Feature/Landmark-based Pronunciation Modeling using Dynamic Bayesian Networks

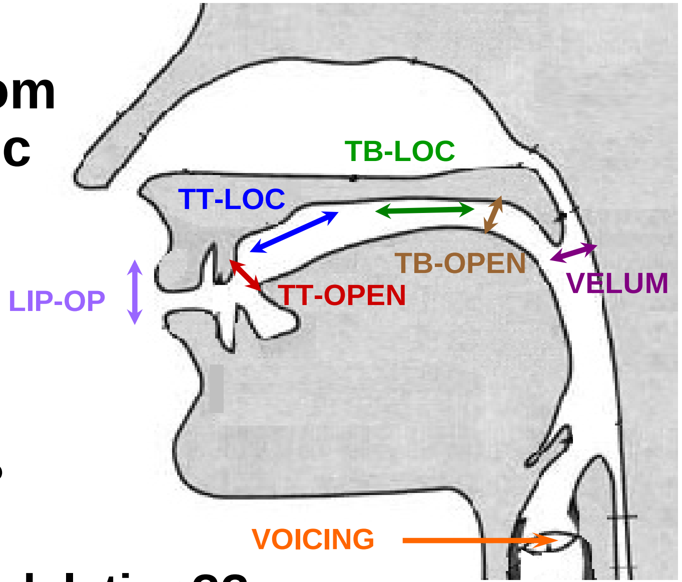
Karen Livescu

Outline

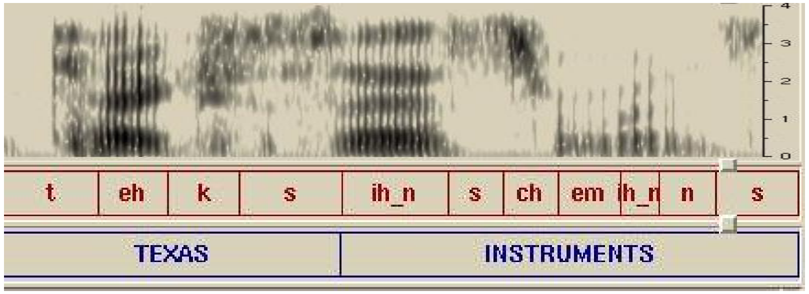
- Motivation
- A feature-based pronunciation model
- Using SVM outputs in the pronunciation model
- WS’04 experiments
- Observations and conclusions

Why feature-based pronunciation modeling?

- **Many pronunciation phenomena can be parsimoniously described as resulting from *asynchrony* and *reduction* of sub-phonetic features**
  - **One set of features based on articulatory phonology** [*Browman & Goldstein 1990*]**:**



- ***warmth* → [w ao r m p th] - Phone insertion?**
- ***I don't know* → [ah dx uh_n ow_n] - Phone deletion??**
- ***several* → [s eh r v ax l] - Exchange of two phones???**
- ***instruments* → [ih_n s ch em ih_n n s]** ***everybody* → [eh r uw ay]**



