

CS 224S / LINGUIST 281

Speech Recognition, Synthesis, and Dialogue

Dan Jurafsky

Lecture 5: Prosodic Processing for TTS

IP Notice: many of the slides in the first half come from two lectures of Jennifer Venditti on intonation (thanks!); lots of other info in these slides comes from Alan Black's and Richard Sproat's lecture notes

Outline

I. Linguistic Background

1. What is Prosody?
2. Thinking about F0
3. Intonational Prominence: Pitch Accents
4. Intonational Boundaries/Phrasing
5. Intonational Tunes

II. Producing Intonation in TTS

1. Predicting Accents
 2. Predicting Boundaries
 3. Predicting Duration
 4. Generating F0
- Advanced: The TOBI Prosodic Transcription Theory

Part I: Linguistic Background

I.1 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.

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.

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)

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]

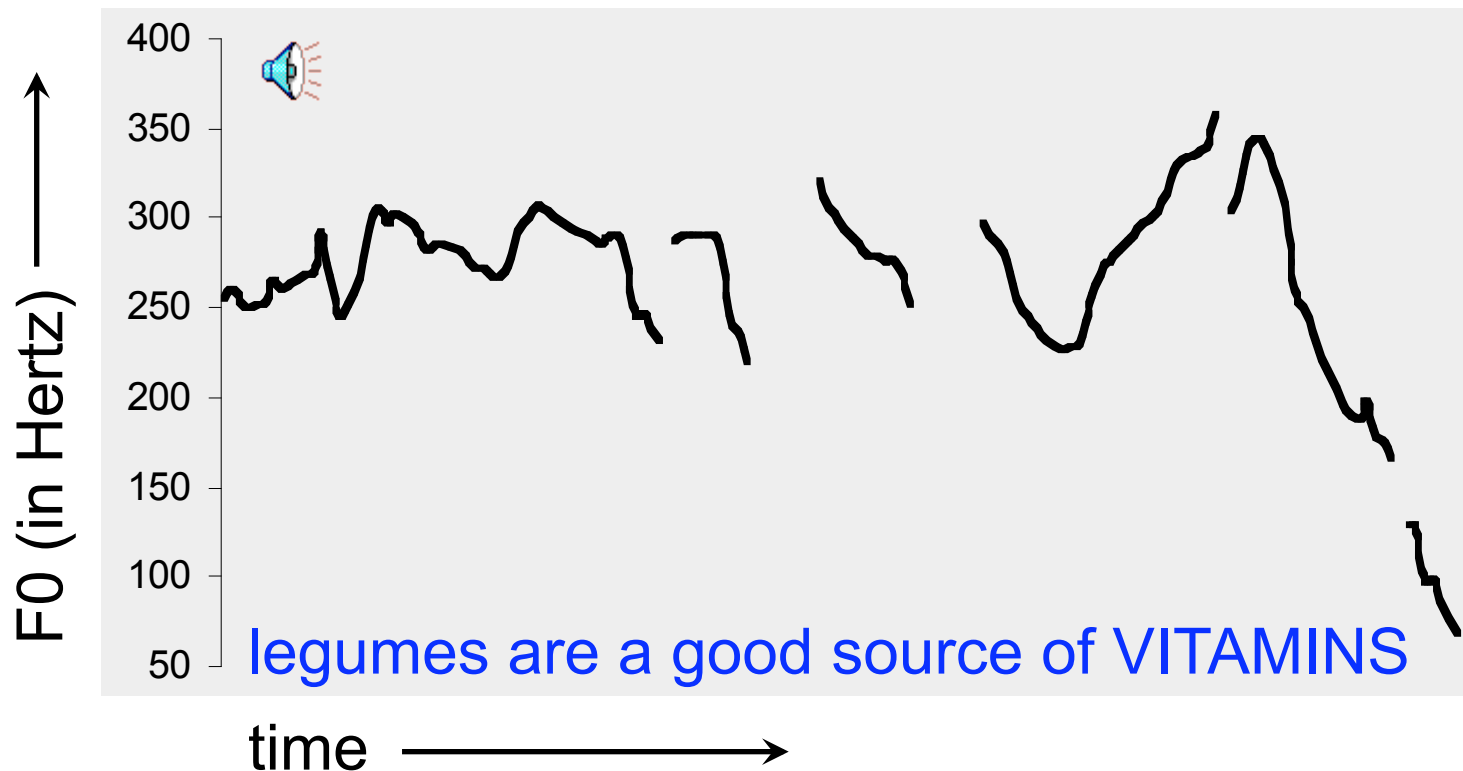
Prosodic Tunes

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

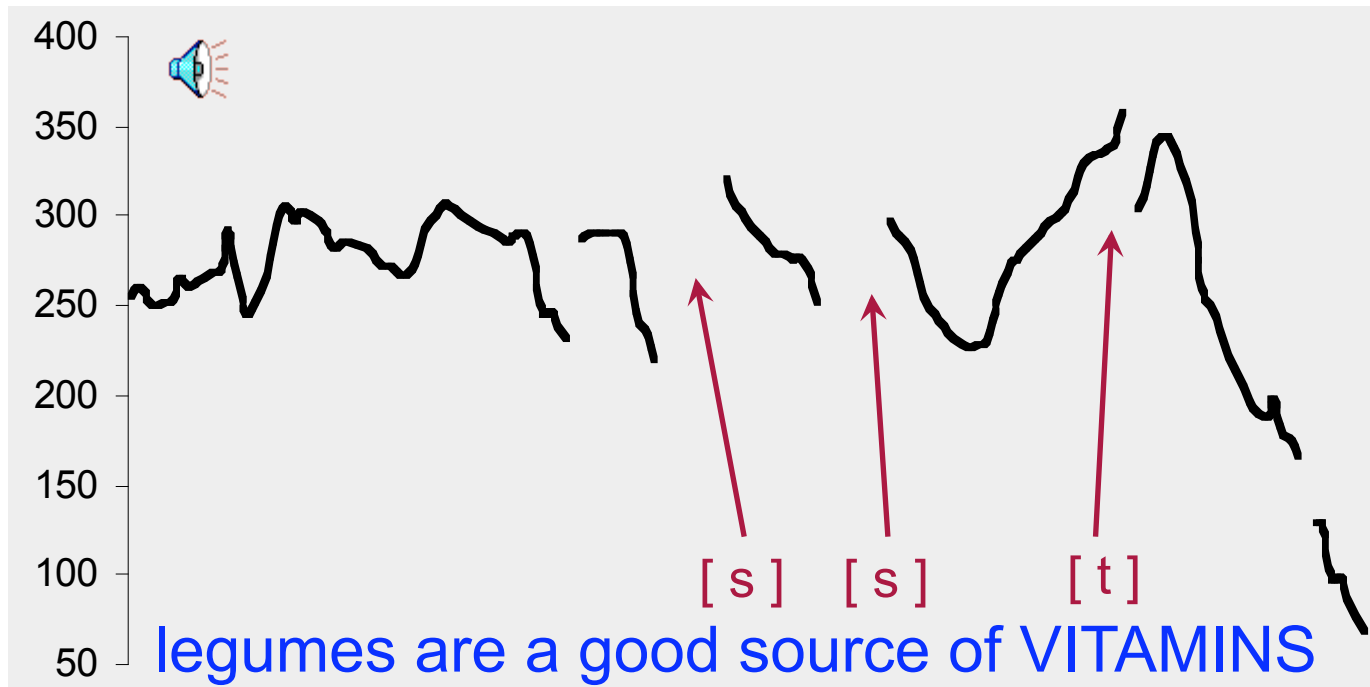
Part I.2

Thinking about F0

