

CS 224S / LINGUIST 281 Speech Recognition and Synthesis

Dan Jurafsky

Lecture 5: Prosodic Processing for TTS, plus brief section on LTS rules

IP Notice: many of these slides come directly from two lectures of Jennifer Venditti on intonation (thanks!); lots of other info in these slides is from Alan Black's excellent TTS lecture notes.

1/24/06

CS 224S Winter 2006

1

Outline

- Letter to Sound Rules
- Thinking about F0
- Accent Placement and Intonational Tunes
- Intonational Phrasing and Disambiguation
- The TOBI Prosodic Transcription Theory
- Producing Intonation in TTS
 - Predicting Accents
 - Predicting Boundaries
 - Predicting Duration
 - Generating F0

1/24/06

CS 224S Winter 2006

2

Part I: Letter-to-Sound Rules

- Festival LTS rules:
(LEFTCONTEXT [ITEMS] RIGHTCONTEXT = NEWITEMS)
- Example:
 - (# [c h] C = k)
 - (# [c h] = ch)
- # denotes beginning of word
- C means all consonants
- Rules apply in order
 - "christmas" pronounced with [k]
 - But word with ch followed by non-consonant pronounced [ch]
 - E.g., "choice"

1/24/06

CS 224S Winter 2006

3

What about stress: practice

- Generally
- Pronounced
- Exception
- Dictionary
- Significant
- Prefix
- Exhale
- Exhalation
- Sally

1/24/06

CS 224S Winter 2006

4

Stress rules in LTS

- English famously evil: one from Allen et al 1987
- $V \rightarrow [1\text{-stress}] / X_C^* \{V_{\text{short}} C C?|V\} \{[V_{\text{short}} C^*|V]\}$
- Where X must contain all prefixes:
- Assign 1-stress to the vowel in a syllable preceding a weak syllable followed by a morpheme-final syllable containing a short vowel and O or more consonants (e.g. difficult)
- Assign 1-stress to the vowel in a syllable preceding a weak syllable followed by a morpheme-final vowel (e.g. oregano)
- etc

1/24/06

CS 224S Winter 2006

5

Modern method: Learning LTS rules automatically

- Induce LTS from a dictionary of the language
- Black et al. 1998
- Applied to English, German, French
- Two steps: alignment and (CART-based) rule-induction

1/24/06

CS 224S Winter 2006

6

Alignment

- Letters: c h e c k e d
- Phones: ch _ eh _ k _ t
- Black et al Method 1:
 - First scatter epsilons in all possible ways to cause letters and phones to align
 - Then collect stats for $P(\text{letter}|\text{phone})$ and select best to generate new stats
 - This iterated a number of times until settles (5-6)
 - This is EM (expectation maximization) alg

1/24/06

CS 224S Winter 2006

7

Alignment

- Black et al method 2
- Hand specify which letters can be rendered as which phones
 - C goes to k/ch/s/sh
 - W goes to w/v/f, etc
- Once mapping table is created, find all valid alignments, find $p(\text{letter}|\text{phone})$, score all alignments, take best

1/24/06

CS 224S Winter 2006

8

Alignment

- Some alignments will turn out to be really bad.
- These are just the cases where pronunciation doesn't match letters:
 - Dept d ih p aa r t m ah n t
 - CMU s iy eh m y uw
 - Lieutenant l eh f t eh n ax n t (British)
- Also foreign words
- These can just be removed from alignment training

1/24/06

CS 224S Winter 2006

9

Building CART trees

- Build a CART tree for each letter in alphabet (26 plus accented) using context of +-3 letters
- # # # c h e c -> ch
- c h e c k e d -> _
- This produces 92-96% correct LETTER accuracy (58-75 word acc) for English

1/24/06

CS 224S Winter 2006

10

Improvements

- Take names out of the training data
- And acronyms
- Detect both of these separately
- And build special-purpose tools to do LTS for names and acronyms

1/24/06

CS 224S Winter 2006

11

Names

- Big problem area is names
- Names are common
 - 20% of tokens in typical newswire text will be names
 - 1987 Donnelly list (72 million households) contains about 1.5 million names
 - Personal names: McArthur, D'Angelo, Jiminez, Rajan, Raghavan, Sondhi, Xu, Hsu, Zhang, Chang, Nguyen
 - Company/Brand names: Infnit, Kmart, Cytoc, Medamicus, Inforte, Aaon, Idexx Labs, Bebe

