

Karen Livescu and James Glass

Spoken Language Systems Group, MIT Computer Science and Artificial Intelligence Laboratory, Cambridge MA 02139 USA

{klivescu,jrg}@sls.csail.mit.edu, http://www.sls.csail.mit.edu/{livescu,glass}

1. Main ideas

Motivation:

- From linguistics:** Speech consists of multiple streams of *semi-independent phonological features*, rather than *phones*
- From speech recognition research:**
 - Phone-based pronunciation modeling has had limited success
 - Spontaneous speech is difficult to describe phonetically (see 2)

Previous work:

- Much research on recognition with feature classifiers
- But *pronunciation model* is still typically phone-based

Contributions:

- Introduction of a general, flexible feature-based pronunciation model, accounting for *feature asynchrony* and *feature substitutions*
- Implementation using dynamic Bayesian networks
- Initial experiments on spontaneous pronunciations

2. Examples

- warmth* → [w ao r m p th] : phone insertion?
- wants* → [w aa_n t s] : phone deletion??
- several* → [s eh r v ax l] : exchange of phones???
- instruments* → [ih_n s ch em ih_n n s] : ????
- everybody* → [eh r uw ay] : !?!?!?!?
- All explainable via feature asynchrony + some feature value substitutions (“undershooting”)

3. The model

baseform dictionary

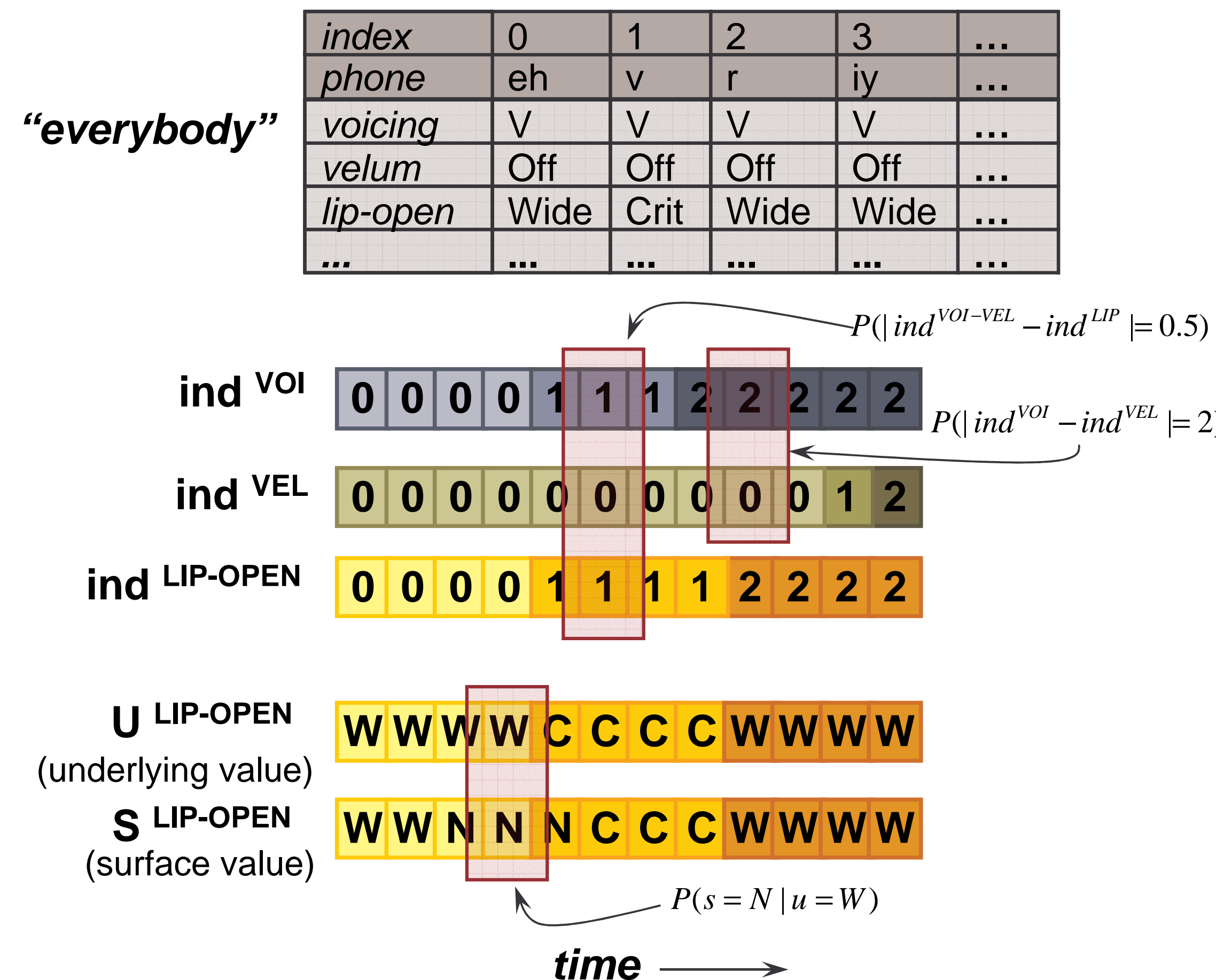
To produce a given utterance, each feature should proceed through a particular sequence of values