Graphic representation of F0

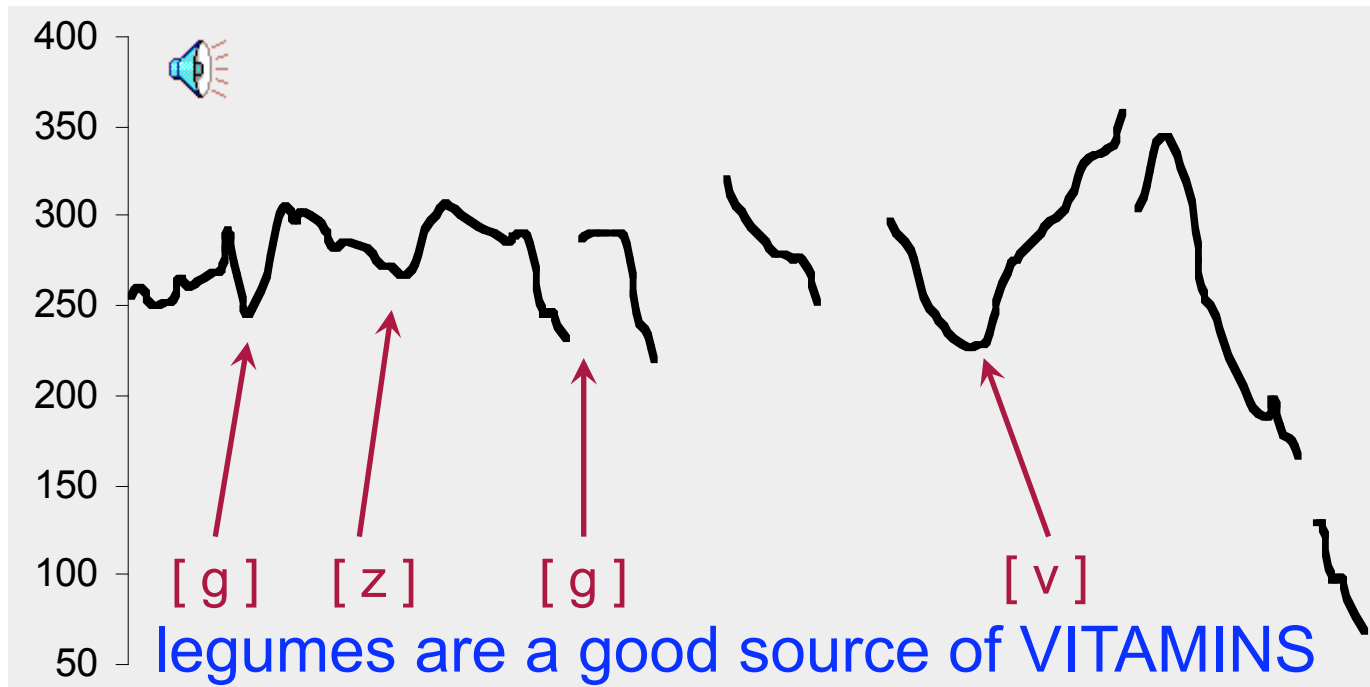


The 'ripples'



F0 is not defined for consonants without vocal fold vibration.

The 'ripples'



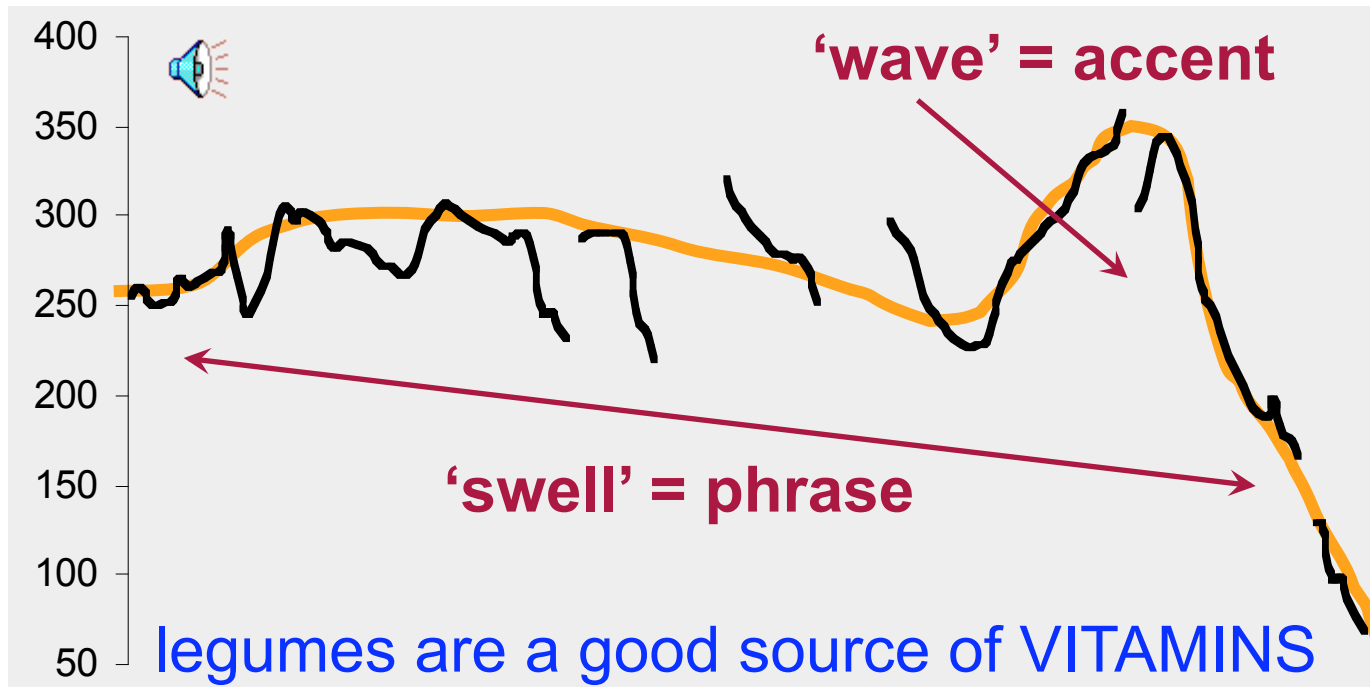
... and F0 can be perturbed by consonants with an extreme constriction in the vocal tract.

Abstraction of the F0 contour



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

The 'waves' and the 'swells'



Part I.3

Prominence: Placement of Pitch Accents

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	x			full vowels
x	x	x	x	x	x	x	syllables
vi	ta	mins	Ca	li	for	nia	

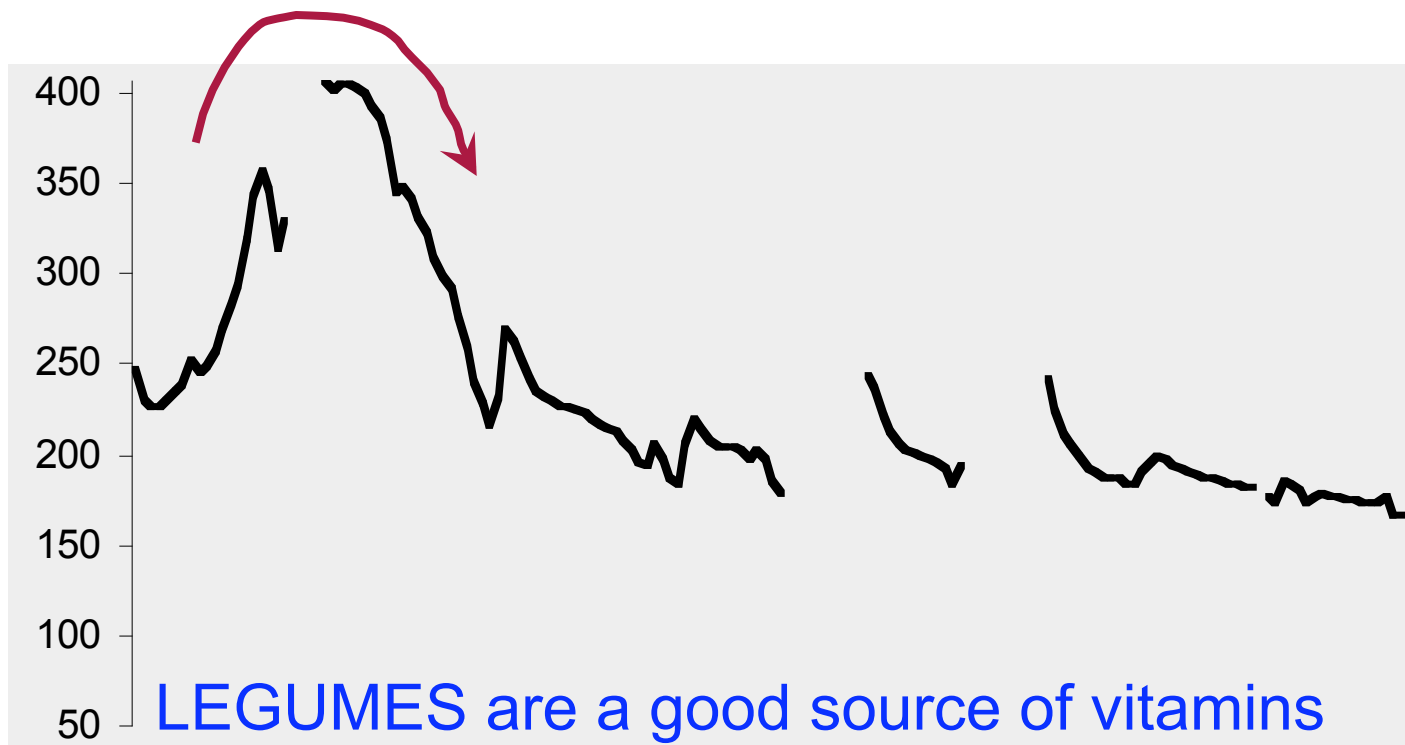
Stress vs. accent (2)

- 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**itamin.
- So we will have to look at both the lexicon and the context to predict the details of prominence
- I'm a little **surPRISED** to hear it
CHARacterized as **upBEAT**

Which word receives an accent?

- It depends on the context.
 - ♦ The 'new' information in the answer to a question is often accented
 - while the 'old' information is usually 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**. 