1/24/06

CS 224S Winter 2006

12

PART II: Intonation

1/24/06

CS 224S Winter 2006

13

Defining Intonation

- Ladd (1996) "Intonational phonology"
- "The use of **suprasegmental phonetic features**
Suprasegmental = above and beyond the segment/phone
 - F0
 - Intensity (energy)
 - Duration
- to convey **sentence-level pragmatic meanings**
 - I.e. meanings that apply to phrases or utterances as a whole, not lexical stress, not lexical tone.

1/24/06

CS 224S Winter 2006

14

Three aspects of prosody

- **Prominence**: some syllables/words are more prominent than others
- **Structure/boundaries**: sentences have prosodic structure
 - Some words group naturally together
 - Others have a noticeable break or disjuncture between them
- **Tune**: the intonational melody of an utterance.

1/24/06

CS 224S Winter 2006

From Ladd (1996) 15

Prosodic Prominence: Pitch Accents

A: What types of foods are a good source of vitamins?

B1: Legumes are a good source of VITAMINS.

B2: LEGUMES are a good source of vitamins.

- Prominent syllables are:
 - Louder
 - Longer
 - Have higher F0 and/or sharper changes in F0 (higher F0 velocity)

1/24/06

CS 224S Winter 2006

16
Slide from Jennifer Venditti

Prosodic Boundaries

I met Mary and Elena's mother at the mall yesterday.
I met Mary and Elena's mother at the mall yesterday.

French [bread and cheese]
[French bread] and [cheese]

1/24/06

CS 224S Winter 2006

17
Slide from Jennifer Venditti

Prosodic Tunes

Legumes are a good source of vitamins.
Are legumes a good source of vitamins?

1/24/06

CS 224S Winter 2006

18
Slide from Jennifer Venditti

TOPIC #1

Thinking about F0

1/24/06 CS 224S Winter 2006 19

Graphic representation of F0

1/24/06 CS 224S Winter 2006 20
Slide from Jennifer Venditti

The 'ripples'

F0 is not defined for consonants without vocal fold vibration.

1/24/06 CS 224S Winter 2006 21
Slide from Jennifer Venditti

The 'ripples'

... and F0 can be perturbed by consonants with an extreme constriction in the vocal tract.

1/24/06 CS 224S Winter 2006 22
Slide from Jennifer Venditti

Abstraction of the F0 contour

Our perception of the intonation contour abstracts away from these perturbations.

1/24/06 CS 224S Winter 2006 23
Slide from Jennifer Venditti

The 'waves' and the 'swells'

'wave' = accent

'swell' = phrase

1/24/06 CS 224S Winter 2006 24
Slide from Jennifer Venditti

TOPIC #2

Accent Placement and Intonational Tunes

1/24/06
CS 224S Winter 2006
25

Stress vs. accent

- Stress* is a structural property of a word — it marks a potential (arbitrary) location for an accent to occur, if there is one.
- Accent* is a property of a word in context — it is a way to mark intonational prominence in order to 'highlight' important words in the discourse.

(x)	(x)	(accented syll)
x	x	stressed syll
x	x	full vowels
x x x	x x x x	syllables
vi ta mins	Ca li for nia	

1/24/06
CS 224S Winter 2006
26

Slide from Jennifer Venditti

Stress vs. accent (2)

- The speaker decides to make the word **vitamin** more prominent by accenting it.
- Lexical stress tell us that this prominence will appear on the first syllable, hence **VI**tamin.

