

---

# **Phonological Models in Automatic Speech Recognition**

**Karen Livescu**

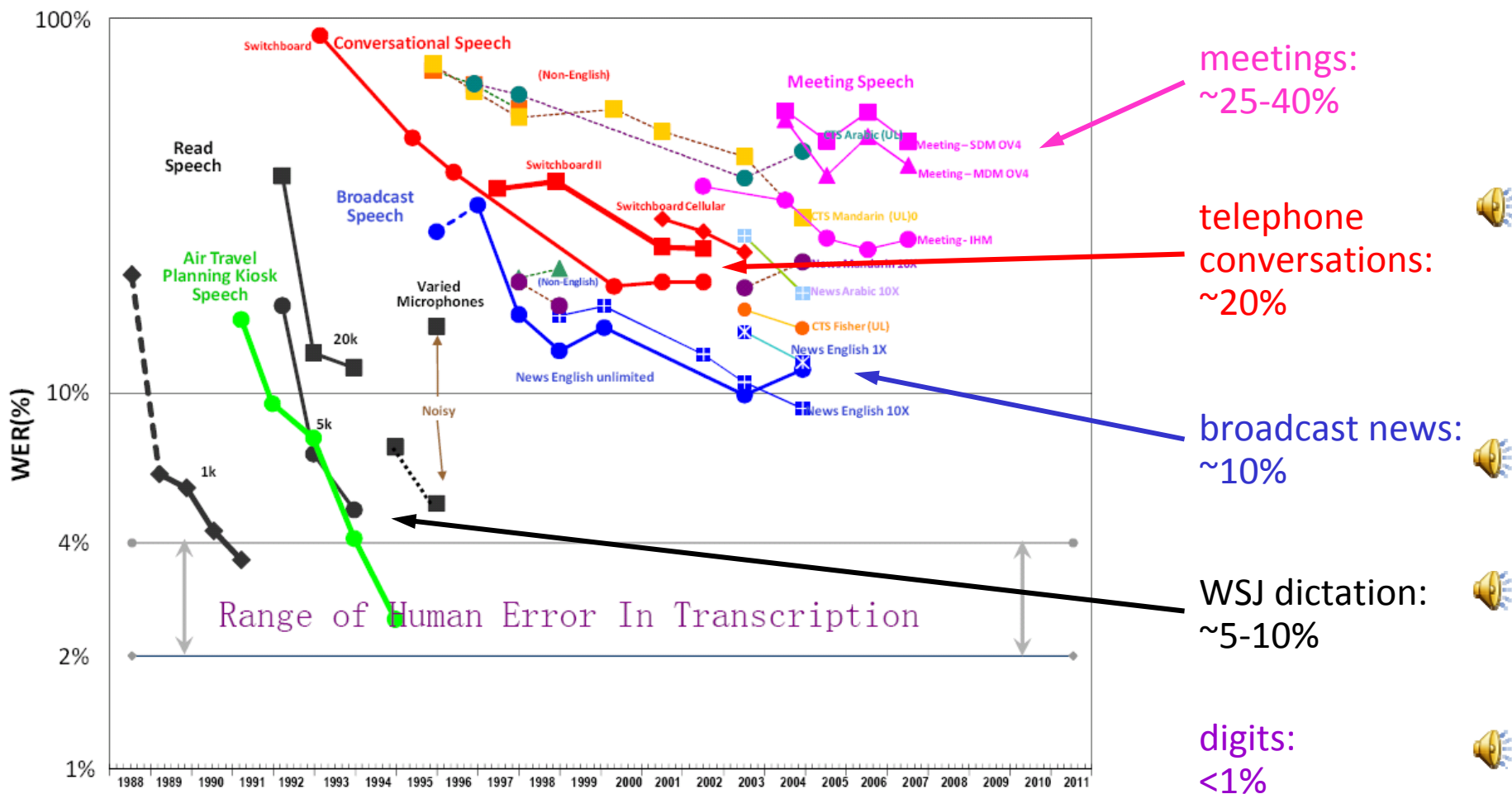
Toyota Technological Institute at Chicago

June 19, 2008

---

# What can automatic speech recognition (ASR) do?

- NIST benchmark evaluation results 1988-2007
- $WER = (\#subs + \#ins + \#del) / \#ref$



[figure from Fiscus et al. '07, "The Rich Transcription 2007 Meeting Recognition Evaluation", <http://www.nist.gov/speech/publications/papers/>]

# What is so difficult about conversational speech?

---

Non-speech (e.g. laughter, sigh) 🗣️

Variable speaking rate 🗣️

Disfluencies (e.g. partial words, hesitations, repeated syllables) 🗣️

Extreme pronunciation variation 🗣️

# Pronunciation variation in conversational speech: Examples

*word*

probably

sense

everybody

don't

*baseform*



p r aa b ax b l iy


s eh n s


eh v r iy b ah d iy

d ow n t

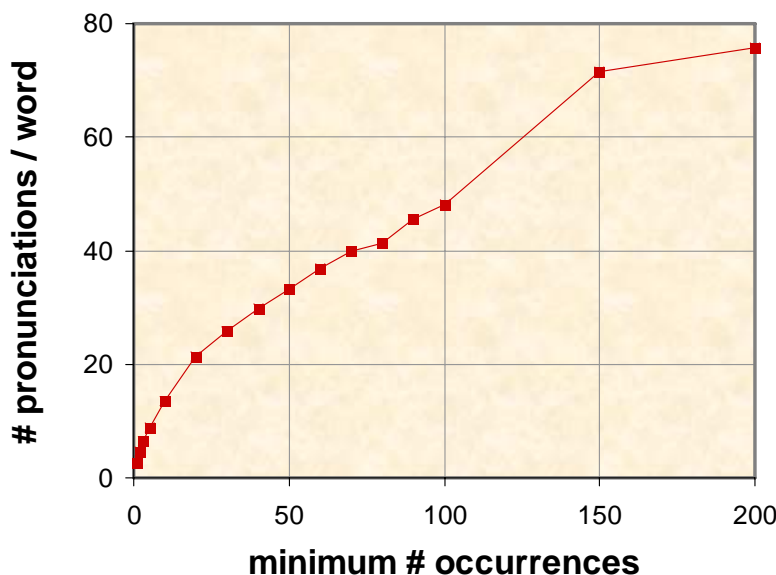
*surface forms*

- (2) p r aa b iy 
- (1) p r ay
- (1) p r aw l uh
- (1) p r ah b iy 
- (1) p r aa l iy
- (1) p r aa b uw
- (1) p ow ih
- (1) p aa iy
- (1) p aa b uh b l iy
- (1) p aa ah iy

- (1) s eh n t s 
- (1) s ih t s

- (1) eh v r ax b ax d iy
- (1) eh v er b ah d iy
- (1) eh ux b ax iy
- (1) eh r uw ay 
- (1) eh b ah iy

- (37) d ow n
- (16) d ow
- (6) ow n
- (4) d ow n t
- (3) d ow t
- (3) d ah n
- (3) ow
- (3) n ax
- (2) d ax n
- (2) ax
- (1) n uw
- ...



