Representing Intonational Variation

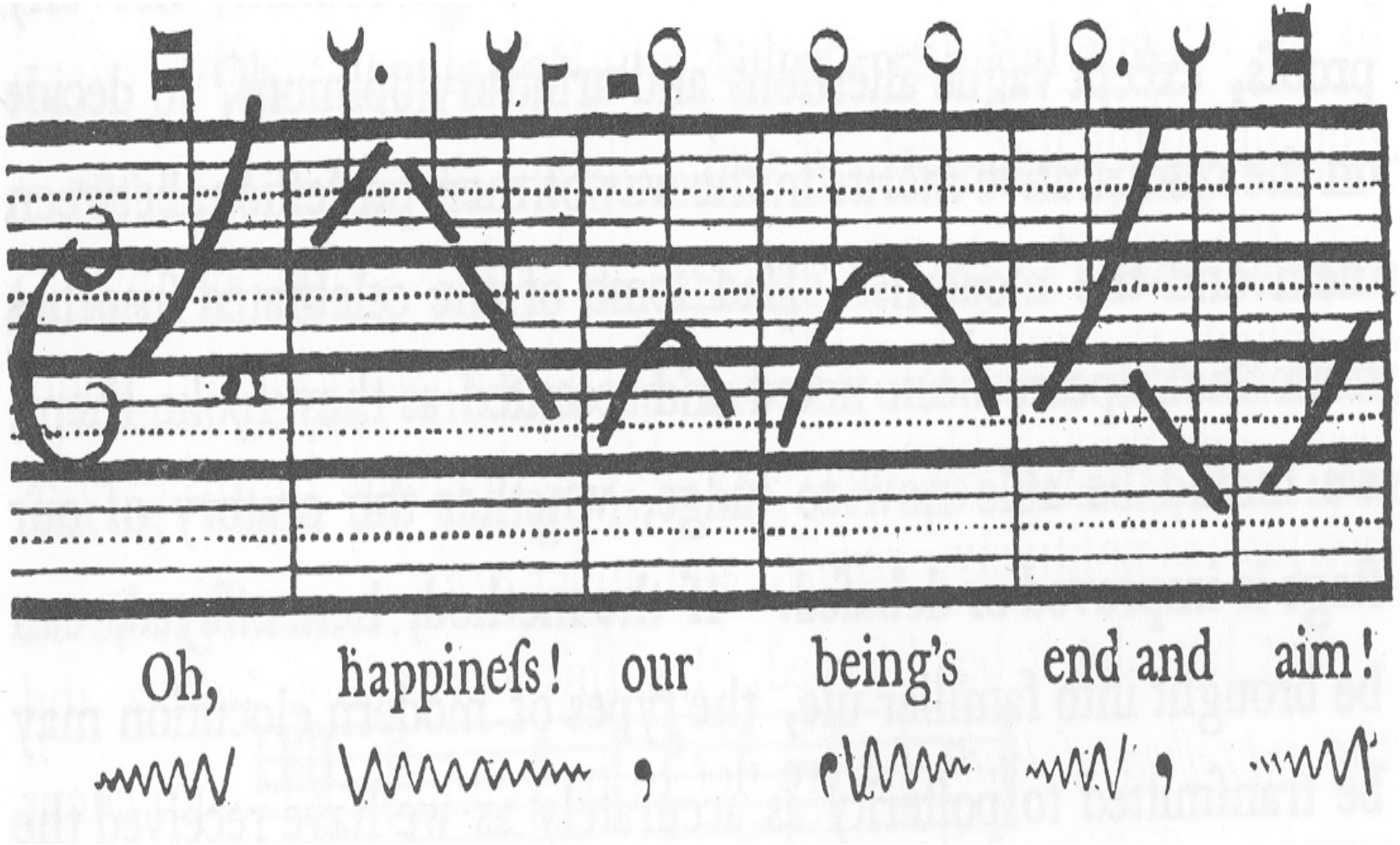
Julia Hirschberg

CS 4706

Today

- How can we represent meaningful speech variation so we can compare utterances? assign in TTS?
  - Expanded vs. compressed pitch range?
  - Louder vs. softer speech?
  - Faster vs. slower speech?
  - Differences in intonational prominence?
  - Differences in intonational phrasing?
  - Differences in pitch contours?