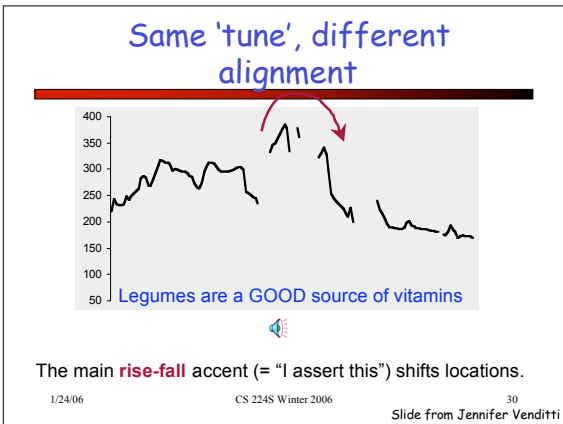
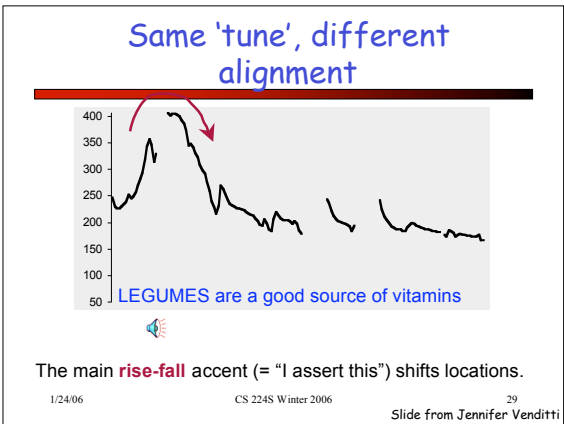
1/24/06
CS 224S Winter 2006
27

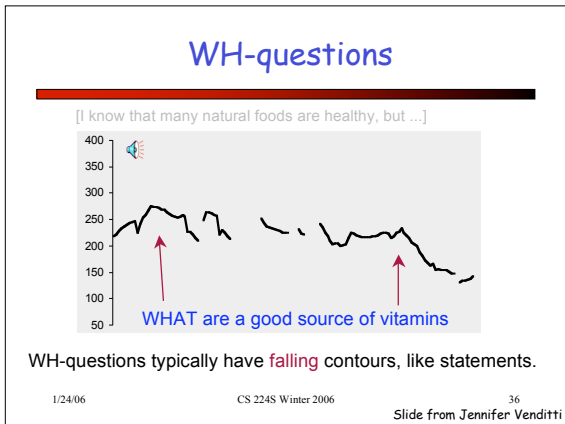
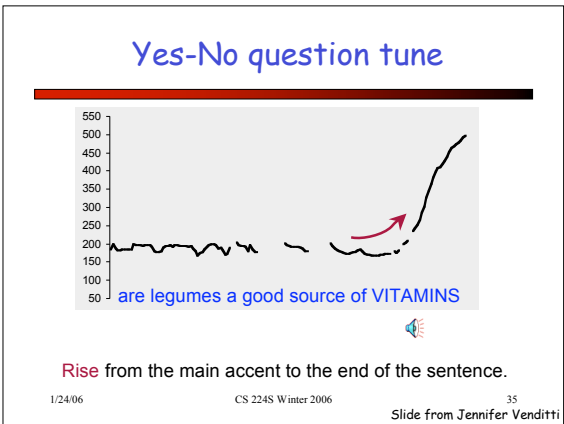
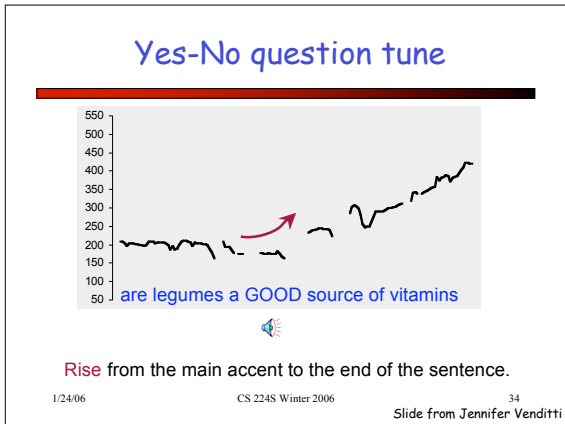
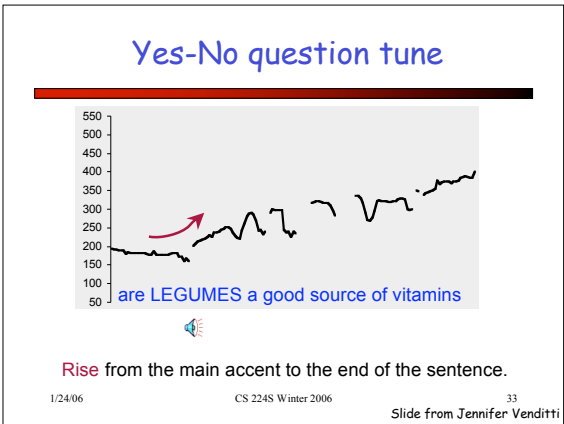
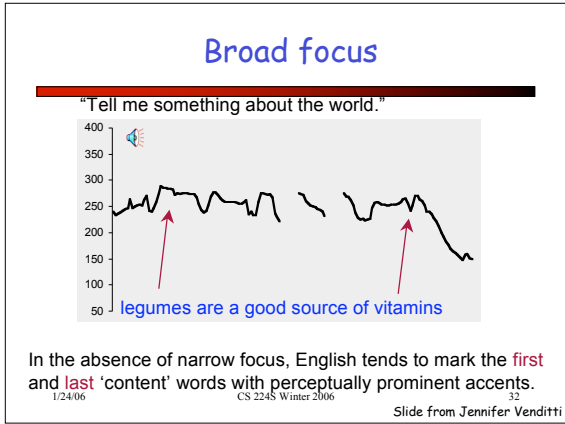
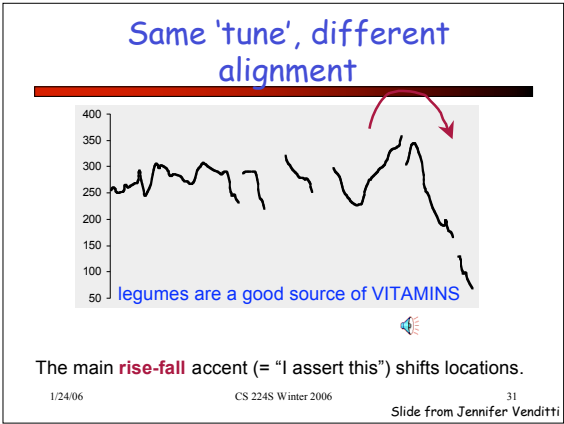
Which word receives an accent?