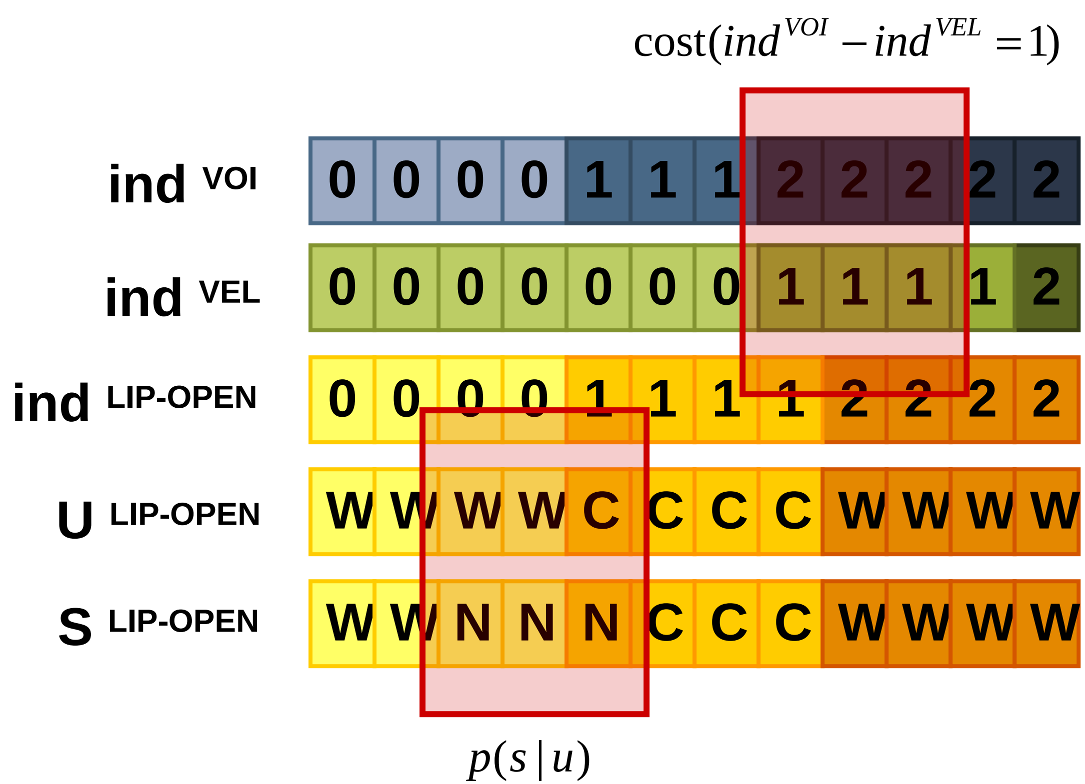
Approach: Main Ideas

baseform dictionary *everybody* →

index	0	1	2	3	…
phone	eh	v	r	iy	…
voicing	V	V	V	V	…
velum	Off	Off	Off	Off	…
lip opening	Wide	Crit	Wide	Wide	…
…	…	…	…	…	…