Same 'tune', different alignment



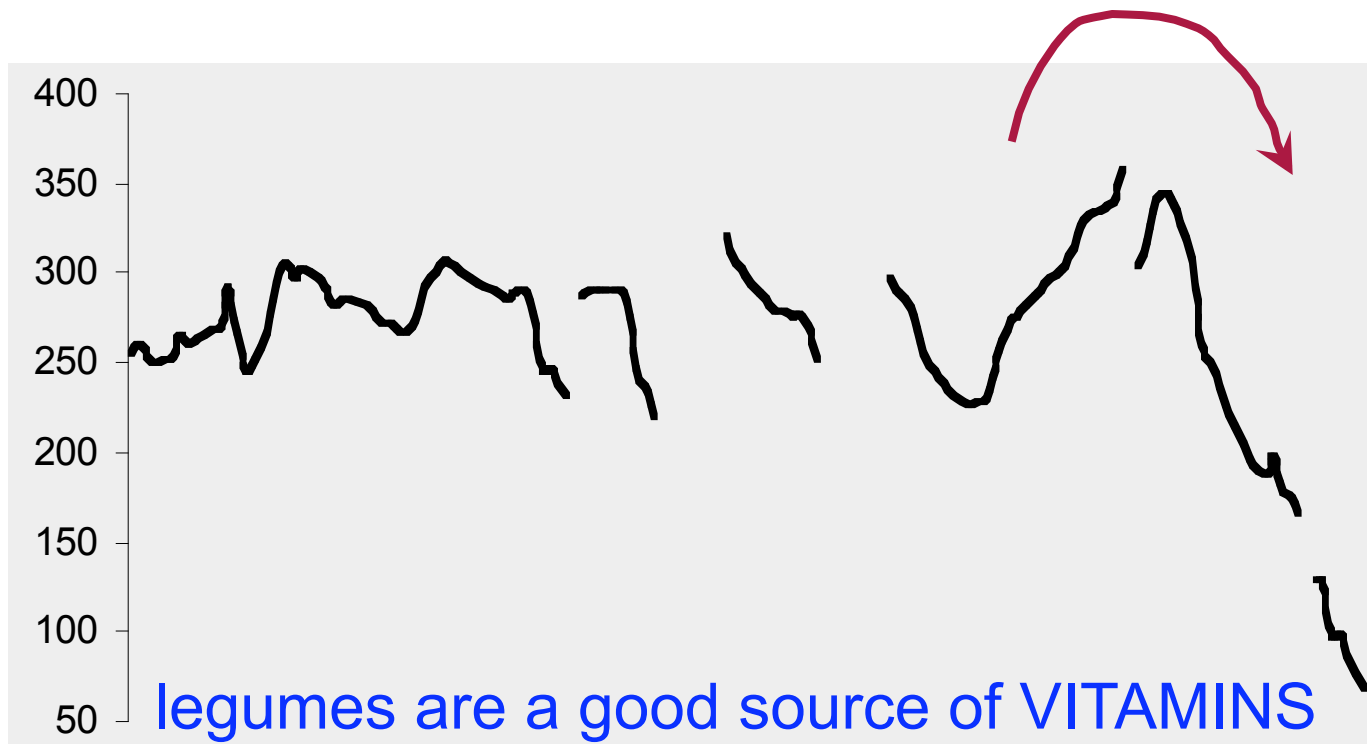
The main **rise-fall** accent (= “I assert this”) shifts locations.

Same 'tune', different alignment



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

Same 'tune', different alignment



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

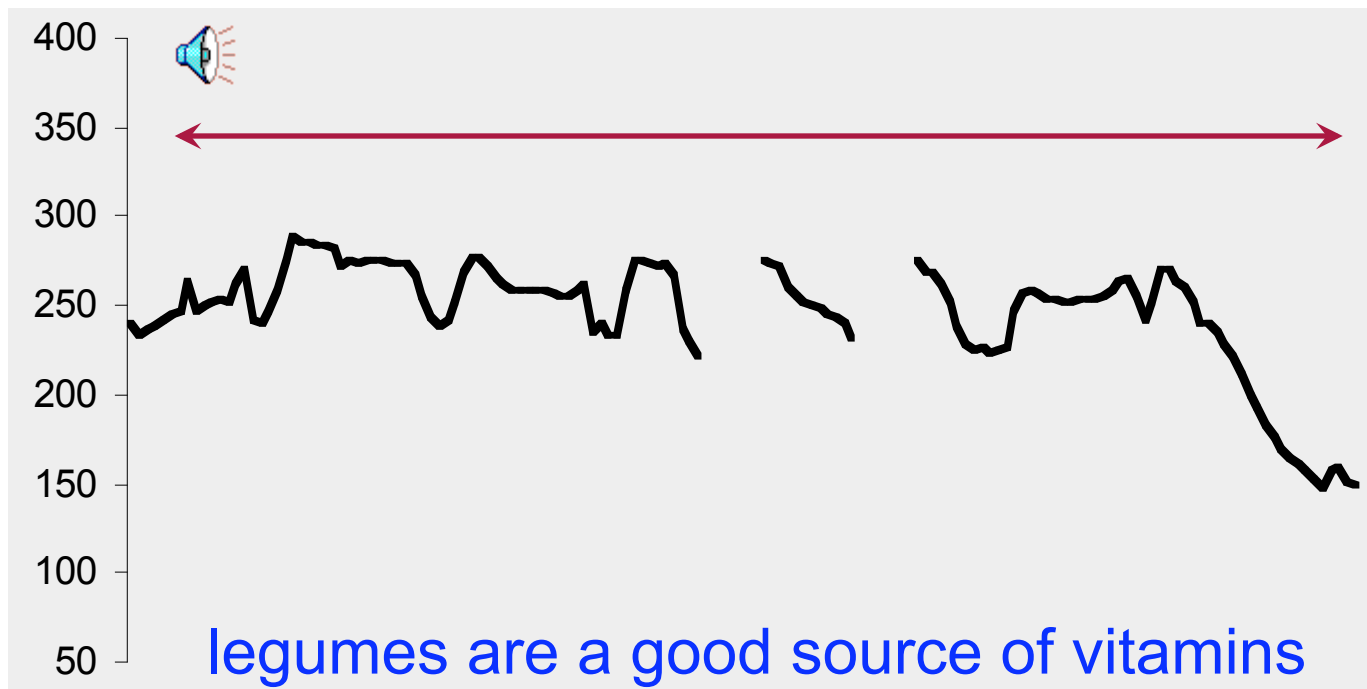
Levels of prominence

- Most phrases have more than one accent
- The last accent in a phrase is perceived as more prominent
 - ♦ Called the **Nuclear Accent**
- Emphatic accents like nuclear accent often used for semantic purposes, such as indicating that a word is contrastive, or the semantic focus.
 - ♦ The kind of thing you uses ***s in IM, or capitalized letters
 - ♦ ' I know **SOMETHING** interesting is sure to happen,' she said to herself.
- Can also have words that are **less** prominent than usual
 - ♦ Reduced words, especially function words.
- Often use 4 classes of prominence:
 - ♦ **Emphatic accent, pitch accent, unaccented, reduced**

Part I.4

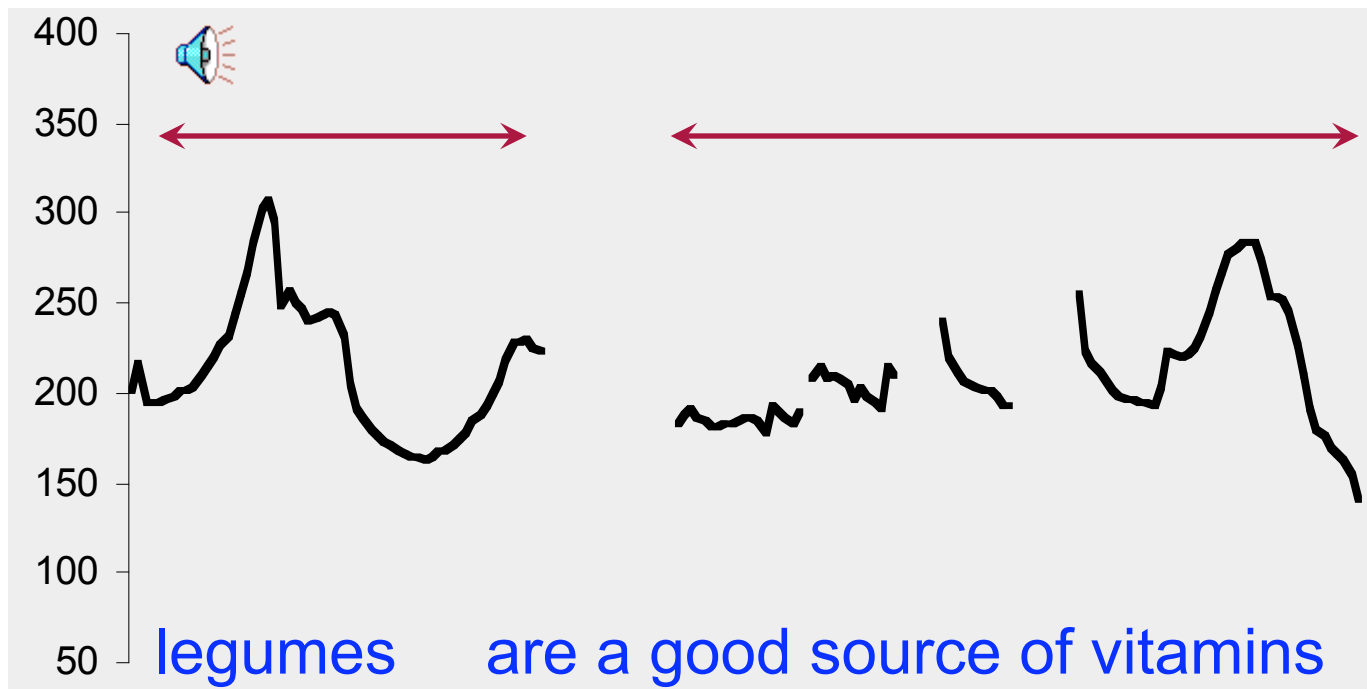
Intonational phrasing/boundaries

A single intonation phrase



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

Multiple phrases



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

- I wanted to go to London, but could only get tickets for France

Phrasing sometimes helps 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.