+ asynchrony

Different features may proceed through their respective sequences at different rates

+ feature substitutions

Surface (actual) feature values may differ from underlying (dictionary) values



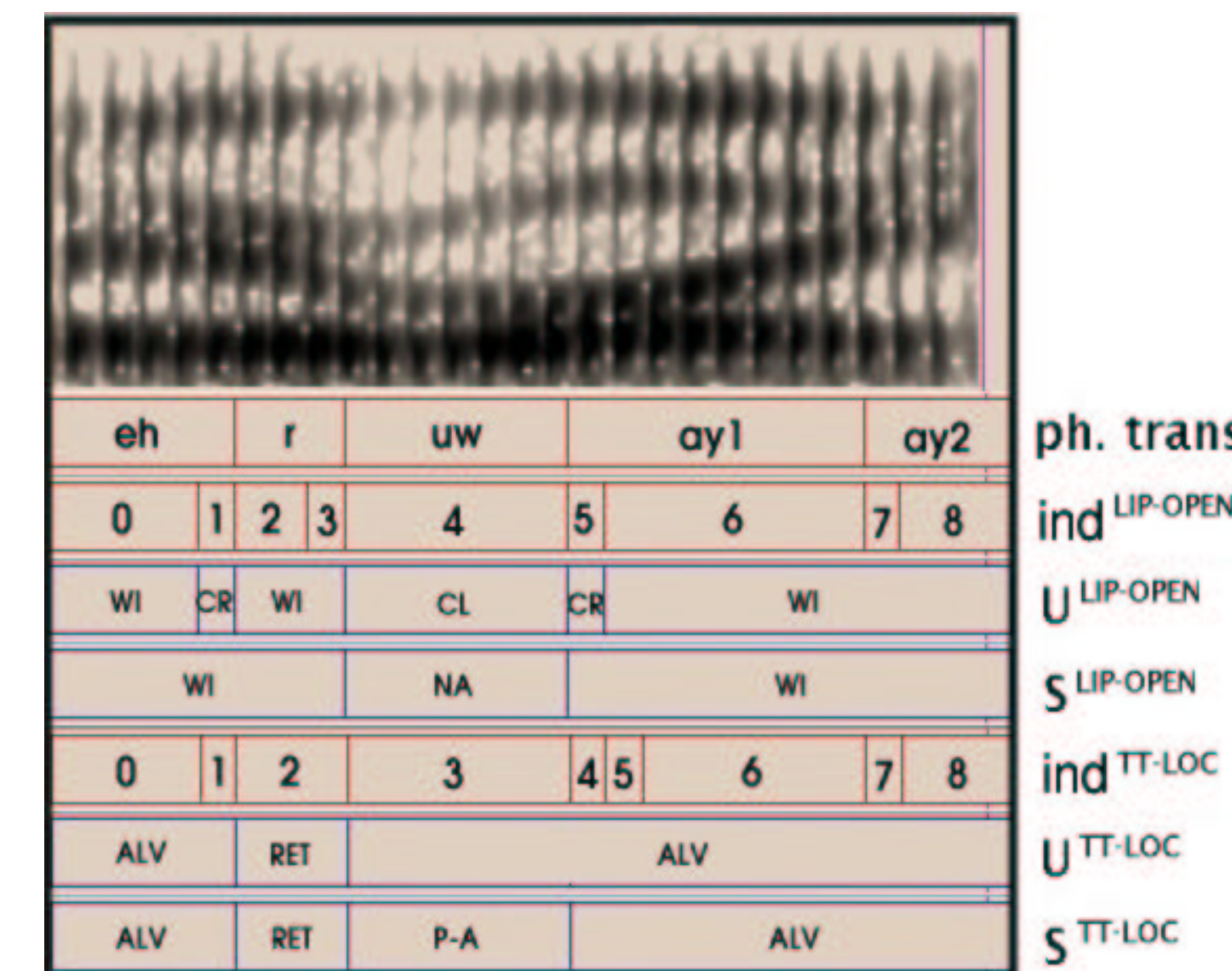
5. Experiments

- Task:** Classify isolated words from the Switchboard corpus, given a detailed phonetic transcription (from ICSI)
 - Convert transcription into feature vectors S_t^i , one per 10ms
 - Find the word $w^* = \text{argmax}_w P(w|S_{1:T}^i)$
- Feature set** based on vocal tract variables of articulatory phonology [1]: lip opening, tongue tip location/opening, tongue body location/opening, velum, glottis
- Experiments implemented in GMTK [2]
- Parameters initialized by hand and trained using EM
- See [3],[4] for additional details

Model	Dev set (165 words)		Test set (236 words)	
	Error rate	Failure rate	Error rate	Failure rate
baseform dict only	60.0	57.6	64.8	62.3
dict + phon. rules	57.0	54.5	63.1	58.5
dict + feature substitutions	35.2	23.0	44.5	29.7
dict + feature subs + asynchrony	28.5	16.4	40.7	24.6
dict + feature subs + asynch + EM	27.9	16.4	40.7	24.6

- Qualitative examination of alignments produced by the model show expected asynchrony effects
