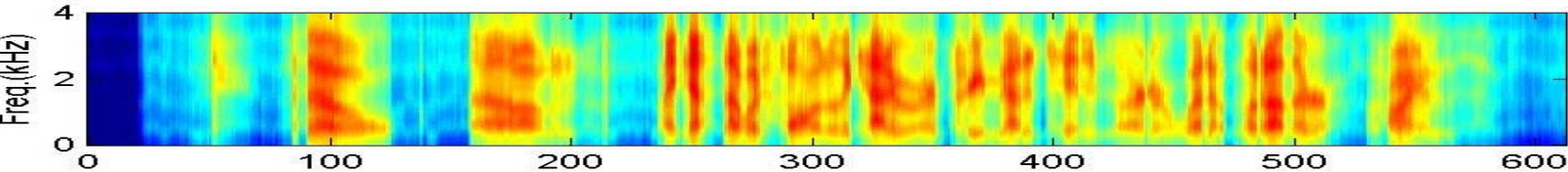
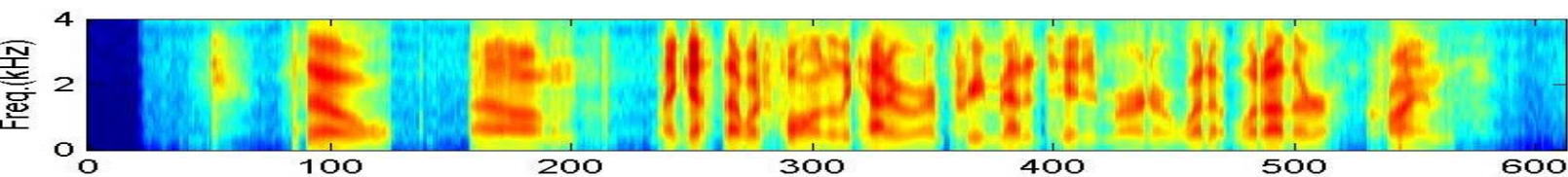
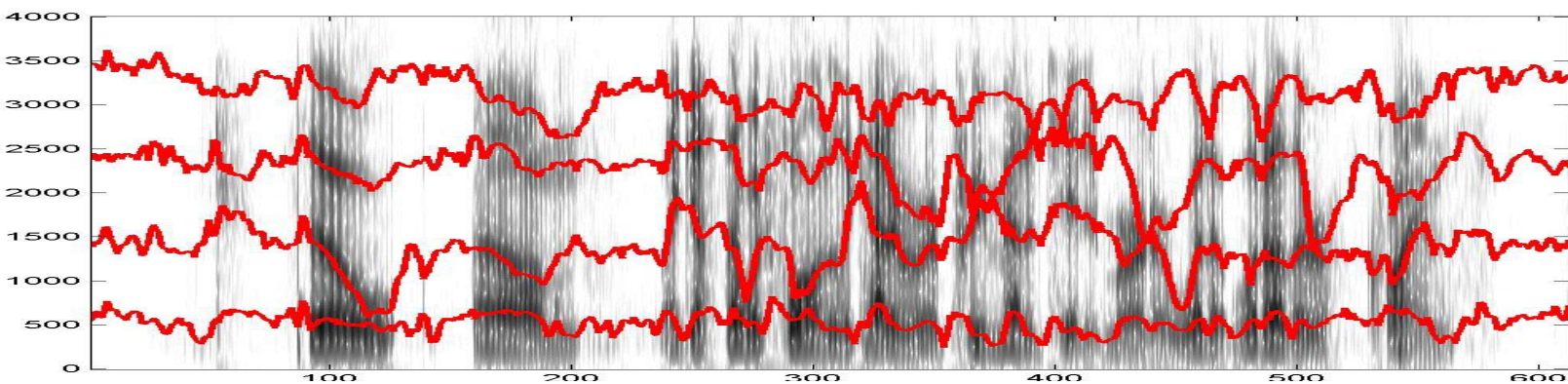


