CS 224S/LING 281
Speech Recognition, Synthesis, and Dialogue

Dan Jurafsky

Lecture 16: Variation and Adaptation

Outline

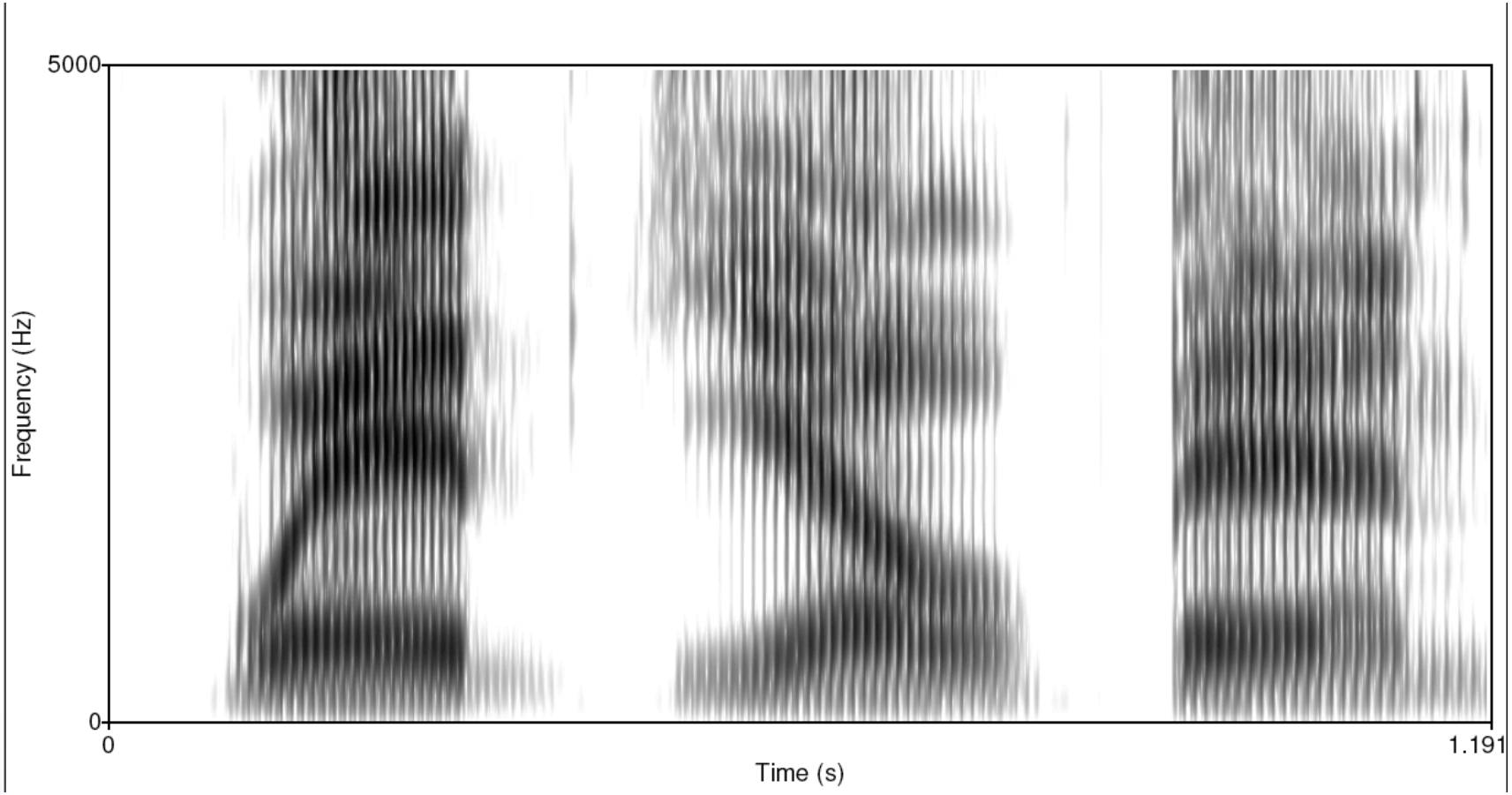
- Variation in speech recognition
- Sources of Variation
- Three classic problems:
  - Dealing with phonetic variation
    - triphones
  - Speaker differences (including accent)
    - Speaker adaptation: MLLR, MAP
  - Variation due to Genre: Conversational Speech
    - Pronunciation modeling issues
    - Unsolved!

Sources of Variability

- Phonetic context
- Environment
- Speaker
- Genre/Task

Most important: phonetic context: different “eh”s

- w eh d y eh l b eh n



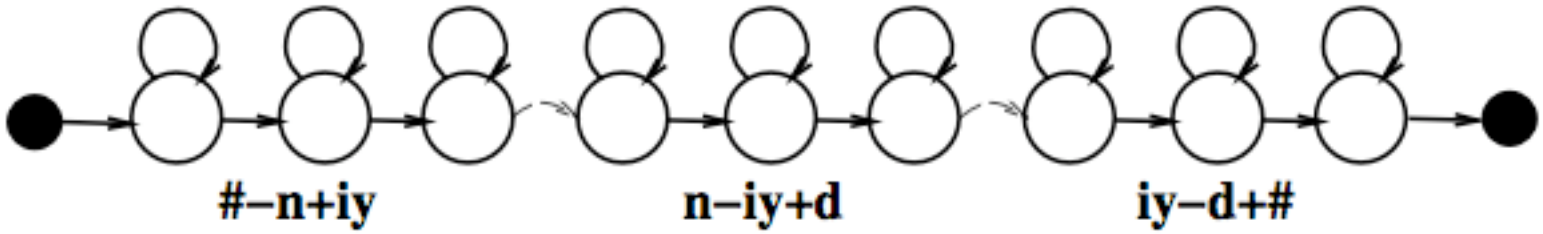
Modeling phonetic context

- The strongest factor affecting phonetic variability is the neighboring phone
- How to model that in HMMs?
- Idea: have phone models which are specific to context.
- Instead of Context-Independent (CI) phones
- We'll have Context-Dependent (CD) phones

CD phones: triphones

- Triphones
- Each triphone captures facts about preceding and following phone
- Monophone:
  - p, t, k
- Triphone:
  - iy-p+aa
  - a-b+c means "phone b, preceding by phone a, followed by phone c"

“Need” with triphone models



Word-Boundary Modeling

- Word-Internal Context-Dependent Models

  'OUR LIST':

  SIL AA+R AA-R L+IH L-IH+S IH-S+T S-T

- Cross-Word Context-Dependent Models

  'OUR LIST':

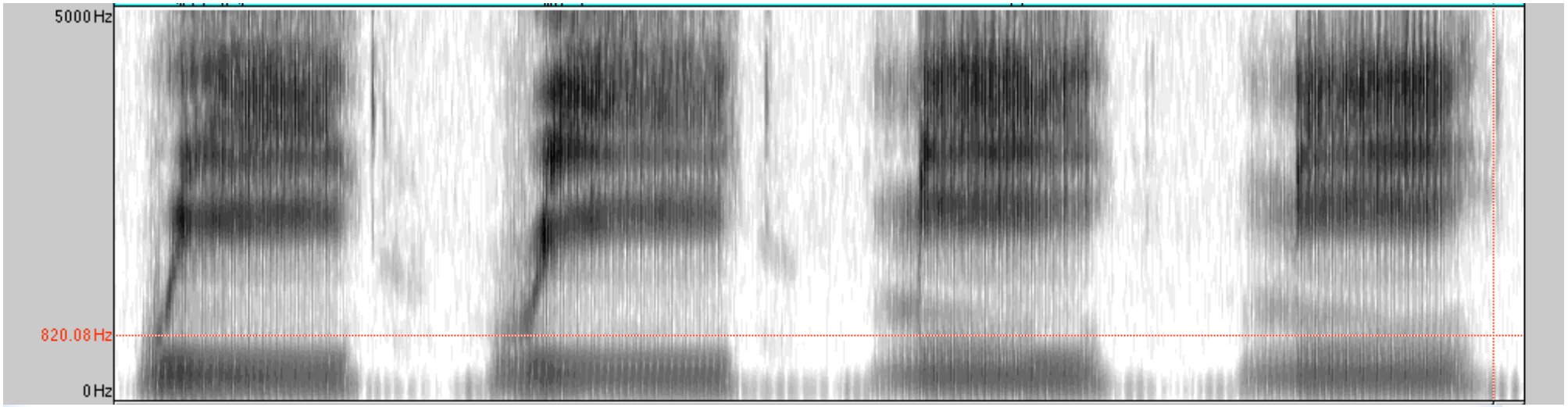
  SIL-AA+R AA-R**+L R-**L+IH L-IH+S IH-S+T S-T**+SIL**

- Dealing with cross-words makes decoding harder! We will return to this.

Implications of Cross-Word Triphones

- Possible triphones: 50x50x50=125,000
- How many triphone types actually occur?
- 20K word WSJ Task, numbers from Young et al
- Cross-word models: need 55,000 triphones
- But in training data only 18,500 triphones occur!
- Need to generalize models.

Modeling phonetic context: some contexts look similar

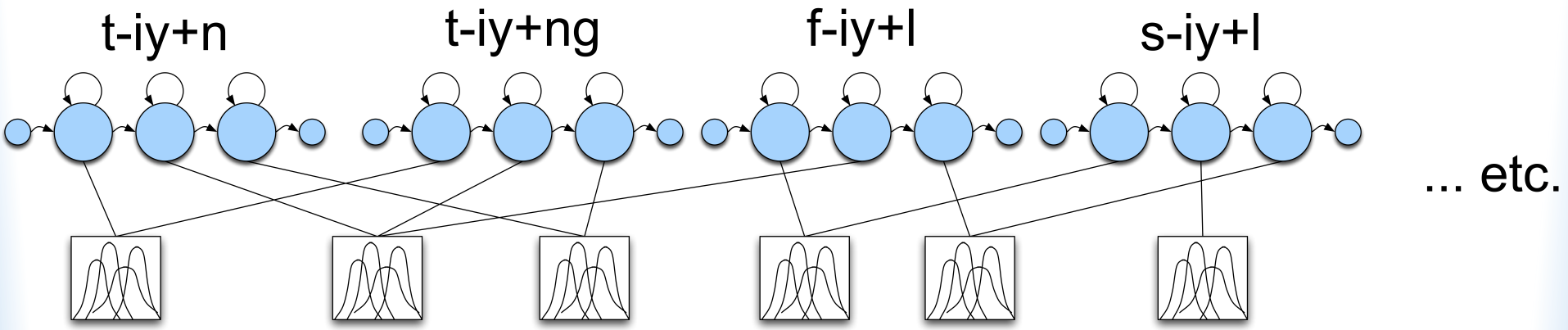


W iy r iy m iy n iy

Solution: State Tying

- Young, Odell, Woodland 1994
- Decision-Tree based clustering of triphone states
- States which are clustered together will share their Gaussians
- We call this "state tying", since these states are "tied together" to the same Gaussian.
- Previous work: generalized triphones
  - Model-based clustering ('model' = 'phone')
  - Clustering at state is more fine-grained