- EM training has minor effect on error rates, but improves rank/score distributions [4]

Example alignment: “everybody” → [eh r uw ay]



4. Implementation

- 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:**

$$\Pr(\text{async}^{1:2} = a) = \Pr(|ind^1 - ind^2| = a)$$

$$\text{checkSync}^{1:2} = 1 \text{ if } |ind^1 - ind^2| = \text{async}^{1:2}$$

given by baseform pronunciations

u	s	0	1	2	3	4	...
0	.7	.2	.1	0	0	0	...
1	0	.7	.2	.1	0	0	...
2	0	0	.7	.2	.1	0	...
...

6. Ongoing/future work

- Integration with landmark-based feature classifiers
- Context-dependent feature distributions
- More complex tasks (multiwords, larger vocabularies)
- Use of articulatory databases
- Possible uses of such a model to learn about speech

References

[1] J. Bilmes and G. Zweig, “The Graphical Models Toolkit: An open source software system for speech and time0series processing,” ICASSP 2002.

[2] C. P. Browman and L. Goldstein, “Articulatory phonology: An overview,” *Phonetica* 49:155-180, 1992.

[3] K. Livescu and J. Glass, “Feature-based pronunciation modeling for speech recognition,” HLT/NAACL 2004.

[4] K. Livescu and J. Glass, “Feature-based pronunciation modeling with trainable asynchrony probabilities,” ICSLP 2004.