

CS 224S / LINGUIST 281 Speech Recognition and Synthesis

Dan Jurafsky

Lecture 10: Variation and Adaptation

IP Notice: Some slides adapted from Bryan Pellom; acoustic modeling material derived from Huang et al.

2/15/06

CS 224S Winter 2006

1

Outline

- Variation in speech recognition
- Sources of Variation
- Three classic problems:
 - Dealing with phonetic variation
 - Speaker/Environment adaptation
 - MLLR, other acoustic adaptation techniques
 - Pretty decent solution
 - Variation due to Genre: Conversational Speech
 - Pronunciation modeling issues
 - Unsolved!

2/15/06

CS 224S Winter 2006

2

Sources of Variability

- Phonetic context
- Environment
- Speaker
- Genre/Task

2/15/06

CS 224S Winter 2006

3

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

2/15/06

CS 224S Winter 2006

4

Sources of Variability: Speaker

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

2/15/06

CS 224S Winter 2006

5

Sources of Variability: Genre/Style/Task

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

2/15/06

CS 224S Winter 2006

6

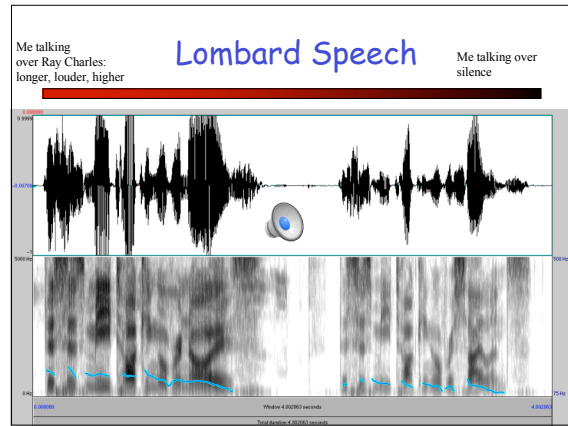
One simple example: The Lombard effect

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

2/15/06

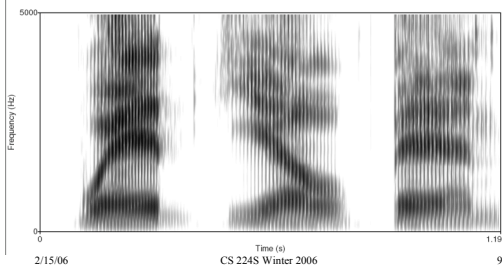
CS 224S Winter 2006

7



Most important: phonetic context: different "eh"s

• w eh d y eh l b eh n



2/15/06

CS 224S Winter 2006

9

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

2/15/06

CS 224S Winter 2006

10

CD phones: triphones

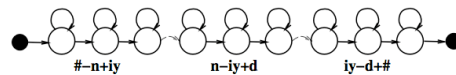
- Triphones
- Each triphone captures facts about preceding and following phone
- Monophone:
 - p, t, k
- Triphone:
 - iy-p+aa
 - a-b+c means "phone b, preceding by phone a, followed by phone c"

2/15/06

CS 224S Winter 2006

11

"Need" with triphone models



2/15/06

CS 224S Winter 2006

12

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.**

2/15/06

CS 224S Winter 2006

13

Implications of Cross-Word Triphones

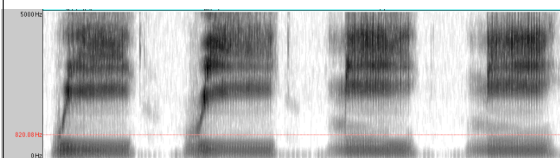
- Possible triphones: $50 \times 50 \times 50 = 125,000$
- How many triphone types actually occur?
- 20K word WSJ Task, numbers from Young et al
- Cross-word models: need 55,000 triphones
- But in training data only 18,500 triphones occur!
- Need to generalize models.

2/15/06

CS 224S Winter 2006

14

Modeling phonetic context: some contexts look similar



W iy r iy m iy n iy

2/15/06

CS 224S Winter 2006

15

Solution: State Tying

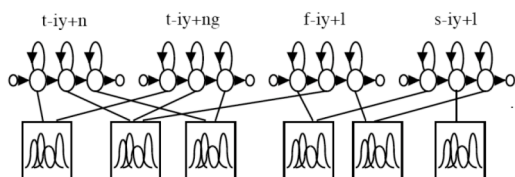
- Young, Odell, Woodland 1994
- Decision-Tree based clustering of triphone states
- States which are clustered together will share their Gaussians
- We call this "state tying", since these states are "tied together" to the same Gaussian.
- Previous work: generalized triphones
 - Model-based clustering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