Joseph Steele, 1775



Language Learning Approaches

- A simpler approach
  - / IS it ↘↗ INteresting /
  - / d'you feel ↗ ANGry? /
  - / WHAT'S the ↘ PROBlem? / (McCarthy, 1991:106)
- How much variation do we need to capture?
  - How detailed?
  - Continuous or categorical features?
  - If categorical, what are the possible classes?

How Do We Decide?

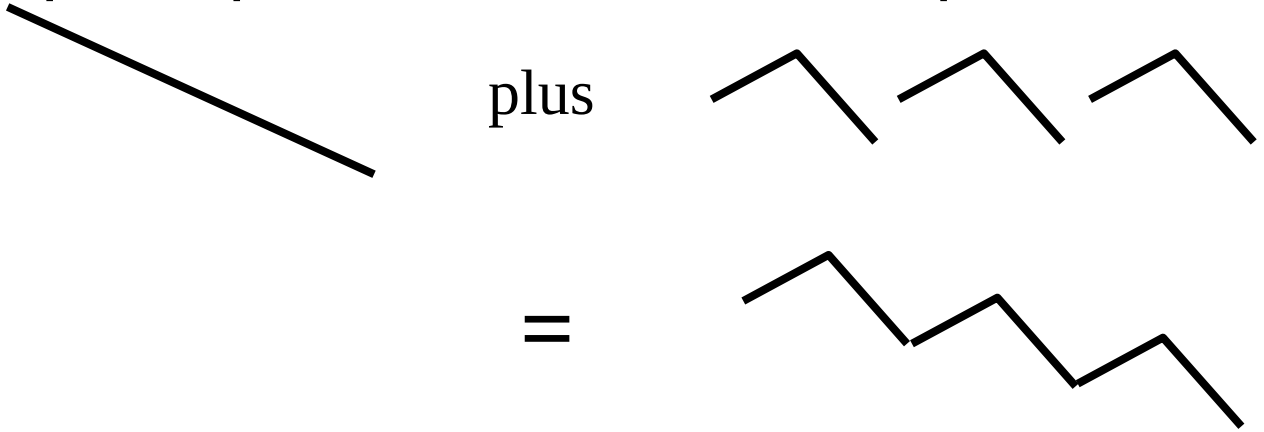
- **Auditory:**
  - Language teachers: what representations can learners understand
- **Acoustic:**
  - Examine the speech signal for critical vs. accidental variation
- **Experimental approaches**
  - Identify potential meaningful variation
  - Design production or perception studies to test
  - E.g. what does a contour ***mean***?

Intonation Models

- ***Superpositional* models** (Fujisaki 1983, Möbius et al. 1993): acoustic/physiological
- ***Linear* or *Tone sequence* models**
  - British school (Kingdon ’58, O’Connor & Arnold ’73, Cruttenden ’97): based on auditory analysis
  - American School (Pierrehumbert ’80, ToBI): mainly acoustic analysis
  - Dutch school (‘t Hart, Collier and Cohen 1990): perceptual data