# Effect of pronunciation variation on ASR performance

Words pronounced non-canonically are more likely to be mis-recognized  
[Fosler-Lussier '99]

Deletions are especially difficult to account for [Jurafsky et al. '01]

Conversational speech is recognized at almost twice the error rate of read speech [Weintraub et al. '96]

Style	Word error rate (%)
Spontaneous conversation	52.6
Read conversational	37.6
Read dictation	28.8

Simulated-data experiments show potential benefit of a good pronunciation model [McAllaster et al. '98]

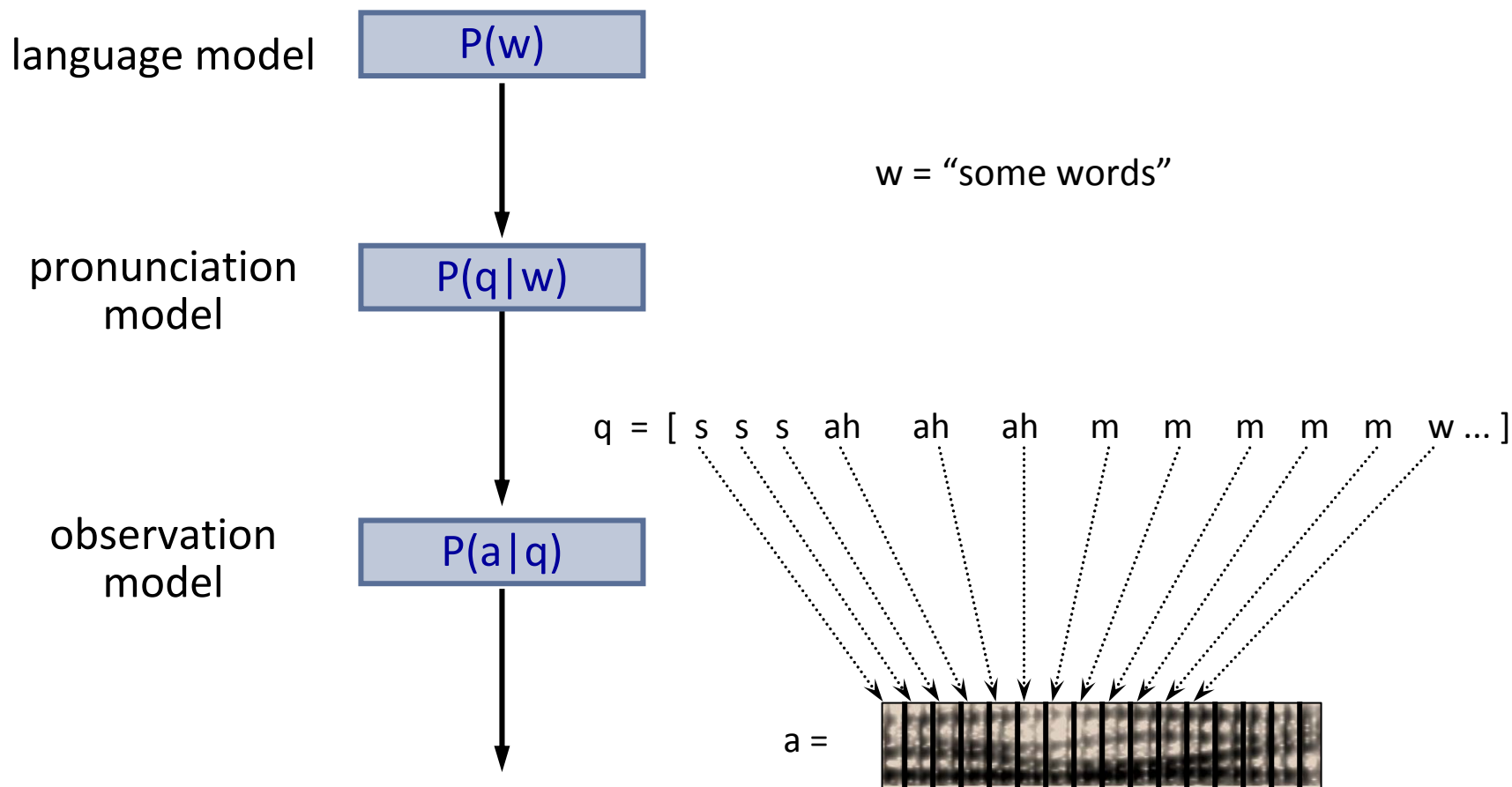
Test data	Word error rate (%)
Real	48.8
Simulated from dictionary	10.8
Simulated from transcription	43.9

# Overview

---

- Preliminaries: Automatic speech recognition (ASR)
- Phone-based pronunciation models
- Non-phonetic alternatives
- Ongoing/future work

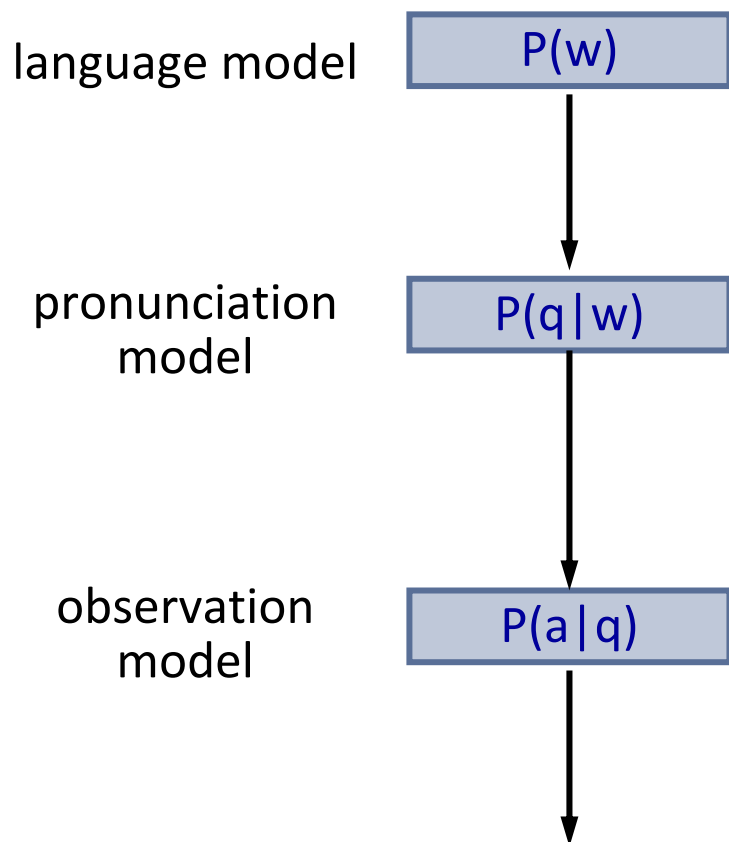
# Speech recognition: The generative statistical setting