Residual
spectrogram

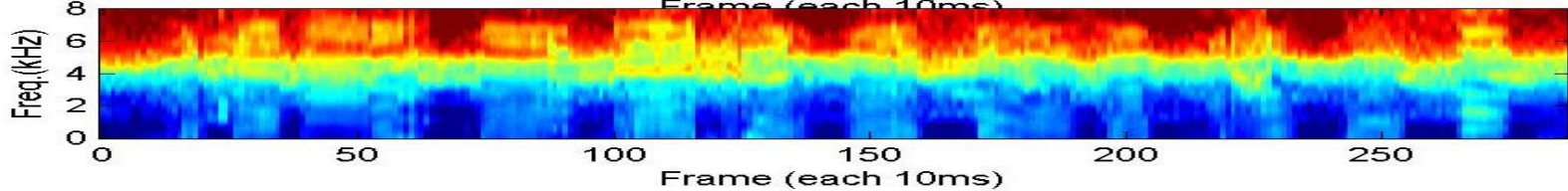
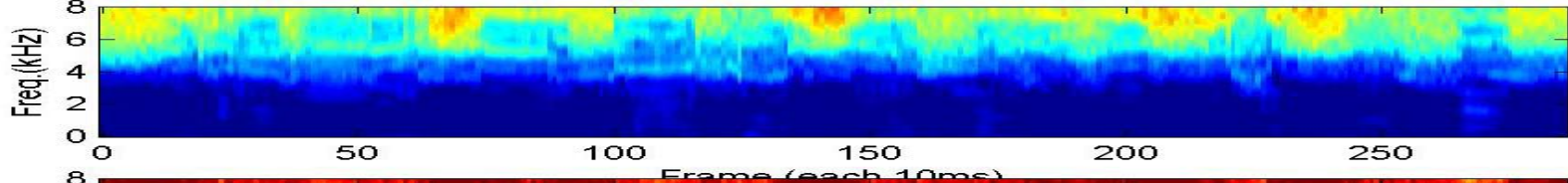
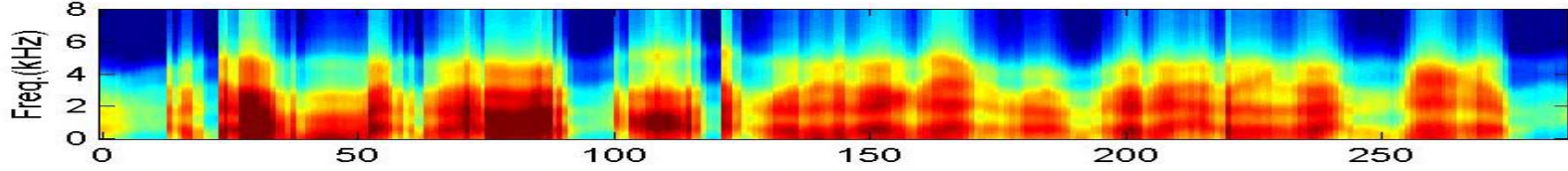
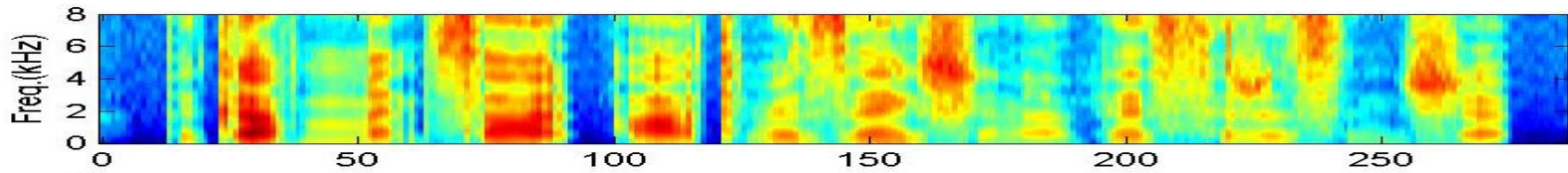
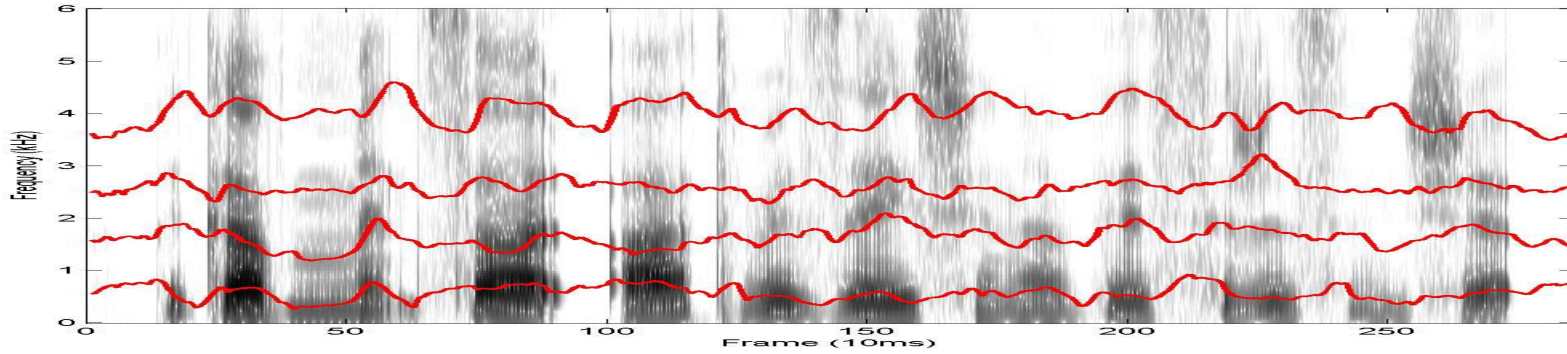


Residual
spectrogram
(enlarged)

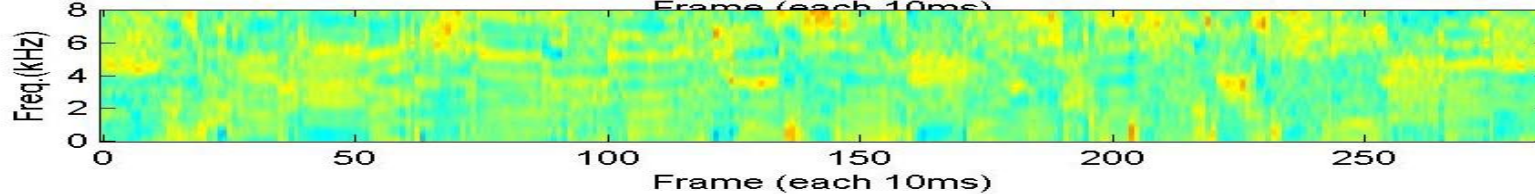
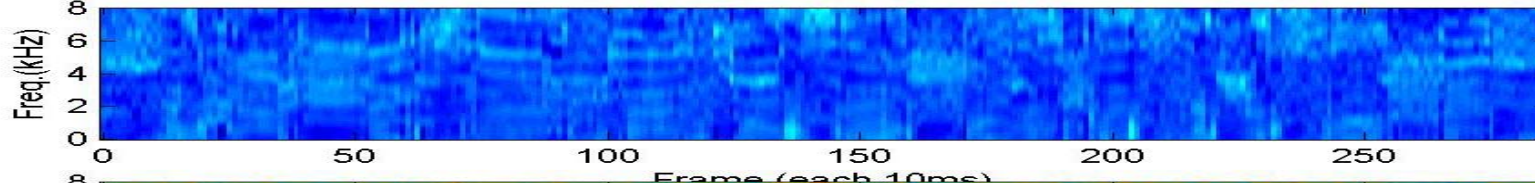
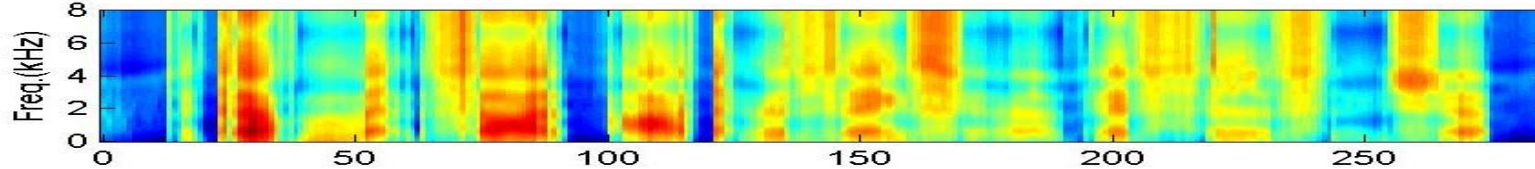
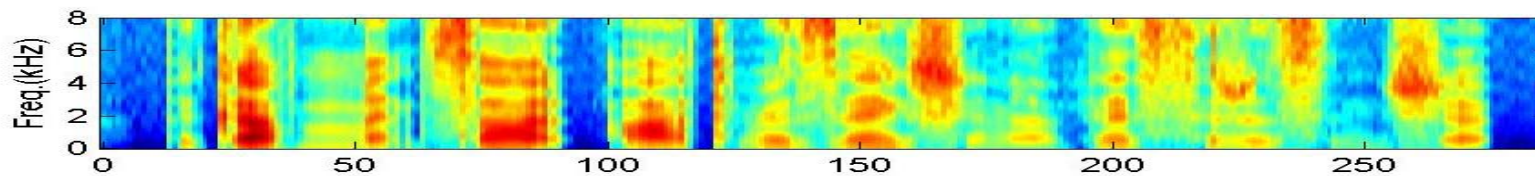
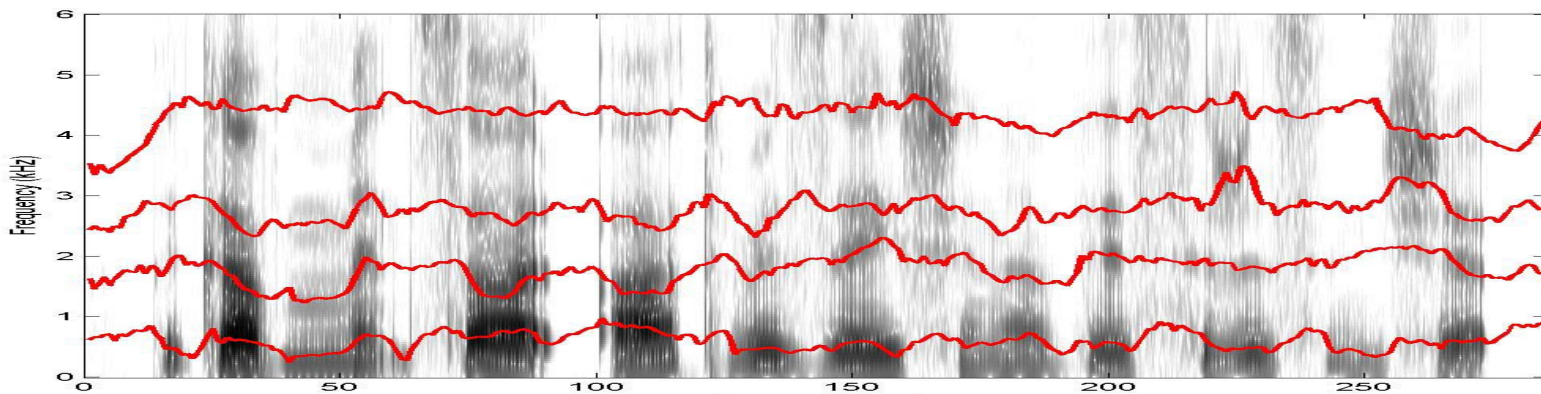
Adding Trainable Parameters



