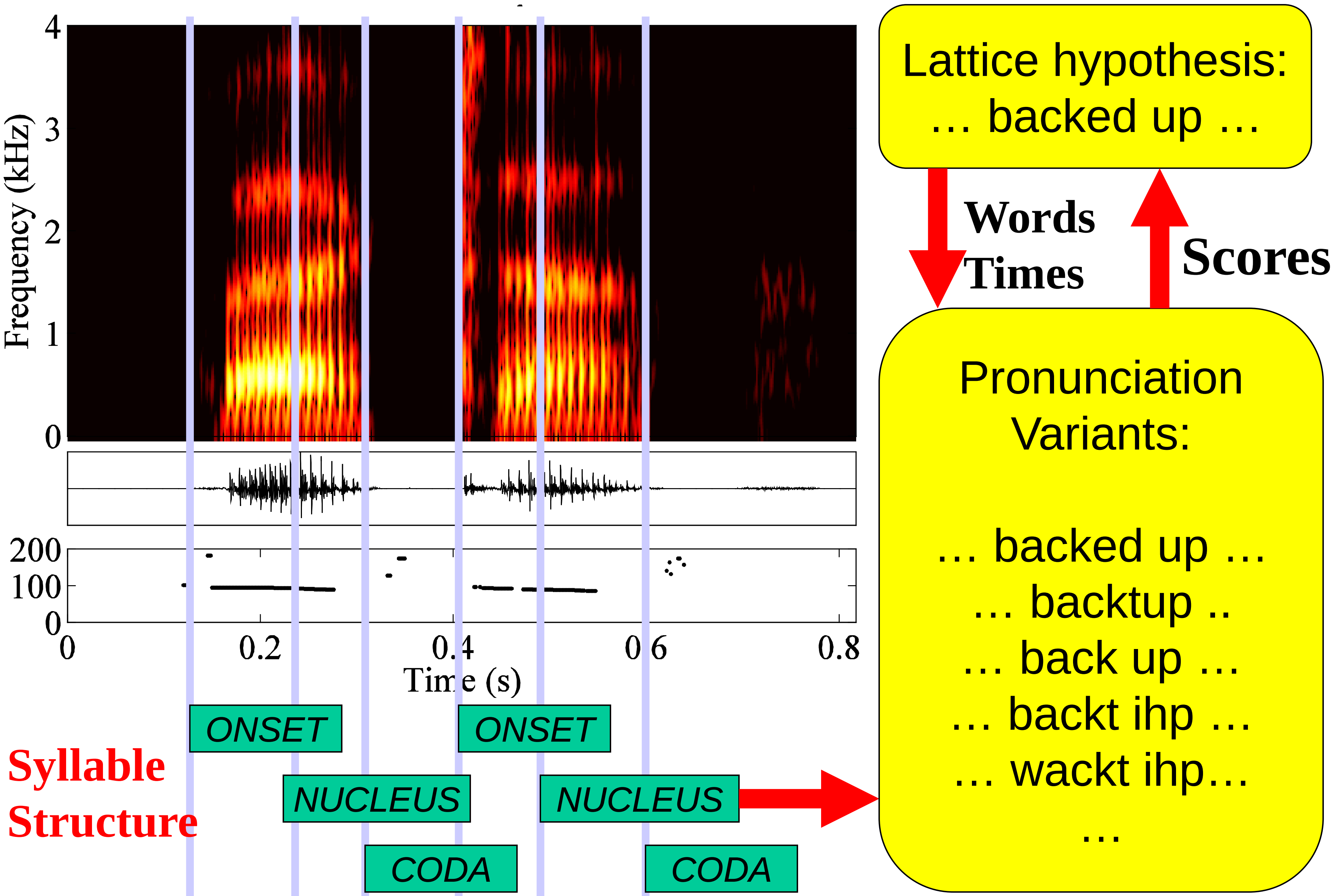
Landmark-Based Speech Recognition

Mark Hasegawa-Johnson
Jim Baker
Steven Greenberg
Katrin Kirchhoff
Jen Muller
Kemal Sonmez
Ken Chen
Amit Juneja
Karen Livescu
Srividya Mohan
Sarah Borys
Tarun Pruthi
Emily Coogan
Tianyu Wang

Goals of the Workshop

- “Acoustic Modeling:”
  - **Manner change landmarks**: 15 binary SVMS
  - **Place of articulation**: currently 33 binary SVMs, dependent on manner (t-50ms,…,t+50ms)
- “Lexical Modeling:”
  - Dictionary implemented using current version of GMTK
  - “Streams” in the dictionary: settings of lips, tongue blade, tongue body, velum, larynx
  - Dependencies in GMTK learn the synchronization among the five articulators.
- Evaluation: Lattice rescoring (EARS RT03)
  - Improve 1-Best WER

Landmark-Based Speech Recognition



Outline

- Scientific Goals of the Workshop
- Resources
  - Speech data
  - Acoustic features
  - Distinctive feature probabilities: trained SVMs
  - Pronunciation models
  - Lattice scoring tools
  - Lattices
- Preliminary Experiments
- Planned Experiments