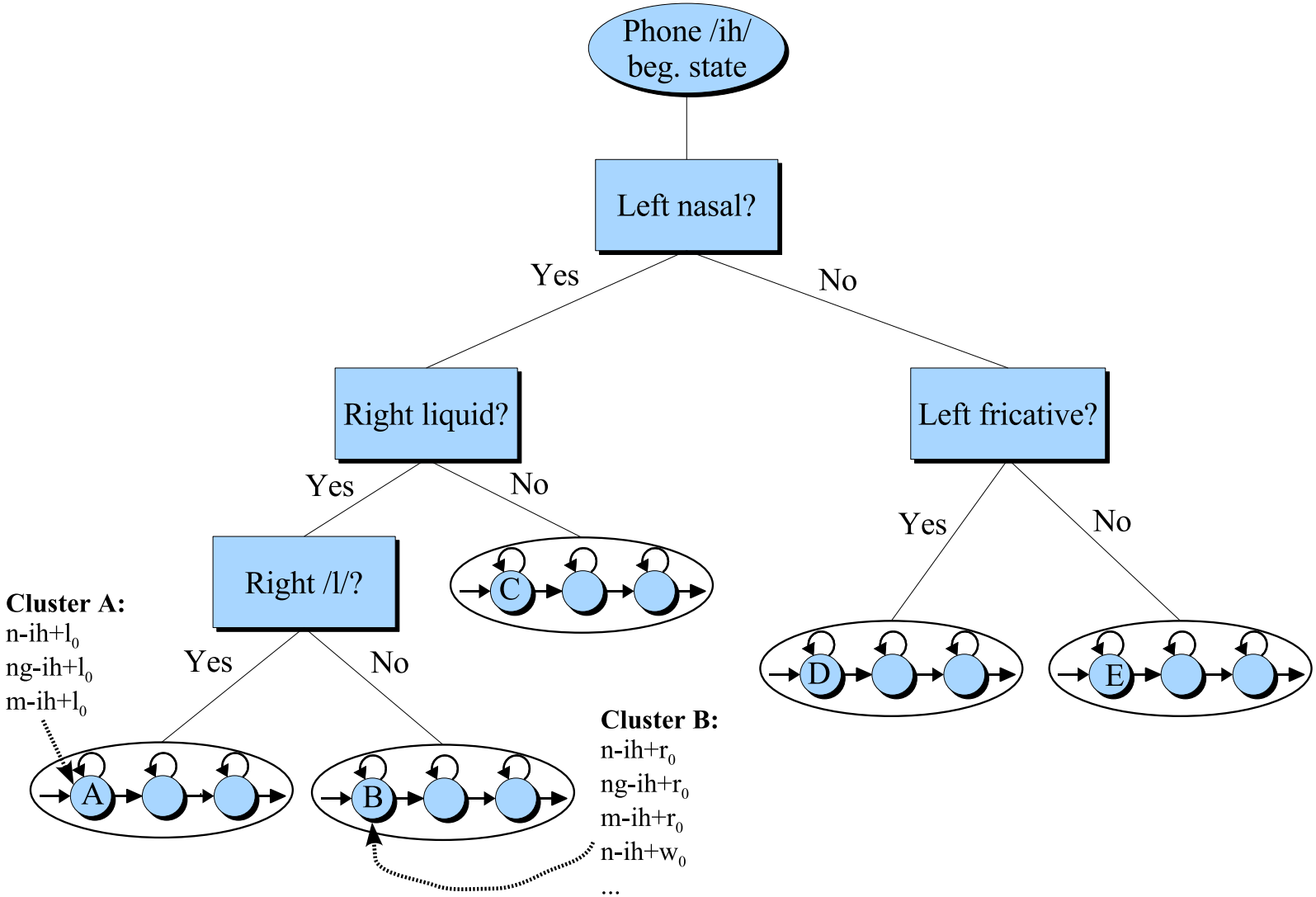
Young et al state tying



State tying/clustering

- How do we decide which triphones to cluster together?
- Use phonetic features (or 'broad phonetic classes')
  - Stop
  - Nasal
  - Fricative
  - Sibilant
  - Vowel
  - lateral

Decision tree for clustering triphones for tying



Decision tree for clustering triphones for tying

Feature	Phones
Stop	b d g k p t
Nasal	m n ng
Fricative	ch dh f jh s sh th v z zh
Liquid	l r w y
Vowel	aa ae ah ao aw ax axr ay eh er ey ih ix iy ow oy uh uw
Front Vowel	ae eh ih ix iy
Central Vowel	aa ah ao axr er
Back Vowel	ax ow uh uw
High Vowel	ih ix iy uh uw
Rounded	ao ow oy uh uw w
Reduced	ax axr ix
Unvoiced	ch f hh k p s sh t th
Coronal	ch d dh jh l n r s sh t th z zh

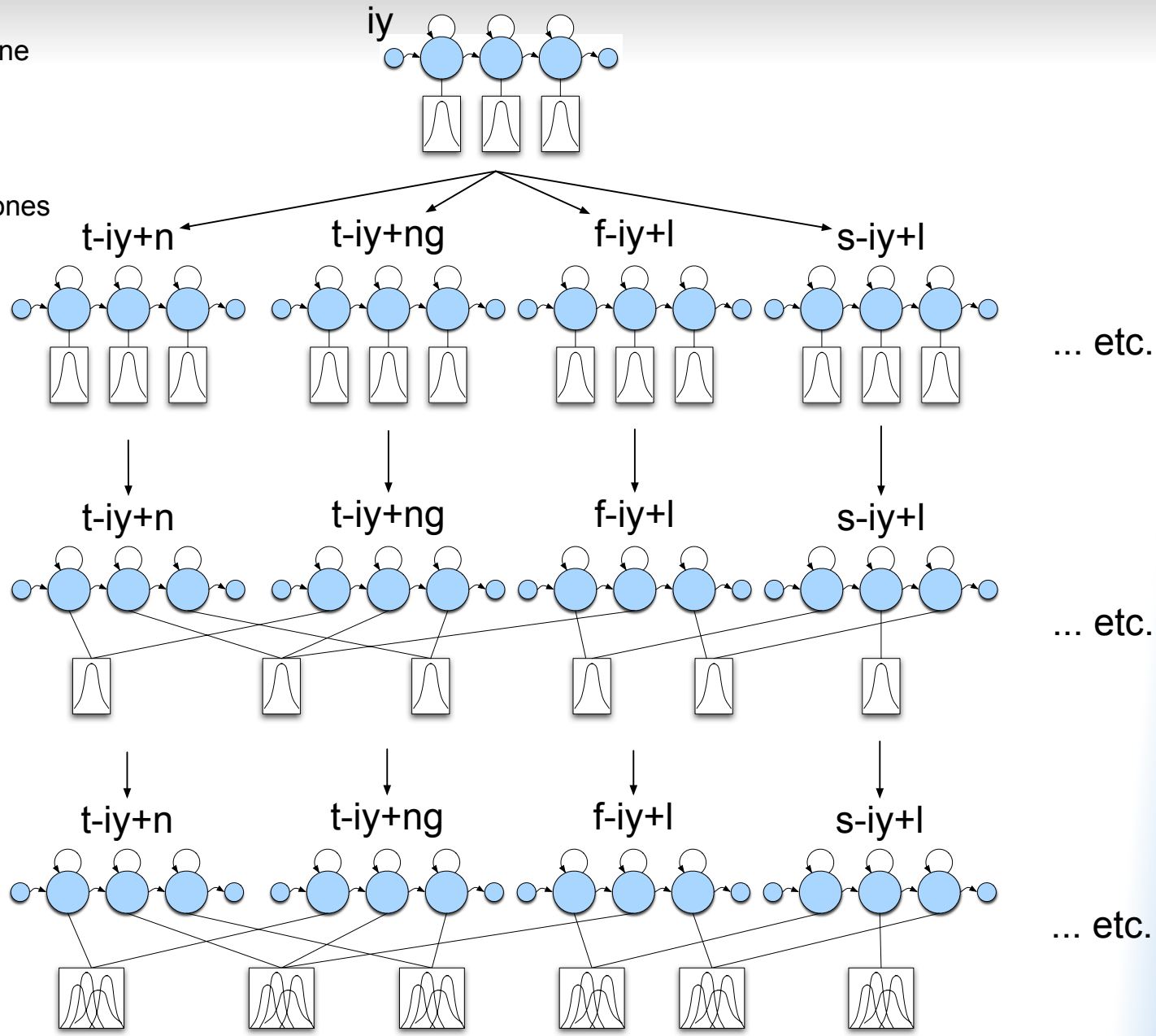
State Tying: Young, Odell, Woodland 1994

(1) Train monophone single Gaussian models

(2) Clone monophones to triphones

(3) Cluster and tie triphones

(4) Expand to GMMs



Summary: Acoustic Modeling for LVCSR.

- Increasingly sophisticated models
- For each state:
  - Gaussians
  - Multivariate Gaussians
  - Mixtures of Multivariate Gaussians
- Where a state is progressively:
  - CI Phone
  - CI Subphone (3ish per phone)
  - CD phone (=triphones)
  - State-tying of CD phone
- Forward-Backward Training
- Viterbi training

The rest of today's lecture

- Variation due to speaker differences
  - Speaker adaptation
    - MLLR
    - MAP
    - Splitting acoustic models by gender
- Speaker adaptation approaches also solve
  - Variation due to environment
    - Lombard speech
    - Foreign accent
      - Acoustic and pronunciation adaptation to accent
- Variation due to genre differences
  - Pronunciation modeling

Speaker adaptation

- The largest source of improvement in ASR bakeoff performance in the last decade. Some numbers from Bryan Pellom's Sonic:

Speech Recognition Task Description	Vocabulary Size	Word Error Rate (without adaptation)	Word Error Rate (with adaptation)
TI-DIGITS (continuous spoken digits)	11	0.4%	0.2%
DARPA Communicator (realtime spoken dialog system, telephone speech related to travel domain)	2.1k	10.9%	--NA--
Wall Street Journal (Nov 1992 5k eval) (dictation task, high-quality microphone speech)	5k	3.9%	3.0%
Wall Street Journal (Nov 1992 20k eval) (dictation task, high-quality microphone speech)	20k	10.0%	8.6%
DARPA/NRL SPINE (spoken dialogs, noisy military environments, microphone speech)	3k	42.2%	31.0%
Switchboard (conversational telephone speech; NIST 2000 eval data, SWB eval results only)	40k	41.9%	31.0%

Acoustic Model Adaptation

- Shift the means and variances of Gaussians to better match the input feature distribution
  - Maximum Likelihood Linear Regression (MLLR)
  - Maximum A Posteriori (MAP) Adaptation
- For both speaker adaptation and environment adaptation
- Widely used!