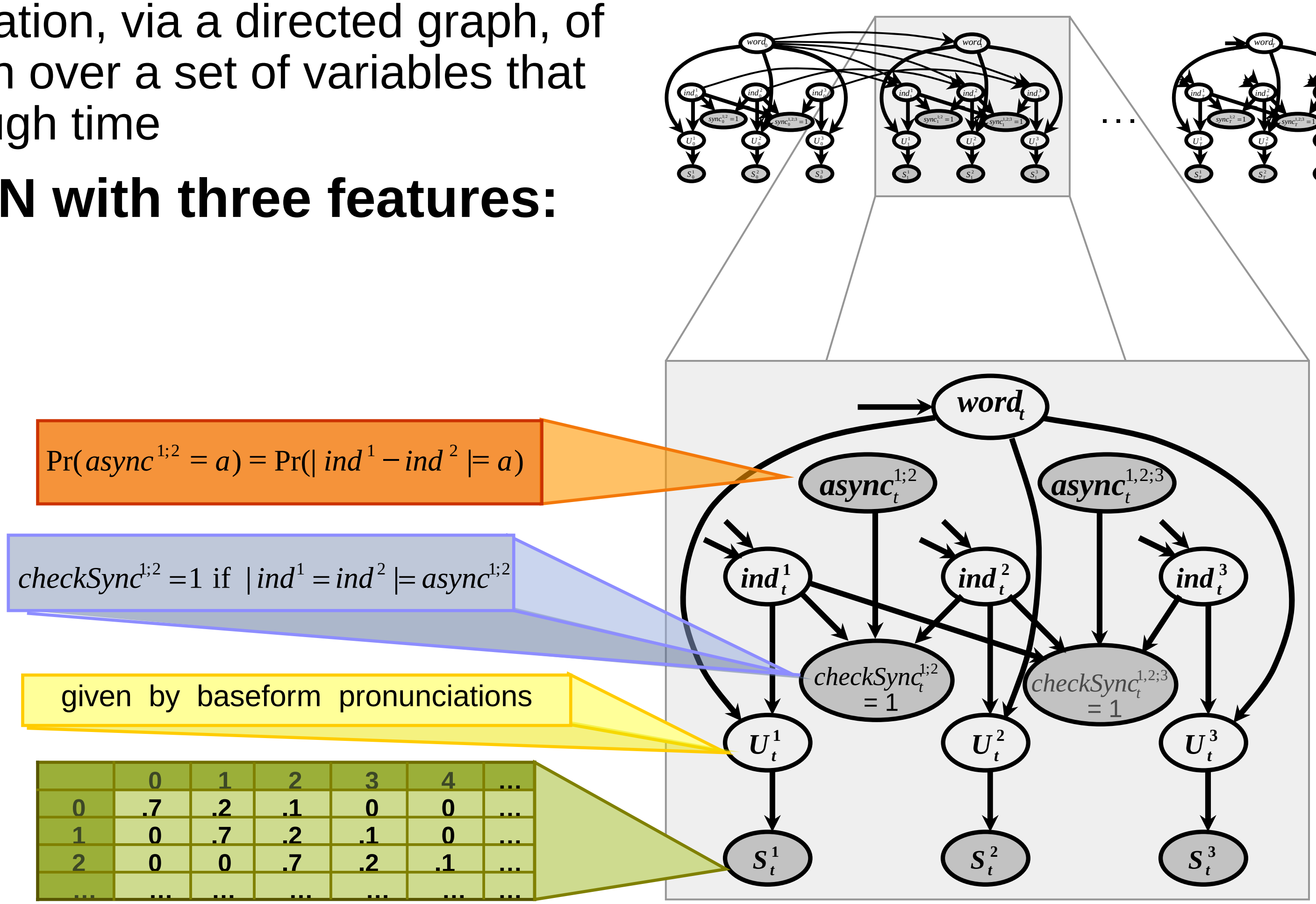
+ asynchrony

+ feature substitutions



A feature-based pronunciation model

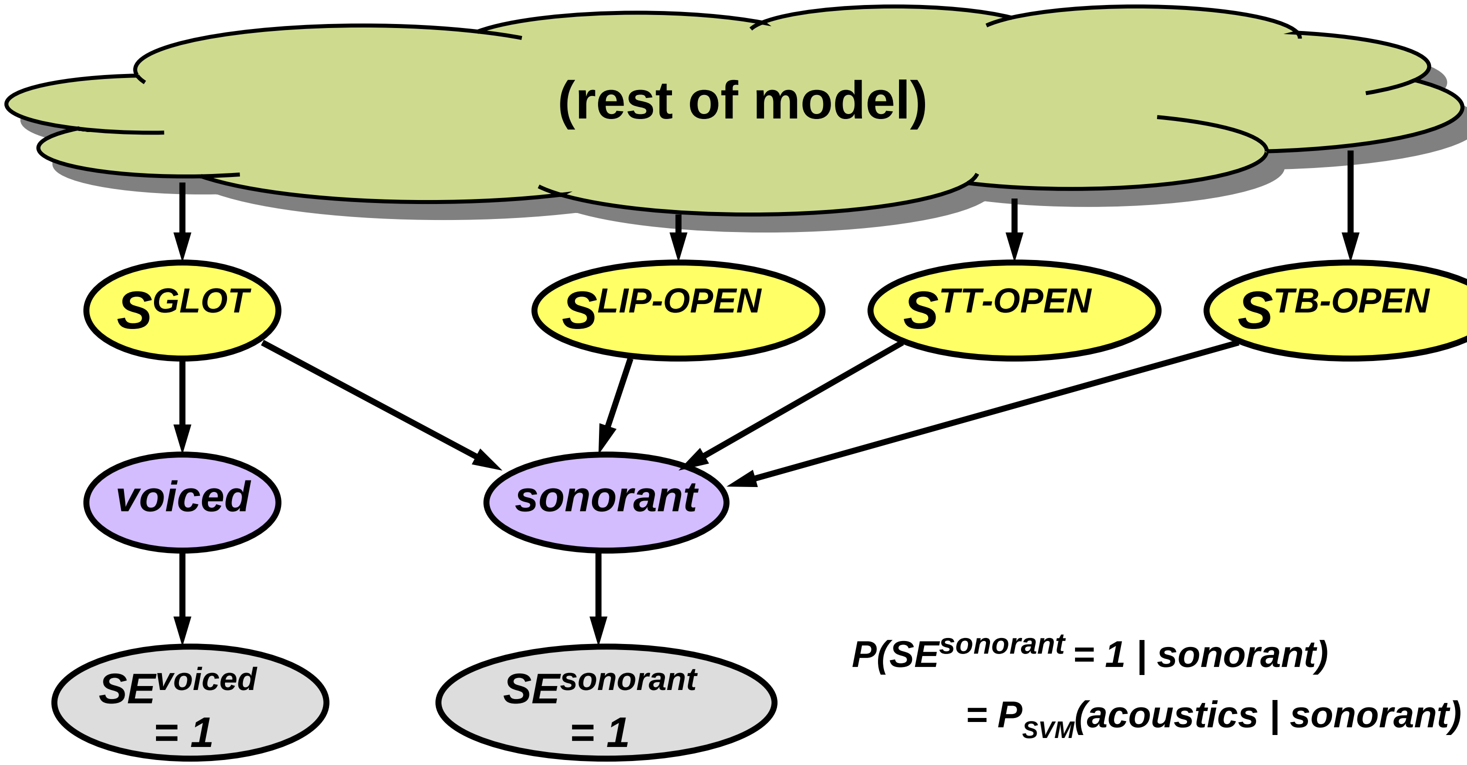
- **The model is implemented as a dynamic Bayesian network (DBN):**
  - A representation, via a directed graph, of a distribution over a set of variables that evolve through time
- **Example DBN with three features:**