Scientific Goals of the Workshop

- Acoustic
  - Learn precise and generalizable models of the acoustic boundary associated with each distinctive feature,
  - … in an acoustic feature space including representative samples of spectral, phonetic, and auditory features,
  - … with regularized learners that trade off training corpus error against estimated generalization error in a very-high-dimensional model space
- Phonological
  - Represent a large number of pronunciation variants, in a controlled fashion, by factoring the pronunciation model into distinct articulatory gestures,
  - … by integrating pseudo-probabilistic soft evidence into a Bayesian network
- Technological
  - A lattice-rescoring pass that reduces WER

Data Resources: Speech Data

	Size	Phonetic Transcr.	Word Lattices
NTIMIT	14hrs	manual	-
WS96&97	3.5hrs	Manual	-
SWB1 WS04 subset	12hrs	auto-SRI	BBN
Eval01	10hrs	-	BBN & SRI
RT03 Dev	6hrs	-	SRI
RT03 Eval	6hrs	-	SRI

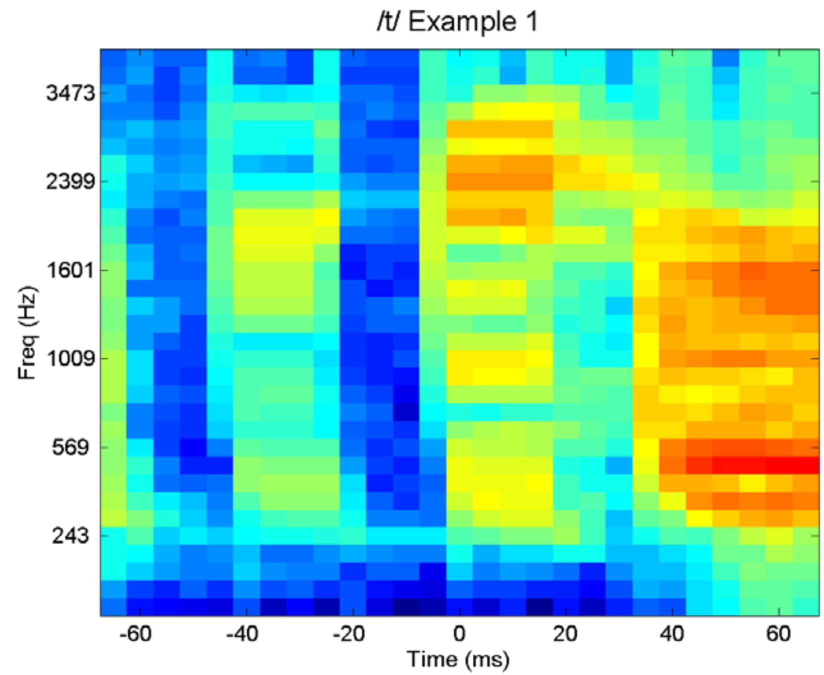
Data Resources: Acoustic Features

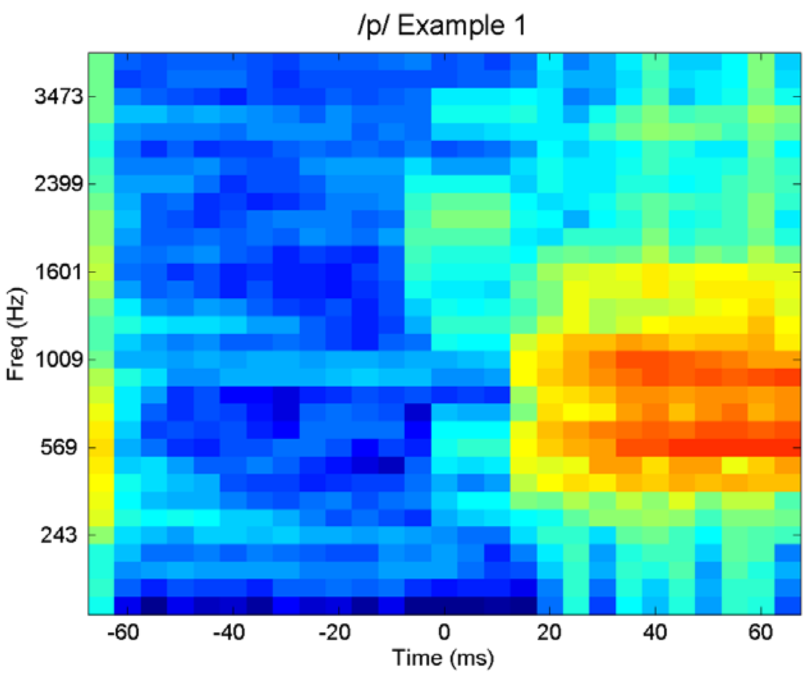
- MFCCs
  - 5ms skip, 25ms window (standard ASR features)
  - 1ms skip, 4ms window (equivalent to calculation of energy, spectral tilt, and spectral compactness once/millisecond)
- Formant frequencies, once/5ms
  - ESPS LPC-based formant frequencies and bandwidths
  - Zheng MUSIC-based formant frequencies, amplitudes, and bandwidths
- Espy-Wilson Acoustic Parameters
  - sub-band aperiodicity, sonorancy, other targeted measures
- Seneff Auditory Model: Mean rate and synchrony
- Shamma rate-place-sweep auditory parameters