- It depends on the context. For example, the 'new' information in the answer to a question is often accented, while the 'old' information usually is not.
 - Q1: What types of foods are a good source of vitamins?
 - A1: **LEGUMES** are a good source of vitamins.
 - Q2: Are legumes a source of vitamins?
 - A2: Legumes are a **GOOD** source of vitamins.
 - Q3: I've heard that legumes are healthy, but what are they a good source of ?
 - A3: Legumes are a good source of **VITAMINS**.

1/24/06
CS 224S Winter 2006
28

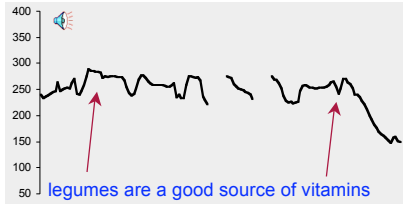
Slide from Jennifer Venditti





Broad focus

"Tell me something about the world."



1/24/06

CS 224S Winter 2006

37

Slide from Jennifer Venditti

Rising statements

"Tell me something I didn't already know."



High-rising statements can signal that the speaker is seeking approval.

1/24/06

CS 224S Winter 2006

38

Slide from Jennifer Venditti

Yes-No question



Rise from the main accent to the end of the sentence.

1/24/06

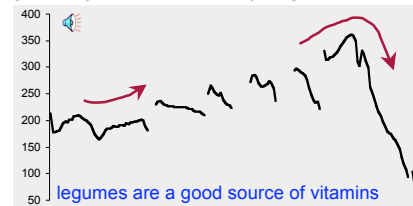
CS 224S Winter 2006

39

Slide from Jennifer Venditti

'Surprise-redundancy' tune

[How many times do I have to tell you ...]



Low beginning followed by a gradual rise to a high at the end.

1/24/06

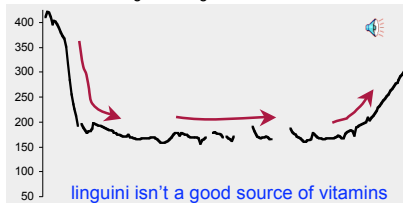
CS 224S Winter 2006

40

Slide from Jennifer Venditti

'Contradiction' tune

"I've heard that linguini is a good source of vitamins."



Sharp fall at the beginning, flat and low, then rising at the end.