Slide from Bryan Pellom

Maximum Likelihood Linear Regression (MLLR)

- Leggetter, C.J. and P. Woodland. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language* 9:2, 171-185.
- Given:
  - a trained AM
  - a small "adaptation" dataset from a new speaker
- Learn new values for the Gaussian mean vectors
  - Not by just training on the new data (too small)
  - But by learning a linear transform which moves the means.

Maximum Likelihood Linear Regression (MLLR)

- Estimates a linear transform matrix (W) and bias vector (ω) to transform HMM model means:

$$\mu_{new} = W_r \mu_{old} + \omega_r$$

- Transform estimated to maximize the likelihood of the adaptation data

Slide from Bryan Pellom

MLLR

- New equation for output likelihood

$$b_j(o_t) = \frac{1}{(2\pi)^{n/2} |\Sigma_j|^{1/2}} \exp\left(-\frac{1}{2} (o_t - (W\mu_j + \omega)) |\Sigma_j|^{-1} (o_t - (W\mu_j + \omega))^T \right)$$

MLLR

- Q: Why is estimating a linear transform from adaptation data different than just training on the data?
- A: Even from a very small amount of data we can learn 1 single transform for all triphones! So small number of parameters.
- A2: If we have enough data, we could learn more transforms (but still less than the number of triphones). One per phone (~50) is often done.

MLLR: Learning A

- Given
  - an small labeled adaptation set (a couple sentences)
  - a trained AM
- Do forward-backward alignment on adaptation set to compute state occupation probabilities $\xi_j(t)$.
- W can now be computed by solving a system of simultaneous equations involving $\xi_j(t)$

MLLR performance on RM
(Leggetter and Woodland 1995)

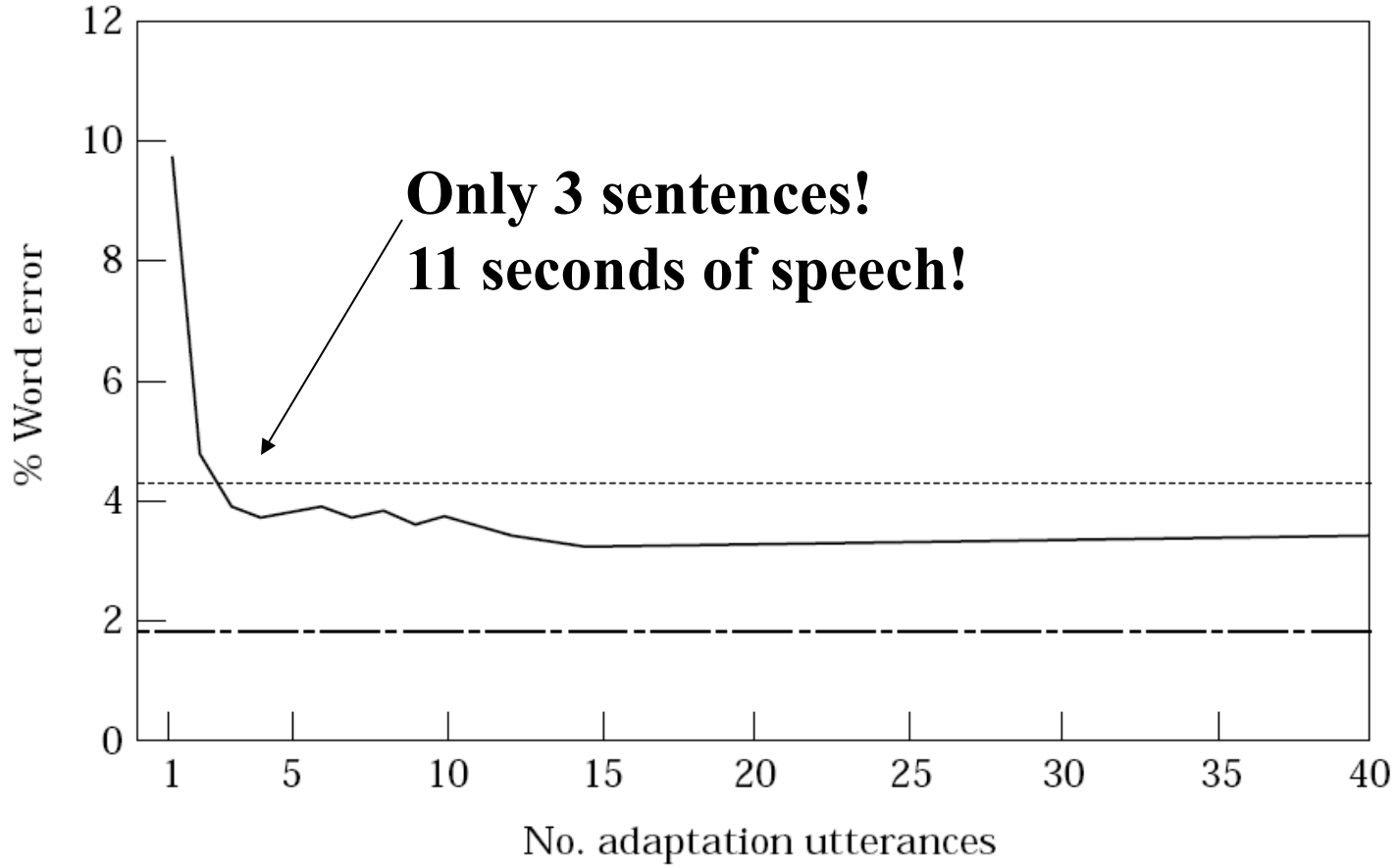


Figure 2. Full matrix maximum likelihood linear regression using global regression class. (·····), Speaker independent; (—-—-—-), speaker dependent; (——), speaker adapted.

MLLR doesn't need supervised adaptation set!

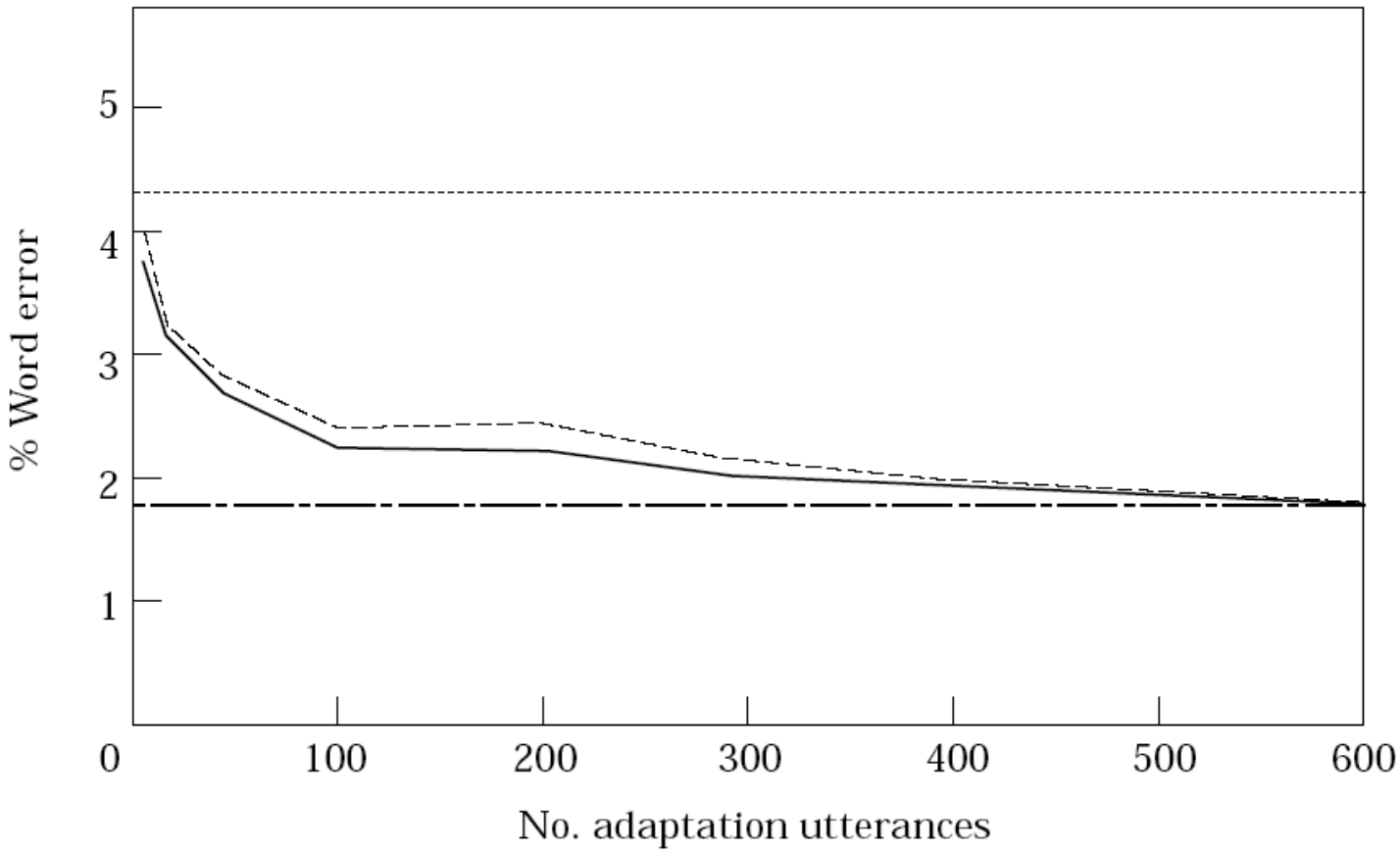


Figure 3. Supervised vs. unsupervised adaptation using maximum likelihood linear regression. (·····), Speaker independent; (—-—-—-), speaker dependent; (——), supervised adapted; (– – –), unsupervised adapted.

