

CS 224S / LINGUIST 236 Speech Recognition and Synthesis

Dan Jurafsky

Lecture 5: Prosodic Processing for TTS

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.

1/18/05

CS 224S Winter 2005

1

Outline

- 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

1/18/05

CS 224S Winter 2005

2

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.

1/18/05

CS 224S Winter 2005

3

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.

1/18/05

CS 224S Winter 2005

From Ladd (1996)

4

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)

1/18/05

CS 224S Winter 2005

5

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]

1/18/05

CS 224S Winter 2005

6

Slide from Jennifer Venditti

Prosodic Tunes

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

1/18/05

CS 224S Winter 2005

7

Slide from Jennifer Venditti

TOPIC #1

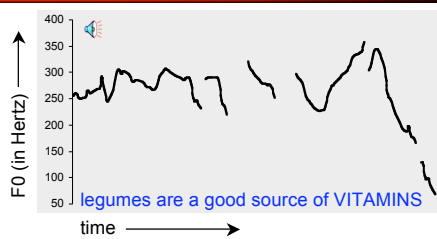
Thinking about F0

1/18/05

CS 224S Winter 2005

8

Graphic representation of F0



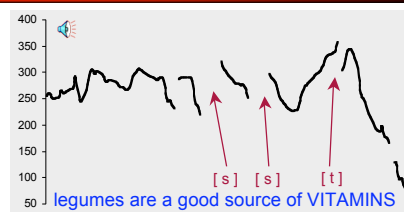
1/18/05

CS 224S Winter 2005

9

Slide from Jennifer Venditti

The 'ripples'



F0 is not defined for consonants without vocal fold vibration.

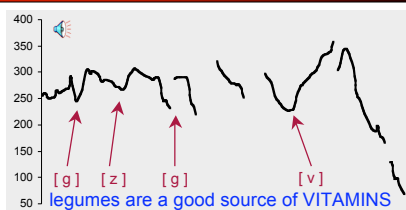
1/18/05

CS 224S Winter 2005

10

Slide from Jennifer Venditti

The 'ripples'



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

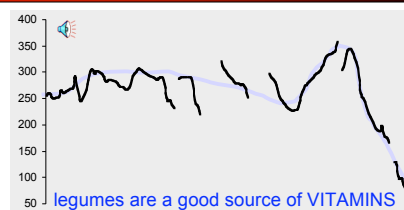
1/18/05

CS 224S Winter 2005

11

Slide from Jennifer Venditti

Abstraction of the F0 contour



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

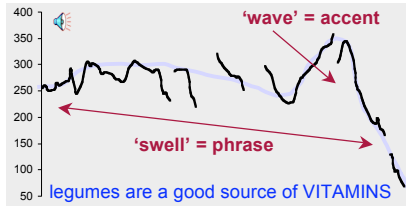
1/18/05

CS 224S Winter 2005

12

Slide from Jennifer Venditti

The 'waves' and the 'swells'



1/18/05

CS 224S Winter 2005

13

Slide from Jennifer Venditti

TOPIC #2

Accent Placement and Intonational Tunes

1/18/05

CS 224S Winter 2005

14

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

1/18/05

CS 224S Winter 2005

15

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.

1/18/05

CS 224S Winter 2005

16

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**.