1/24/06

CS 224S Winter 2006

41

Slide from Jennifer Venditti

TOPIC #3

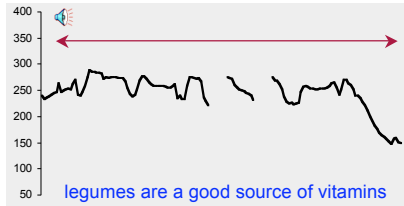
Intonational phrasing and disambiguation

1/24/06

CS 224S Winter 2006

42

A single intonation phrase



Broad focus statement consisting of one intonation phrase (that is, one intonation tune spans the whole unit).

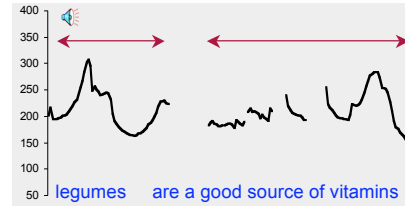
1/24/06

CS 224S Winter 2006

43

Slide from Jennifer Venditti

Multiple phrases



Utterances can be 'chunked' up into smaller phrases in order to signal the importance of information in each unit.

1/24/06

CS 224S Winter 2006

44

Slide from Jennifer Venditti

Phrasing can disambiguate

- Global ambiguity:

The old men and women stayed home.

Sally saw the man with the binoculars.

John doesn't drink because he's unhappy.

1/24/06

CS 224S Winter 2006

45

Slide from Jennifer Venditti

Phrasing can disambiguate

- Global ambiguity:

The old men and women stayed home.

The old men % and women % stayed home.

Sally saw % the man with the binoculars.

Sally saw the man % with the binoculars.

John doesn't drink because he's unhappy.

John doesn't drink % because he's unhappy.

1/24/06

CS 224S Winter 2006

46

Slide from Jennifer Venditti

Phrasing can disambiguate

- Temporary ambiguity:

When Madonna sings the song ...

1/24/06

CS 224S Winter 2006

47

Slide from Jennifer Venditti

Phrasing can disambiguate

- Temporary ambiguity:

When Madonna sings the song is a hit.

1/24/06

CS 224S Winter 2006

48

Slide from Jennifer Venditti

Phrasing can disambiguate

- Temporary ambiguity:

When Madonna sings % the song is a hit.

When Madonna sings the song % it's a hit.

[from Speer & Kjelgaard (1992)]

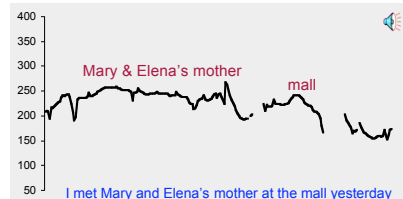
1/24/06

CS 224S Winter 2006

49

Slide from Jennifer Venditti

Phrasing can disambiguate



One intonation phrase with relatively flat overall pitch range.

1/24/06

CS 224S Winter 2006

50

Slide from Jennifer Venditti

Phrasing can disambiguate



Separate phrases, with expanded pitch movements.

1/24/06

CS 224S Winter 2006

51

Slide from Jennifer Venditti

TOPIC #4

The TOBI Intonational Transcription Theory

1/24/06

CS 224S Winter 2006

52

ToBI: Tones and Break Indices

- Pitch accent tones
 - H* "peak accent"
 - L* "low accent"
 - L+H* "rising peak accent" (contrastive)
 - L*+H "scooped accent"
 - H+H* downstepped high
- Boundary tones
 - L-L% (final low; Am Eng. Declarative contour)
 - L-H% (continuation rise)
 - H-H% (yes-no question)
- Break indices
 - 0: clitics, 1, word boundaries, 2 short pause
 - 3 intermediate intonation phrase
 - 4 full intonation phrase/final boundary.