	0	1	2	3	4	...
0	.7	.2	.1	0	0	...
1	0	.7	.2	.1	0	...
2	0	0	.7	.2	.1	...
...	...	...	...	...	...	...

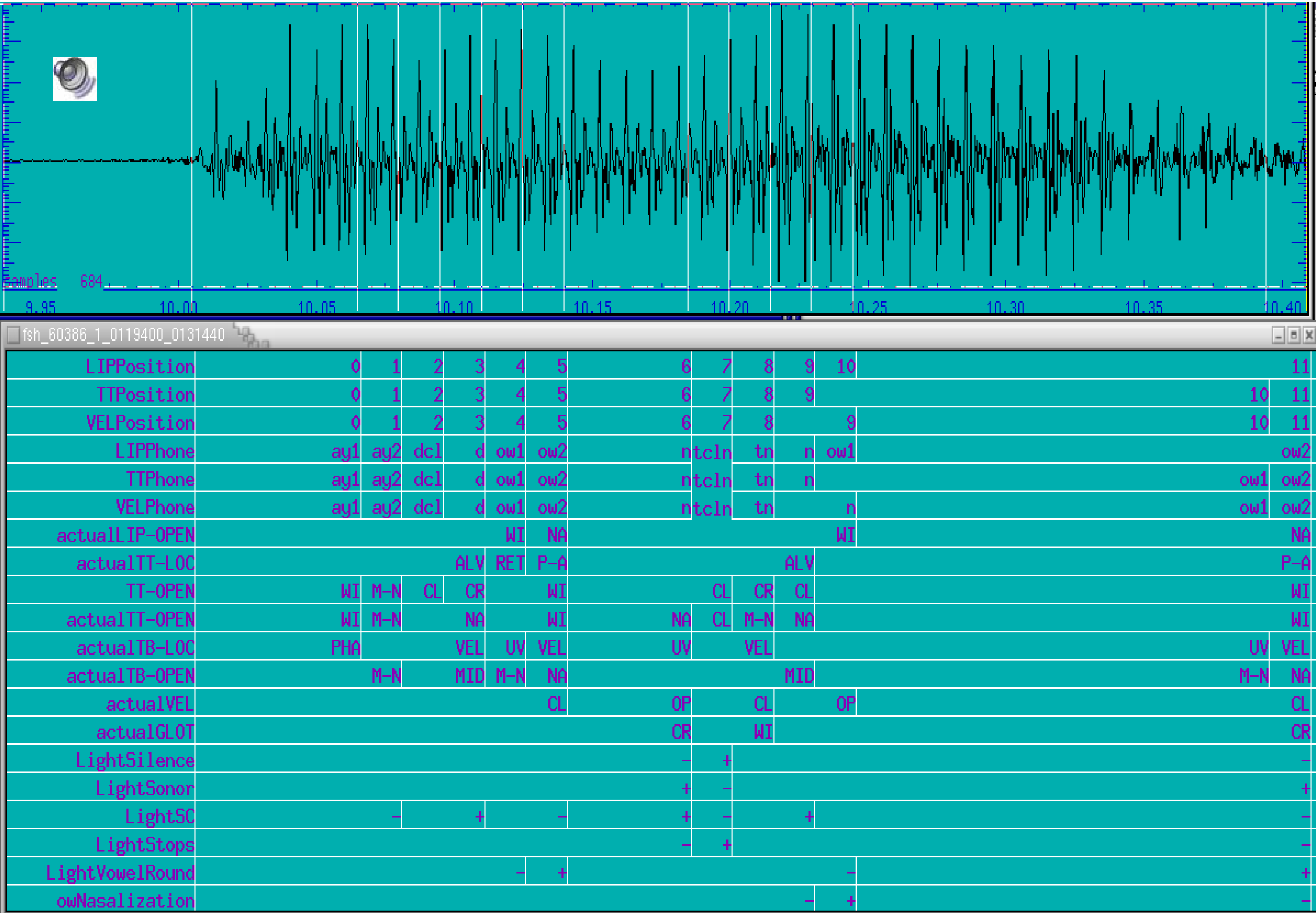
Combining SVM outputs with the DBN

- **Task 1: Converting between articulatory features and SVM distinctive features (DFs)**
  - **Method: Add DBN variables corresponding to DFs, and add deterministic mappings from surface articulatory variables to DFs**
- **Task 2: Incorporating SVM output probabilities**
  - **Method: Soft evidence – similar in spirit to HMM/ANNs**



$$P(SE^{sonorant} = 1 \mid sonorant) = P_{SVM}(acoustics \mid sonorant)$$

Example alignment using SVM/DBN



Design decisions

- **What kind of SVM outputs should be used in the DBN?**
  - **Method 1 (EBS/DBN): Generate landmark segmentation with EBS using manner SVMs, then apply place SVMs at appropriate points in the segmentation**
    - * Force DBN to use EBS segmentation
    - * Allow DBN to stray from EBS segmentation, using place/voicing SVM outputs whenever available
  - **Method 2 (SVM/DBN): Apply all SVMs in all frames, allow DBN to consider all possible segmentations**
    - * In a single pass
    - * In two passes: (1) manner-based segmentation; (2) place+manner scoring
- **How should we take into account the distinctive feature hierarchy?**
- **How do we avoid "over-counting" evidence?**
- **How do we train the DBN (feature transcriptions vs. SVM outputs)?**

A chronology of DBN/SVM rescoring experiments

- **For each lattice edge:**
  - SVM probabilities computed over edge duration and used as soft evidence in DBN
  - DBN computes a score $S \propto P(\text{word} \mid \text{evidence})$
  - Final edge score is a weighted interpolation of baseline scores and EBS/DBN or SVM/DBN score