Maximum A Posteriori Adaptation (MAP)

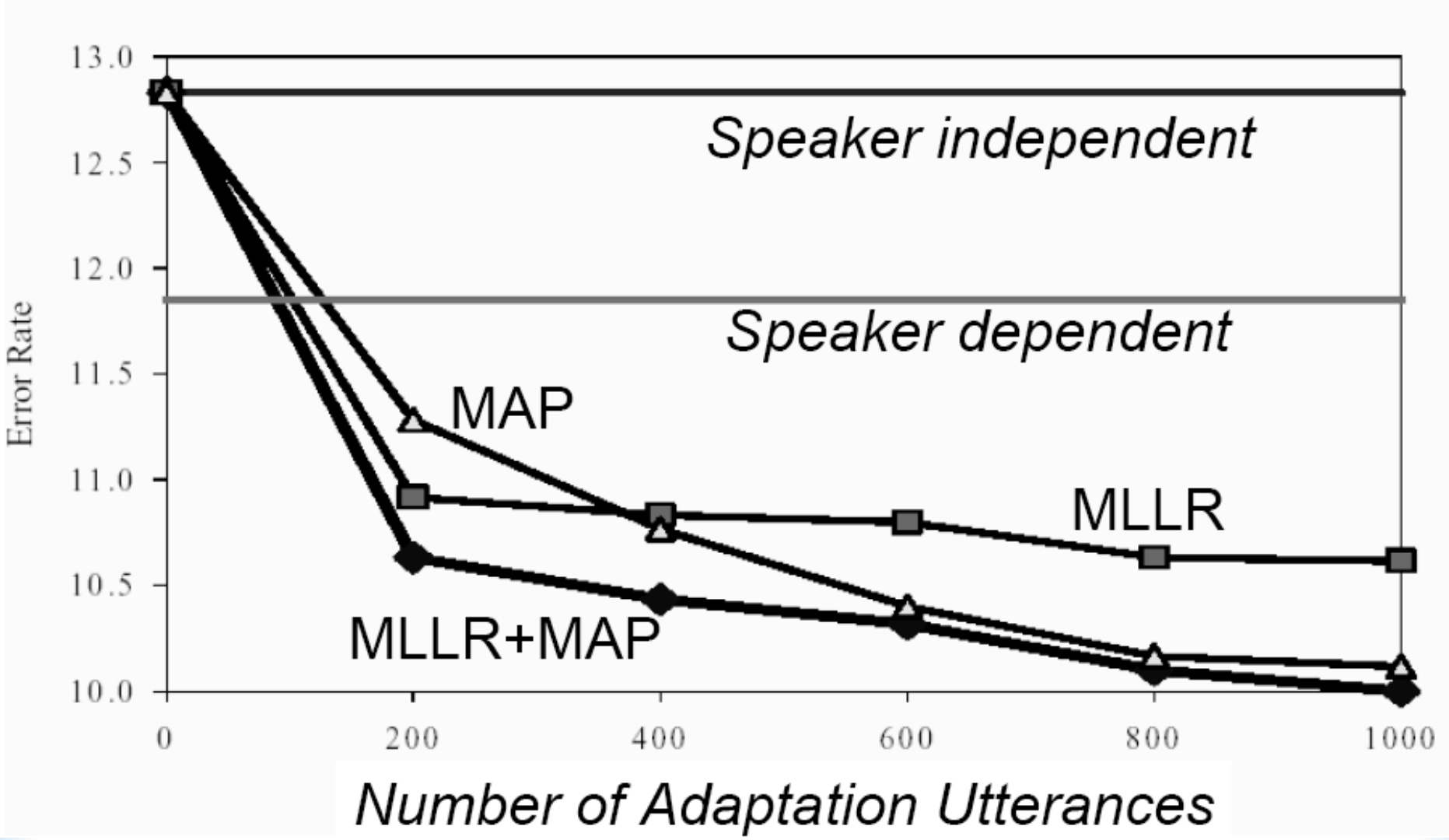
- **MAP Adaptation can only be applied Gaussians that are "seen" in the test data,**

$$\mu_{new} = \frac{\hat{N}}{\hat{N} + \alpha} \hat{m}_{obs} + \frac{\alpha}{\hat{N} + \alpha} \mu_{old}$$

$\hat{N}$	Number of frames of adaptation data
α	Weight for prior estimate of old mean
$\hat{m}_{obs}$	Mean vector of adaptation data assigned to Gauss.

Slide from Bryan Pellom

Performance of MLLR and MAP



Slide from Bryan Pellom after Huang et al

Summary

- MLLR: works on small amounts of adaptation data
- MAP: Maximum A Posterior Adaptation
  - Works well on large adaptation sets
- Acoustic adaptation techniques are quite successful at dealing with speaker variability
- If we can get 10 seconds with the speaker.

Sources of Variability: Environment

- Noise at source
  - Car engine, windows open
  - Fridge/computer fans
- Noise in channel
  - Poor microphone
  - Poor channel in general (cellphone)
  - Reverberation
- Lots of research on noise-robustness
  - Spectral subtraction for additive noise
  - Cepstral Mean Normalization
  - Microphone arrays

Do you hear what the wise man says?

Note corrections in the notes about recognition accuracy!!!

SNR = -18 dB 

SNR = -14 dB 

SNR=-10 dB 

Human digit recognition exceeded 75% accuracy in half of utterances (tests on headphones)

SNR = -6 dB 

SNR = -2 dB 

SNR = 2 dB 

Machine digit recognition exeeded 75% accuracy in half of utterances (Palomäki & Brown submitted)

Slide from Kalle Palomäki, Ulpu Remes, Mikko Kurimo

What is additive noise?

- Sound pressure for two non coherent sources

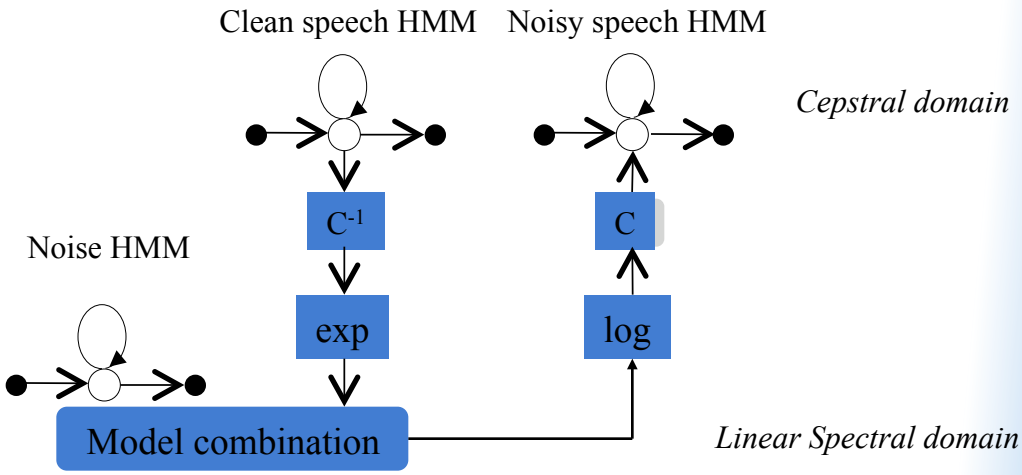
$$p^2 = p_s^2 + p_n^2$$

- p_s : speech
- p_n : noise source
- p : mixture of speech and noise sources

Slide from Kalle Palomäki, Ulpu Remes, Mikko Kurimo

Parallel Model Combination

- For Additive Noise
  - Best: train models with exact same noisy speech as test set - impossible
  - Instead: Collect noise in test, from the silence, generate a model
  - Combine the noise model and the clean-speech models in real-time
- Basic Approaches
  - performed on model parameters in cepstral domain
  - noise and signal are additive in linear domain rather than the cepstral domain, so transforming the parameters back to linear domain for combination
  - Can modify both the means and variances
- Parameters :
  - the clean speech models
  - a noise model



Slide from Li Lin Shan 李琳山

Sources of Variability: Speaker

- Gender
- Dialect/Foreign Accent
- Individual Differences
  - Physical differences
  - Language differences (“idiolect”)

Sources of Variability: Genre/ Style/Task

- Read versus conversational speech
- Lombard speech
- Domain (Booking flights versus managing stock portfolio)
- Emotion

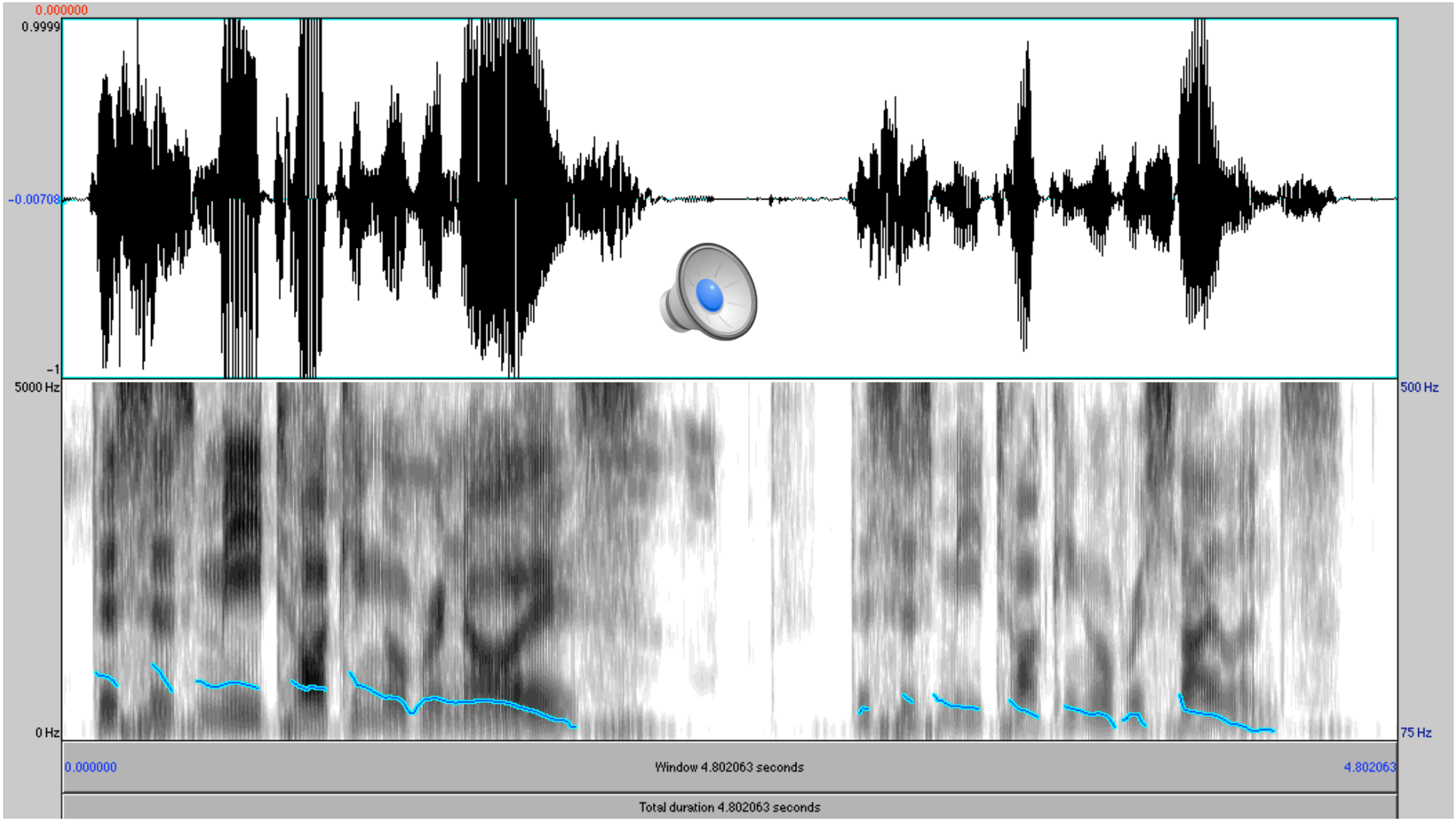
One simple example: The Lombard effect

- Changes in speech production in the presence of background noise
- Increase in:
  - Amplitude
  - Pitch
  - Formant frequencies
- Result: intelligibility increases