Recognition  $\equiv$

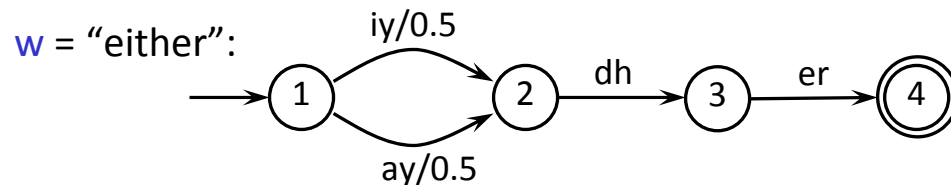
$$w^* = \operatorname{argmax}_w P(w|a)$$
$$= \operatorname{argmax}_w P(a|w) P(w)$$
$$= \operatorname{argmax}_w P(w) \sum_q P(q|w) P(a|q)$$

# Speech recognition: The generative statistical setting

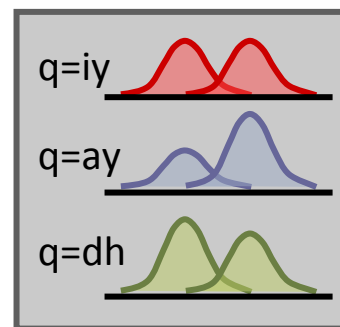


e.g. n-gram:

$$P(w = w_1, w_2, \dots, w_k) = \prod_i P(w_i | w_{i-1}, w_{i-2}, \dots, w_{i-(n-1)})$$



$P(a_i | q_i)$



# Overview

---

- Preliminaries: Automatic speech recognition (ASR)
- Phone-based pronunciation models
- Non-phonetic alternatives
- Ongoing/future work

# Phone-based pronunciation modeling

Lexicon is expanded with substitution, deletion, and insertion rules as in derivational phonology [Chomsky & Halle '68]



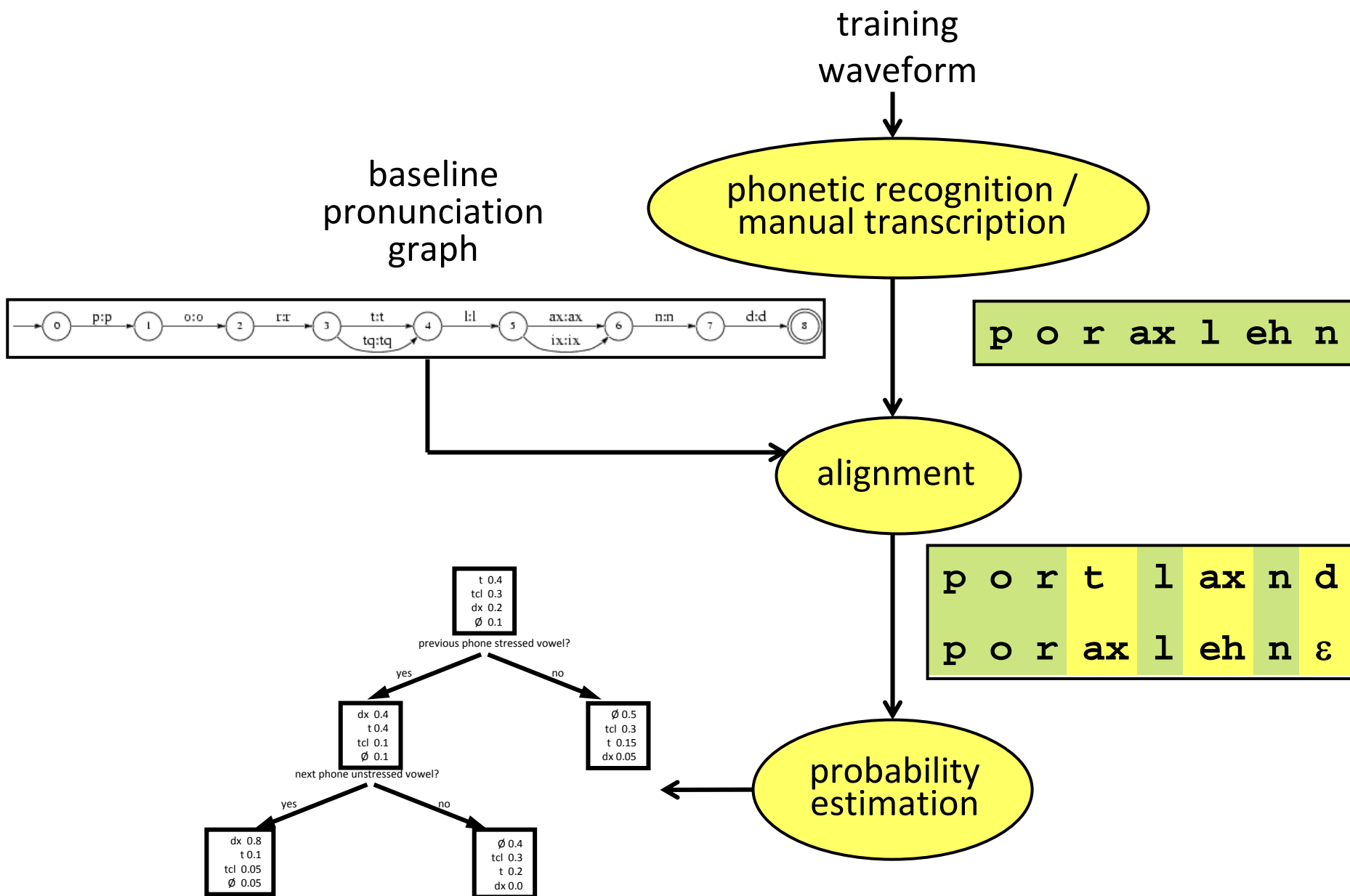
Transformation rules are of the form  $u \rightarrow s / u_L \_ u_R; p$ , e.g.