Model-Driven Production (TIMIT --- wideband speech)

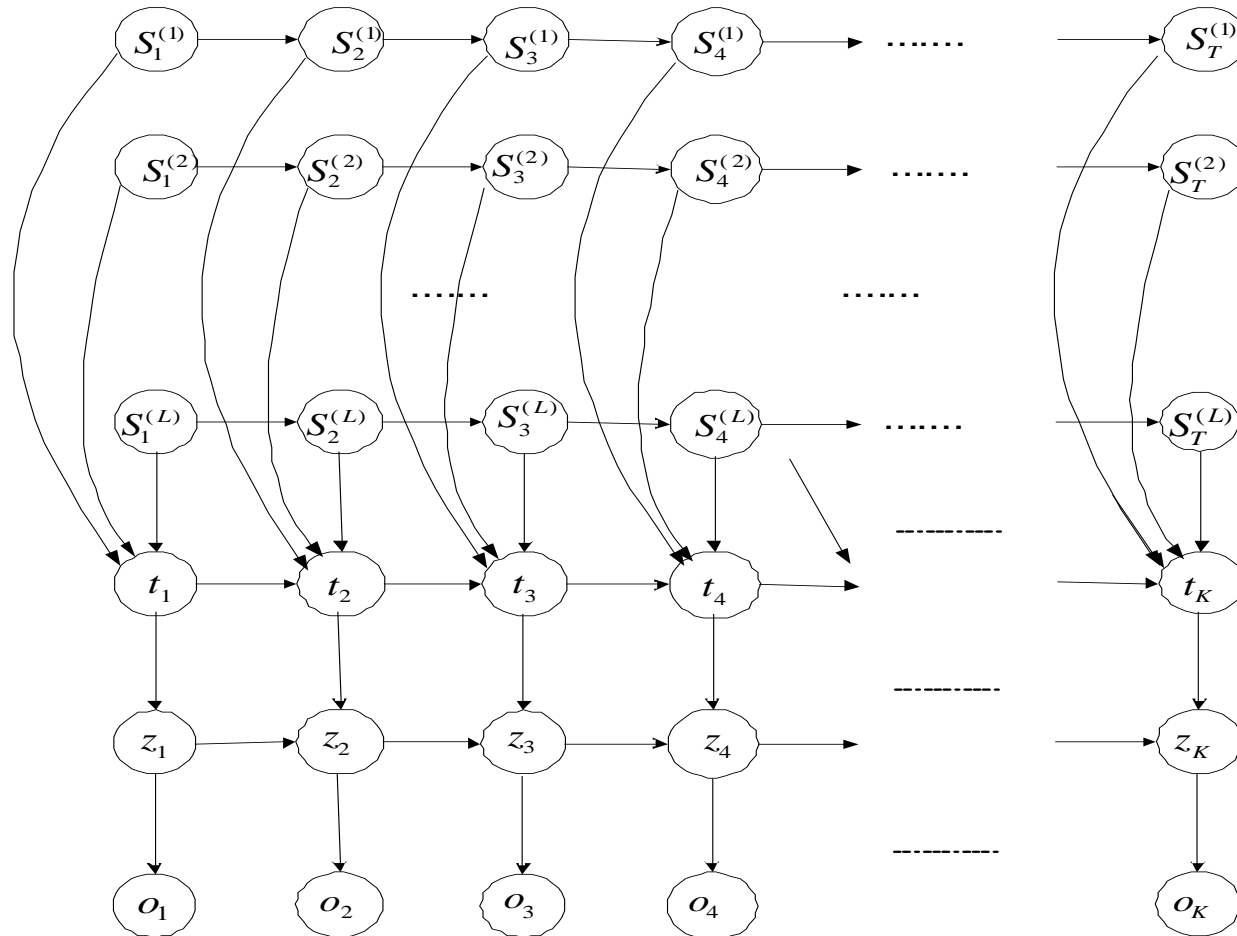


Adding Trainable Parameters (TIMIT)