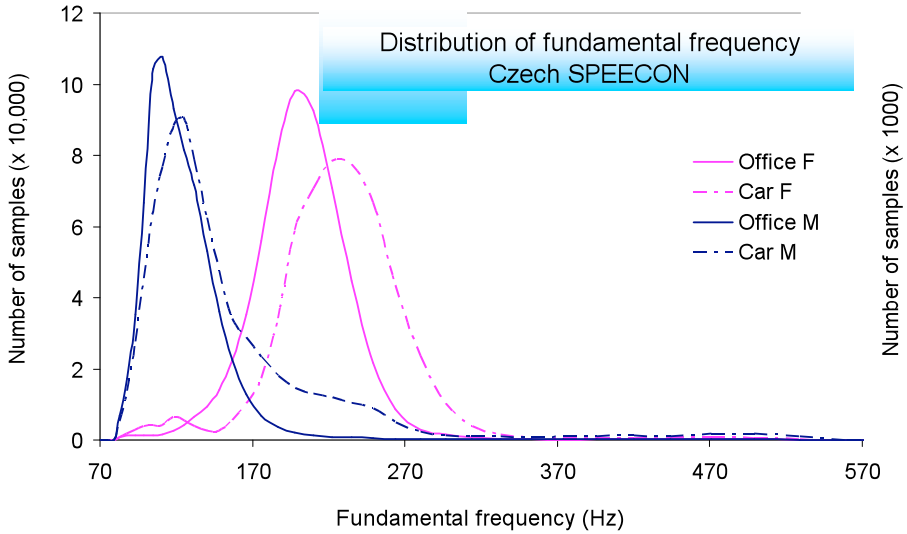
Lombard Speech

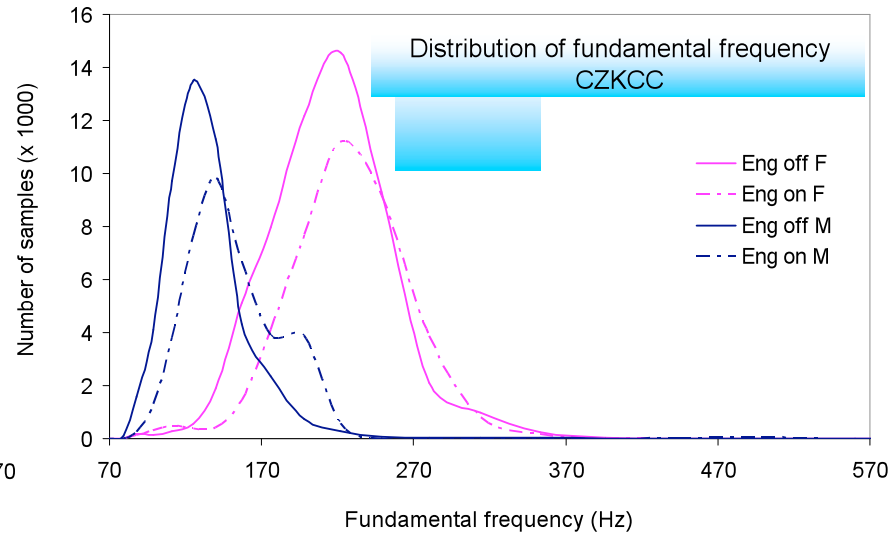
Me talking
over Ray Charles:
longer, louder, higher

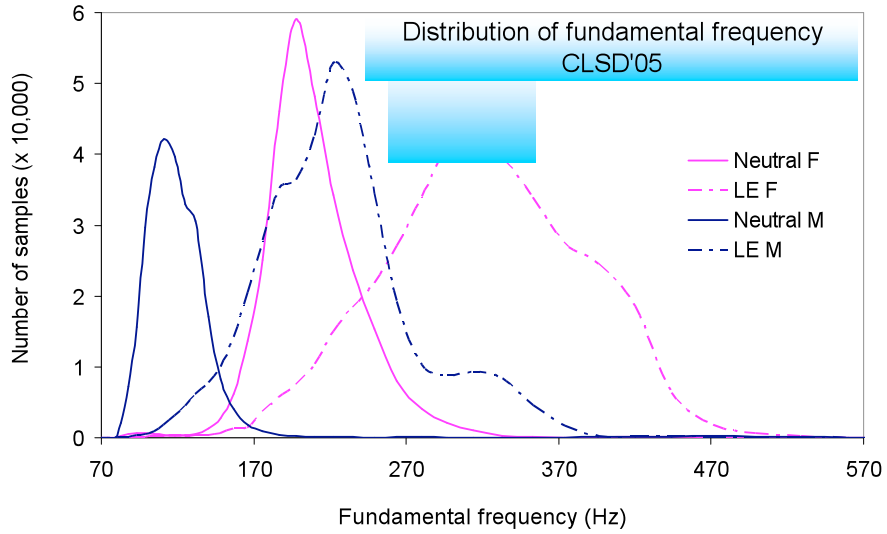
Me talking over
silence



Analysis of Speech Features under LE
Fundamental Frequency

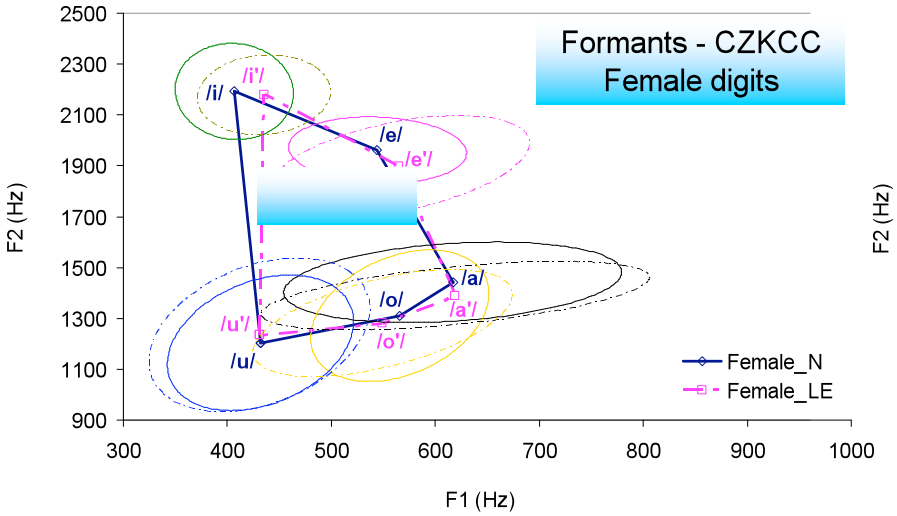


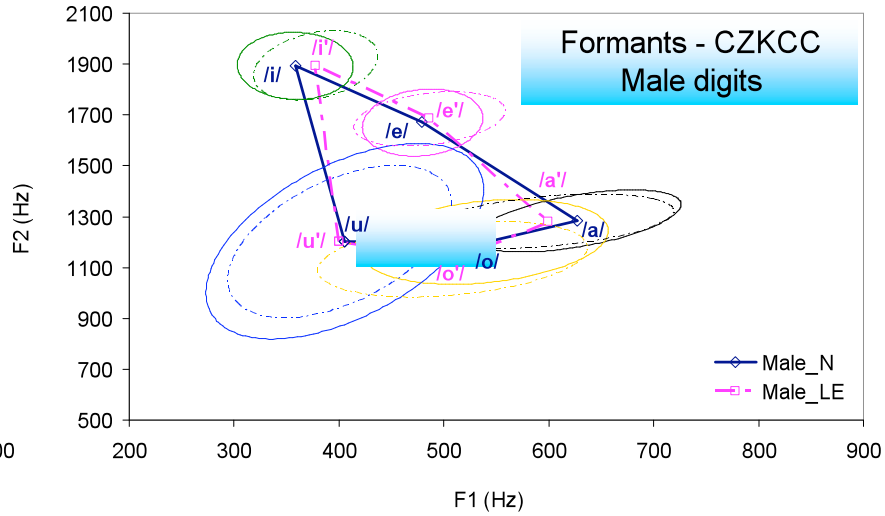


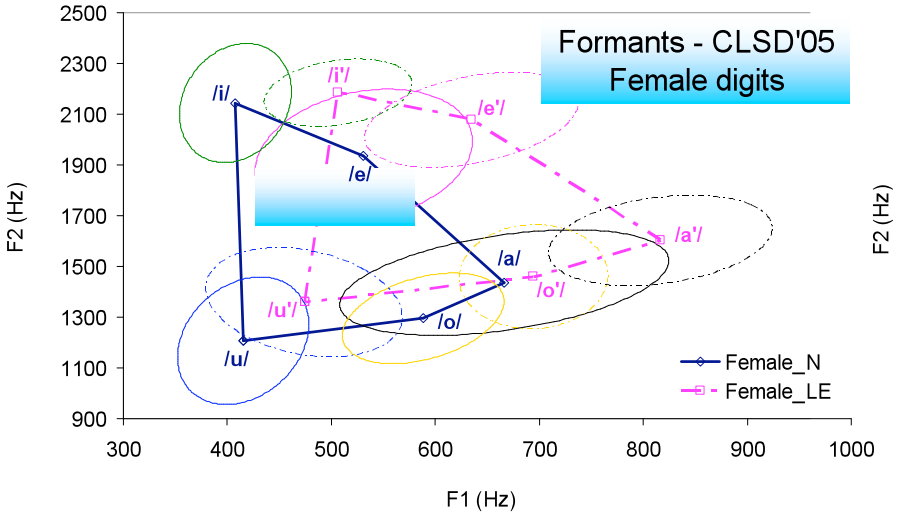


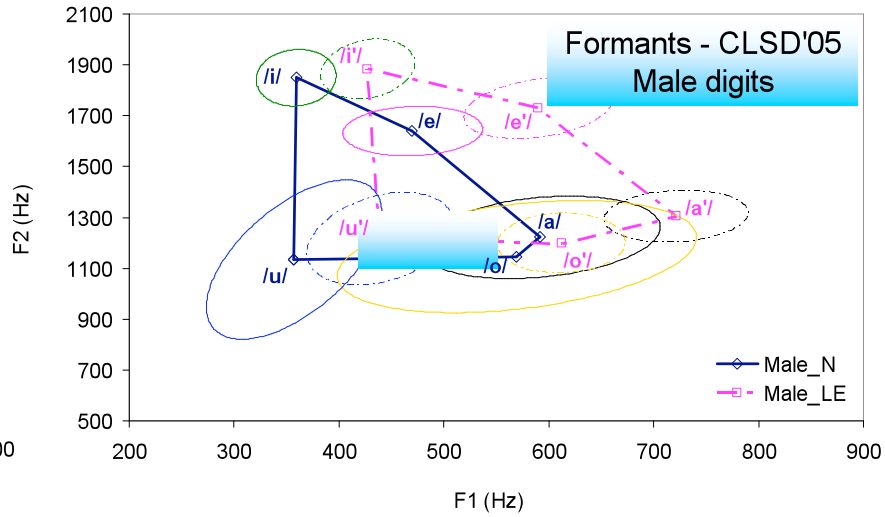
Slides from John Hansen

Analysis of Speech Features under LE
Formant Locations









Slides from John Hansen

How to deal with Lombard Effect? Adaptation

➢Model Adaptation

- Often effective when only limited data from given conditions are available
- **Maximum Likelihood Linear Regression (MLLR)** – if limited amount of data per class, acoustically close classes are grouped and transformed together

$$\boldsymbol{\mu}'_{MLLR} = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