Superpositional models

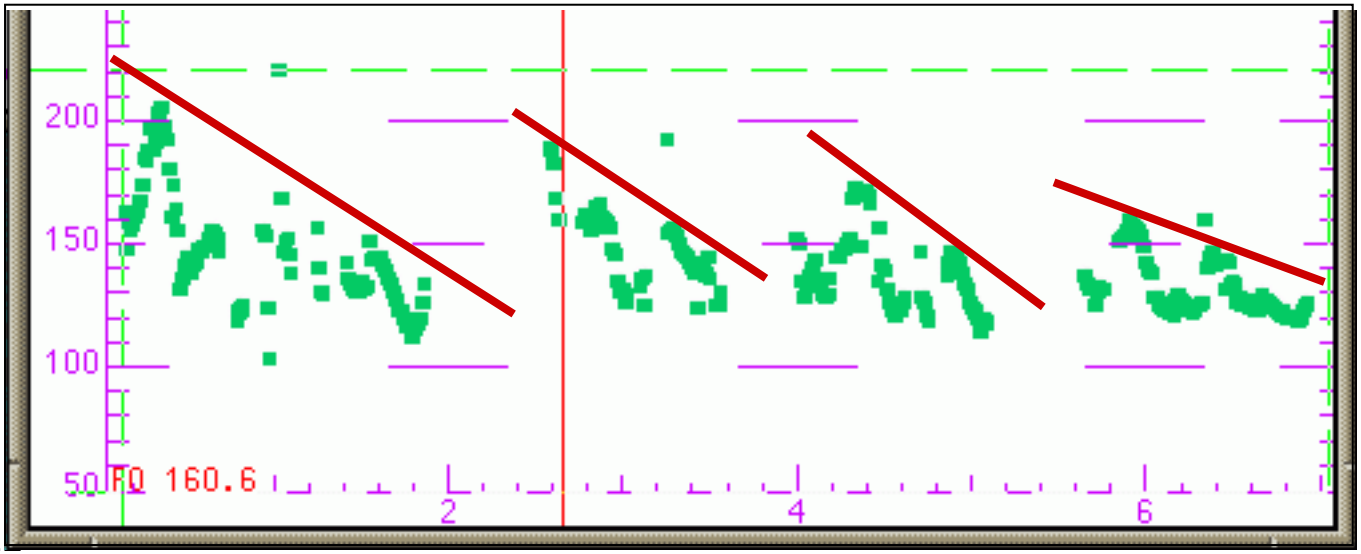
- Pitch pattern of intonation modeled with two components: phrase component and accent component.
- Phrase has basic shape, and pitch movements for individual accents are superimposed over basic shape:



Apples, oranges and tomatoes

Good for modeling utterance-level trends

- Declination: downtrend in f0 over the course of an utterance
- Successful in speech synthesis for languages like Japanese (little variation in accent type, e.g.)



Disadvantages

- Disadvantages
  - Too rigid: All contours must be modeled with an accent and a phrase component
  - Many SAE contours cannot be captured easily
    - Cannot distinguish prominence types
    - Cannot capture differences in phrase endings

- No account of different accent types, or variations in phrase endings
- No notation system which allows users to share observations from large speech corpora or to compare contours
- Used primarily for synthesis

Tone Sequence Models

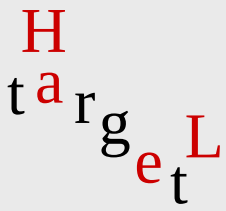
- Intonation generated from sequences of categorically different, phonologically distinctive tones
- Basic unit of intonational description: intonation phrase (tone unit, breath group)
  - Delimited by pauses, phrase-final lengthening, pitch
- Syllables may be stressed or accented
  - Accent aligned with primary stress -- *telephone*
  - Indicated by F0, duration, intensity, voice quality