- Epenthetic stop insertion:  $\emptyset \rightarrow t / n \_ s; 0.5$
- Flapping:  $t \rightarrow dx / V' \_ V; 0.7$

Rules are derived from

- Linguistic knowledge  
[Zue et al. '75, Cohen '89, Tajchman et al. '95, Finke & Waibel '97, Hazen et al. '02, Seneff & Wang '05]
- Data  
[Chen '90, Riley & Ljolje '95, Byrne et al. '97, Riley et al. '99, Fosler-Lussier '99]

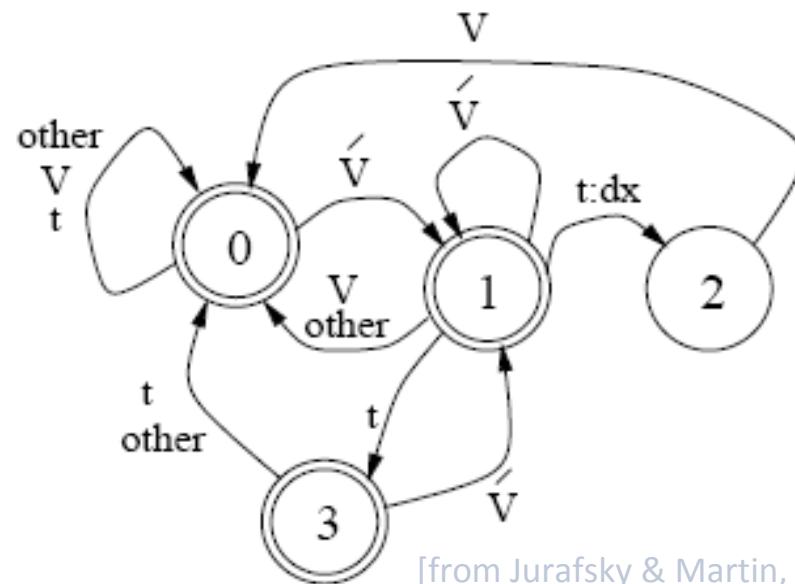
# Learning phonological rules from data



# Finite-state representation of phonological rules

- Rewrite rules of the form  $u \rightarrow s / u_L \_ u_R$  can be represented as finite-state transducers (FSTs) [Johnson '72]

- Example: /t/ flapping rule  $t \rightarrow dx / V' \_ V$



[from Jurafsky & Martin,  
*Speech and Language  
Processing*, '00]

- Multiple ordered rules  $F_1, F_2, \dots$  can be combined into a single FST via composition  $F_1 \circ F_2 \circ \dots$

## Phone-based pronunciation modeling: Some results

Model	Task	Impact on WER (%)
Rule learning from manual transcriptions + retraining [Riley et al. '99]	Broadcast news	12.7 → 10.0
	Switchboard	44.7 → 43.8
Decision trees + dynamic lexicon [Fosler-Lussier '99]	Broadcast news	21.4 → 20.4
Knowledge-based rules + FST weight learning [Hazen et al. '02]	Weather queries	12.1 → 11.0

- Roughly 1-3% WER improvement across tasks
- Significant improvements on difficult tasks, but not as large as expected
- “Implicit” pronunciation modeling with one pronunciation per word [Hain '02]
  - Observation model accounts for remaining variability
  - Similar performance to multi-pronunciation dictionaries
- State of the art uses one/a few fixed pronunciations per word

# Overview

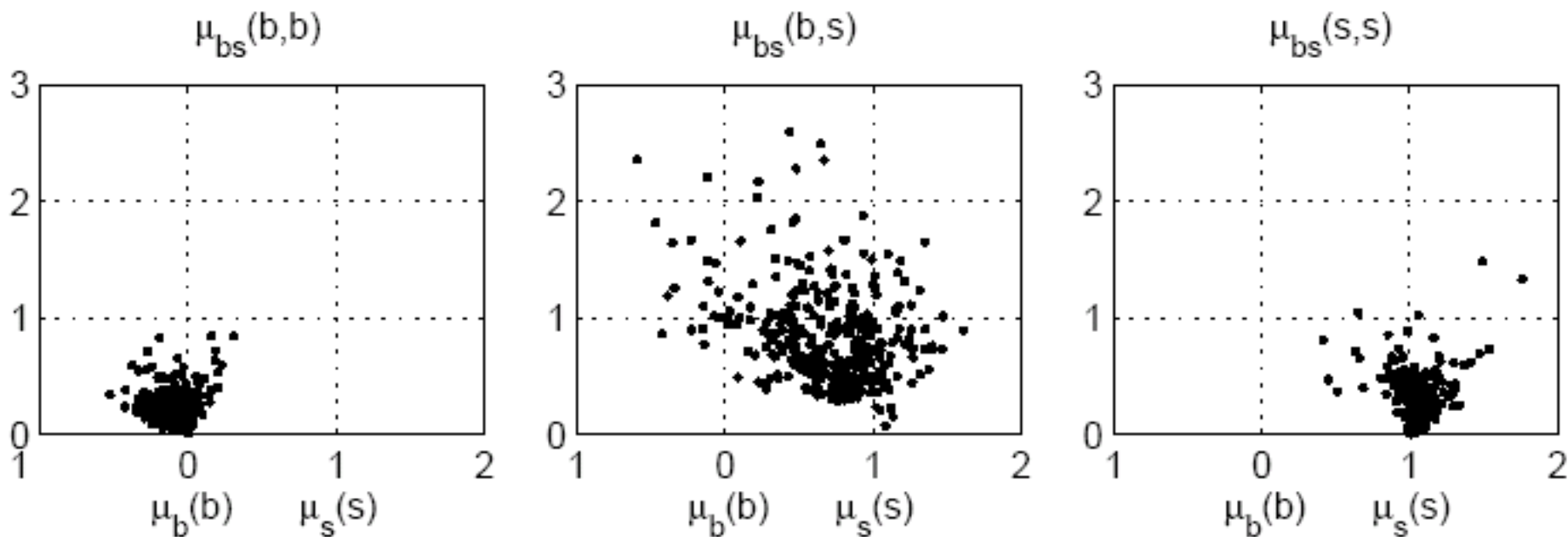
---

- Preliminaries: Automatic speech recognition (ASR)
- Phone-based pronunciation models
- **Non-phonetic alternatives**
- Ongoing/future work

# The argument against the phone

Pronunciation changes are gradual

- If  $\text{had} \rightarrow [\text{h ae d}] \rightarrow [\text{h eh d}]$  then “had” is confusable with “head”
- Is  $[\text{ae}] \rightarrow [\text{eh}]$  really happening?
- No: [figure from Saraclar & Khudanpur, *Speech Communication*, '04]



# Automatically derived units & syllables

---

Automatically derived sub-word units [Holter & Svendsen '97, Bacchiani & Ostendorf '99, Varadarajan et al. '08]

- Learned by segmentation + clustering of the acoustics
- Lexicon built by aligning word segments with learned units

Syllable units [Ganapathiraju et al. '01, Sethy & Narayanan '03]

- Motivation: Reduction phenomena reported to occur within syllable boundaries
- Human transcribers label syllables more easily than phones [Fosler-Lussier et al. '99]
- States not shared across syllables → “had” and “head” are always different

Both approaches have impressive results on small-vocabulary tasks (~1/3 reduction in WER), but are not directly applicable to infrequent words/syllables