- **Maximum a posteriori approach (MAP)** – initial models are used as informative priors for the adaptation

$$\boldsymbol{\mu}'_{MAP} = \frac{N}{N+\tau}\bar{\boldsymbol{\mu}} + \frac{\tau}{N+\tau}\boldsymbol{\mu}$$

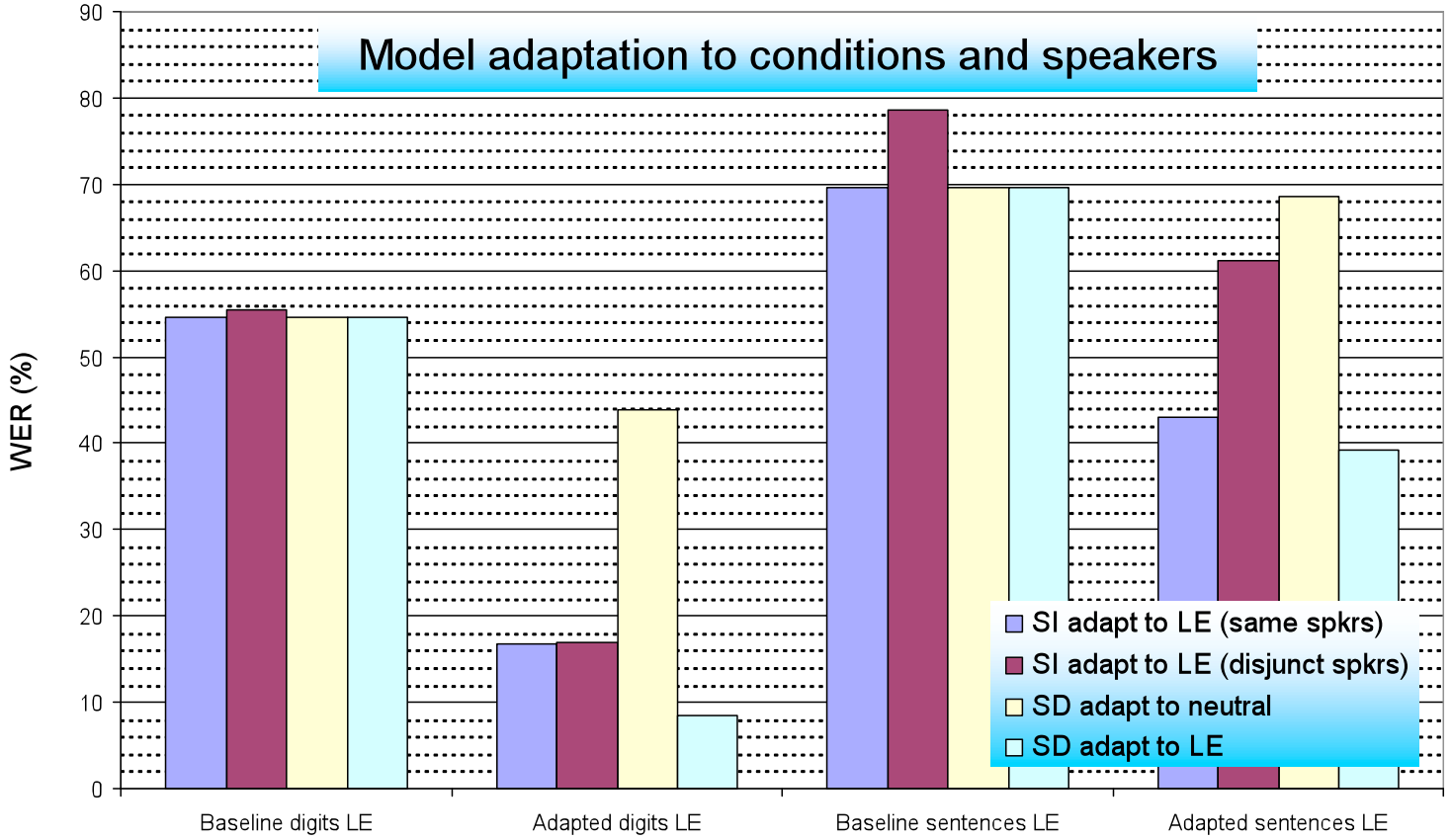
➢Adaptation Procedure

- First, neutral speaker-independent (SI) models transformed by MLLR, employing clustering (binary regression tree)
- Second, MAP adaptation – only for nodes with sufficient amount of adaptation data

Slides from John Hansen

Adaptation for Lombard Effect

- Adaptation Schemes
  - Speaker-independent adaptation (SI) – group dependent/independent
  - Speaker-dependent adaptation (SD) – to neutral/LE



Slides from John Hansen

Variation due to task/genre

- Probably largest remaining source of error in current ASR
- I.e., is an unsolved problem
- Maybe one of you will solve it!

Conversational Speech Genre effects

- Switchboard corpus
- I was like, “It’s just a stupid bug!”
- ax z l ay k ih s jh ah s t ey s t uw p ih b ah g

Variation due to the conversational genre

- Weintraub, Taussig, Hunicke-Smith, Snodgrass. 1996. Effect of Speaking Style on LVCSR Performance.
- SRI collected a spontaneous conversational speech corpus, in two parts:
  - 1. Spontaneous Switchboard-style conversation on an assigned topic
    - Here's an example from Switchboard, just to give a flavor
  - A reading session in which participants read transcripts of their own conversations
    - 2. As if they were dictating to a computer
    - 3. As if they were having a conversation

How do 3 genres affect WER?

- WER on exact same words:

Speaking Style	*Word Error*
Read Dictation	28.8%
Read Conversational	37.6%
Spontaneous Conversation	52.6%

Weintraub et al conclusions

- Speaking style is a large factor in what makes conversational speech hard
  - It's not the LM: words were identical
- Even "simulated natural" speech is harder than read speech
- "Natural" conversational speech is harder still.
- Speaking style is due to the AM:
  - Pronunciation model
  - Output likelihoods
- This kind of variation not captured by current triphone systems

Source of variation

- Acoustic variation
- Pronunciation variation
  - HMMs built from pronunciation dictionary
  - What if strings of phones don't match phones in dictionary!

- ax z l ay k ih s jh ah s t ey s t uw p ih b ah g

- I was: ax z
- It's: ih s

Pronunciation variation in conversational speech is source of error

- Saraclar, M, H. Nock, and S. Khudanpur. 2000. Pronunciation modeling by sharing Gaussian densities across phonetic models. Computer Speech and Language 14:137-160.
- "Cheating experiment" or "Oracle experiment"
  - In general, asks how well one could do if one had some sort of oracle giving perfect knowledge.
- Switchboard task
- 1) Extracted the actual pronunciation of each word in test set
  - Run phone recognition on test speech
  - Align this phone string with reference word transcriptions for test set
  - Extract observed pronunciation of each word
- Many of these pronunciations different than canonical pronunciation

Saraclar et al. 2000

- Now we have an "alternative" pronunciation for many words in test set.
- Now enhance the pronunciation dictionary used during recognition in two ways:
  - 1) Create "global oracle dictionary":
    - Add new pronunciations for any words in test set to static pronunciation dictionary
  - 2) Create "per-sentence oracle dictionary":
    - Create a new dictionary for each sentence with the new pronunciations seen in that sentence.

Saraclar et al results

- Use the 2 dictionaries to rescore lattices

Speaking Style	*Word Error*
Baseline SWBD system	**47%**
Static Global Oracle Dictionary	**38%**
Per-sentence Oracle Dictionary	**27%**
Lattice error rate	**13%**

Implications

- If you knew (roughly) which pronunciation to use for each sentence
- Could cut WER from 47% to 27%!

What kinds of pronunciation variation?

- Bernstein, Baldwin, Cohen, Murveit, Weintraub (1986)
- Conversational speech is faster than read speech in words/second
  But is similar to read speech in phones/second!
- In spontaneous speech
  - It's not that each phone is shorter
  - Rather, phones are deleted
- Fosler et al (1996) on switchboard
  - 12.5% of phones deleted
  - Other phones altered
  - Only 67% of phones same as canonical

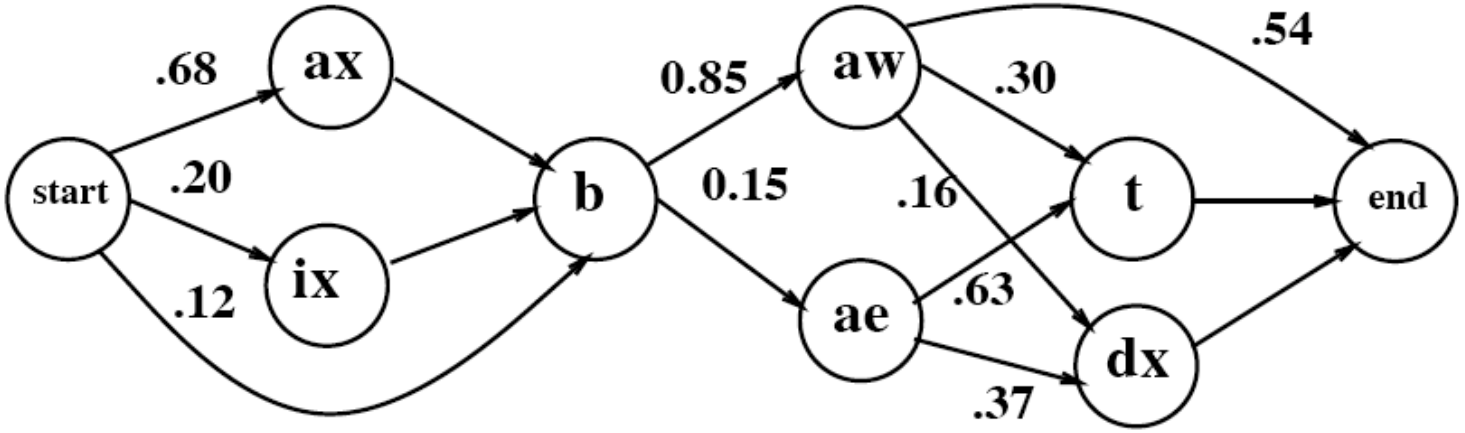
SWBD pronunciation of "because" and "about"