1/18/05

CS 224S Winter 2005

17

Slide from Jennifer Venditti

Same 'tune', different alignment



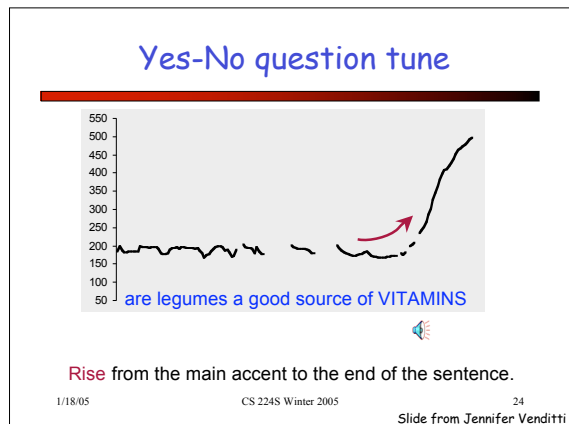
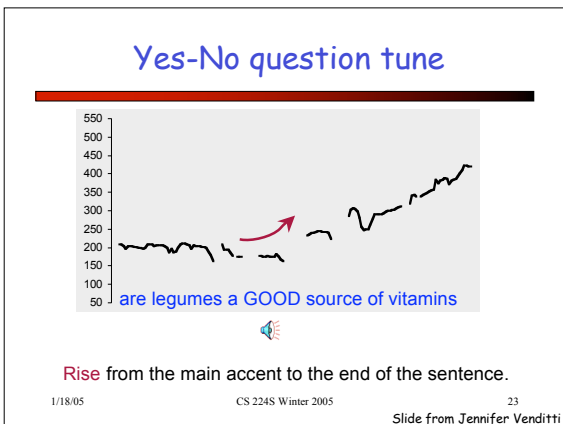
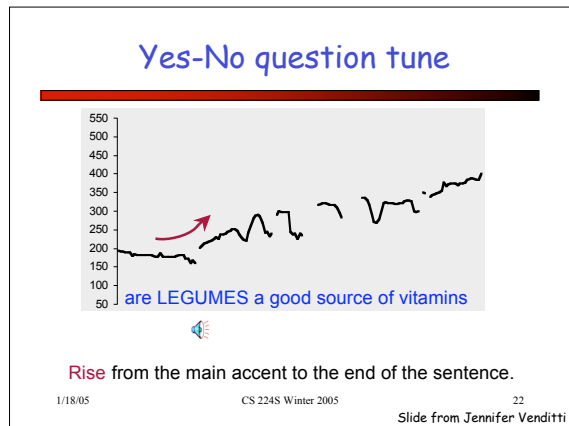
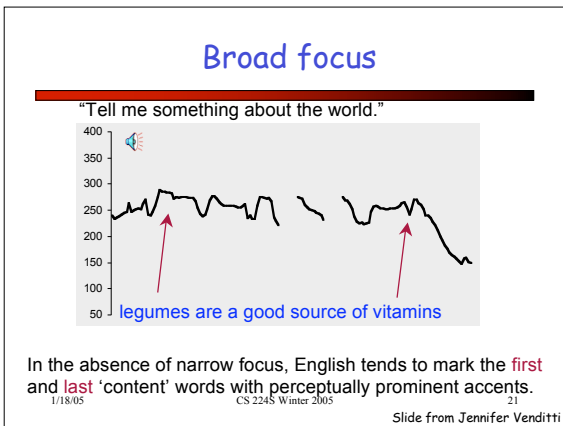
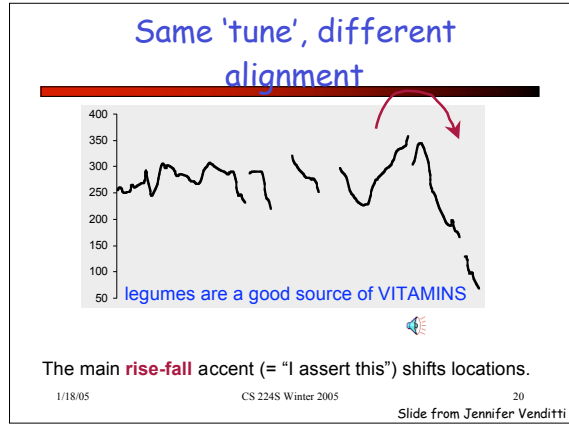
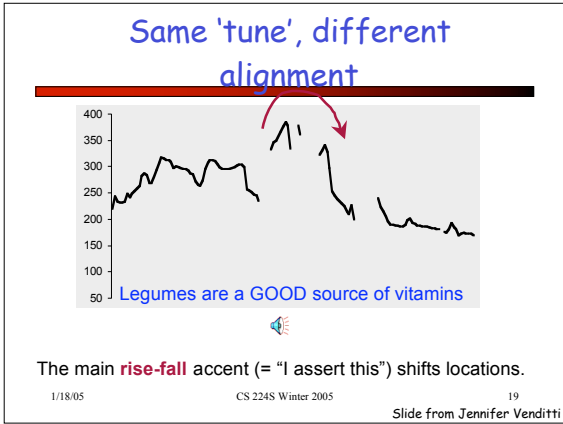
The main **rise-fall** accent (= "I assert this") shifts locations.