Types of Tone-sequence Models

Type 1: based on ***pitch movements***

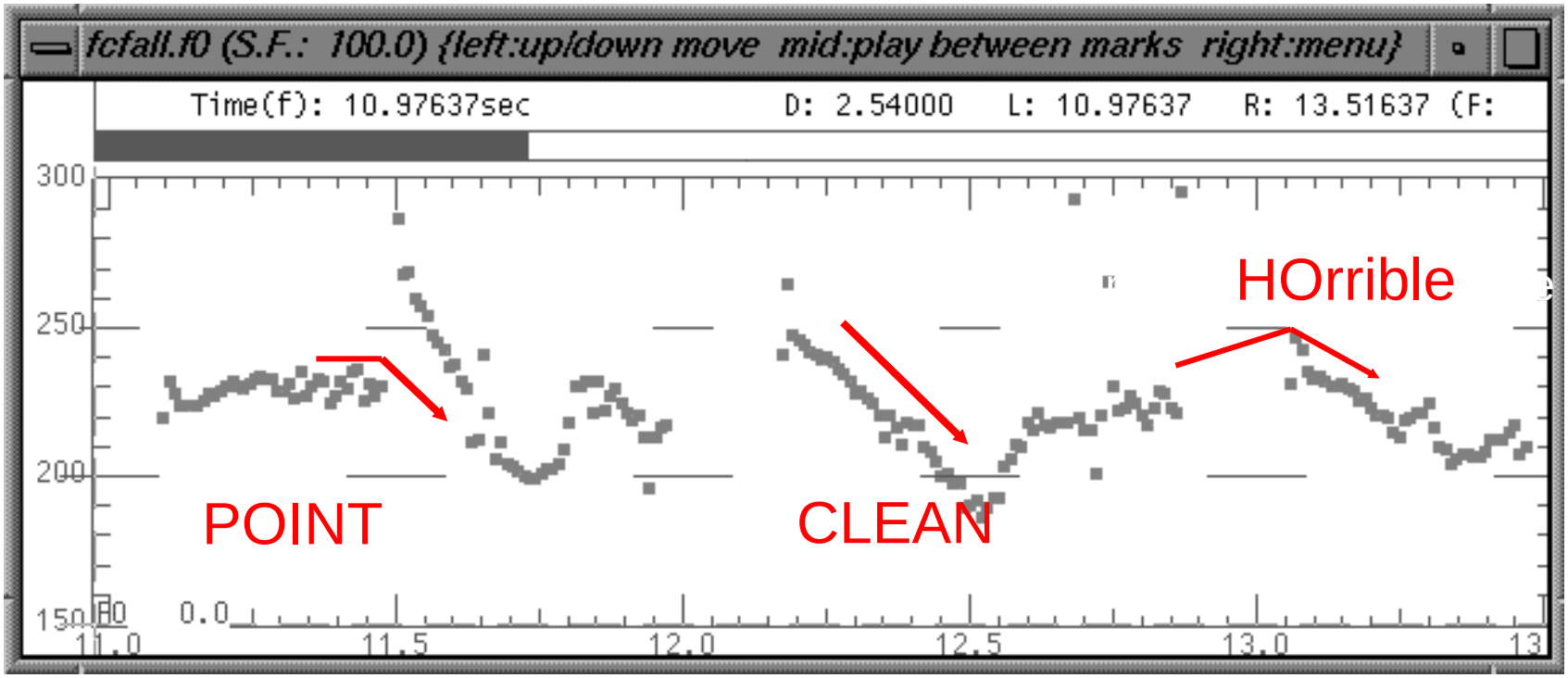
The British School
The Dutch School

Type 2: based on ***pitch levels***



The American School

An example

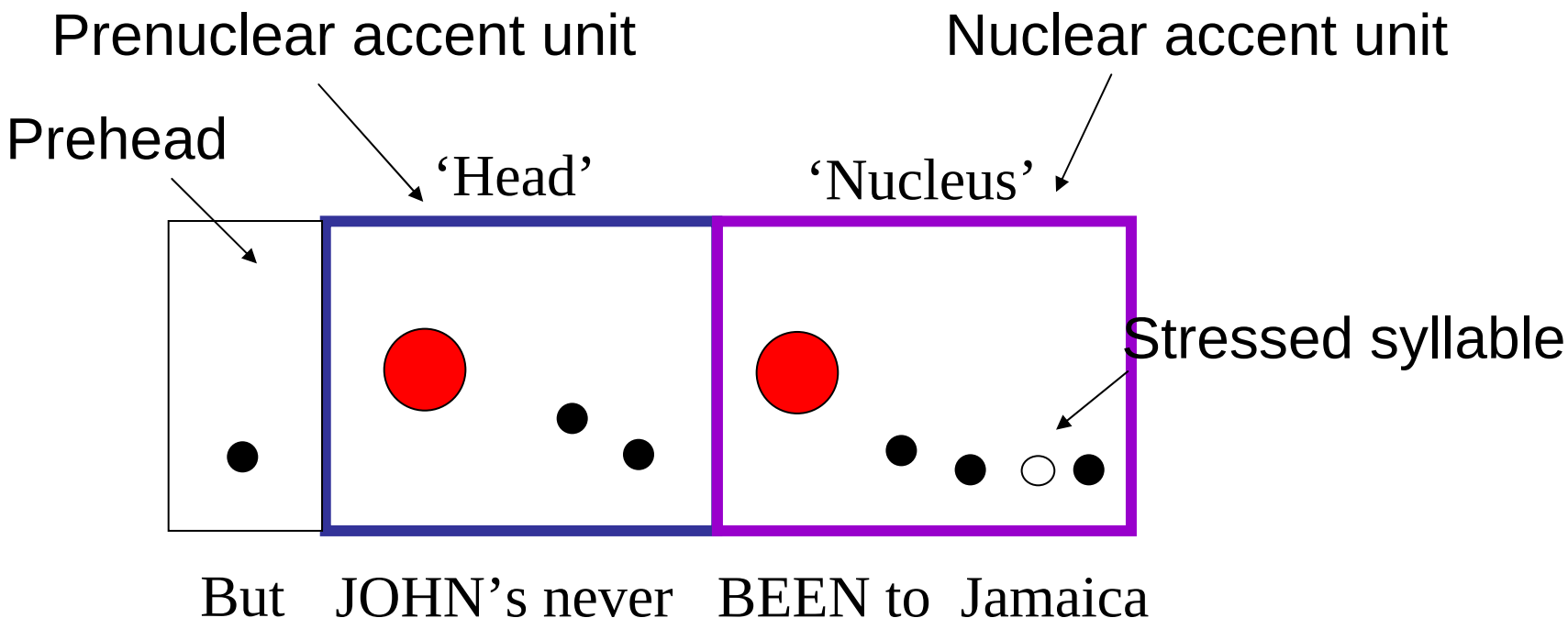