Ongoing work: “Decoder” development

- Decoder is equipped with structural knowledge of the “encoder” --- Nonlinear switching state-space model:



Ongoing work: “Decoder” development

- Inference: compute posterior probability of linguistic units (discrete hidden variables)
- Decoding: find optimal linguistic unit sequence (goal of ASR) that maximizes the posterior probability.
- This problem is intractable for the above “encoder” model (i.e. exponential in computation)
- Solution:
 - use “variational approximation” in E-step of the EM algorithm
 - and use “temporal landmarks” detected by the auditory processor to further constrain the search

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
- Much can be improved using structured knowledge of articulation-centric speech production and advanced learning algorithms

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
 - need to take into account in NG ASR for conversational speech
- Scientific background and computational framework for our recent MSR EARS research

End
&
Backup Slides

segments IPA (TIMIT)	example words	Lips	Tongue-Blade	Tongue-Dorsum	Velum	Larynx
b (bcl-b)	bee	L_{lab}				X_v
d (dcl-d)	day		B_{lab}			X_v
g (gcl-g)	geese			D_{lab}		X_v
p ^h (pcl-p)	pea	L_{lab}				X_{asp}
t ^h (tcl-t)	tea		B_{lab}			X_{asp}
k ^h (kcl-k)	key			D_{lab}		X_{asp}
ɾ (q)	bat					X_{glo}
r (dx)	city		B_{flap}			
f (f)	fin	L_{crst}				
θ (th)	thin		B_{crst}			
s (s)	sea		B_{crst}			
ʃ (sh)	she		B_{crst}			
v (v)	van	L_{crst}				X_v
ð (dh)	then		B_{crst}			X_v
z (z)	zone		B_{crst}			X_v
ʒ (zh)	azure		B_{crst}			X_v
h (hh)	hay					X_{asp}
tʃ (ch)	choke		$B_{lab} \oplus B_{crst}$			
dʒ (jh)	joke		$B_{lab} \oplus B_{crst}$			X_v
m (m, em)	man, bottom	L_{lab}			V	X_v
n (n, en)	noon, button		B_{lab}		V	X_v
ŋ (ng, eng)	sing, Washington			D_{lab}	V	X_v
r̄ (nx)	money		B_{flap}		V	X_v
l (l)	lay		B_l			X_v
ɫ (l, el)	all		B_l	D_L		X_v
r (r)	ray	L_r	B_r			X_v
ɹ (r)	are		B_r	D_R		X_v
w (w)	way	L_w		D_w		X_v
j (y)	yacht			D_j		X_v
i (iy)	beet			D_i		X_v
ɪ (ih)	bit			D_I		X_v
e (eh)	bet			D_E		X_v
æ (ae)	bat			$D_{æ}$		X_v
ə (ax, ax-h)	about			D_{ax}		X_v
ɜ (er, axr)	bird			D_{er}		X_v
u (uw, ux)	boot	L_u		D_u		X_v
ʊ (uh)	book	L_U		D_U		X_v
ɔ (ao)	bought	L_o		D_o		X_v
ʌ (ah)	but			D_{ah}		X_v
ɑ (aa)	bother			D_{aa}		X_v
e ^j (ey)	bait			$D_e \oplus D_j$		X_v
a ^j (ay)	bite			$D_a \oplus D_j$		X_v
ɔ ^j (oy)	boy	$L_o \oplus \square$		$D_o \oplus D_j$		X_v
ɔ ^w (ow)	boat	$L_o \oplus L_w$		$D_o \oplus D_w$		X_v
ɑ ^w (aw)	bout	$\square \oplus L_w$		$D_a \oplus D_w$		X_v

Table 1: Feature specification system for American English.

	Stage I	Stage II	Stage III	Stage IV
Input	Distinctive-feature representation of an utterance	Discrete articulatory state sequence	Segmental target sequence	Articulatory or vocal tract resonance vector
Mediating process	Temporal overlapping Mechanism	Symbolic-to- numerical mapping	Explicit or recursive trajectory modeling	Static, numerical-to-numerical, nonlinear mapping
Output	Overlapping articulatory gestures, represented by a set of discrete meta-states	Segmental target sequence, represented as a left-to-right, constrained switching random process	Continuous, smooth, and target-directed trajectories for articulatory or vocal tract resonance variables	Acoustic or auditory feature vector computable or measurable directly from speech waveforms
Domain	Phonology	Interface between phonology and phonetics	Phonetics	Phonetics
Properties	Account for partial or full sound deletion or modification for the pronunciation variation in casual speech; Also account for contextual variation at the pronunciation level	Account for compensatory articulation, or different ways of activating articulators to achieve similar acoustic effects or auditory perception; Targets are used as the control signal directing the dynamic system governing speech articulation	Account for variability of speech due to reduced speaking effort or increased speaking rate (phonetic reduction) and due to increased effort (e.g., Lombard effect); Also account for coarticulation at the physical level due to inertia in articulation	Account for differences in different speakers' speech production organs and the distorting effects due to acoustic environments

Table: Four major stages in the architectural design of a novel speech recognizer

Phonetic Reduction Illustration

$$\mathbf{z}_n = 2\gamma_s \mathbf{z}_{n-1} - \gamma_s^2 \mathbf{z}_{n-2} + (1 - \gamma_s)^2 \mathbf{T}_s + \mathbf{w}_n$$

yo-yo (formal)