1/24/06

CS 224S Winter 2006

53

Examples of the TOBI system

- I don't eat beef.
 - L* L* L*L-L%
 - Marianna made the marmalade.
 - H* L-L%
 - L* H-H%
 - "I" means insert.
 - H* H* H*L-L%
- 1 H*L- H*L-L%
- 3

1/24/06

CS 224S Winter 2006

54

Slide from Lavoie and Podesva

ToBI

- <http://www.ling.ohio-state.edu/~tobi/>
- **ToBI for American English**
 - http://www.ling.ohio-state.edu/~tobi/ame_tobi/
- Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: a standard for labelling English prosody. In *Proceedings of ICSLP92*, volume 2, pages 867-870
- Pitrelli, J. F., Beckman, M. E., and Hirschberg, J. (1994). Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *ICSLP94*, volume 1, pages 123-126
- Pierrehumbert, J., and J. Hirschberg (1990) The meaning of intonation contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, and M. E. Pollack, eds., *Plans and Intentions in Communication and Discourse*, 271-311. MIT Press.
- Beckman and Elam. Guidelines for ToBI Labelling. Web.

1/24/06

CS 224S Winter 2006

55

TOPIC #5

PRODUCING INTONATION IN TTS

1/24/06

CS 224S Winter 2006

56

Intonation in TTS

- 1) **Accent:** Decide which words are accented, which syllable has accent, what sort of accent
- 2) **Boundaries:** Decide where intonational boundaries are
- 3) **Duration:** Specify length of each segment
- 4) **F0:** Generate F0 contour from these

1/24/06

CS 224S Winter 2006

57

TOPIC #5a

Predicting pitch accent

1/24/06

CS 224S Winter 2006

58

Factors in accent prediction

- **Contrast**
 - Legumes are poor source of **VITAMINS**
 - No, legumes are a **GOOD** source of vitamins

 - I think **JOHN** or **MARY** should go
 - No, I think **JOHN AND MARY** should go

1/24/06

CS 224S Winter 2006

59

But it's more than just contrast

- **List intonation:**
 - I went and saw **ANNA**, **LENNY**, **MARY**, and **NORA**.

1/24/06

CS 224S Winter 2006

60

In fact, accents are common!

- A Broadcast News example from Hirschberg (1993)
- SUN MICROSYSTEMS INC, the UPSTART COMPANY that HELPED LAUNCH the DESKTOP COMPUTER industry TREND TOWARD HIGH powered WORKSTATIONS, was UNVEILING an ENTIRE OVERHAUL of its PRODUCT LINE TODAY. SOME of the new MACHINES, PRICED from FIVE THOUSAND NINE hundred NINETY five DOLLARS to seventy THREE thousand nine HUNDRED dollars, BOAST SOPHISTICATED new graphics and DIGITAL SOUND TECHNOLOGIES, HIGHER SPEEDS AND a CIRCUIT board that allows FULL motion VIDEO on a COMPUTER SCREEN.

1/24/06

CS 224S Winter 2006

61

Factors in accent prediction

- **Part of speech:**
 - Content words are usually accented
 - Function words are rarely accented
 - Of, for, in on, that, the, a, an, no, to, and but or will may would can her is their its our there is am are was were, etc

1/24/06

CS 224S Winter 2006

62

Factors in accent prediction

- **Word Order**
- **Preposed items are accented more frequently**
- **TODAY** we will **BEGIN** to **LOOK** at **FROG** anatomy.
- We will **BEGIN** to **LOOK** at **FROG** anatomy today.

1/24/06

CS 224S Winter 2006

63

Factors in Accent Prediction

- **Information Status:**
- **New versus old information.**
- **Old information is not deaccented**
- **There are LAWYERS, and there are GOOD lawyers**
- **EACH NATION DEFINES its OWN national INTERST.**
- **I LIKE GOLDEN RETRIEVERS, but MOST dogs LEAVE me COLD.**

1/24/06

CS 224S Winter 2006

64

Complex Noun Phrase Structure

- Sproat, R. 1994. English noun-phrase accent prediction for text-to-speech. Computer Speech and Language 8:79-94.
- **Proper Names, stress on right-most word**
 - New York **CITY**; Paris, **FRANCE**
- **Adjective-Noun combinations, stress on noun**
 - Large **HOUSE**, red **PEN**, new **NOTEBOOK**
- **Noun-Noun compounds: stress left noun**
 - **HOT**dog (food) versus **HOT DOG** (overheated animal)
 - **WHITE** house (place) versus **WHITE HOUSE** (made of stucco)
- **examples:**
 - **MEDICAL** Building, **APLPE** cake, cherry **PIE**.
 - What about: Madison avenue, park street,
- **Some Rules:**
 - Furniture+Room -> **RIGHT** (e.g., kitchen **TABLE**)
 - Proper-name + Street -> **LEFT** (e.g. **PARK** street)