2/15/06

CS 224S Winter 2006

16

Young et al state tying



2/15/06

CS 224S Winter 2006

17

State tying/clustering

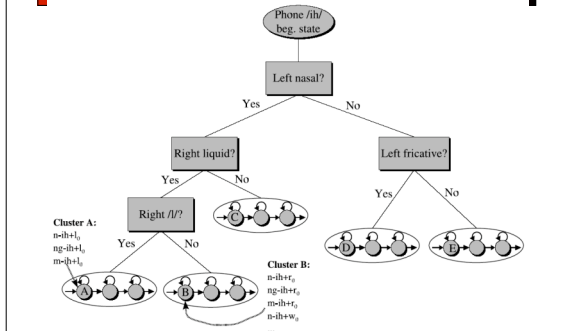
- How do we decide which triphones to cluster together?
- Use **phonetic features** (or 'broad phonetic classes')
 - Stop
 - Nasal
 - Fricative
 - Sibilant
 - Vowel
 - lateral

2/15/06

CS 224S Winter 2006

18

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 ŋ
Fricative	ch dh f j h 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 oy ow uh
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 th z zh

2/15/06

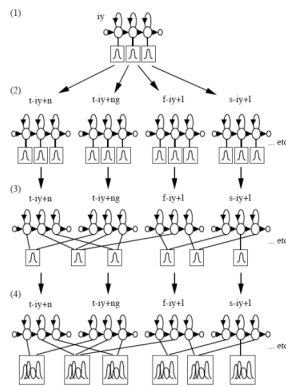
CS 224S Winter 2006

20

State Tying:

Young, Odell, Woodland 1994

- The steps in creating CD phones.
- Start with monophone, do EM training
- Then clone Gaussians into triphones
- Then build decision tree and cluster Gaussians
- Then clone and train mixtures (GMMs)



2/15/06

CS 2:

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

2/15/06

CS 224S Winter 2006

22

The rest of today's lecture

- Variation due to speaker differences
 - Speaker adaptation
 - MLLR
 - VTLN
 - Splitting acoustic models by gender
 - Foreign accent
 - Acoustic and pronunciation adaptation to accent
- Variation due to genre differences
 - Pronunciation modeling

2/15/06

CS 224S Winter 2006

23

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/NL 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%

2/15/06

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!

2/15/06

CS 224S Winter 2006
Slide from Bryan Pellom

25

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.

2/15/06

CS 224S Winter 2006

26

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

2/15/06

CS 224S Winter 2006
Slide from Bryan Pellom

27

MLLR

- New equation for output likelihood

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

2/15/06

CS 224S Winter 2006

28

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.

2/15/06

CS 224S Winter 2006

29

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 $\xi_j(t)$

2/15/06

CS 224S Winter 2006

30

MLLR performance on RM (Leggetter and Woodland 1995)

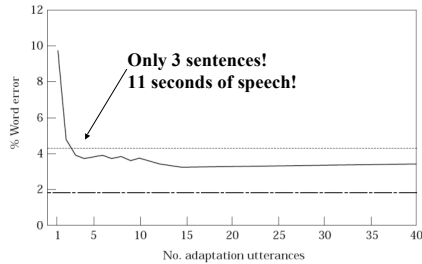


Figure 2. Full matrix maximum likelihood linear regression using global regression class. (.....), Speaker independent; (-.-.-.-), speaker dependent; (—), speaker adapted.

31

MLLR doesn't need supervised adaptation set!

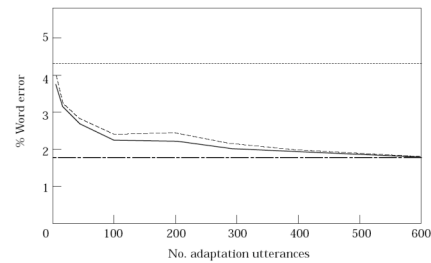


Figure 3. Supervised vs. unsupervised adaptation using maximum likelihood linear regression. (.....), Speaker independent; (-.-.-.-), speaker dependent; (—), supervised adapted; (---), unsupervised adapted.

2/15/0

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.

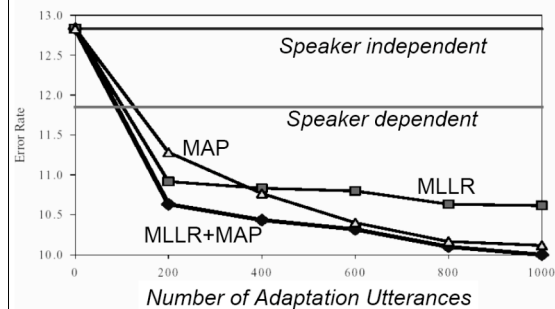
2/15/06

CS 224S Winter 2006

33

Slide from Bryan Pellom

Performance of MLLR and MAP



2/15/06

CS 224S Winter 2006

34

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.

2/15/06

CS 224S Winter 2006

35

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!

2/15/06

CS 224S Winter 2006

36

Switchboard example in Praat

2/15/06

CS 224S Winter 2006

37

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



2/15/06

CS 224S Winter 2006

38

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

2/15/06

CS 224S Winter 2006

39

How do 3 genres affect WER?

- WER on exact same words:

<i>Speaking Style</i>	<i>Word Error</i>
Read Dictation	28.8%
Read Conversational	37.6%
Spontaneous Conversation	52.6%

2/15/06

CS 224S Winter 2006

40

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

2/15/06

CS 224S Winter 2006

41

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

2/15/06

CS 224S Winter 2006

42

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

2/15/06

CS 224S Winter 2006

43

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.