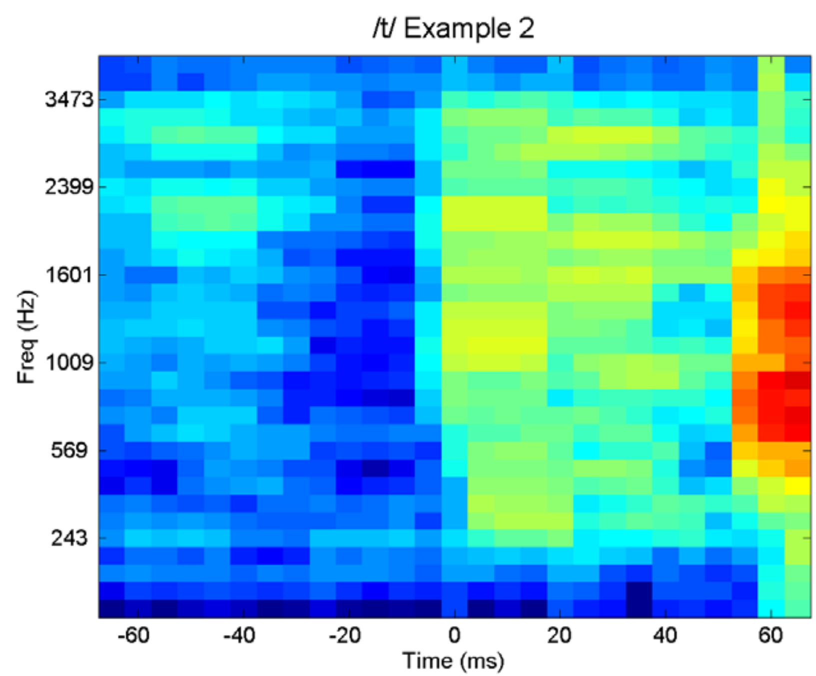
Background: a Distinctive Feature definition

- Distinctive feature = a binary partition of the phonemes
- Landmark = Change in the value of an “Articulator-Free Feature” (a.k.a. manner feature)
  - +speech to –speech, –speech to +speech
  - consonantal, continuant, sonorant, syllabic
- “Articulator-Bound Features” (place and voicing): SVMs are only trained at landmarks
  - Primary articulator: lips, tongue blade, or tongue body
  - Features of primary articulator: anterior, strident
  - Features of secondary articulator: voiced

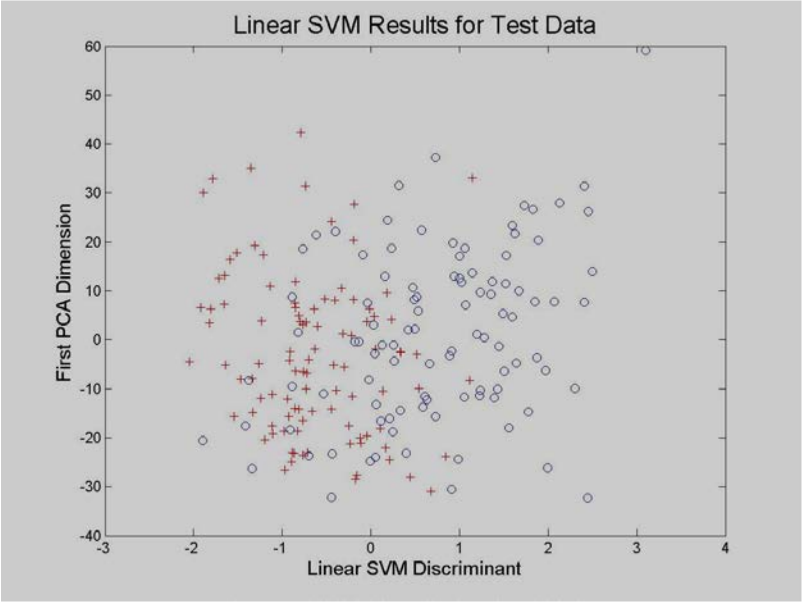
Place of Articulation: cued by the WHOLE PATTERN of spectral change over time within 150ms of a landmark