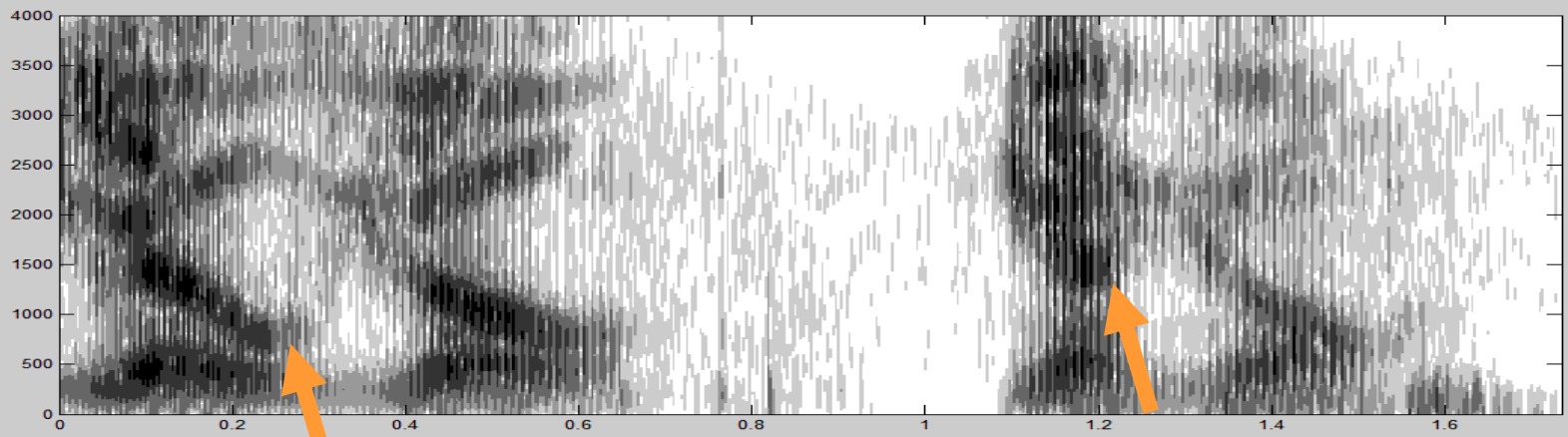
yo-yo (casual)

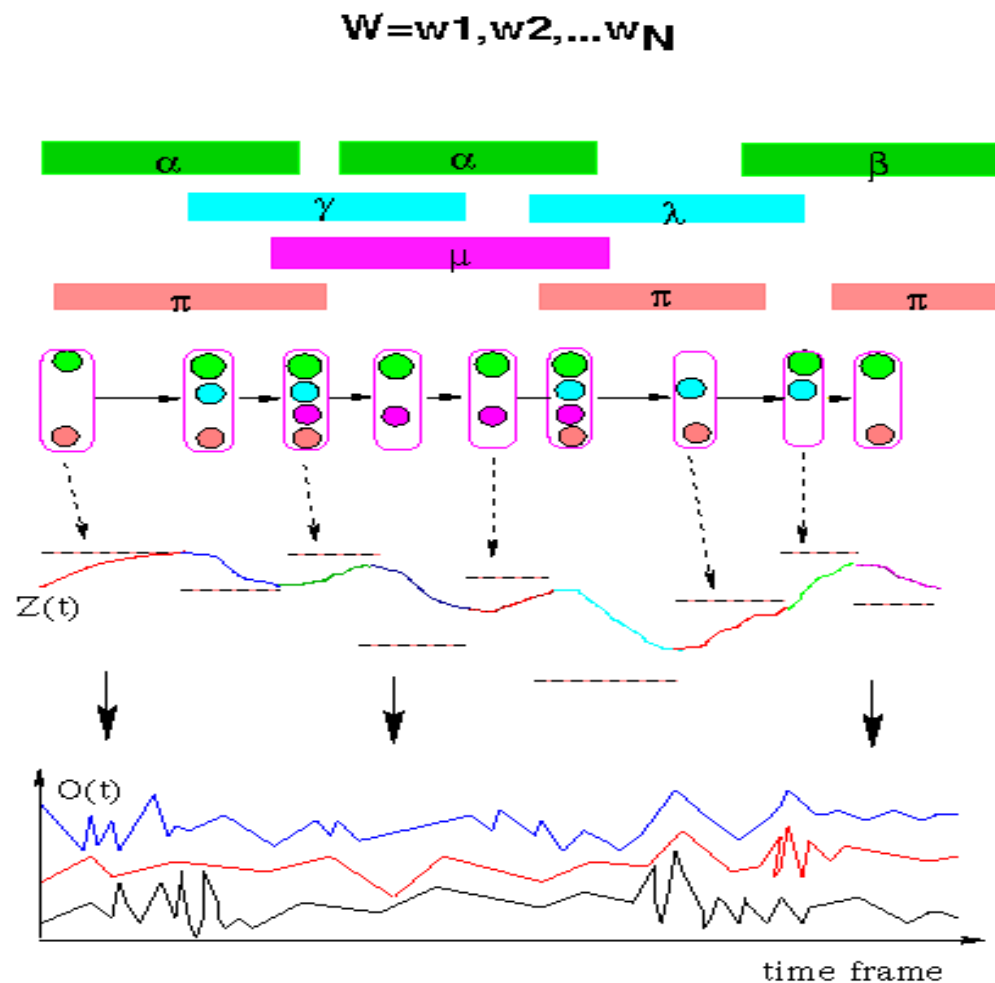
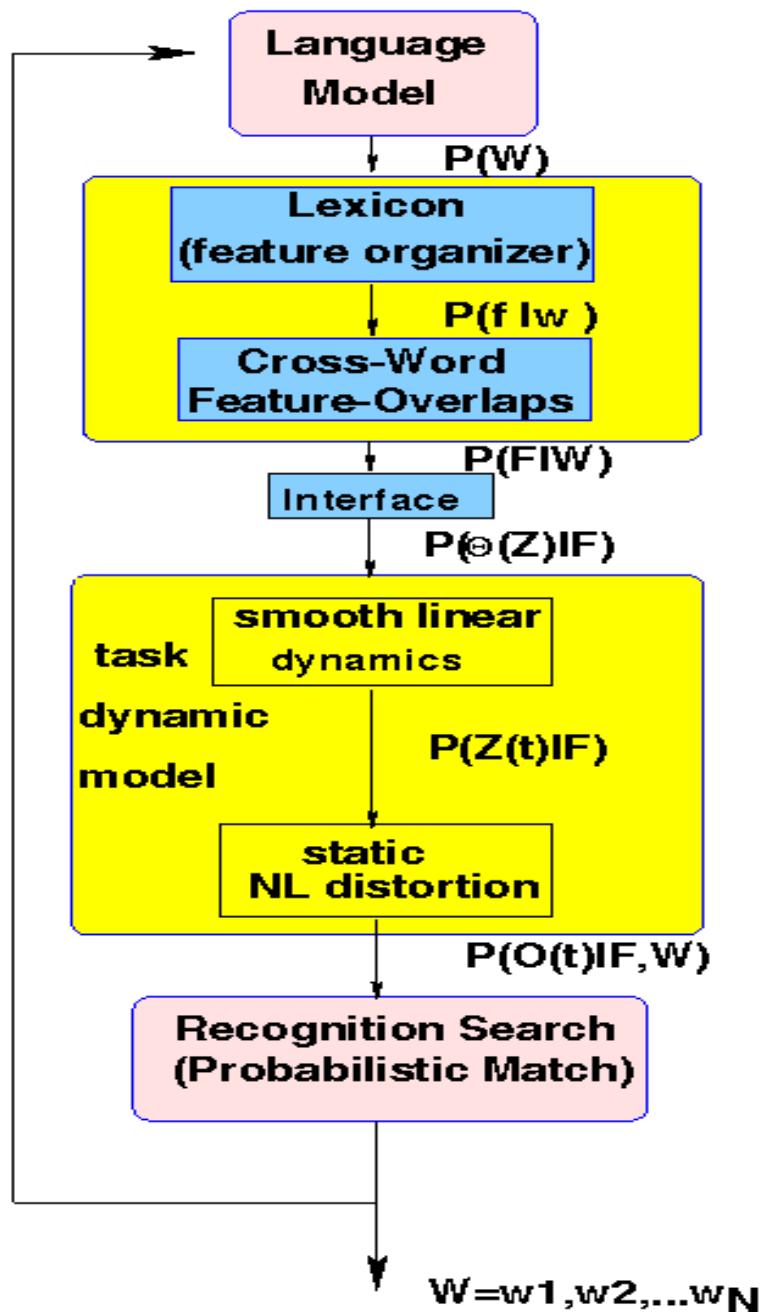
Compute your Vocal Tract Resonances in your running speech