From ICSI labels (Greenberg et al. 1996, Greenberg 1999)

because			about		
IPA	**ARPAbet**	**%**	**IPA**	**ARPAbet**	**%**
[bikʌz]	[b iy k ah z]	27%	[əbaʊ]	[ax b aw]	32%
[bɨkʌz]	[b ix k ah z]	14%	[əbaʊt]	[ax b aw t]	16%
[kʌz]	[k ah z]	7%	[baʊ]	[b aw]	9%
[kəz]	[k ax z]	5%	[ʌbaʊ]	[ix b aw]	8%
[bɨkəz]	[b ix k ax z]	4%	[ɨbaʊt]	[ix b aw t]	5%
[bɪkʌz]	[b ih k ah z]	3%	[ɨbæ]	[ix b ae]	4%
[bəkʌz]	[b ax k ah z]	3%	[əbæɾ]	[ax b ae dx]	3%
[kʊz]	[k uh z]	2%	[baʊɾ]	[b aw dx]	3%
[ks]	[k s]	2%	[bæ]	[b ae]	3%
[kɨz]	[k ix z]	2%	[baʊt]	[b aw t]	3%

Pronunciation modeling methods

- Allophone networks (Cohen 1989)

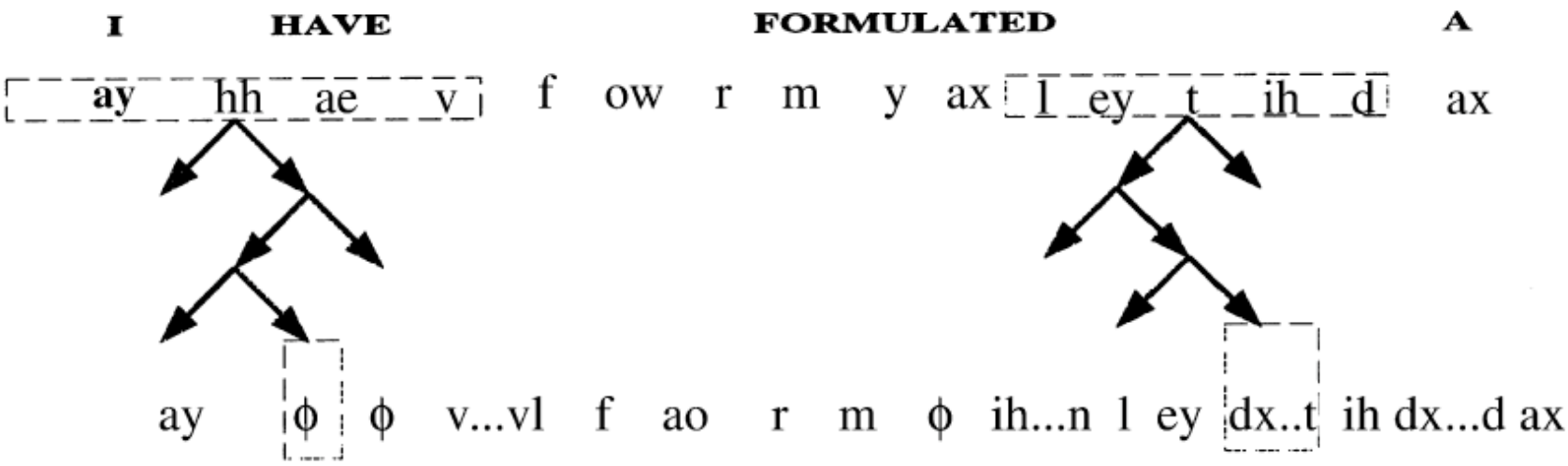


- C
- C
- F

H)

Pronunciation modeling methods

- Decision trees (Riley et al 1999, inter alia)
- Take phonetically hand-labeled data
- Building decision tree to predict "surface" form from "dictionary" form



Problem with all these methods

- They don't seem to work!
  - Phone-based decision trees (Riley et al 1999)
  - Phonological Rules (Tajchman et al. 1995)
  - Adding multiple pronunciations (Saraclar 1997, Tajchman et al. 1995)
- Why not?

Why don't these "use the phonetic context" methods work for pronunciation modeling?

- Error analysis experiment (Jurafsky et al 2001)
- Idea:
  - Give a triphone recognizer iteratively more training data
  - Look at what kinds of pronunciation variation it gets better at
  - Look at what kinds of pronunciation variation it doesn't get better at
  - A kind of "error-analysis-of-the-learning-curve"

The idea: compare forced alignment scores from two different lexicons

- One is 'canonical' dictionary pronunciation
- One is 'cheating' or 'surface' lexicon of actual pronunciation
  - Just like Saraclar et al. 2000
  - A collection of 2780 lexicons
  - One for each sentence in test set
  - Pronunciation for each word taken from ICSI hand-labels (Greenberg et al. 1996) converted to triphones

Example: two lexicons for "That is right"

Word	Canonical lexicon	Surface lexicon
that	dh ae t	dh ae
is	ih z	s
right	r ay t	r ay

Which sentences were handled by a simple lexicon after more training?

- Run forced alignment twice for each of 2780 sentences
  - Surface lexicon from phonetic transcriptions
  - Canonical lexicons from dictionary
- For each sentence, which which lexicon has higher likelihood: SURFACE or CANONICAL
- Now can look at what kind of sentences get higher scores with which lexicons

Which sentences were handled by a simple lexicon after more training?

- Stage 1: bootstrap acoustic models
- Stage 2: more training of acoustic models on SWBD
  - Look at sentences that
    - fit SURFACE lexicon better at stage 1
    - fit CANONICAL lexicon better at stage 2
  - In other words, sentences whose score with canonical lexicon improved after triphones had more exposure to data

Which sentences were handled by a simple lexicon after more training?

- We thus compared the following sets of sentences:
  - 1. "GOT BETTER": 807 sentences that began with higher scores from SURFACE lexicon, but ended up with higher scores from CANONICAL lexicon
  - 2. "STAYED": 1047 sentences that began with higher scores from SURFACE lexicon and stayed that way
- Our question:
  - what kinds of variation
  - caused certain sentences (in GOT BETTER) to improve with a canonical lexicon as their triphones see more data,
  - but others (STAYED) do not improve

Question 1: Are sentences with syllable deletions hard for triphones to model?

	Stayed	Got better
% Syllables deleted	3.3%	1.8%

- Result: "GOT BETTER" sentences had less deletion ($p < .05$)
- Conclusion: Syllable deletion is not well modeled by simply having more training data for the triphones

Question 2: Are sentences with phone substitutions hard for triphones to model?

- "Assimilation", "coarticulation"
- Would you /w uh d y uw/ -> /w uh d jh uw/
- Is /ih z/ -> /ih s/
- Because /b iy k ah z/, /b ix k ah z/, /b ix k ax z/, /b ax k ah z/

	Stayed	Got better
% of Phone substitutions	7.0%	7.2%

- Result: No difference
- Conclusion: triphones may do OK at modeling phone substitutions

Previous pronunciation modeling methods only model PHONETIC variability

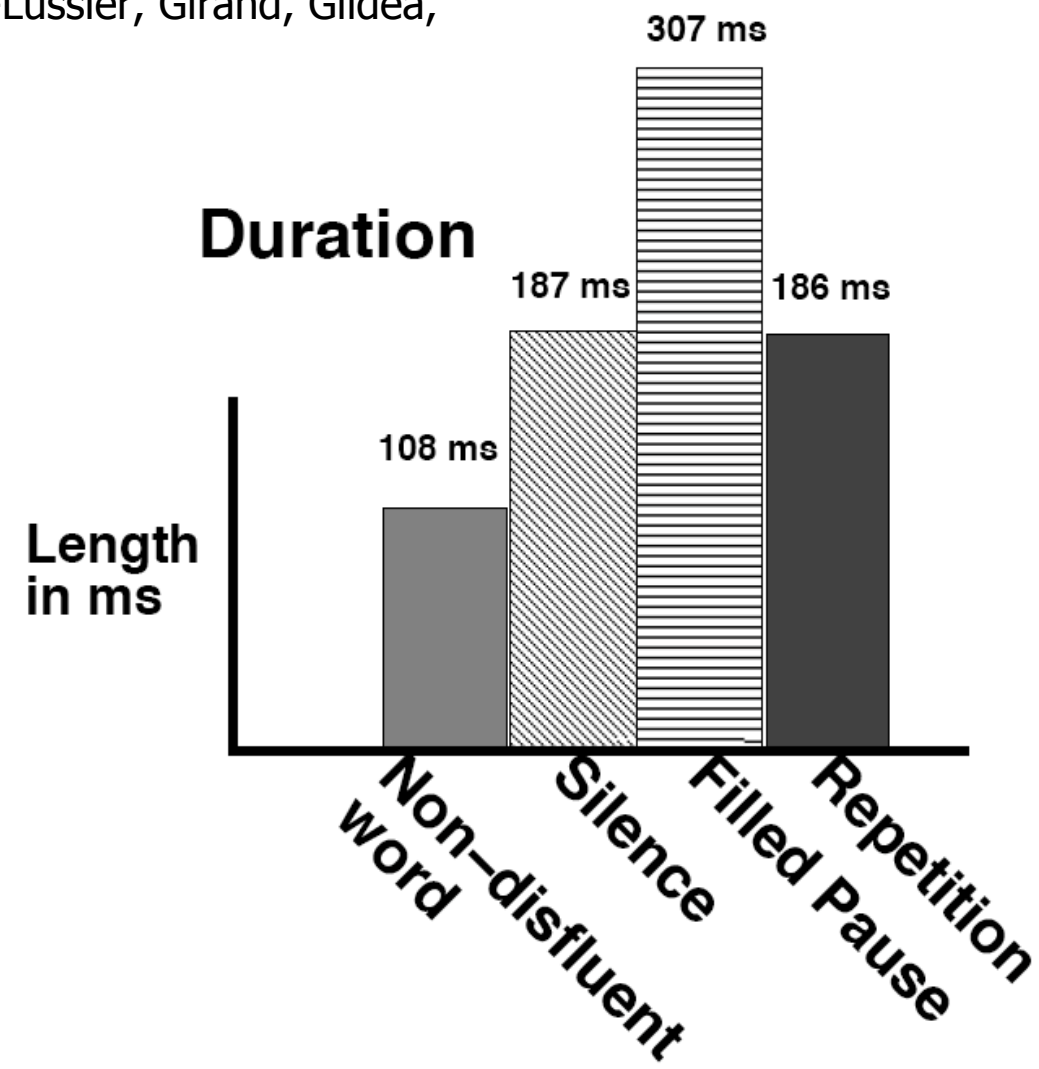
- Summary:
  - Massive deletion of phones/syllables is not solved by triphones
  - Other kinds of phonetic variation are solved by triphones
- Previous methods only capture phonetic variability due to neighboring phones
- But triphones already capture this!
- The difficult variability is caused by other factors

What causes massive deletion/ shortening/lengthening?