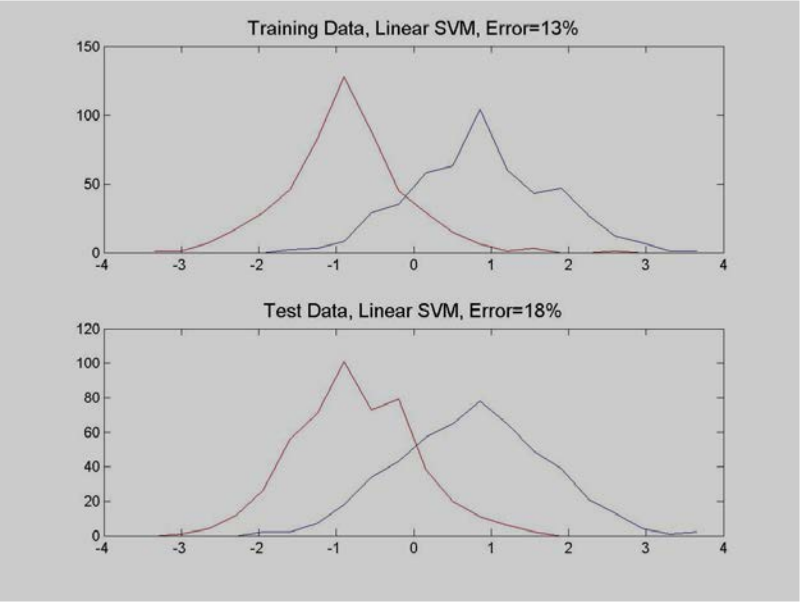




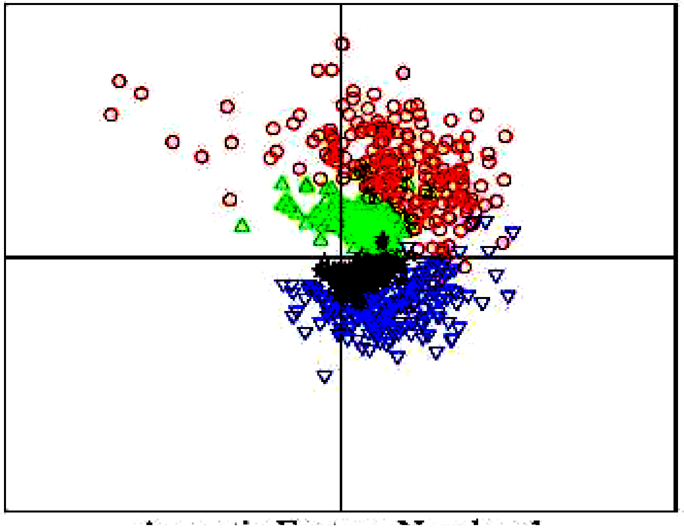


Software Resources: SVMs trained for binary distinctive feature classification

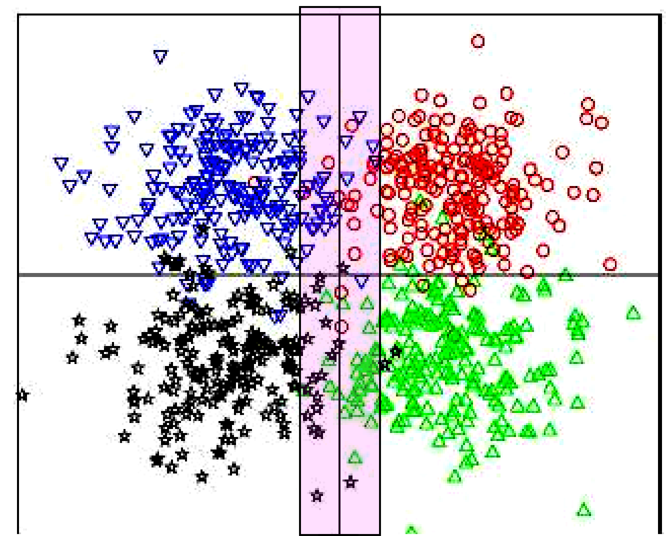




Software Resources: Posterior probability estimator based on SVM discriminant

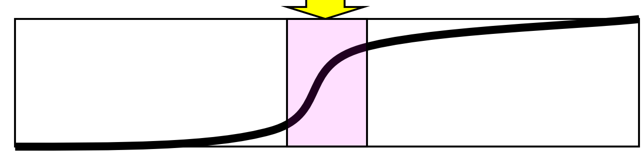


Kernel: Transform to Infinite-Dimensional Hilbert Space

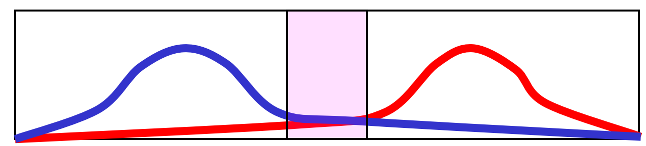


SVM Extracts a Discriminant Dimension

(SVM Discriminant Dimension = argmin(error(margin)+1/width(margin))



(Niyogi & Burges, 2002: Posterior PDF = Sigmoid Model in Discriminant Dimension)