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.

Phrasing sometimes helps disambiguate

- **Temporary ambiguity:**

When Madonna sings the song ...

Phrasing sometimes helps disambiguate

- **Temporary ambiguity:**

When Madonna sings the song is a hit.

Phrasing sometimes helps disambiguate

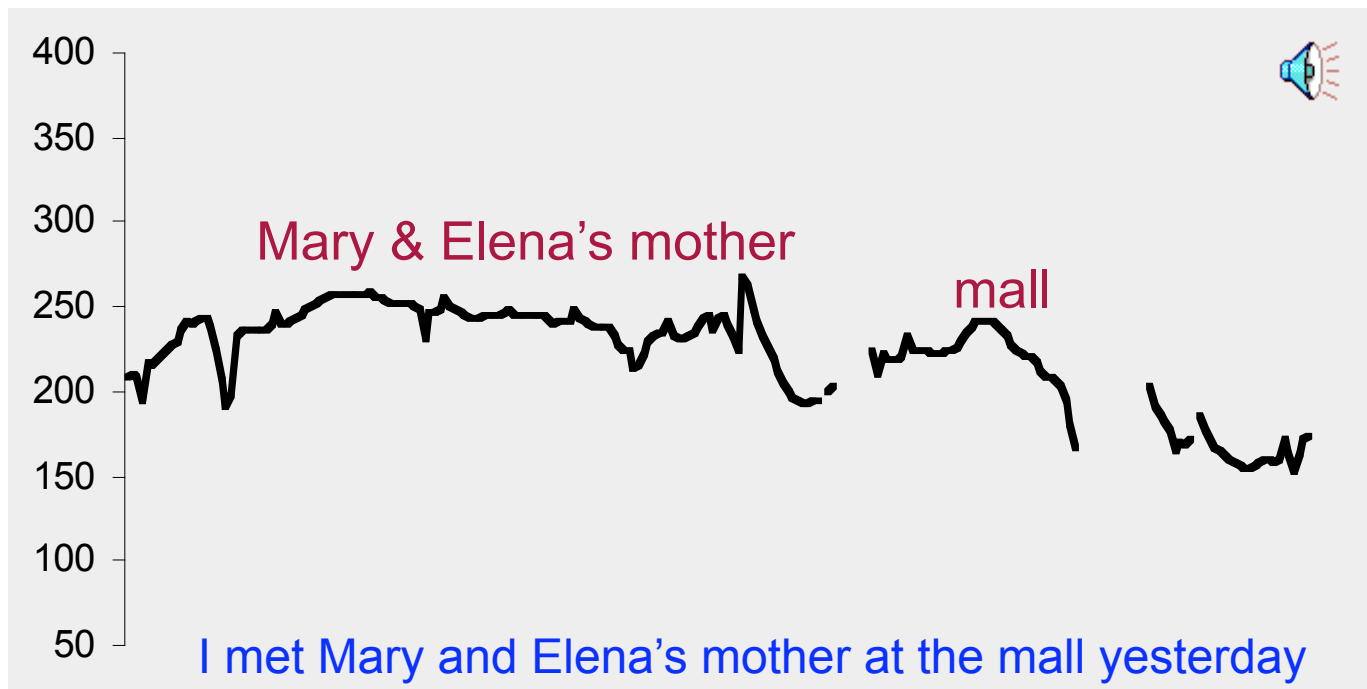
- **Temporary ambiguity:**

When Madonna sings % the song is a hit.

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

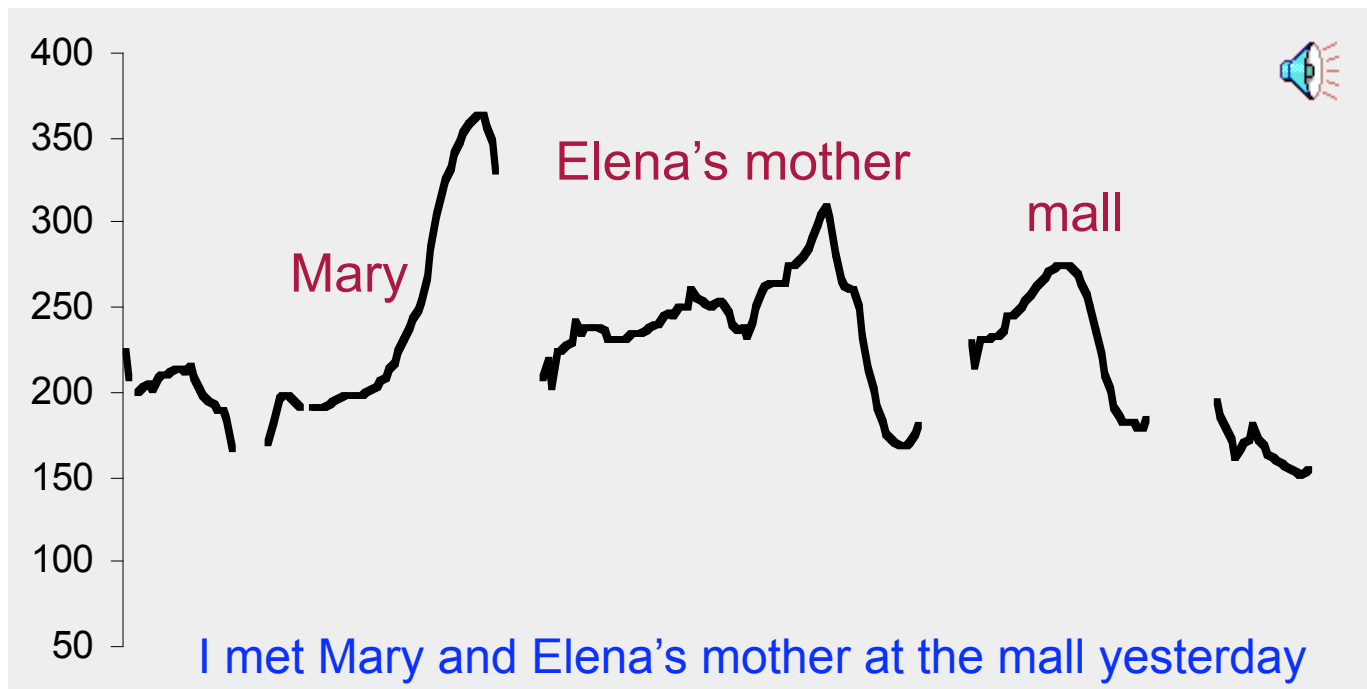
[from Speer & Kjelgaard (1992)]

Phrasing sometimes helps disambiguate



One intonation phrase with relatively flat overall pitch range.

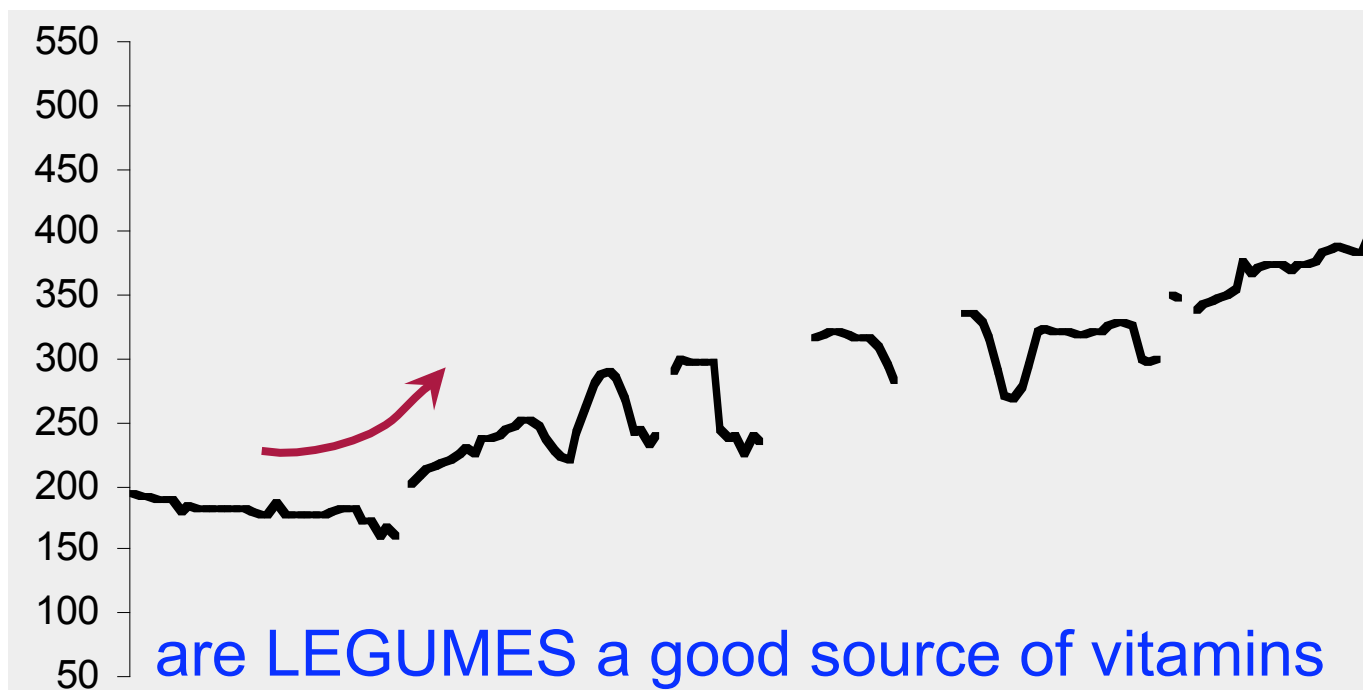
Phrasing sometimes helps disambiguate



Separate phrases, with expanded pitch movements.