File



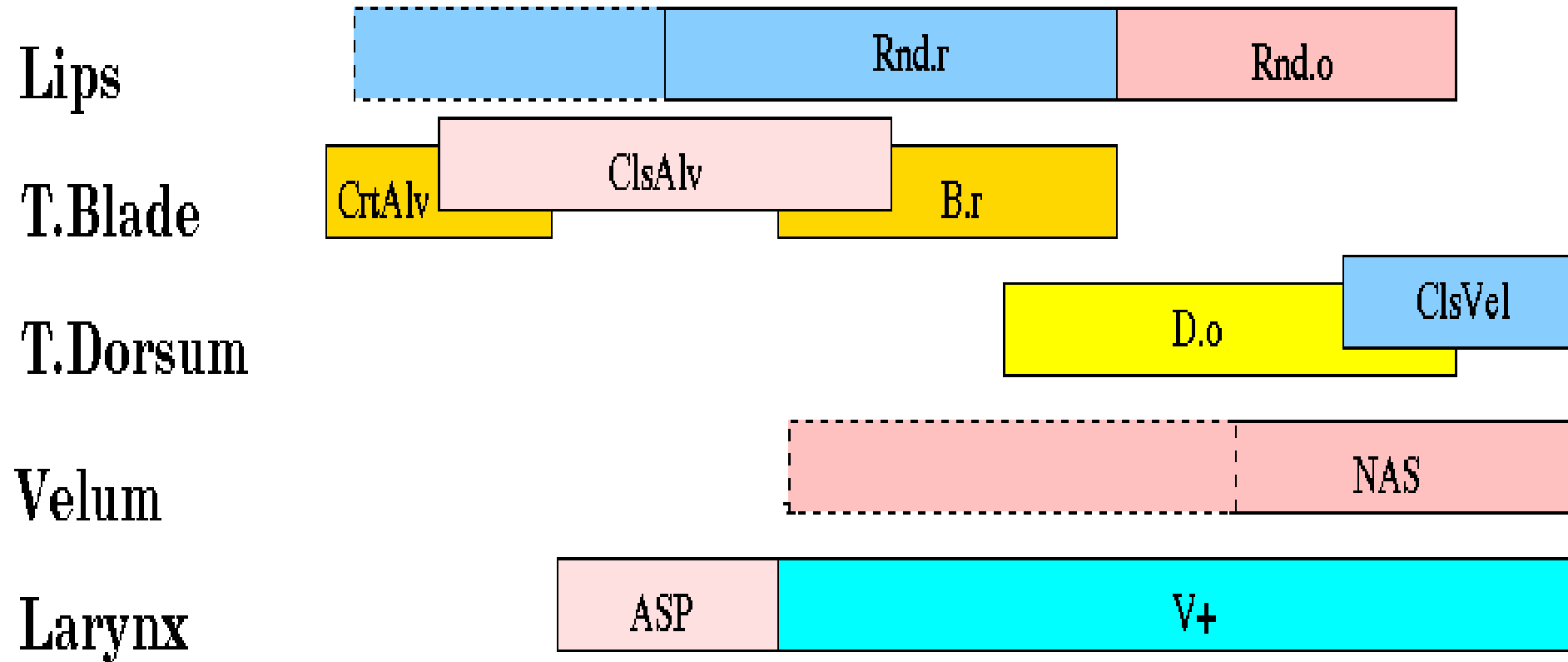
- Record Audio
- Play Audio
- Track Formants
- Refine Tracker
- Load Models
- Save Models