An Equivalent Model: Likelihoods = Gaussian in Discriminant Dimension

Data Resources: 33-track Distinctive Feature Probs, NTIMIT, ICSI, 12hr, RT03

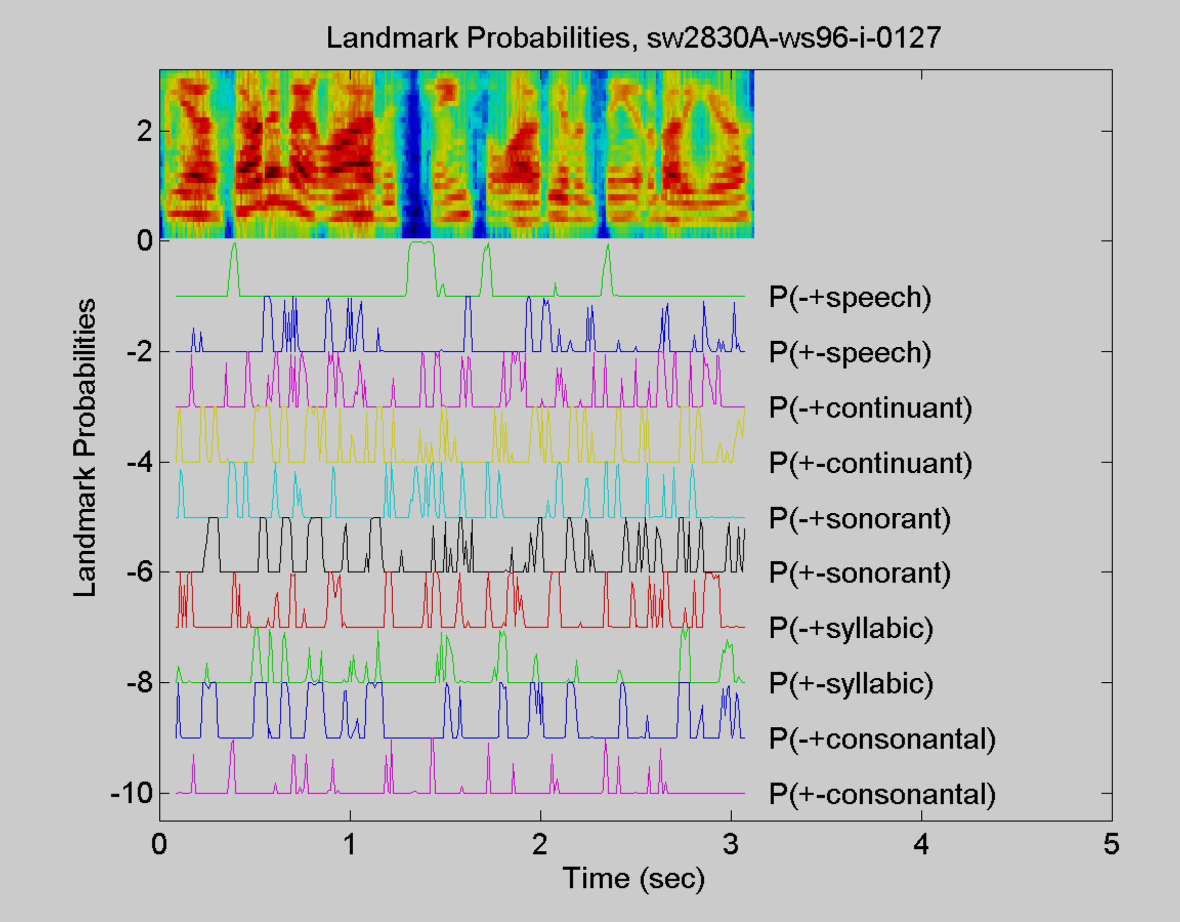
2000-dimensional acoustic feature vector

SVM

Discriminant $y_i(t)$

Sigmoid or Histogram

Posterior probability of distinctive feature $p(d_i(t)=1 \mid y_i(t))$



Lexical Resources: Landmark-Based Lexicon

- Merger of English Switchboard and Callhome dictionaries
- Converted to landmarks using Hasegawa-Johnson's perl transcription tools

Landmarks in blue, Place and voicing features in green.

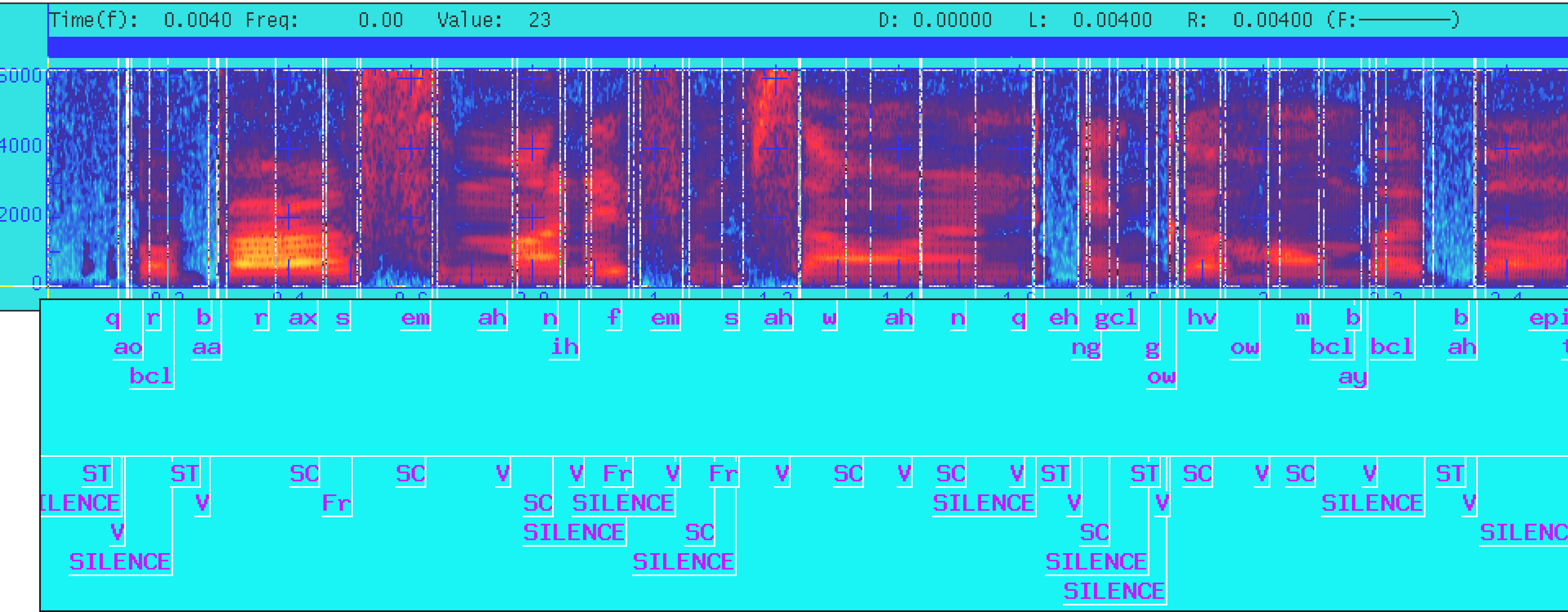
AGO(0.441765)	**+syllabic +reduced +back**	**AX**
	+–continuant +– sonorant +velar +voiced	**G closure**
	–+continuant –+sonorant +velar +voiced	**G release**
	+syllabic –low –high +back +round +tense	**OW**
AGO(0.294118)	**+syllabic +reduced –back**	**IX**
	–+ continuant –+sonorant +velar +voiced	**G closure**
	–+continuant –+sonorant +velar +voiced	**G release**
	+syllabic –low –high +back +round +tense	**OW**