1/18/05

CS 224S Winter 2005

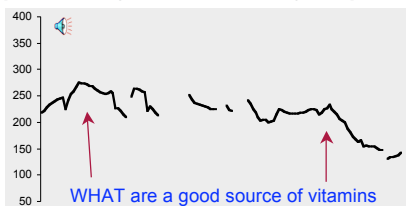
18

Slide from Jennifer Venditti



WH-questions

[I know that many natural foods are healthy, but ...]



WH-questions typically have **falling** contours, like statements.

1/18/05

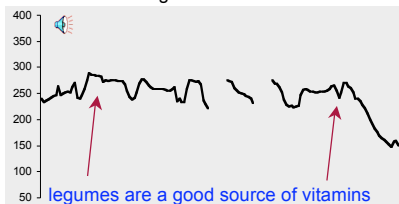
CS 224S Winter 2005

25

Slide from Jennifer Venditti

Broad focus

"Tell me something about the world."



1/18/05

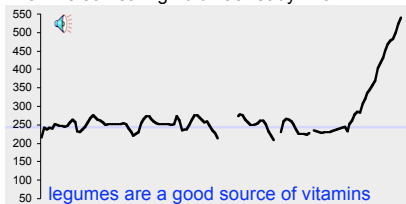
CS 224S Winter 2005

26

Slide from Jennifer Venditti

Rising statements

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



[... does this statement qualify?]

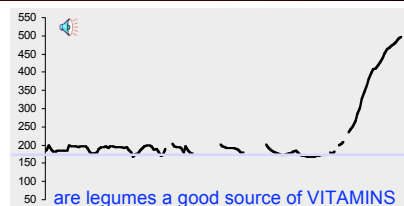
High-rising statements can signal that the speaker is seeking approval.

1/18/05

27

Slide from Jennifer Venditti

Yes-No question



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

1/18/05

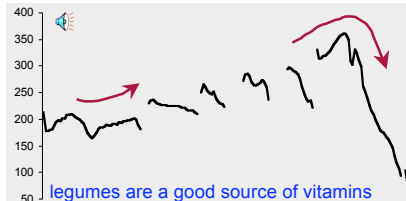
CS 224S Winter 2005

28

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.

1/18/05

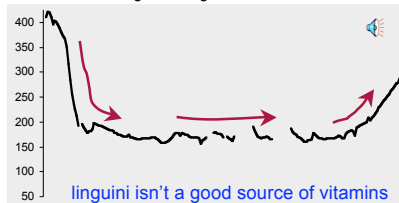
CS 224S Winter 2005

29

Slide from Jennifer Venditti

'Contradiction' tune

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



[... how could you think that?]

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

1/18/05

CS 224S Winter 2005

30

Slide from Jennifer Venditti

TOPIC #3

Intonational phrasing and disambiguation

1/18/05
CS 224S Winter 2005
31

A single intonation phrase

legumes are a good source of vitamins

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

1/18/05
CS 224S Winter 2005
32

Slide from Jennifer Venditti

Multiple phrases

legumes are a good source of vitamins

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

1/18/05
CS 224S Winter 2005
33

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.

1/18/05
CS 224S Winter 2005
34

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.

1/18/05
CS 224S Winter 2005
35

Slide from Jennifer Venditti

Phrasing can disambiguate

- Temporary ambiguity:
 - When Madonna sings the song ...

1/18/05
CS 224S Winter 2005
36

Slide from Jennifer Venditti

Phrasing can disambiguate

- Temporary ambiguity:

When Madonna sings the song is a hit.

1/18/05

CS 224S Winter 2005

37

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)]

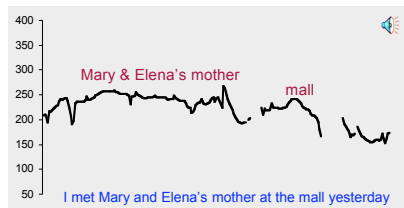
1/18/05

CS 224S Winter 2005

38

Slide from Jennifer Venditti

Phrasing can disambiguate



One intonation phrase with relatively flat overall pitch range.

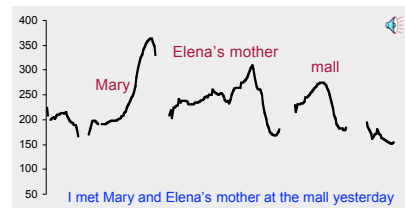
1/18/05

CS 224S Winter 2005

39

Slide from Jennifer Venditti

Phrasing can disambiguate



Separate phrases, with expanded pitch movements.