Part I.4 Intonational tunes

Yes-No question tune



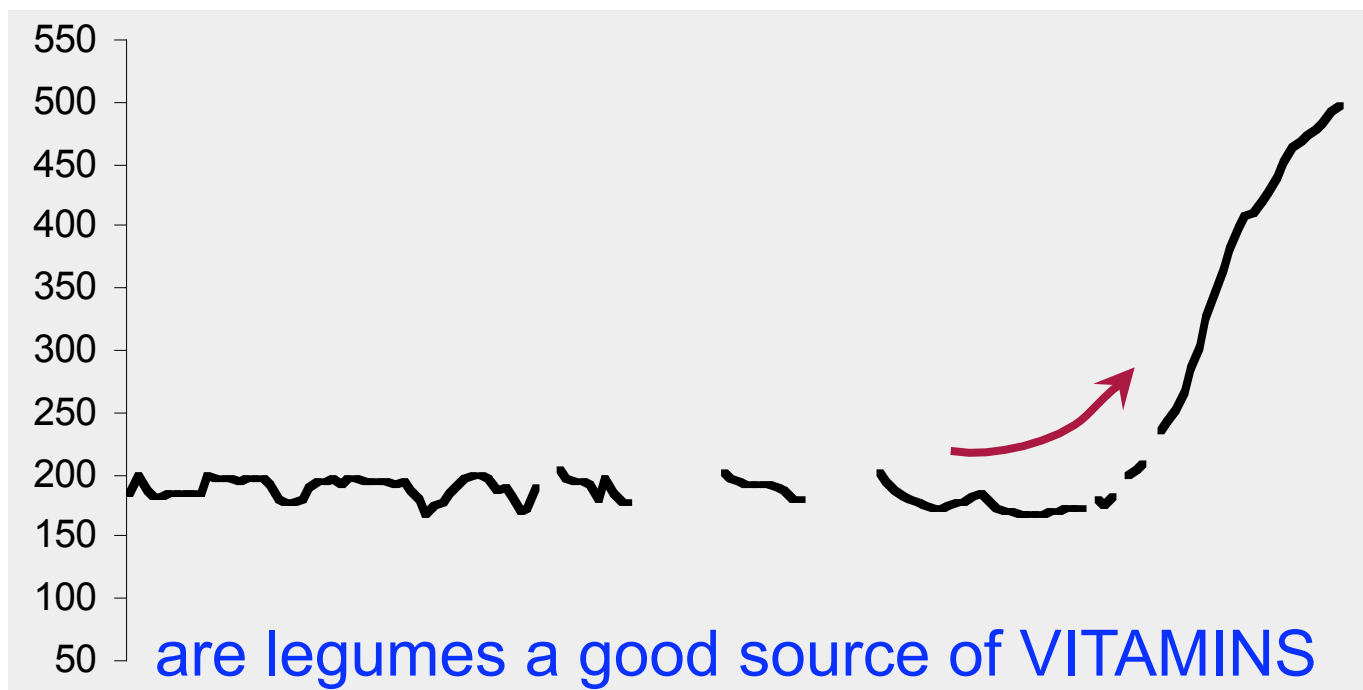
Rise from the main accent to the end of the sentence.

Yes-No question tune



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

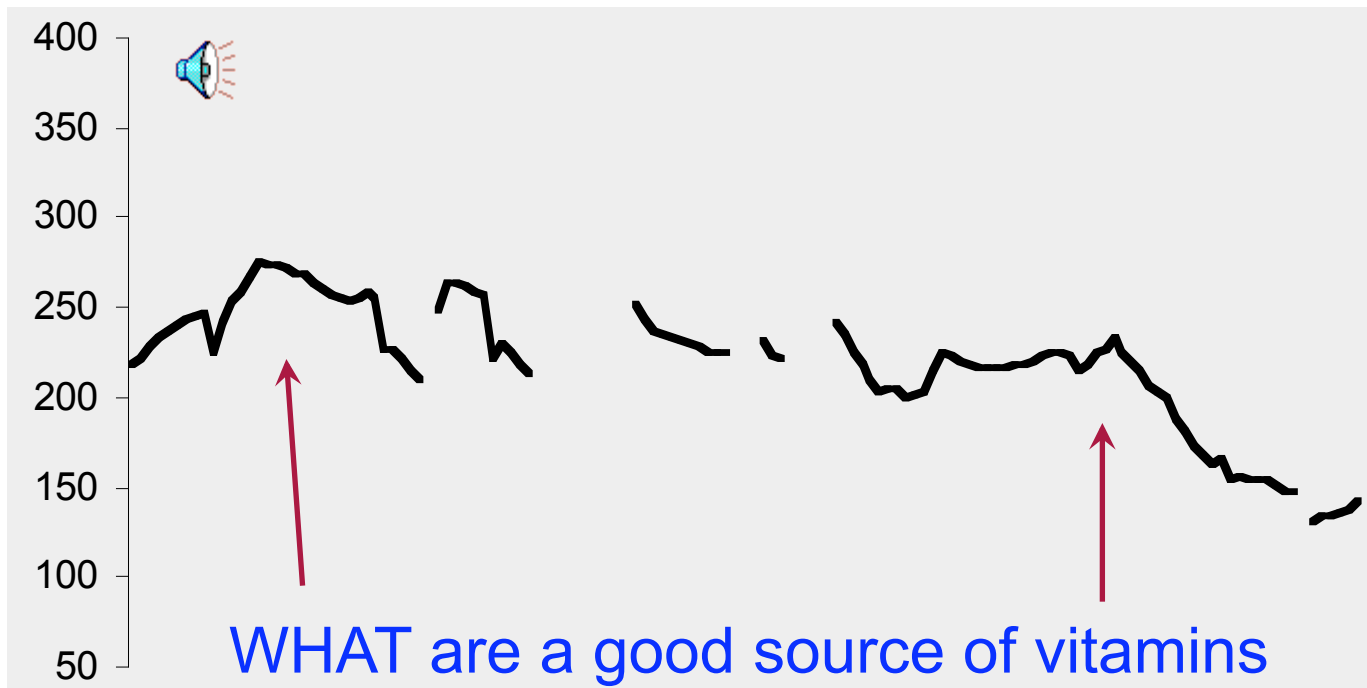
Yes-No question tune



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

WH-questions

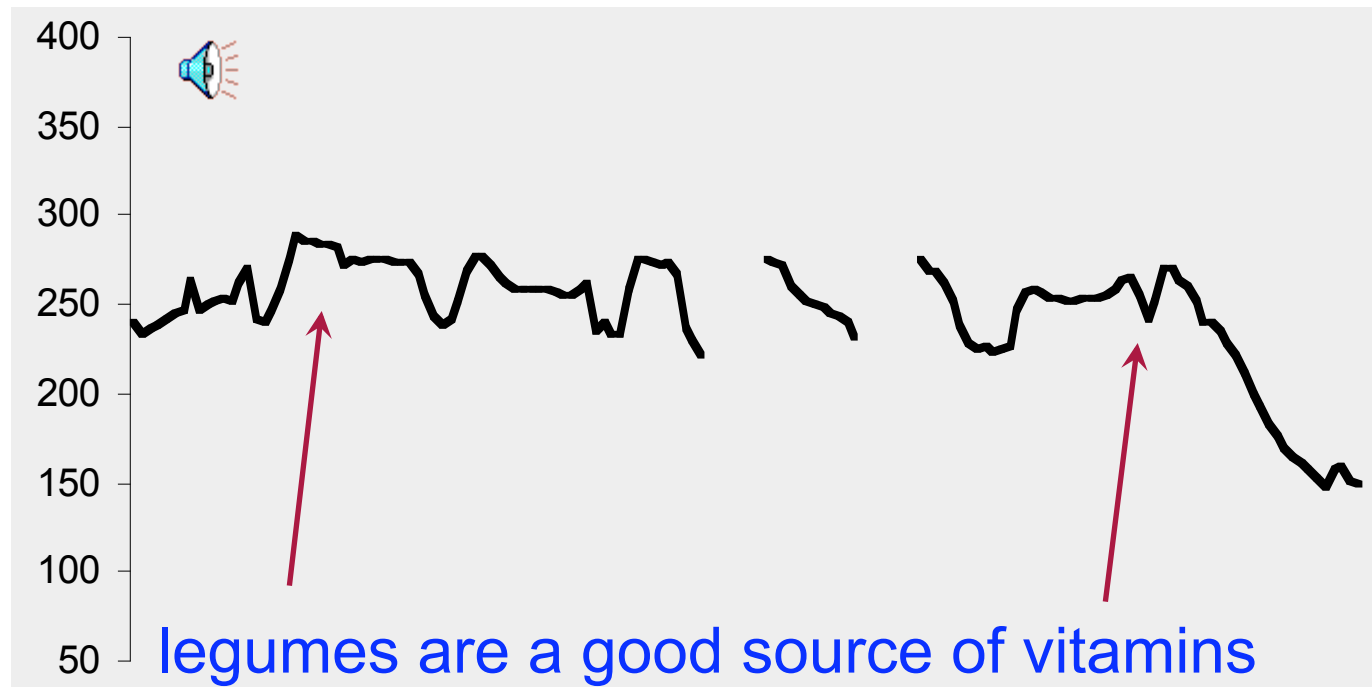
[I know that many natural foods are healthy, but ...]