There's a point where you have to clean it and I think it's horrible...

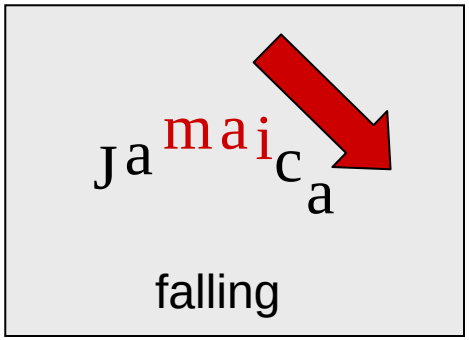
Intonation Phrases

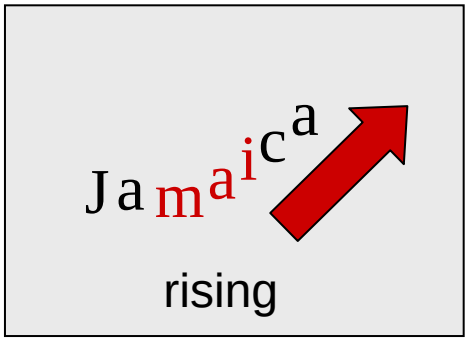
- Internal structure
  - Determined by location of accents in an IP
  - Each accent defines the **beginning** of a prosodic constituent