# Two paths toward progress

---

Adapt syllable/automatic-unit models for larger vocabularies

Look to phonology again

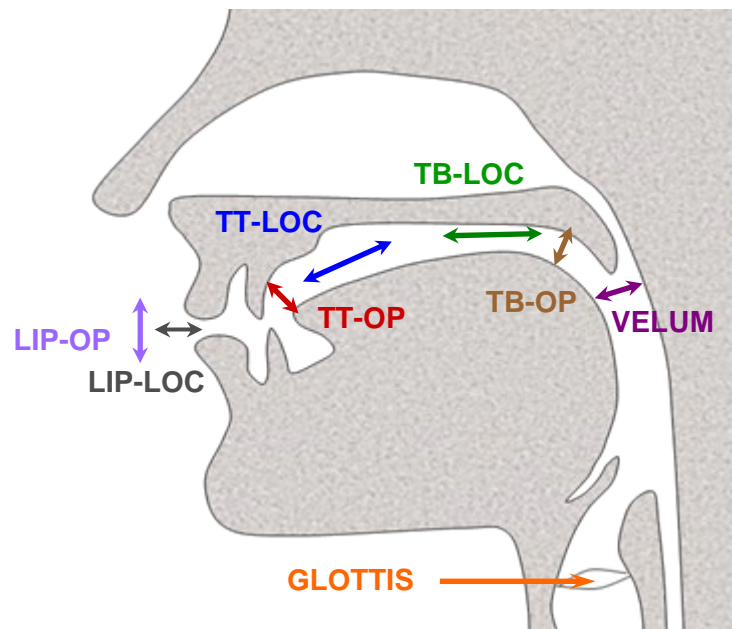
- This time, autosegmental/articulatory phonology

# Articulatory features as subword units

Inspired by ideas in phonology

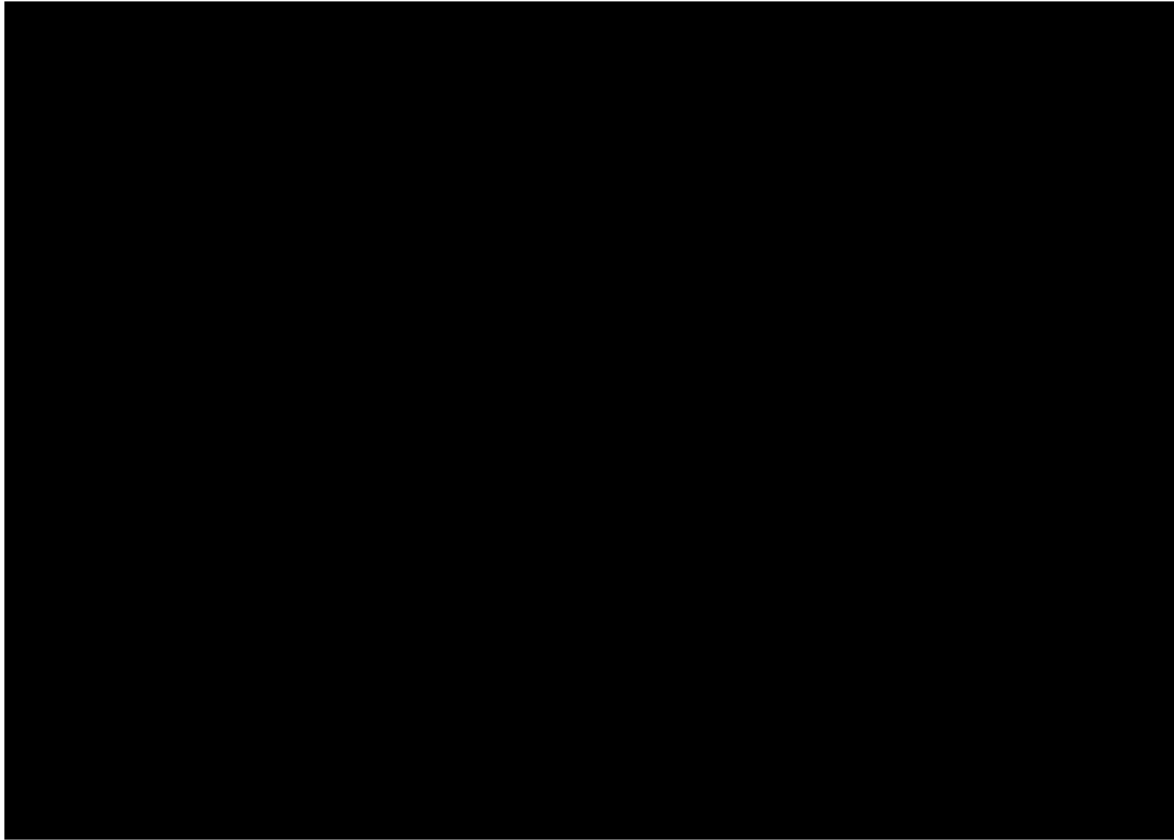
- Autosegmental phonology [Goldsmith '76]: Phonetic representation consists of multiple tiers of segments, with some constraints (“associations”) among them
- Articulatory phonology [Browman & Goldstein '92]:
  - Tiers consist of articulatory gestures, with “phasing relations”
  - Surface realizations stray from dictionary via (1) asynchrony and (2) gesture reduction

feature	values
LIP-LOC	protruded, labial, dental
LIP-OP	closed, critical, narrow, wide
TT-LOC	dental, alveolar, palato-alveolar, retroflex
TB-LOC	palatal, velar, uvular, pharyngeal
TT-OP, TB-OP	closed, critical, narrow, mid-narrow, mid, wide
GLO	closed (glottal stop), critical (voiced), open (voiceless)
VEL	closed (non-nasal), open (nasal)



## The argument against the phone (2)

---



[X-ray video from Speech Communication Group, MIT]

# The argument against the phone (3)

*sense* → [s eh n t s] - Phone insertion ?

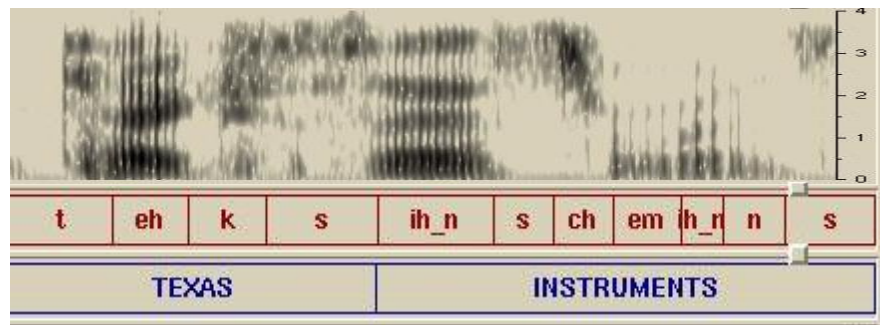
*wants* → [w aa\_n t s] - Phone deletion ??

*sense* → [s ih\_n t s] - Phone deletion + substitution??

*several* → [s eh r v ax l] - Exchange of two phones ?!?!?