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

Broad focus

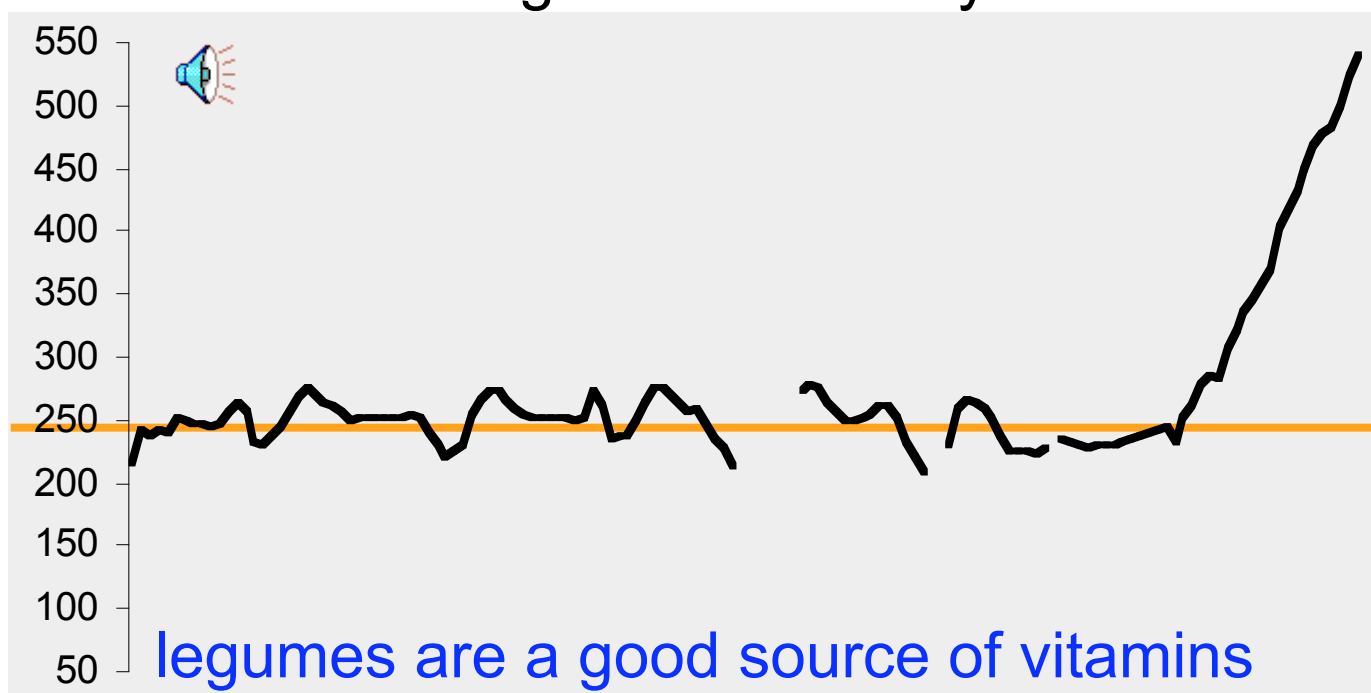
“Tell me something about the world.”



In the absence of narrow focus, English tends to mark the **first** and **last** 'content' words with perceptually prominent accents.

Rising statements

“Tell me something I didn’t already know.”

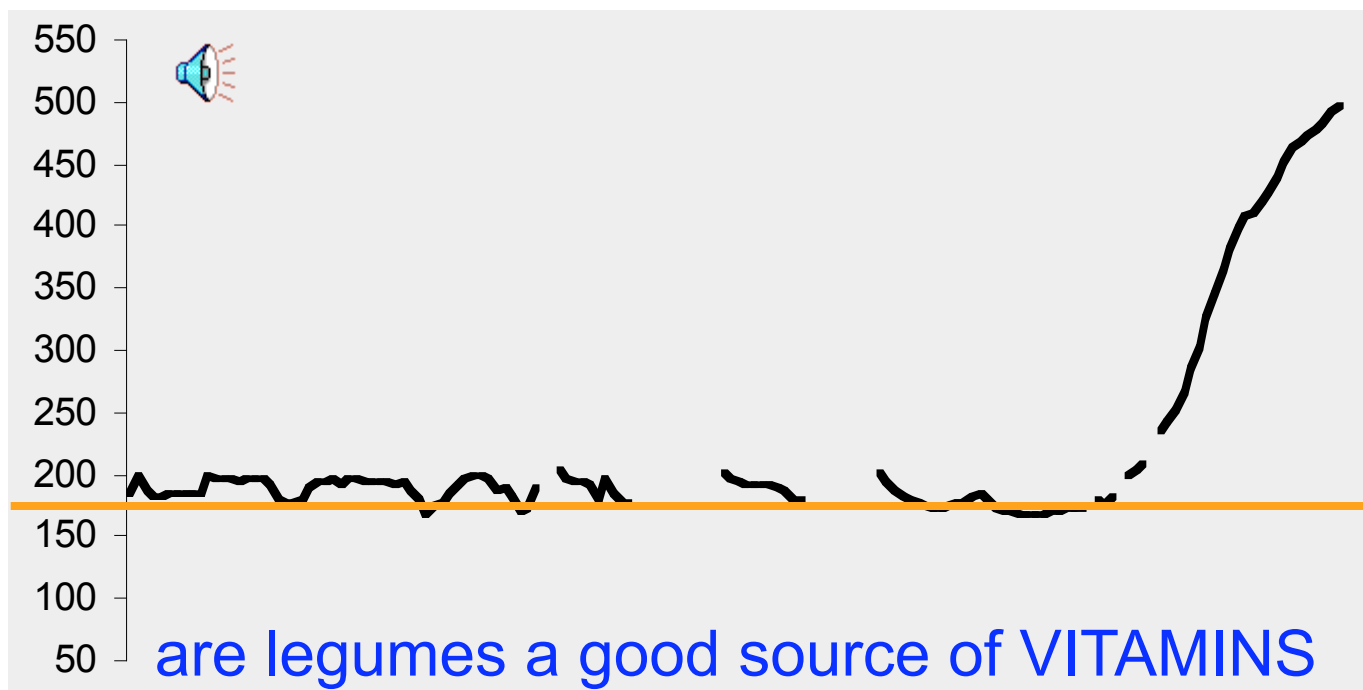


[... does this statement qualify?]