2/15/06

CS 224S Winter 2006

44

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%

2/15/06

CS 224S Winter 2006

45

Implications

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

2/15/06

CS 224S Winter 2006

46

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

2/15/06

CS 224S Winter 2006

47

SWBD pronunciation of "because" and "about"

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

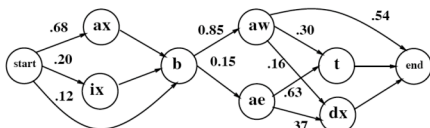
because			about		
IPA	ARPAbet	%	IPA	ARPAbet	%
[bɪkəz]	[b iy k ah z]	27%	[əbau]	[ax b aw]	32%
[bɪkəz]	[b ix k ah z]	14%	[əbaut]	[ax b aw t]	16%
[kəz]	[k ah z]	7%	[bau]	[b aw]	9%
[kəz]	[k ax z]	5%	[ɒbau]	[ix b aw]	8%
[bɪkəz]	[b ix k ax z]	4%	[ɪbaut]	[ix b aw t]	5%
[bɪkəz]	[b ih k ah z]	3%	[ɪbæ]	[ix b æ]	4%
[bəkəz]	[b ax k ah z]	3%	[əbæɹ]	[ax b æ dx]	3%
[kuz]	[k uh z]	2%	[baʊɹ]	[b aw dx]	3%
[ks]	[k s]	2%	[bæ]	[b æ]	3%
[kɪz]	[k ix z]	2%	[baut]	[b aw t]	3%

2

48

Pronunciation modeling methods

- Allophone networks (Cohen 1989)



- Or by automata-induction from strings (Wooters and Stolcke 1994)
- Probabilities from forced-alignment on training data

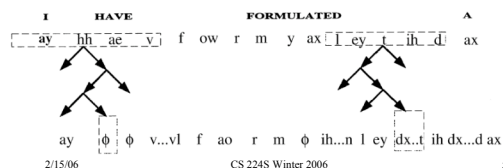
2/15/06

CS 224S Winter 2006

49

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



2/15/06

CS 224S Winter 2006

50

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?

2/15/06

CS 224S Winter 2006

51

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"

2/15/06

CS 224S Winter 2006

52

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

2/15/06

CS 224S Winter 2006

53

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

2/15/06

CS 224S Winter 2006

54

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

2/15/06

CS 224S Winter 2006

55

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

2/15/06

CS 224S Winter 2006

56

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

2/15/06

CS 224S Winter 2006

57

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

2/15/06

CS 224S Winter 2006

58

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

2/15/06

CS 224S Winter 2006

59

Previous pronunciation modeling methods only model PHONETIC variability

- Previous methods only capture phonetic variability due to neighboring phones
- But triphones already capture this!
- So phonetic variability is already well-captured by triphone models
- The difficult variability is caused by other factors

2/15/06

CS 224S Winter 2006

60

Some non-phonetic predictive factors for pronunciation change

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

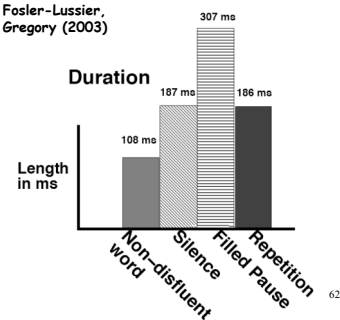
2/15/06

CS 224S Winter 2006

61

Effect of disfluencies on pronunciation

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

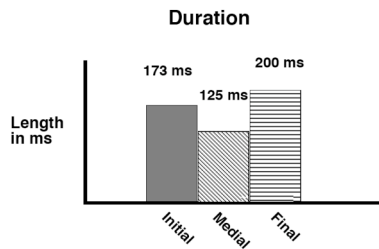


2/15/06

62

Effect of position in utterance on pronunciation

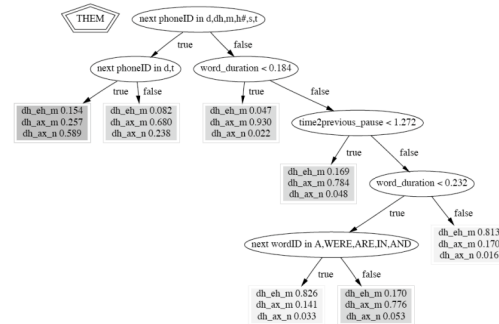
- Bell, Jurafsky, Fosler-Lussier, Girand, Gildea, Gregory (2003)



2/15/06

63

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!!

2/15/06

CS 224S Winter 2006

65

Accent example in Praat

2/15/06

CS 224S Winter 2006

66

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%

2/15/06

CS 224S Winter 2006

67

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:

2/15/06

CS 224S Winter 2006

68

MLLR and MAP for foreign accent

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

2/15/06

CS 224S Winter 2006

69

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

2/15/06

CS 224S Winter 2006

70

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

2/15/06

CS 224S Winter 2006

71

Summary

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

2/15/06

CS 224S Winter 2006

72