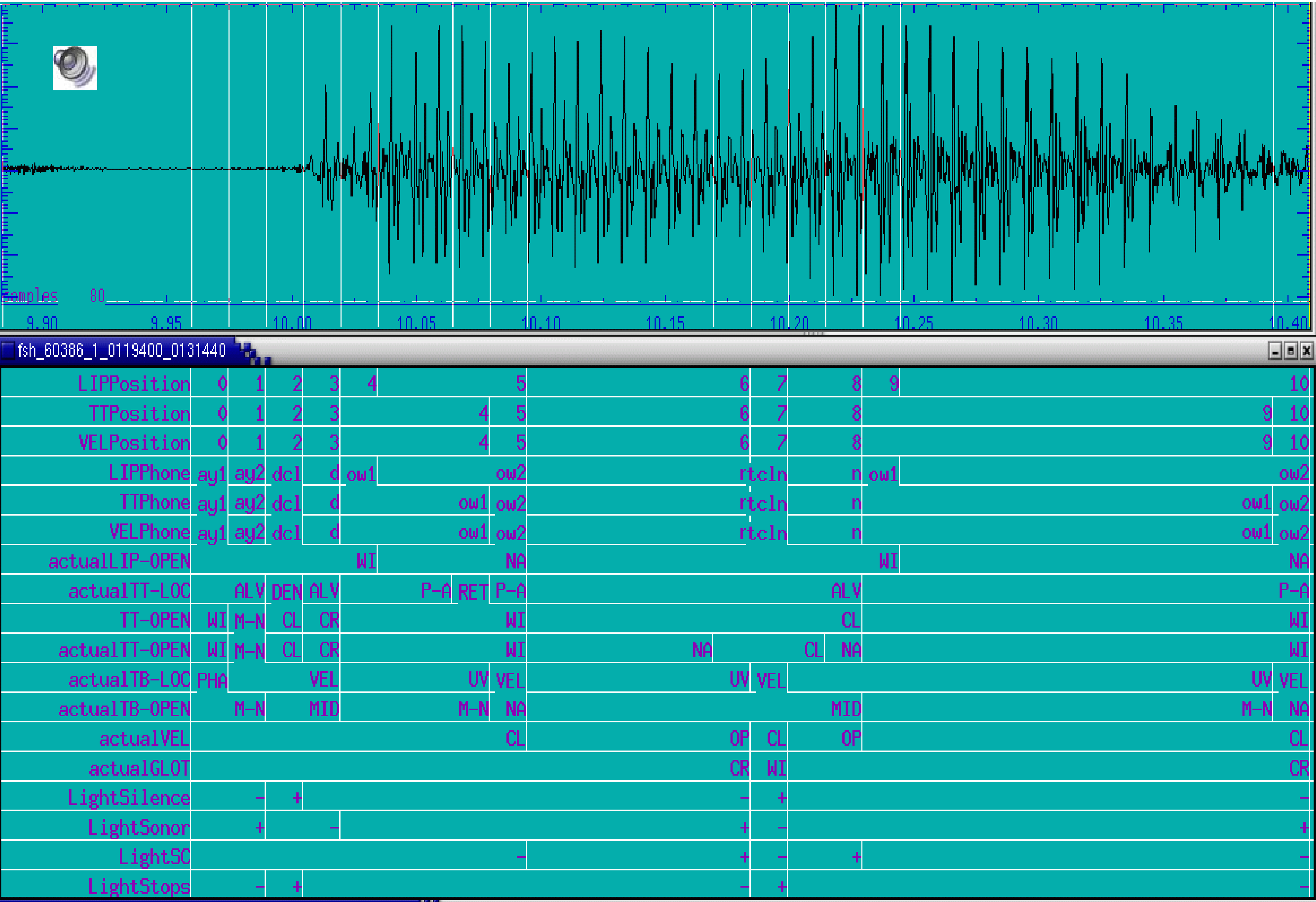
Date	*Experimental setup*	*3-speaker WER (# errors)*	*RT03 dev WER*
- ∞	Baseline	27.7 (550)	26.8
Jul31_0	EBS/DBN, "hierarchically-normalized" SVM output probabilities, DBN trained on subset of ICSI transcriptions	27.6 (549)	26.8
Aug1_19	+ improved silence modeling	27.6 (549)	
Aug2_19	EBS/DBN, unnormalized SVM probs + fricative lip feature	27.3 (543)	26.8
Aug4_2	+ DBN trained using SVM outputs	27.3 (543)	
Aug6_20	+ full feature hierarchy in DBN	27.4 (545)	
Aug7_3	+ reduction probabilities depend on word frequency	27.4 (544)	
Aug8_19	+ retrained SVMs + nasal classifier + DBN bug fixes	27.4 (544)	
Aug11_19	SVM/DBN, 1 pass	*Miserable failure!*	
Aug14_0	SVM/DBN, 2 pass	27.3 (542)	
Aug14_20	SVM/DBN, 2 pass, using only high-accuracy SVMs	27.2 (541)	

Some complicating factors...

- **Practicalities:**
  - Inaccurate word boundaries in lattices
  - Very short words
  - Pauses, laughter, non-words

- **More general issues:**
  - Relative weighting of soft evidence vs. articulatory variables
  - Over-counting of evidence largely not addressed
  - SVM/DBN rescoring complicated by context-dependent SVM training

The word boundary problem



Some conclusions

- **No major error rate improvements yet... BUT:**
- **The SVM/DBN system produces reasonable analyses of reduction and coarticulation in spontaneous speech**
- **EM parameter learning produces reasonable distributions**
- **Many ideas for future work, e.g.:**
  - **Further analysis of the current system**
    - Error analysis
    - Computational complexity analysis
  - **More context-dependent modeling (based on syllable structure, stress accent, position in word, speaker clustering)**
  - **Investigation of the usefulness of different features**
  - **Better understanding of the mathematical issues of feature hierarchies in landmark-based recognition**
  - **Exploration of soft evidence in DBNs for ASR in general**