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

Slide from Jennifer Venditti

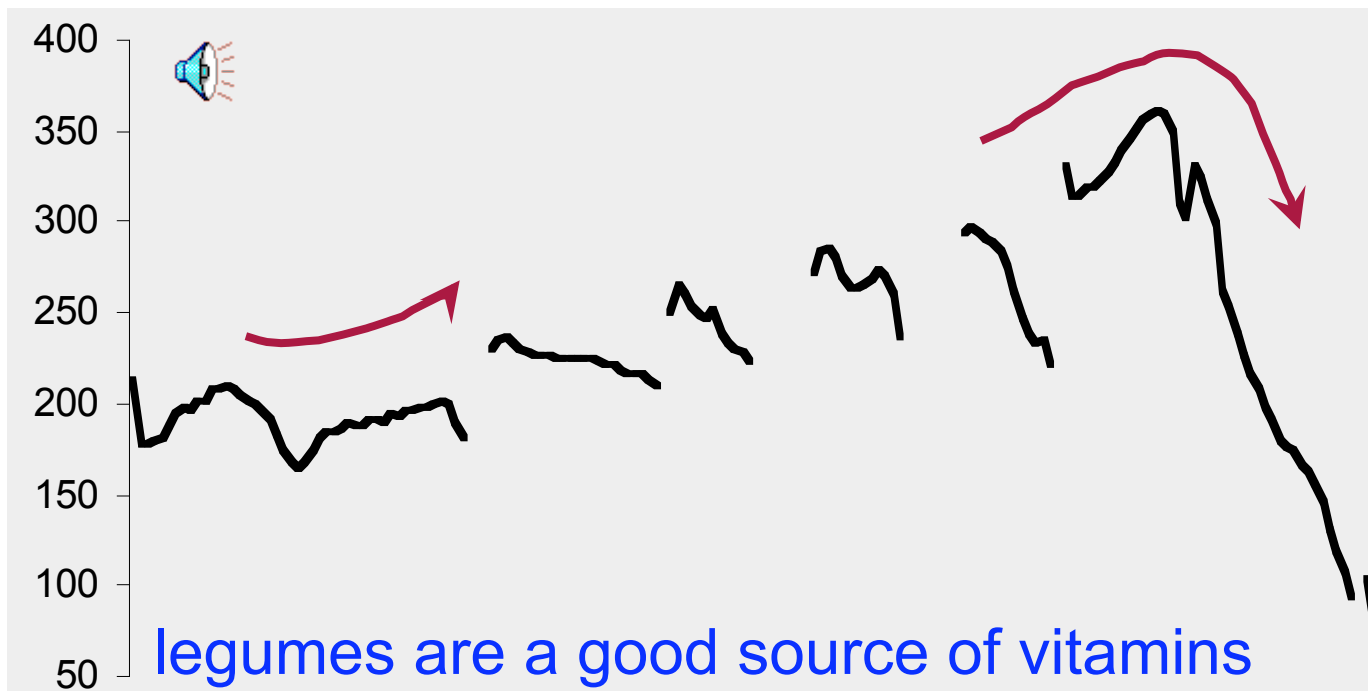
Yes-No question



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

'Surprise-redundancy' tune

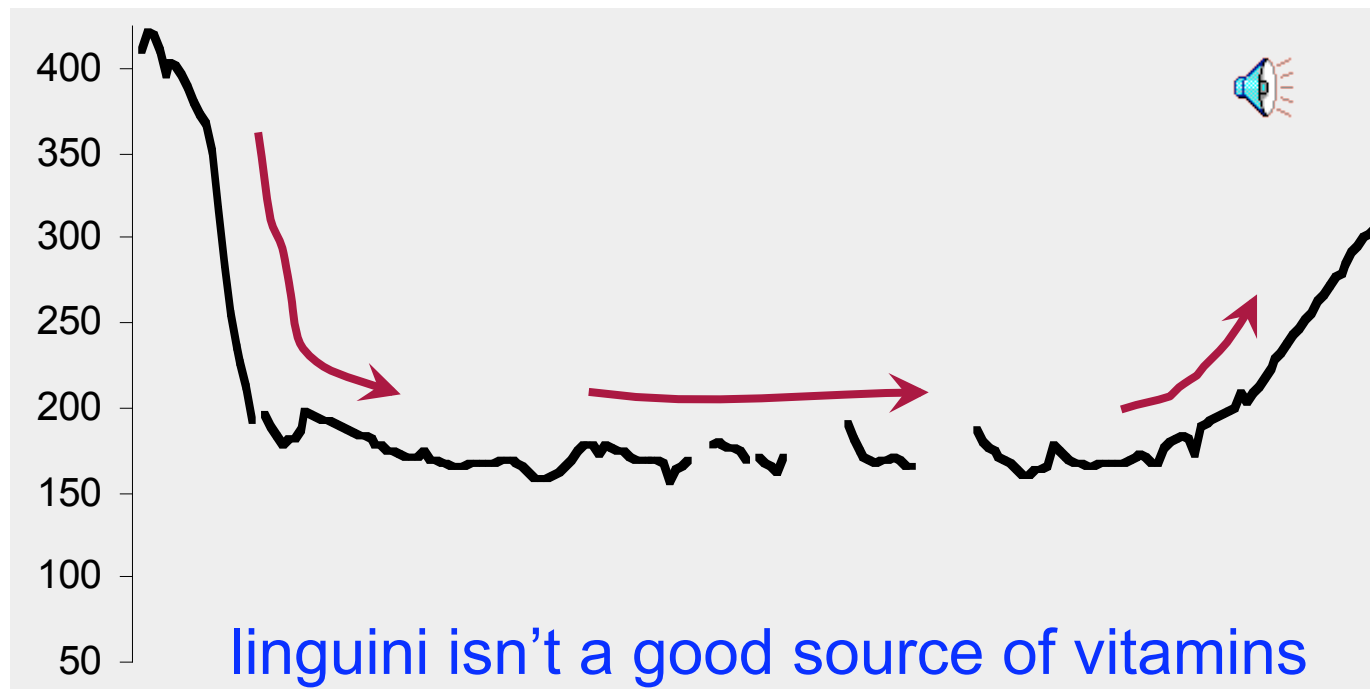
[How many times do I have to tell you ...]



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

'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.

Part 2: Using Intonation in TTS

- 1) **Prominence/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

TOPIC II.1

Predicting pitch accent

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

Simplest possible algorithm for pitch accent assignment

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

But not just function/content:

- 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.

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

List intonation

- I went and saw ANNA, LENNY, MARY, and NORA.

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.

Information Status

- New versus old information.
- Old information is deaccented
- There are LAWYERS, and there are GOOD lawyers
- EACH NATION DEFINES its OWN national INTEREST.
- I LIKE GOLDEN RETRIEVERS, but MOST dogs LEAVE me COLD.



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:
 - ♦ MEDICAL Building, APPLE 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)

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

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

State of the art

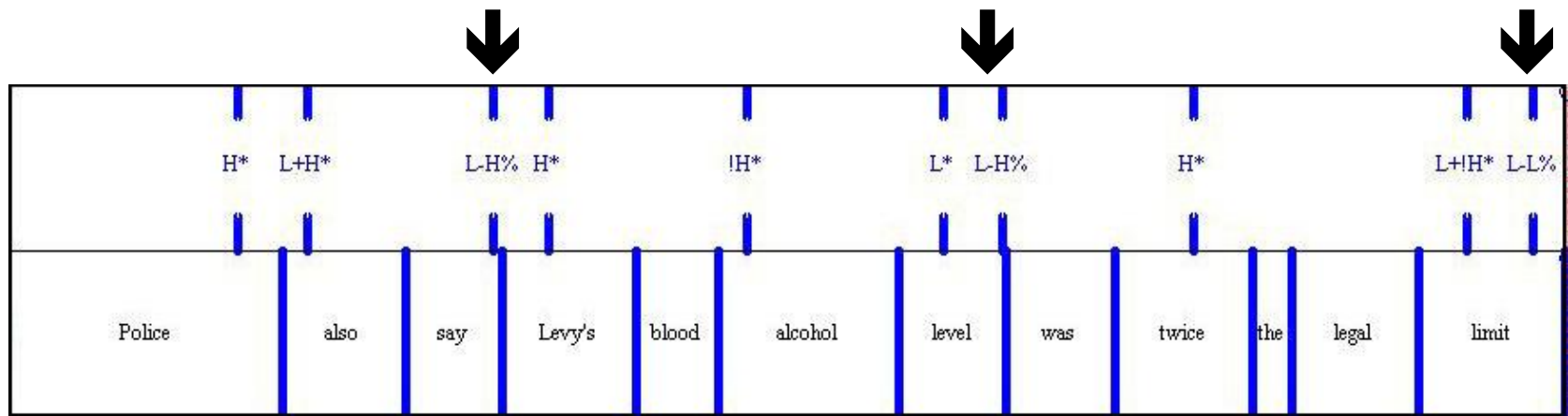
- Hand-label large training sets
- Use CART, SVM, CRF, etc to predict accent
- Lots of rich features from context
- Classic lit:
 - ♦ Hirschberg, Julia. 1993. Pitch Accent in context: predicting intonational prominence from text. *Artificial Intelligence* 63, 305-340

TOPIC II.2

Predicting boundaries

Predicting Boundaries

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



More examples

- 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.