*Texas Instruments*

→ [t eh k s ih\_n s ch em ih\_r n s]



*everybody* → [eh r uw ay]

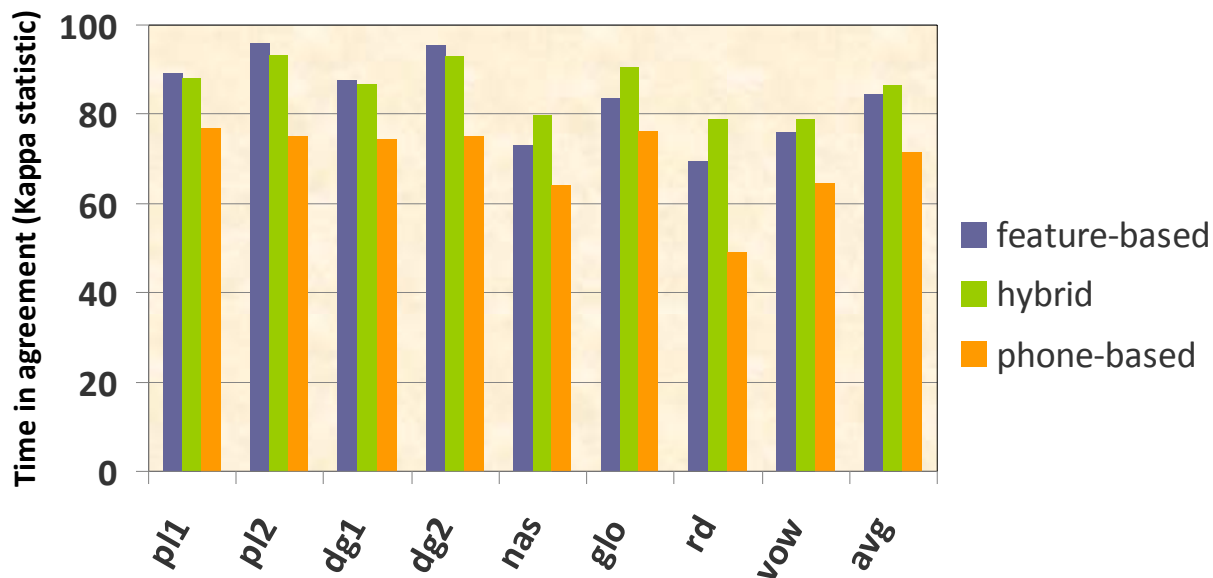


## The argument against the phone (4)

Even humans have difficulty with phonetic transcription [Ostendorf '99, Fosler-Lussier et al. '99]

- “Deleted” phones are sometimes still perceived
- Inter-transcriber disagreement is high (~25% string error) [Saraclar '04]

Feature-level transcription may be more reliable [Livescu et al. '07]



# Revisiting sense → [s eh n t s], [s ih<sub>n</sub> t s]

dictionary

<i>feature</i>	<i>values</i>			
<b>GLO</b>	open	critical		open
<b>VEL</b>	closed		open	closed
<b>TB</b>	mid / uvular	mid / palatal	mid / uvular	
<b>TT</b>	critical / alveolar	mid / alveolar	closed / alveolar	critical / alveolar
<b>phone</b>	s	eh	n	s

surface variant #1

<i>feature</i>	<i>values</i>			
<b>GLO</b>	open	critical	open	
<b>VEL</b>	closed	open	closed	
<b>TB</b>	mid / uvular	mid / palatal	mid / uvular	
<b>TT</b>	critical / alveolar	mid / alveolar	closed / alveolar	critical / alveolar
<b>phone</b>	s	eh	n	t

surface variant #2

<i>feature</i>	<i>values</i>			
<b>GLO</b>	open	critical	open	
<b>VEL</b>	closed	open	closed	
<b>TB</b>	mid / uvular	mid-nar / palatal	mid / uvular	
<b>TT</b>	critical / alveolar	mid-nar / alveolar	closed / alveolar	critical / alveolar
<b>phone</b>	s	ih	t	s



# Articulatory feature models: Main Ideas

baseform  
dictionary

“everybody” →

index	0	1	2	3	...
phone	eh	v	r	iy	...
GLO	crit	crit	crit	crit	...
LIPS	wide	crit	nar	wide	...
...	...	...	...	...	...

+  
asynchrony

index<sup>GLO</sup>

0	0	0	0	1	1	1	2	2	2	2	2
---	---	---	---	---	---	---	---	---	---	---	---

index<sup>LIPS</sup>

0	0	0	0	1	1	1	1	1	2	2	2
---	---	---	---	---	---	---	---	---	---	---	---

+  
feature  
substitutions

target<sup>LIPS</sup>

W	W	W	W	C	C	C	C	C	N	N	N
---	---	---	---	---	---	---	---	---	---	---	---

actual<sup>LIPS</sup>

W	W	N	N	N	C	C	C	C	N	N	N
---	---	---	---	---	---	---	---	---	---	---	---

# Articulatory feature models: Initial approaches

## Finite-state models with “product” state space

[Erler & Freeman '96; Deng et al. '97; Richardson & Bilmes '03]

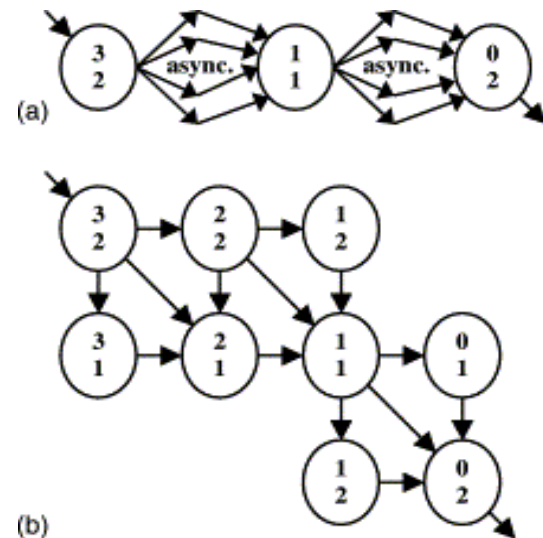
- Each state is a vector of feature values
- Asynchrony among features allowed between target articulations

## Two-pass models [Huckvale '94, Blackburn '96, Reetz '98]

- 1<sup>st</sup> pass: Feature classification
- 2<sup>nd</sup> pass: Decoding word sequence from features

## A modeling problem

- Finite-state models don't take advantage of known independence properties
- Two-pass models assume too much independence



[from Richardson & Bilmes, *Speech Communication*, '03]