1/18/05

CS 224S Winter 2005

40

Slide from Jennifer Venditti

TOPIC #4

The TOBI Intonational Transcription Theory

1/18/05

CS 224S Winter 2005

41

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.

1/18/05

42

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



1/18/05

CS 224S Winter 2005

43

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.

1/18/05

CS 224S Winter 2005

44

TOPIC #5

PRODUCING INTONATION IN TTS

1/18/05

CS 224S Winter 2005

45

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

1/18/05

CS 224S Winter 2005

46

TOPIC #5a

Predicting pitch accent

1/18/05

CS 224S Winter 2005

47

Factors in accent prediction

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

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

1/18/05

CS 224S Winter 2005

48

But it's more than just contrast

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

1/18/05

CS 224S Winter 2005

49

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.**

1/18/05

CS 224S Winter 2005

50

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

1/18/05

CS 224S Winter 2005

51

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.**

1/18/05

CS 224S Winter 2005

52

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.**

1/18/05

CS 224S Winter 2005

53

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**
 - HOTdog (food) versus HOT DOG (overheated animal)
 - WHITE house (place) versus WHITE HOUSE (made of stucco)
- **examples:**
 - Madison AVENUE, park STREET, MEDICAL building
 - APPLE cake, cherry PIE
- **Some Rules:**
 - Furniture+Room -> RIGHT (e.g., kitchen TABLE)
 - Proper-name + Street -> LEFT (e.g., PARK street)

1/18/05

CS 224S Winter 2005

54

Simplest possible algorithm for pitch accent assignment

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

1/18/05

CS 224S Winter 2005

55

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

1/18/05

CS 224S Winter 2005

56

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 [LONG] on people skills, he may be [SHORT] on money

1/18/05

CS 224S Winter 2005

57

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

1/18/05

CS 224S Winter 2005

58

Hot issues in my lab now

- Project possibilities! And publishable!
 - Very little training data available. Could use unsupervised or semi-supervised methods?
 - We have good models of accent prediction from acoustics + text; how to combine to bootstrap on unsupervised text?
 - Cross-corpus issues in accent prediction
 - How to integrate better accent models into the unit selection search algorithms of Festival?
 - Prediction of reduced or weak forms
 - [ax] for "of", [dh ax] for "the", [dh] for "that"

1/18/05

CS 224S Winter 2005

59

TOPIC #5b

Predicting boundaries

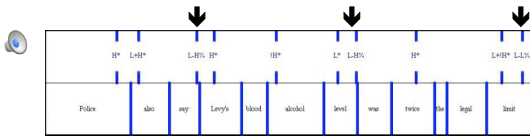
1/18/05

CS 224S Winter 2005

60

Predicting Boundaries

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



1/18/05

CS 224S Winter 2005

61

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.

1/18/05

CS 224S Winter 2005

62

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))))))
```

1/18/05

CS 224S Winter 2005

63

More complex features

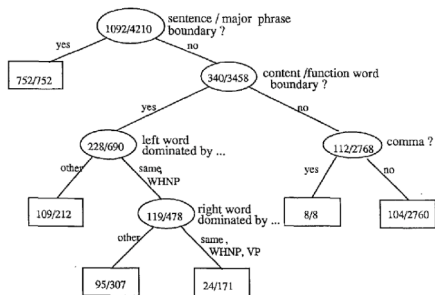
- Ostendorf and Veilleux
- English: boundaries are more likely between content words and function words
- Syntactic structure (parse trees)
 - Largest syntactic category dominating preceding word but not succeeding word
 - How many syntactic units begin/end between words
- Type of function word to right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

1/18/05

CS 224S Winter 2005

64

Ostendorf and Veilleux CART



1/18/05

CS 224S Winter 2005

65

TOPIC #5c

Predicting duration

1/18/05

CS 224S Winter 2005

66

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:

1/18/05

CS 224S Winter 2005

67

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

1/18/05

CS 224S Winter 2005

68

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

1/18/05

CS 224S Winter 2005

69

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))))))
```

1/18/05

CS 224S Winter 2005

70

TOPIC #5d

F0 Generation

1/18/05

CS 224S Winter 2005

71

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")

1/18/05

CS 224S Winter 2005

72

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))))))
```

1/18/05

CS 224S Winter 2005

73

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

1/18/05

CS 224S Winter 2005

74

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

1/18/05

CS 224S Winter 2005

75

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.

1/18/05

CS 224S Winter 2005

76