Software Resources: DP smoothing of SVM outputs

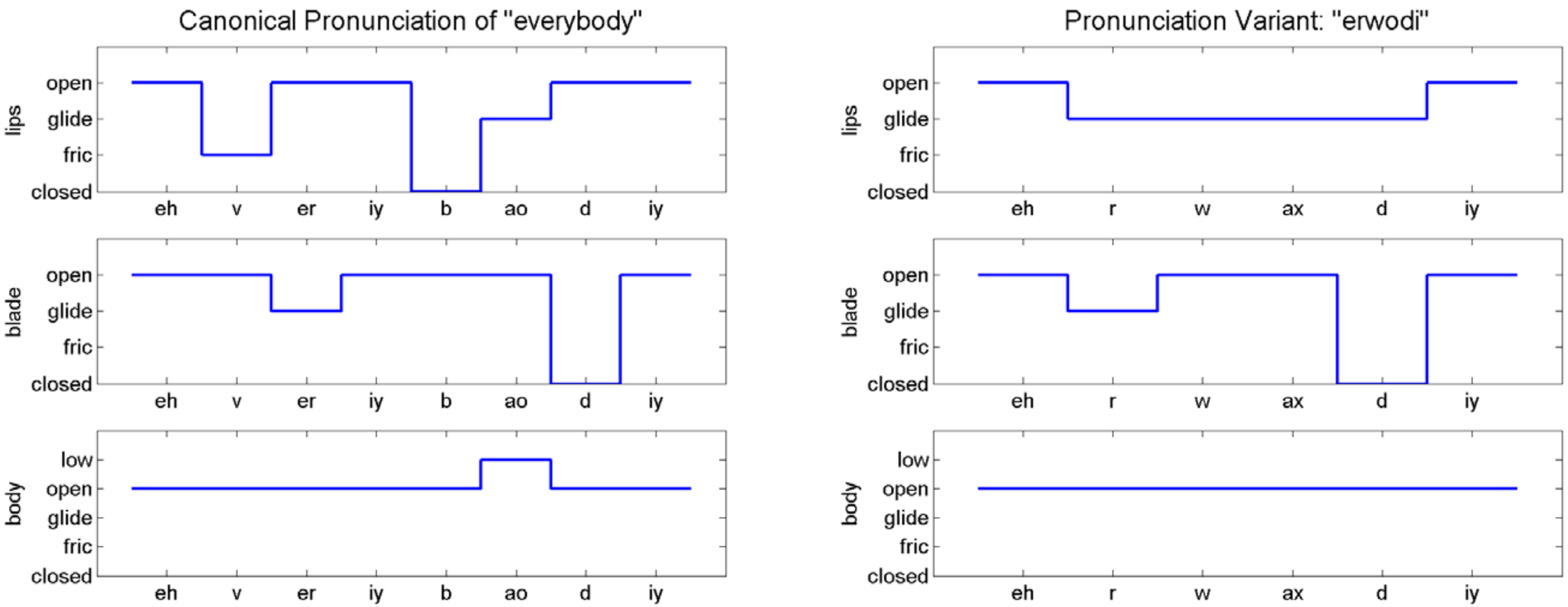
- Maximize $\Pi_i\ p(\ \text{features}(t_i) \mid X(t_i)\ )\ p(t_{i+1}-t_i \mid \text{features}(t_i))$
- Forced alignment mode:

  computes p(word | acoustics); rescores the word lattice
- Manner class recognition mode:

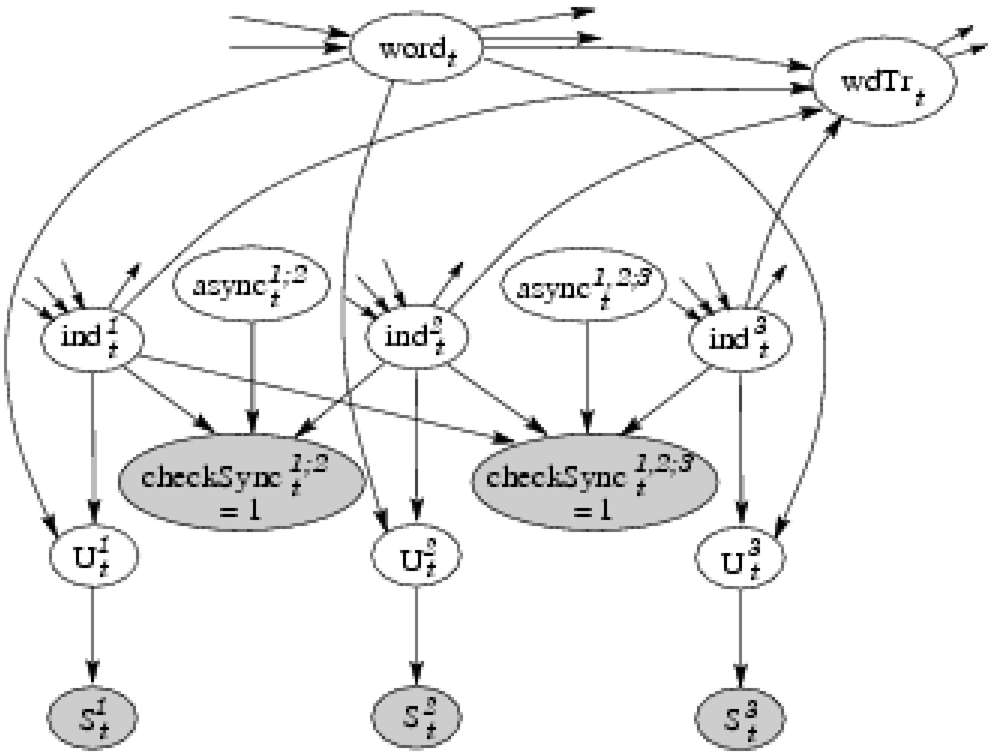
  smooths SVM output; preprocessor for the DBN



Software Resources: Dynamic Bayesian Network model of pronunciation variability



DBN model: a bit more detail



- word_t: word ID at frame #t
- wdTr_t: word transition?
- ind_t^i: which gesture, from the canonical word model, should articulator i be trying to implement?
- $\text{async}_t^{i;j}$: how asynchronous are articulators i and j?
- U_t^i: canonical setting of articulator #i
- S_t^i: surface setting of articulator #i

Lattice Rescoring Resources: SRILM, finite state toolkit, GMTK

- Lattice annotation: each word carries multiple scores
  - Original language model
  - Original acoustic model
  - DP-smoothed SVM scores
  - DBN scores
- Integration: weighted sum of log probabilities?
  - N-best lists vs. lattices
  - Stream-weight optimization: amoeba search?
  - How many different scores? Bayesian justification?

Lattices

- RT03: lattices from SRI
  - 72 conversations (12 hours), Fisher & Switchboard
  - Development test and Evaluation subcorpora
  - Devel set: WER=24.1%
- EVAL01: lattices from BBN
  - 60 conversations (10 hours), Switchboard
  - Evaluation corpus only
  - WER=23.5%

Lattices: Analysis

- RT03 development test lattices:
  - SUB=13.4%, INS=2.2%, DEL=8.5%
  - Function words account for most substitutions:
    - it→that,99 (1.78%); the→a,68 (1.22%); a→the,68 (1.03%)
    - and→in,64 (1.15%); that→the,40 (0.72%); the→that,35 (0.63%)
  - Percent of word substitutions involving the following errors:
    - Insertions of Onset 23%, Vowel 15%, Coda 13%
    - Deletions of Onset 29%, Vowel 17%, Coda 3%
    - Place Error of Onset 9.6%, Vowel 15.8%, Coda 9.6%
    - Manner Error of Onset 20.1%, Coda 20.2%

Lattice Rescoring Experiment, Week 0:

- Unconstrained DP-smoothing of SVM outputs,
- … integrated using a DBN that allows asynchrony of constrictions, but not reduction,
- … used to compute a new “SVM-DBN” score for each word in the lattice,
- … added to the existing language model and acoustic model scores (with stream weight of 1.0 for the new score)

"Week 0" Lattice Rescoring Results & Examples

- Reference transcription:

yeah I bet that restaurant was but what how did the food taste

- Original lattice, WER=76%:

yeah but that that's what I was traveling with how the school safe

- SVM-DBN acoustic scores replace original acoustic scores, WER=69%:

yeah yeah but that restrooms problems with how the school safe

- Analysis (speculative, with just one lattice…):
  - **SVM improves syllable count:**
    - **"yeah I but" → "yeah yeah but" vs. "yeah but"**
  - **SVM improves recognition of consonants:**
    - **"restaurant"→"restrooms" vs. "that's what I"**
  - **SVM currently has NO MODEL of vowels**
  - **In this case, the net result is a drop in WER**

Schedule of Experiments: Current Problem Spots

- Combination of language model, HMM acoustic model, and SVM-DBN acoustic model scores!!!
  - Solving this problem may be enough to get a drop in WER!!!
- Pronunciation variability vs. DBN computational complexity
  - Current model: asynchrony allowed, but not reductions, e.g., stop→glide
  - Current computational complexity ~720XRT
  - Extra flexibility (e.g., stop→glide reductions) desirable but expensive
- Accuracy of SVMs:
  - Landmark detection error, S+D+I: 20%
  - Place classification error, S: 10-35%
  - Already better than GMM, but still worse than human listeners. Is it already good enough? Can it be improved?