# Articulatory feature models: Recent work

---

Articulatory approaches require more flexible probability models

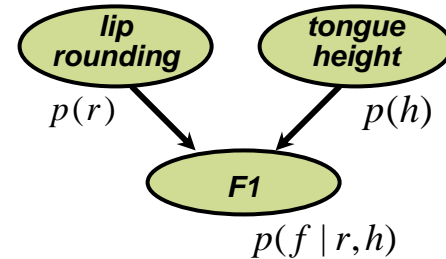
One solution: dynamic Bayesian networks

- Allows the factorization of the state into multiple variables
- Can represent independence assumptions exactly
- Recently gaining popularity in ASR [Zweig '98, Bilmes '99, JHU WS01/04/06]
- At least one ASR-oriented toolkit available (GMTK) [Bilmes '02]

## Aside: Bayesian networks

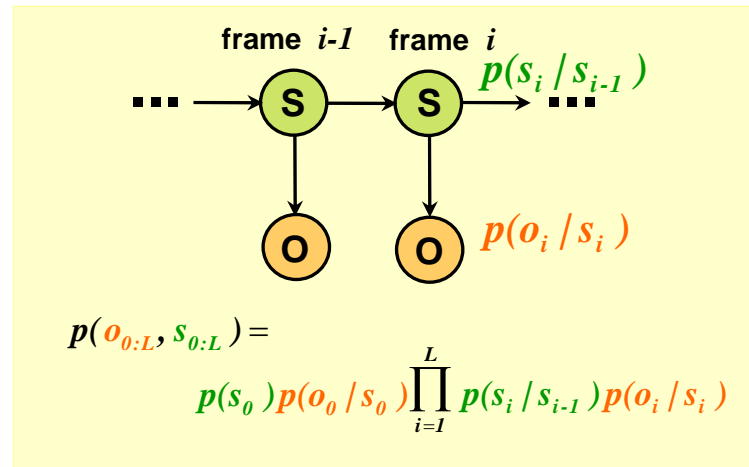
Bayesian network (BN): Directed-graph representation of a distribution over a set of variables

- Graph node  $\Leftrightarrow$  variable + its distribution given parents
- (Lack of) graph edges  $\Leftrightarrow$  independencies
- Joint distribution = product of local distributions



Dynamic Bayesian network (DBN): BN with a repeating structure

Example:  
hidden Markov model  
(HMM)



Uniform algorithms for (among other things)

- Finding the most likely values of some variables, given the rest (analogous to Viterbi algorithm for HMMs)
- Learning model parameters via expectation-maximization

# Approach: Main Ideas

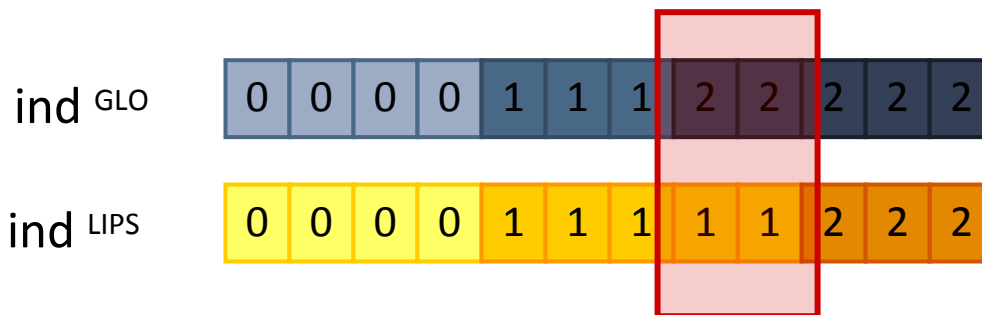
baseform  
dictionary

“everybody” →

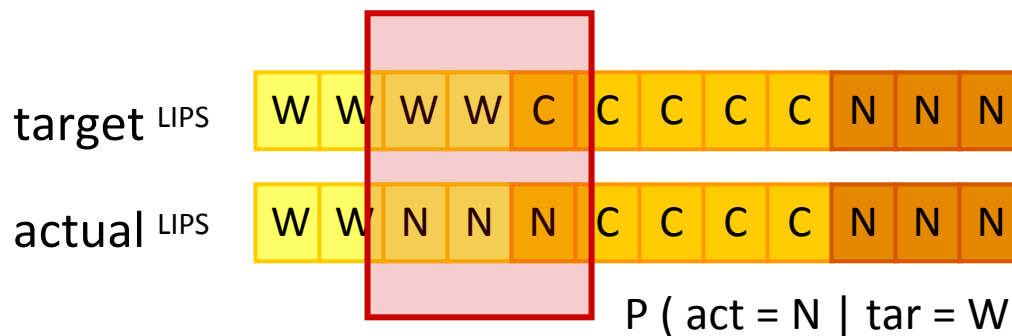
index	0	1	2	3	...
phone	eh	v	r	iy	...
GLO	crit	crit	crit	crit	...
LIPS	wide	crit	nar	wide	...
...	...	...	...	...	...

$$P ( | \text{index}^{\text{GLO}} - \text{index}^{\text{LIPS}} | = 1 )$$

+  
asynchrony



+  
feature  
substitutions





# Model parameters

- Phone-to-feature mapping

<i>phone</i>	<i>GLOT</i>	<i>VEL</i>	<i>LIP-OPEN</i>	<i>TT-OPEN</i>	...
aa	V (1)	CL (1)	WI (1)	WI (1)	...
m	V (1)	OP (1)	CL (1)	CL (.2), CR (.2), NA (.2), M-N (.2)	...
...	...	...	...	...	...

- Soft synchrony constraints  $P(\text{async}^{A;B})$
- Feature substitution probabilities

LIP-OPEN

<i>u \ s</i>	<i>CL</i>	<i>CR</i>	<i>NA</i>	<i>WI</i>
<i>CL</i>	.8	.15	.05	0
<i>CR</i>	0	.8	.2	0
<i>NA</i>	0	0	.8	.2
<i>WI</i>	0	0	0	1

GLOT

<i>u \ s</i>	<i>V</i>	<i>VL</i>
<i>V</i>	1	0
<i>VL</i>	0	1

- Transition probabilities
  - In each frame, the probability of transitioning to the next state in the word
- Maximum-likelihood parameter values learned via Expectation-Maximization