- Neighboring disfluencies
- Hyperarticulation
- Rate of speech (syllables/second)
- Prosodic boundaries (beginning and end of utterance)
- Word predictability (LM probability)

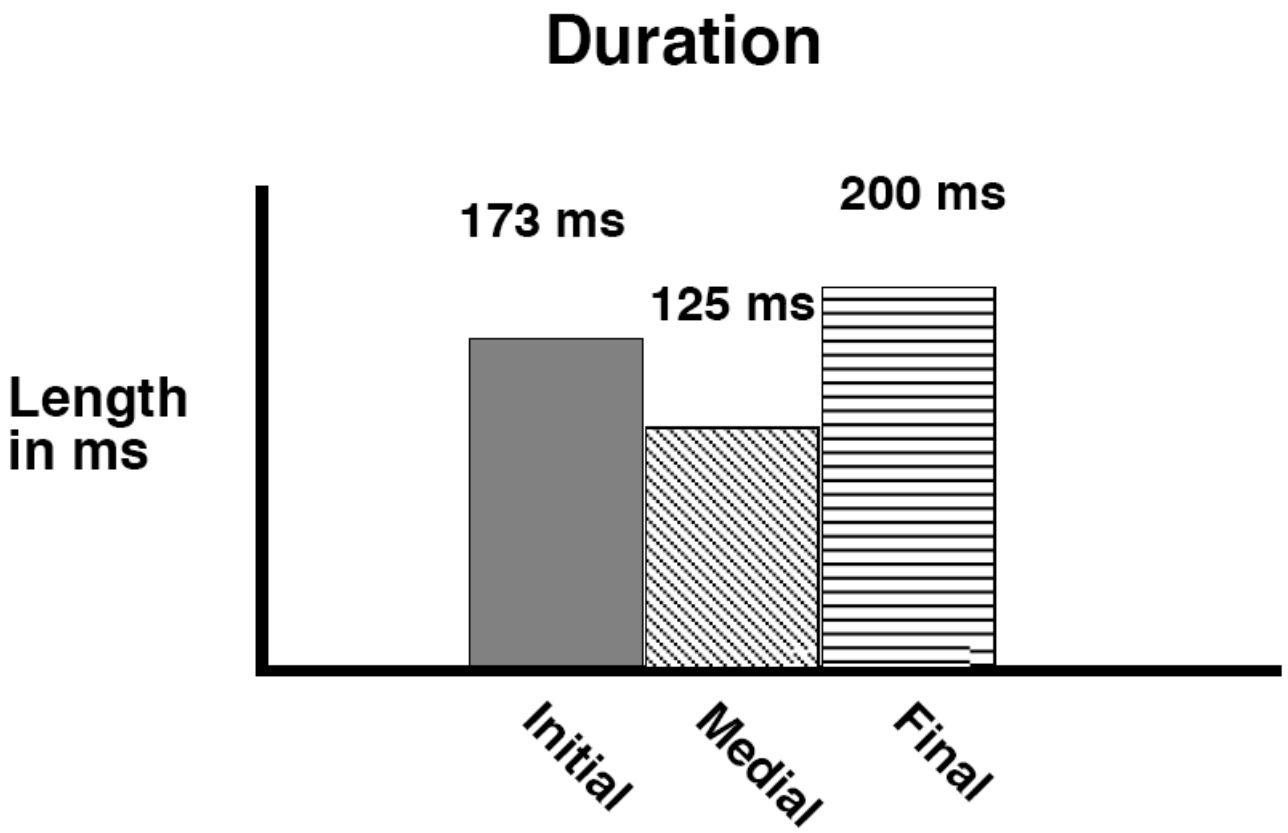
Effect of disfluencies on pronunciation

- Bell, Jurafsky, Fosler-Lussier, Girand, Gildea, Gregory (2003)

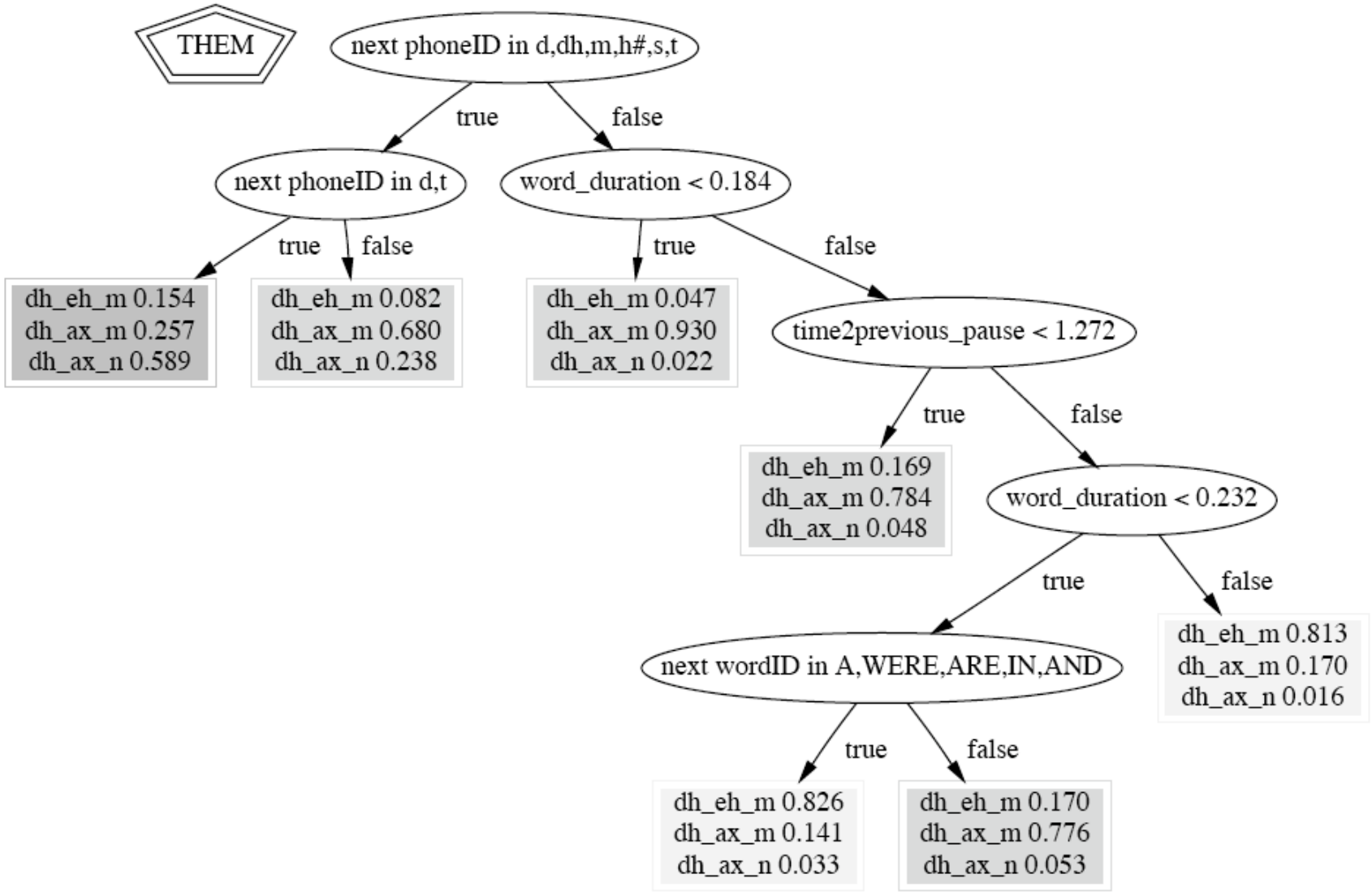


Effect of position in utterance on pronunciation

- Bell, Jurafsky, Fosler-Lussier, Girand, Gildea, Gregory (2003)



Adding pauses, neighboring words into pronunciation model - Fosler-Lussier (1999)



Variation due to (foreign or regional) accent

- Sample (old) result from Byrne et al
  - Strongly Spanish-accented conversational data
  - Baseline recognizer performance
    - Train on SWBD, test on SWBD: 42%
  - On Spanish-accented data
    - Train on SWBD, MLLR on accented data: 66%
- These are old numbers
- But the basic idea is still the same
- Accent is an important cause of error in current recognizers!!

Accents: An experiment

- A word by itself 
- The word in context 

Acoustic Adaptation to Foreign Accent

- Train on accented data
  - Wang, Schultz, Waibel (2003) VERBMOBIL
    - Training on 52 minutes German-accented English WER=43.5%
    - Training on 34 hours of native English (same domain) WER=49.3%
- Pool accented + unaccented data
    - Training on 34 hours (native) + 52 minutes (accented) WER=42.3%
- Interpolating with "oracle" weight
    - WER=36.0%

Acoustic Adaptation to Foreign Accent

- Train on native speech, run a few additional forward-backward iterations on non-native speech
  - Mayfield-Tomokiyo and Waibel (2001)
  - Japanese-accented English in VERBMOBIL
  - 63%: Native English training only:
  - 53%: Pooling accented + native data:
  - 48%: Native English training + 2 EM passes on accented data:

MLLR and MAP for foreign accent

- Combine MLLR and MAP
- Most successful approach
- Most people use now.

Pronunciation modeling in current CTS recognizers

- Use single pronunciation for a word
- How to choose this pronunciation?
  - Generate many pronunciations
  - Forced alignment on training set
  - Do some merging of similar pronunciations
  - For each pronunciation in dictionary
    - If it occurs in training, pick most likely pronunciation
    - Else learn some mappings from seen pronunciations, apply these to unseen pronunciations

Another way of capturing variation in pronunciation in CTS: Multiwords

- Finke and Waibel, Stolcke et al (2000)
- Grab frequently occurring bigram/trigrams
- Going to, a lot of, want to
- Hand-write a pronunciation for each
  - 1300 "multiwords", 1800 pronunciations
- A lot of: 3 pronunciations
  - REDUCED: ax l aa dx ax
  - CANONICAL: ax l ao t ah v
  - CANONICAL WITH PAUSES ax - l ao t - ah v
- Retrain language model with 1300 new multiwords

Summary

- Lots of sources of variation.
  - Noise
    - Model-based methods
    - Spectral subtraction,
    - Cepstral mean normalization
    - Microphone arrays
  - Speaker variation
    - VTLN
    - MLLR
    - MAP
- Open problems
  - Genre variation
    - Especially human-human conversation, meetings, etc
  - Pronunciation modeling in general
  - Language model adaptation: some recent work on this