Feature overlapping: Box diagram

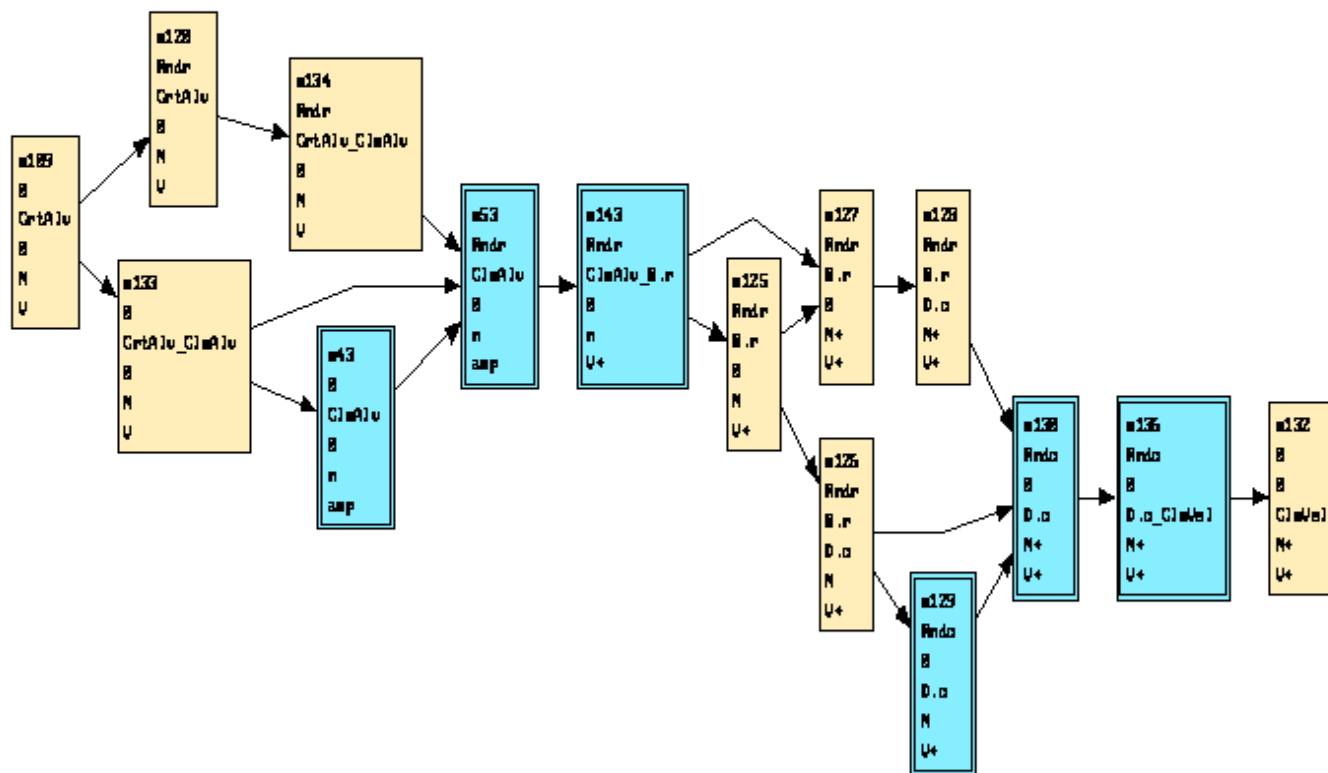
- *Long-range, optional feature spreading in Lip-Rounding & Nasalization*
- *Word: strong*

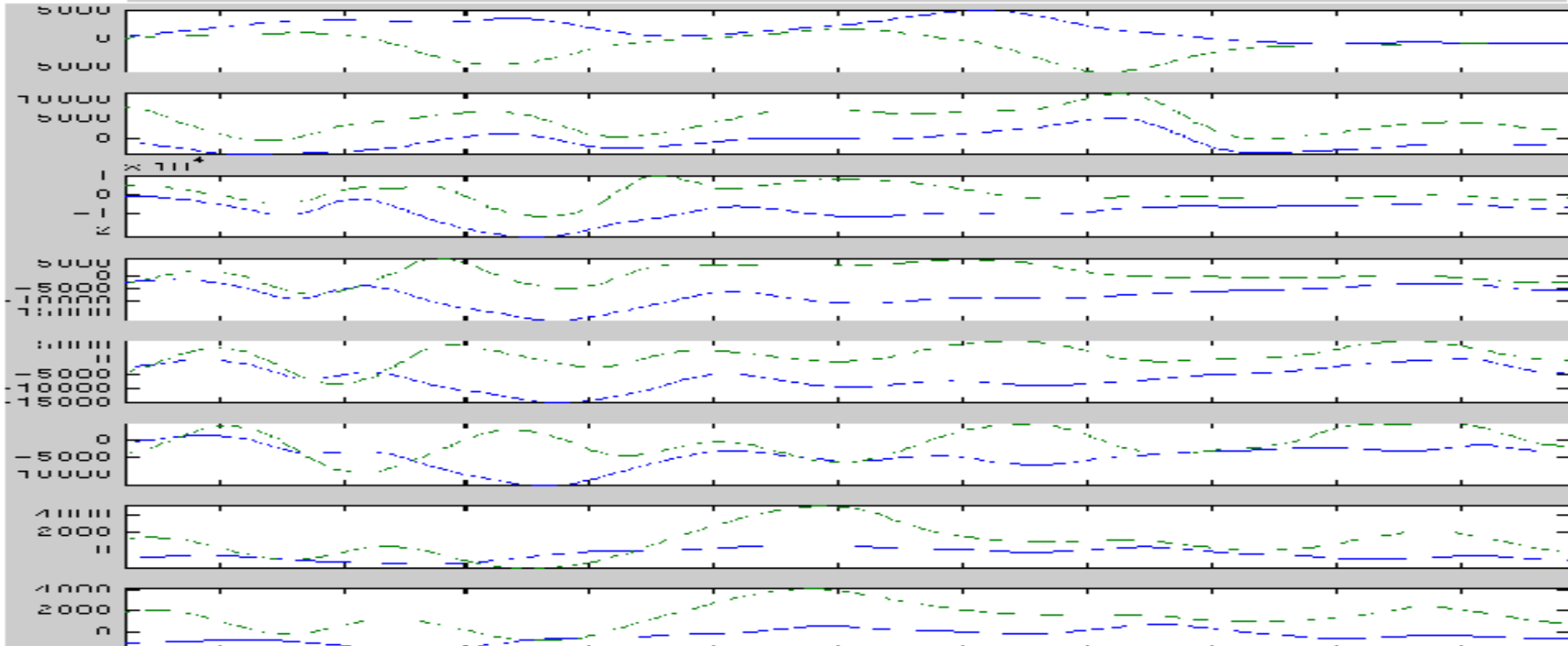
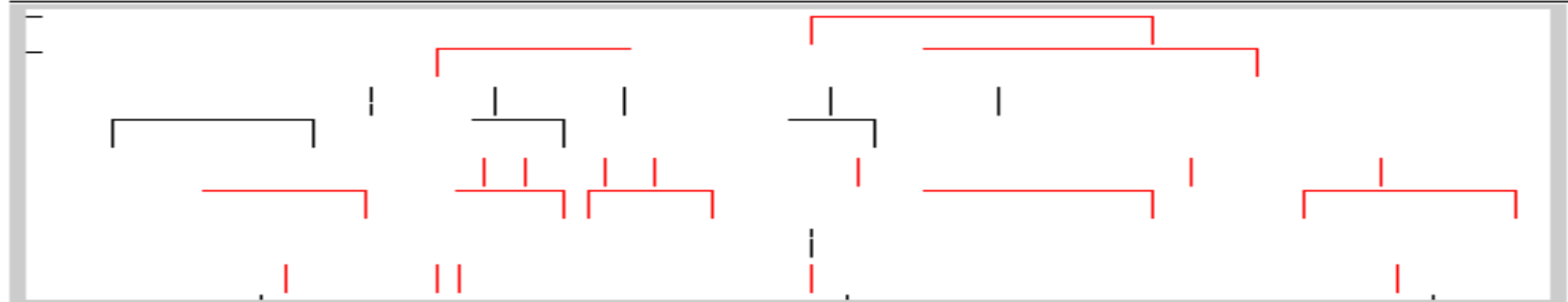
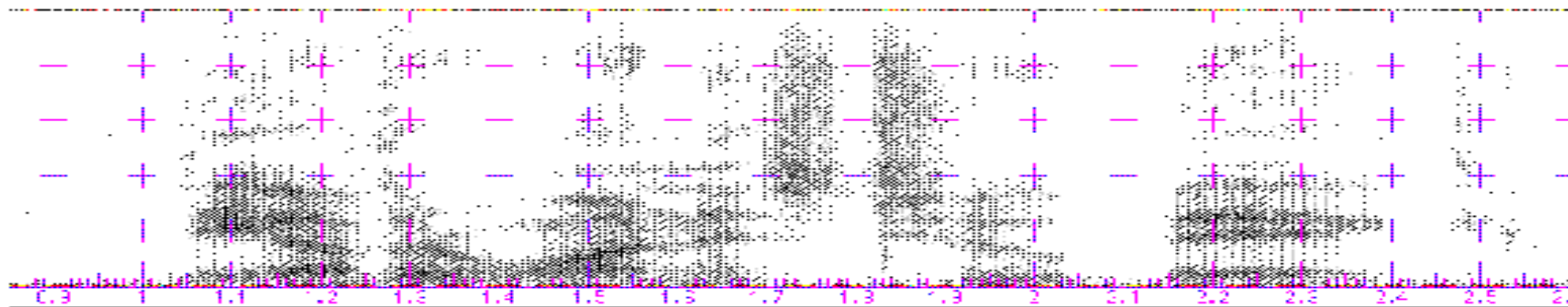