Schedule of Experiments

- July 12 targets:
  - SVMs using all acoustic observations
  - Write scripts to automatically generate word scores and annotate n-best list
  - N-best-list streamweight training
  - Complete rescoring experiment for RT03-development n-best lists
- July 19 targets:
  - Error-analysis-driven retraining of SVMs
  - Error-analysis-driven inclusion of closure-reduction arcs into the DBN
  - Second rescoring experiment
  - Error-analysis driven selection of experiments for weeks 3-4
- August 7 targets:
  - Ensure that all acoustic features and all distinctive feature probabilities exist for RT03/Evaluation
  - Final experiments to pick best novel word scores
- August 10 target: Rescoring pass using RT03/evaluation lattices
- August 16 target: Dissect results: what went right? What went wrong?

Summary

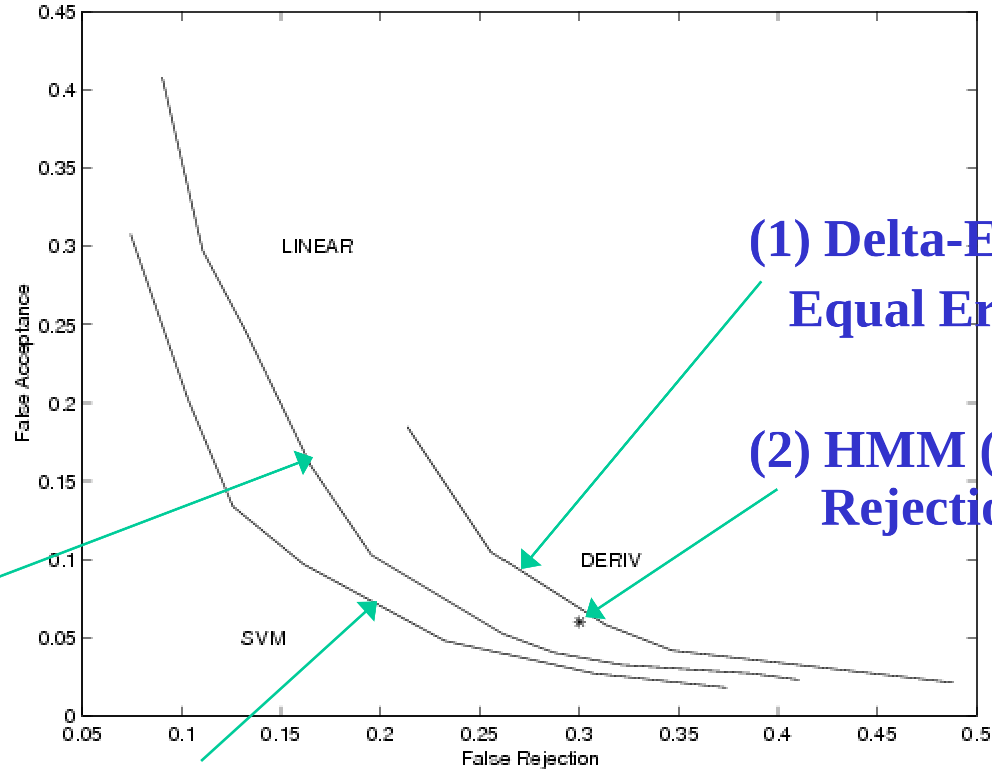
- Acoustic modeling:
  - **Target problem:** **2000-dimensional** observation space
  - **Proposed method**: **regularized learner** (SVM) to explicitly control tradeoff between training error & generalization error
  - **Resulting constraints:**
    - Choice of binary distinctions is important: choose **distinctive features**
    - Choice of time alignment is important: train place SVMs at **landmarks**
- Lexical modeling:
  - **Target problem**: increase flexibility of pronunciation model **without over-generating** pronunciation variants
  - **Proposed method**: factor the probability of pronunciation variants into misalignment & reduction probabilities of 5 **hidden articulators**
  - **Resulting constraints:**
    - Choice of factors is important: choose **articulatory factors**
    - Integration of **SVMs into Bayesian model** is an interesting problem
- Lattice rescoring:
  - **Target problem**: integrate **word-level side information** into a lattice
  - **Proposed method**: **amoeba search** optimization of stream weights
  - **Potential problems**: amoeba search may only work for **N-best lists**

Extra Slides

Stop Detection using Support Vector Machines

False Acceptance vs. False Rejection Errors, TIMIT, per 10ms frame

SVM Landmark Detector: Half the Error of an HMM



(1) Delta-Energy ("Deriv"): Equal Error Rate = 0.2%

(2) HMM (*): False Rejection Error=0.3%

(3) Linear SVM: EER = 0.15%

(4) Kernel SVM: Equal Error Rate=0.13%

Niyogi & Burges, 1999, 2002

Manner Class Recognition Accuracy in TIMIT (errors per phoneme)

	13 Mixture HMM (22716 parameters) (Borys & Hasegawa-Johnson)	Landmark SVM (**160** parameters) (Juneja & Espy-Wilson)
speech vs. silence	80.2%	94.1
+vocalic vs. –vocalic	77.8	78.9
+sonorant vs. –sonorant	77.8	93.4
+continuant vs. –continuant	77.0	93.7
Vowel vs. Glide vs. Nasal vs. Stop vs. Fricative	73.5	**79.8**

Place Classification Accuracy, RBF SVMs observing MFCCs+formants in 110ms window at consonant release

	TIMIT	NTIMIT	ICSI Switchboard
Lips-stop	95.0%	90.5	83.1
Blade-stop	85.1	83.3	63.7
Body-stop	87.2	88.1	82.1
Lips-fric			89.9
Anterior			87.8
Strident			82.3