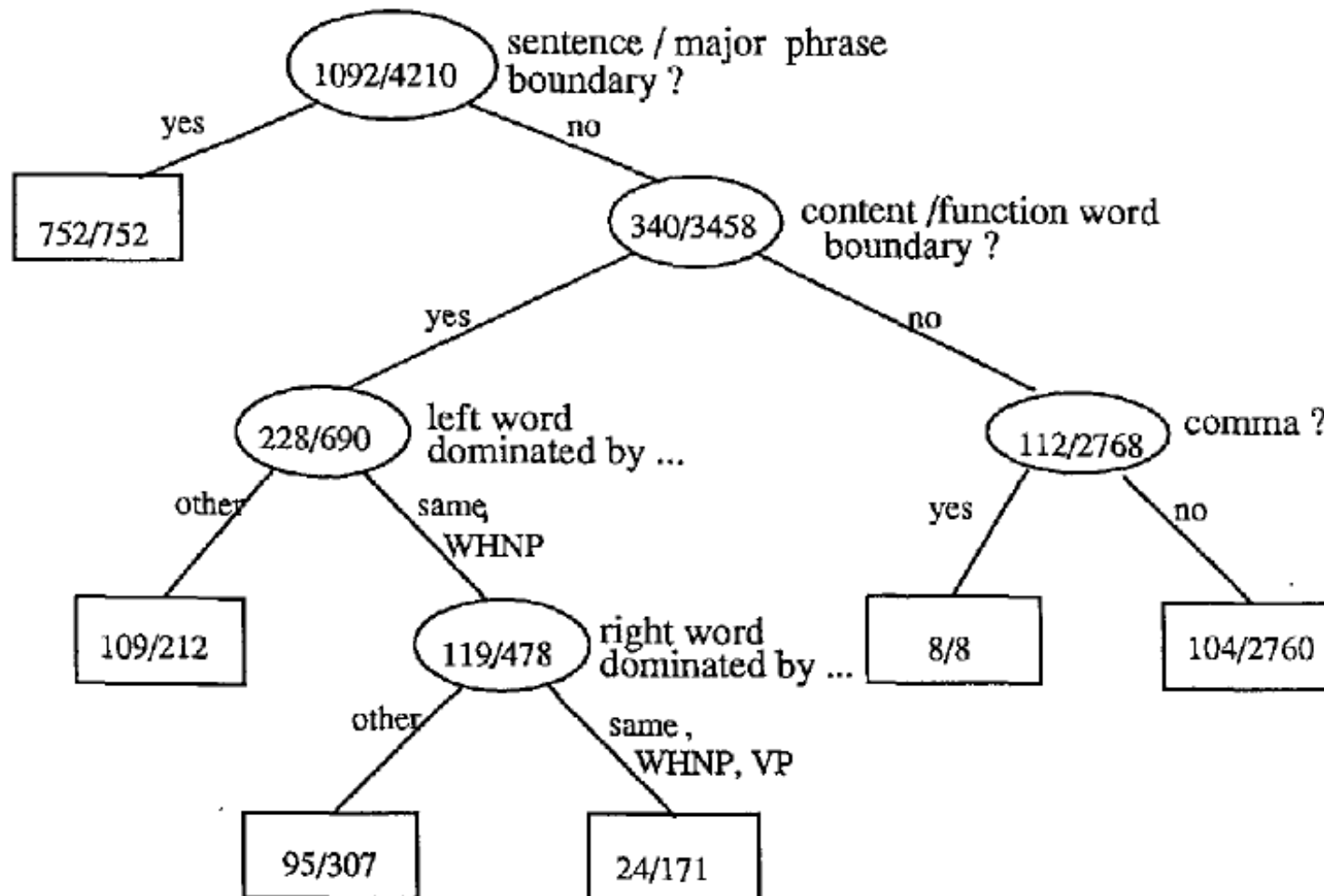
Simplest CART

```
(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))))))
```

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

Ostendorf and Veilleux CART



TOPIC II.3

Predicting duration

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

Klatt duration rules

Models how context-neutral duration of a phone lengthened/shortened by context

- ♦ While staying above a min duration d_{\min}
- Prepausal lengthening:
 - ♦ The vowel or syllabic consonant in the syllable before a pause is lengthened by 1.4
- Non-phrase-final shortening
 - ♦ Segments which are not phrase-final are shortened by 0.6. Phrase-final postvocalic liquids and nasals are lengthened by 1.4
- Unstressed shortening
 - ♦ Unstressed segments are more compressible, so their minimum duration d_{\min} is halved, and are shortened by .7 for most phone types.
- Lengthening for accent
 - ♦ A vowel which bears accent is lengthened by 1.4
- Shortening in clusters
 - ♦ A consonant followed by a consonant is shortened by 0.5
- Pre-voiceless shortening
 - ♦ Vowels are shortened before a voiceless plosive by 0.7

Klatt duration rules

- Klatt formula for phone durations:

$$d = d_{\min} + \prod_{i=1}^N f_i \times (\bar{d} - d_{\min})$$

- Festival: 2 options
 - ♦ Klatt rules
 - ♦ Use labeled training set with Klatt features to train CART
 - Identity of the left and right context phone
 - Lexical stress and accent values of current phone
 - Position in syllable, word, phrase
 - Following pause

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

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))))))
```

TOPIC II.4

F0 Generation

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”)

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)
```

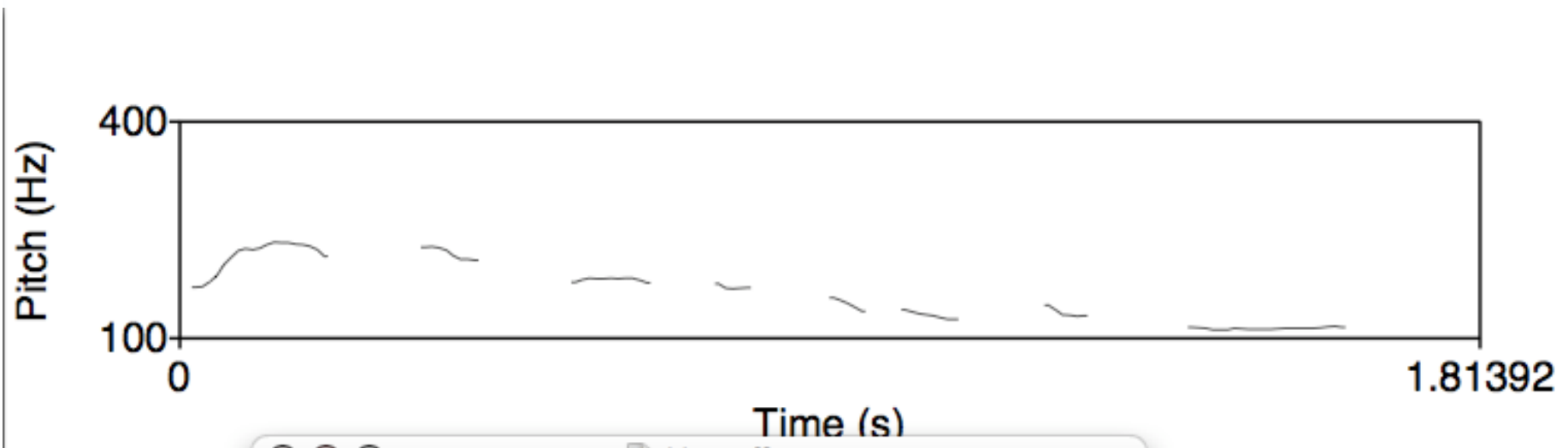
```
        (list end 100))))))
```

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
- F0 is generally defined relative to **pitch range**
 - ♦ A speaker's pitch range is the range between
 - **Baseline frequency**: lowest freq in a particular utterance
 - **Topline frequency**: highest freq in a particular utterance

Declination

- F0 tends to decline throughout a sentence



Advanced:
Intonational
Transcription Theories:
ToBI and Tilt

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.

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



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.

Tilt

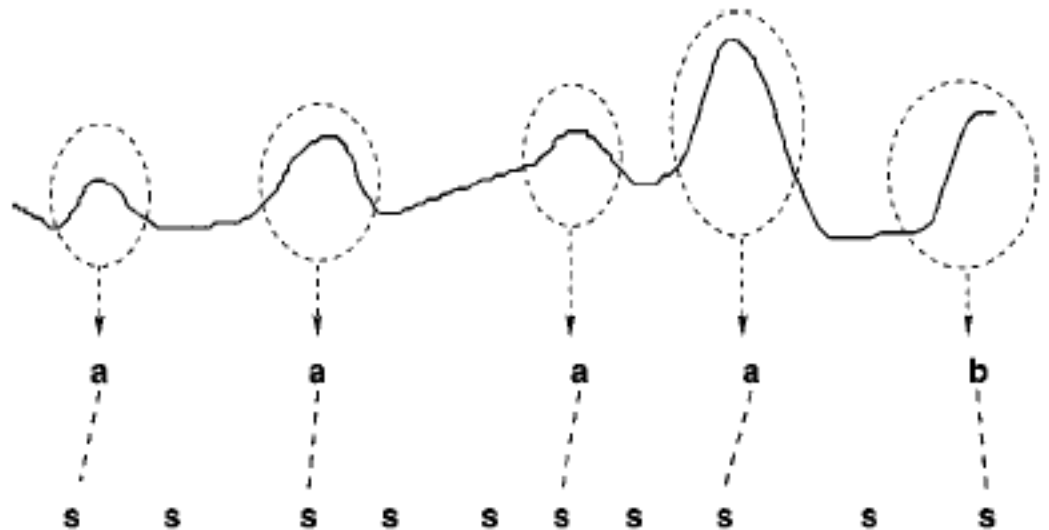
- Taylor (2000)
- Like ToBI,
 - ♦ has sequences of intonational events
 - ♦ Accents and boundary tones
- Unlike ToBI,
 - ♦ not based on discrete classes
 - ♦ Continuous parameters representing F0 shape

Tilt

- Each accent in tilt has a (possibly zero) **rise** followed by a (possibly zero) **fall**

$$\text{tilt} = \frac{\text{tilt}_{\text{amp}} + \text{tilt}_{\text{dur}}}{2} = \frac{|A_{\text{rise}}| - |A_{\text{fall}}|}{|A_{\text{rise}}| + |A_{\text{fall}}|} + \frac{D_{\text{rise}} - D_{\text{fall}}}{D_{\text{rise}} + D_{\text{fall}}}$$

- Tilt** values:
 - ◆ 0 equal rise and fall
 - ◆ 1.0 rise
 - ◆ -1.0 fall
 - ◆ -0.5 rise w/larger fall



Summary

- Linguistics of Prosody
 - ♦ Prominence
 - ♦ Boundaries
 - ♦ Tune
- Producing Intonation in TTS
 - ♦ Predicting Accents
 - ♦ Predicting Boundaries
 - ♦ Predicting Duration
 - ♦ Generating F0

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.
- .