strong (/ s t r a o n g /)

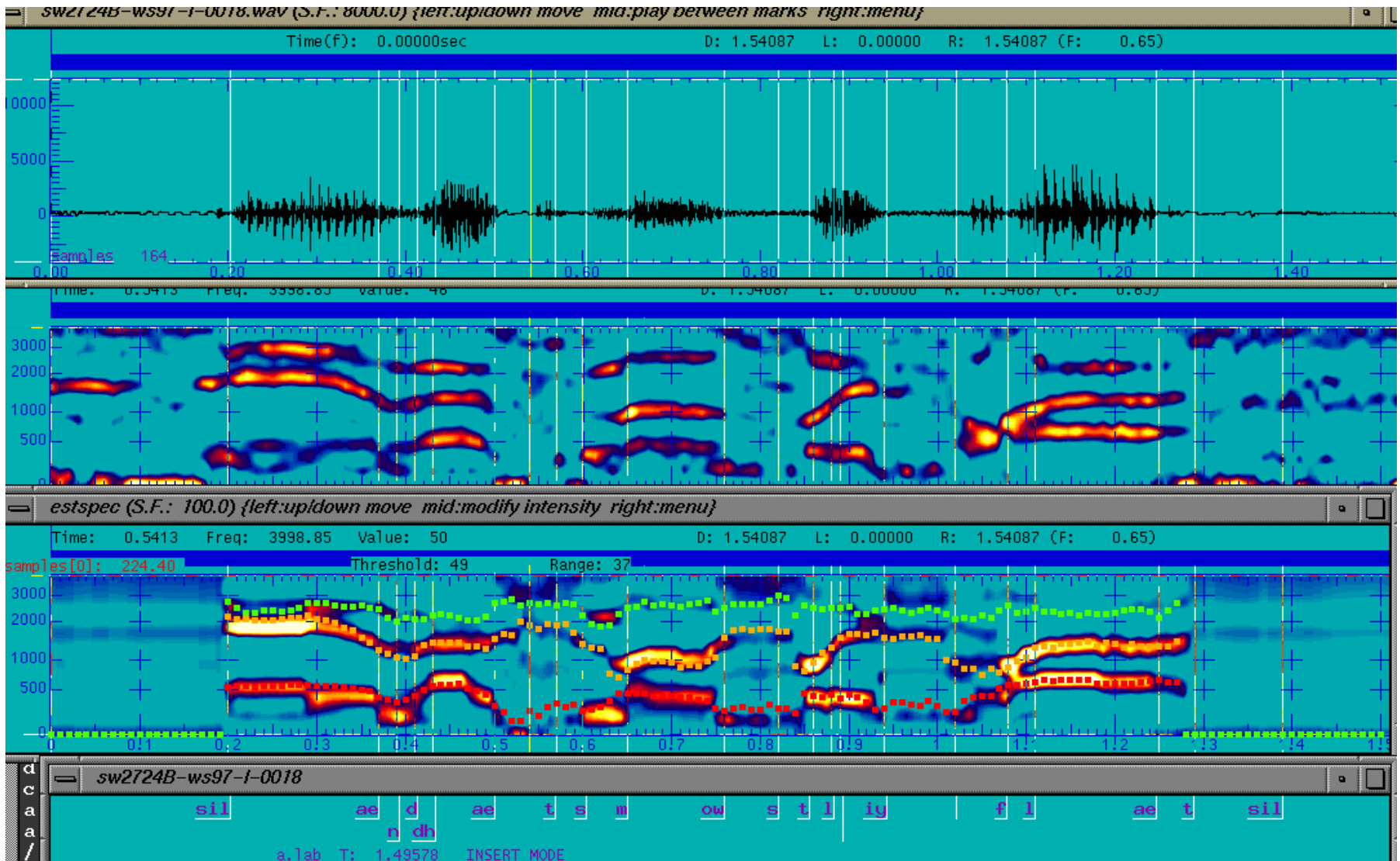
Feature Overlapping: HMM State-Transition Graph

- Example word *strong* (/ s t r o ŋ /)





Example: hidden dynamics (Vocal Tract Resonance)



Feature Overlapping: Box-Diagram & State Graph

- Example word *display* (/ d I - s p l e y /)

Lips

ClsLab

T. Blade

ClsAlv

CrtAlv

B.l

T. Dorsum

D.ih

D.c

D.j

Velum

Larynx

V+

ASP

V+

display (/d I- s p l e y/)