1/24/06

CS 224S Winter 2006

65

Simplest possible algorithm for pitch accent assignment

```
(set! simple_accent_cart_tree
'
(
(R:SylStructure.parent.gpos is content)
( (stress is 1)
((Accented))
((NONE))
)
)
)
```

1/24/06

CS 224S Winter 2006

66

Other features

- POS
- POS of previous word
- POS of next word
- Stress of current, previous, next syllable
- Unigram probability of word
- Bigram probability of word
- Position of word in sentence

1/24/06

CS 224S Winter 2006

67

Advanced features

- Accent is often deflected away from a word due to **focus** on a neighboring word.
- Could use syntactic parallelism to detect this kind of contrastive focus:
 -driving [**FIFTY** miles] an hour in a [**THIRTY** mile] zone
 - [**WELD**] [**APPLAUDS**] mandatory recycling. [**SILBER**] [**DISMISSES**] recycling goals as meaningless.
 - ...but while Weld may be [**ONE**] on people skills, he may be [**SHORT**] on money

1/24/06

CS 224S Winter 2006

68

State of the art

- Hand-label large training sets
- Use CART, SVM, CRF, etc to predict accent
- Lots of rich features from context
- Classic lit:
 - Hirschberg, Julia. 1993. Pitch Accent in context: predicting intonational prominence from text. *Artificial Intelligence* 63, 305-340

1/24/06

CS 224S Winter 2006

69

TOPIC #5b

Predicting boundaries

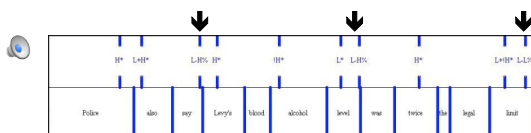
1/24/06

CS 224S Winter 2006

70

Predicting Boundaries

- Intonation phrase boundaries
 - Intermediate phrase boundaries
 - Full intonation phrase boundaries



1/24/06

CS 224S Winter 2006

71

More examples

- From Ostendorf and Veilleux. 1994 "Hierarchical Stochastic model for Automatic Prediction of Prosodic Boundary Location", *Computational Linguistics* 20:1
- Computer phone calls, || which do everything | from selling magazine subscriptions || to reminding people about meetings || have become the telephone equivalent | of junk mail. ||
- Doctor Norman Rosenblatt, || dean of the college | of criminal justice at Northeastern University, || agrees.||
- For WBUR, || I'm Margo Melnicove.

1/24/06

CS 224S Winter 2006

72

Simplest CART

```
(set! simple_phrase_cart_tree
'
((lisp_token_end_punc in ("?" "." ":"))
 (BB))
((lisp_token_end_punc in ("'" "\"" " "
";"))
 (B))
((n.name is 0) ;; end of utterance
 ((BB))
 ((NB))))))
```

1/24/06

CS 224S Winter 2006

73

More complex features

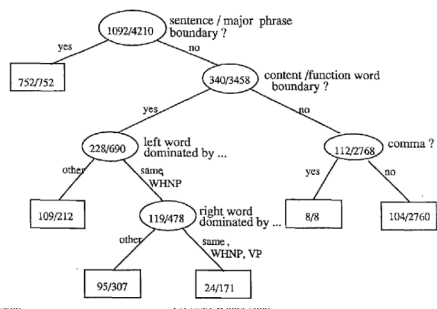
- Ostendorf and Veilleux
- English: boundaries are more likely between content words and function words
- Syntactic structure (parse trees)
 - Largest syntactic category dominating preceding word but not succeeding word
 - How many syntactic units begin/end between words
- Type of function word to right
- Capitalized names
- # of content words since previous function word

1/24/06

CS 224S Winter 2006

74

Ostendorf and Veilleux CART



1/24/06

CS 224S Winter 2006

75

TOPIC #5c

Predicting duration

1/24/06

CS 224S Winter 2006

76

Duration

- Simplest: fixed size for all phones (100 ms)
- Next simplest: average duration for that phone (from training data). Samples from SWBD in ms:

- aa	118	b	68
- ax	59	d	68
- ay	138	dh	44
- eh	87	f	90
- ih	77	g	66
- Next Next Simplest: add in phrase-final and initial lengthening plus stress:

1/24/06

CS 224S Winter 2006

77

Duration in Festival (2)

- Klatt duration rules. Modify duration based on:
 - Position in clause
 - Syllable position in word
 - Syllable type
 - Lexical stress
 - Left+right context phone
 - Prepausal lengthening
- Festival: 2 options
 - Klatt rules
 - Use labeled training set with Klatt features to train CART

1/24/06

CS 224S Winter 2006

78

Duration: state of the art

- Lots of fancy models of duration prediction:
 - Using Z-scores and other clever normalizations
 - Sum-of-products model
 - New features like word predictability
 - Words with higher bigram probability are shorter

1/24/06

CS 224S Winter 2006

79

Duration in Festival

```
(set! spanish_dur_tree
'
((R:SylStructure.parent.R:Syllable.p.syl_break >
1) ;; clause initial
((R:SylStructure.parent.stress is 1)
((1.5))
((1.2)))
((R:SylStructure.parent.syl_break > 1) ;;
clause final
((R:SylStructure.parent.stress is 1)
((2.0))
((1.5)))
((R:SylStructure.parent.stress is 1)
((1.2))
((1.0))))))
```

1/24/06

CS 224S Winter 2006

80

TOPIC #5d

F0 Generation

1/24/06

CS 224S Winter 2006

81

F0 Generation

- Generation in Festival
 - F0 Generation by rule
 - F0 Generation by linear regression
- Some constraints
 - F0 is constrained by accents and boundaries
 - F0 declines gradually over an utterance ("declination")

1/24/06

CS 224S Winter 2006

82

F0 Generation by rule

- Generate a list of target F0 points for each syllable
- Here's a rule to generate a simple H* "hat" accent (with fixed = speaker-specific F0 values):

```
(define (targ_func1 utt syl)
  "(targ_func1 UTT STREAMITEM)
Returns a list of targets for the given syllable."
  (let ((start (item.feats syl 'syllable_start))
        (end (item.feats syl 'syllable_end)))
    (if (equal? (item.feats syl
      "R:Intonation.daughter1.name") "Accented")
      (list
        (list start 110)
        (list (/ (+ start end) 2.0) 140)
        (list end 100))))))
```

1/24/06

CS 224S Winter 2006

83

F0 generation by regression

- Supervised machine learning again
- We predict: value of F0 at 3 places in each syllable
- Predictor features:
 - Accent of current word, next word, previous
 - Boundaries
 - Syllable type, phonetic information
 - Stress information
- Need training sets with pitch accents labeled

1/24/06

CS 224S Winter 2006

84

Summary

- Thinking about F0
- Accent Placement and Intonational Tunes
- Intonational Phrasing and Disambiguation
- The TOBI Prosodic Transcription Theory
- Producing Intonation in TTS
 - Predicting Accents
 - Predicting Boundaries
 - Predicting Duration
 - Generating F0

1/24/06

CS 224S Winter 2006

85

Jennifer Venditti's References

Jennifer's list of introductory readings on intonational form and function:

- Bolinger, D. (1972) *Intonation* [introduction and chapter 1]. Penguin Books, Ltd.
- Ladd, D.R. (1996) *Intonational Phonology*. Cambridge Univ. Press.
- Kadmon, N. (2001) *Formal Pragmatics* [chapter 12]. Blackwell Publ.
- Beckman, M. & J. Pierrehumbert (1986) Intonational structure in Japanese and English. *Phonology Yearbook* 3: 255-309.
- Pierrehumbert, J. & Hirschberg (1990) The meaning of intonational contours in interpretation of discourse. In Cohen, et al. (eds.) *Intentions in Communication*. MIT Press.

1/24/06

CS 224S Winter 2006

86

Names

- Methods:
 - Can do morphology (Walters -> Walter, Lucasville)
 - Can write stress-shifting rules (Jordan -> Jordanian)
 - Rhyme analogy: Plotsky by analogy with Trostsky (replace tr with pl)
 - Liberman and Church: for 250K most common names, got 212K (85%) from these modified-dictionary methods, used LTS for rest.
 - Can do automatic country detection (from letter trigrams) and then do country-specific rules

1/24/06

CS 224S Winter 2006

87