# Where will the data for parameter learning come from?

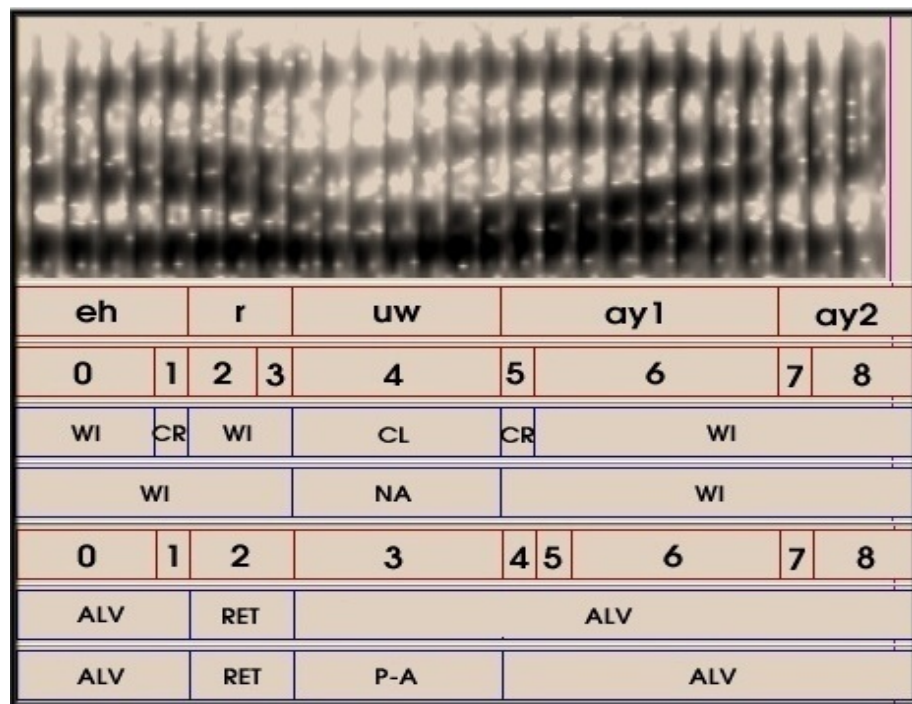
---

- Manual transcriptions
- Articulatory measurements (EMA, X-ray microbeam, MRI, ...)
- Nowhere!

# Lexical access experiments

## “Recognition” of Switchboard words

- Given manual transcription
- Phone-based model: 66% coverage, 54% accuracy
- Feature-based model: 75% coverage, 61% accuracy



*everybody* → [ eh r uw ay ]

ph. trans.

ind<sup>LIP-OPEN</sup>

U<sup>LIP-OPEN</sup>

S<sup>LIP-OPEN</sup>

ind<sup>TT-LOC</sup>

U<sup>TT-LOC</sup>

S<sup>TT-LOC</sup>

hyp. state seq.

hyp. targets

input

## What works?

Vowel nasalization & rounding; nasal + stop → nasal, some schwa deletions

## What doesn't work?

Some deletions; vowel retroflexion; alveolar + [y] → palatal

# Overview

---

- Preliminaries: Automatic speech recognition (ASR)
- Phone-based pronunciation models
- Non-phonetic alternatives
- Ongoing/future work

## Ongoing work

---

Not a complete recognizer – need observation model  $P(a | q)$ , where  $q$  = hidden variables

- Gaussian mixture distribution conditioned on all features [Livescu et al. '07]
- Separate observation model per feature  $P(a | \text{voicing})$ ,  $P(a | \text{lips})$ , ... [Livescu et al. '03, '07]
- Posterior-based models  $P(\text{voicing} | a)$ ,  $P(\text{lips} | a)$ , ... [Hasegawa-Johnson et al. '05, Cetin et al. '07]
  - Applied to lipreading, improves accuracy over viseme-based models [Saenko et al. '05, '06]

Additional ongoing work: cross-word modeling, audio-visual speech recognition [Hasegawa-Johnson et al. '07]

## Concluding remarks

---

Speech recognition has borrowed much from phonology

- Derivational phonology → phonetic rule-based pronunciation modeling
- Autosegmental/articulatory phonology → feature-based modeling

The best sub-word representation is unlikely to be the phone

- Syllables, acoustically defined units, articulatory features

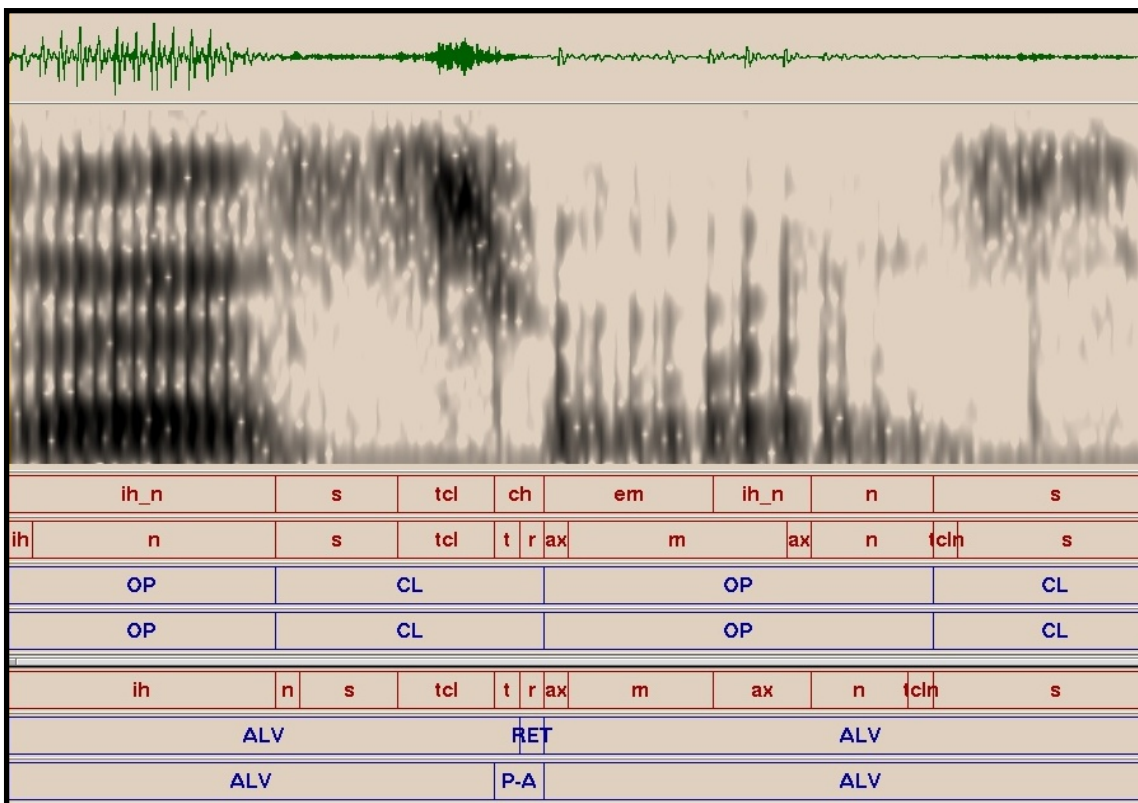
A time of transition for pronunciation modeling

- New approaches may require new statistical/machine learning tools
- Graphical models provide a natural framework

# Concluding questions

Can we use speech recognition models to learn something about speech?

*instruments* → [ ih s ch em ih n s ]



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

transcription

ph<sup>VEL</sup>

U<sup>VEL</sup>

S<sup>VEL</sup>

ph<sup>TT-LOC</sup>

U<sup>TT-LOC</sup>

S<sup>TT-LOC</sup>

How much reduction can occur?

How do these depend on the speaker, dialect, language impairment, ... ?

How do model scores relate to human perceptual judgments?