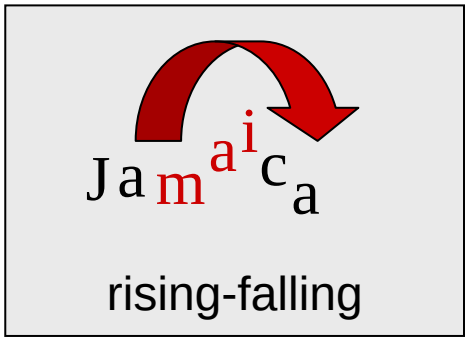
British School

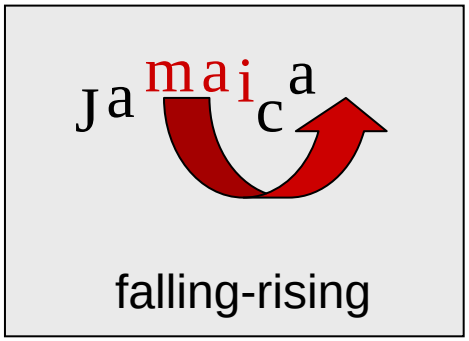


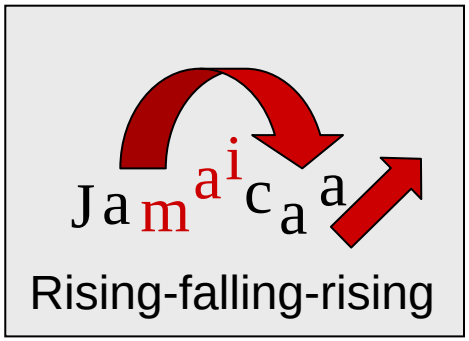
Six nuclear choices in English

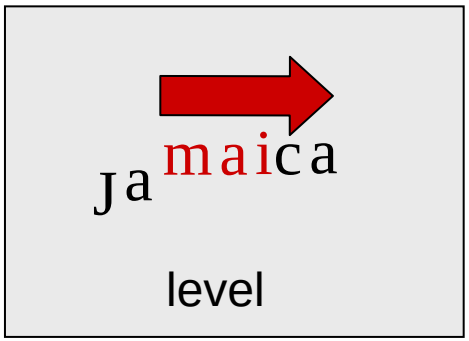










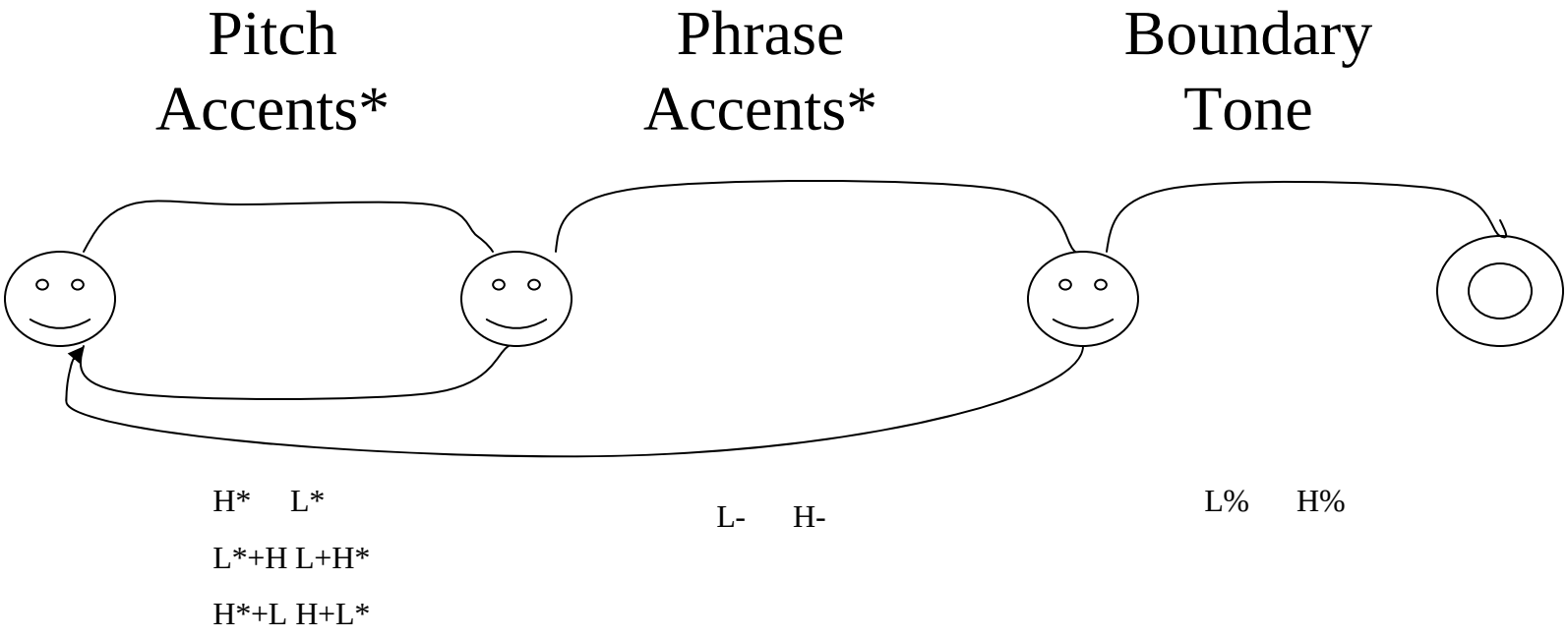


The American School

- American school-type models make a distinction between accents (what makes a particular word prominent) and boundary tones (how a phrase ends)
- Autosegmental metrical or two-tone models
- Only two tones, which may be combined
  - H = high target
  - L = low target

Pierrehumbert 1980

- Contours = pitch accents, phrase accents, boundary tones



Price, Ostendorf et al

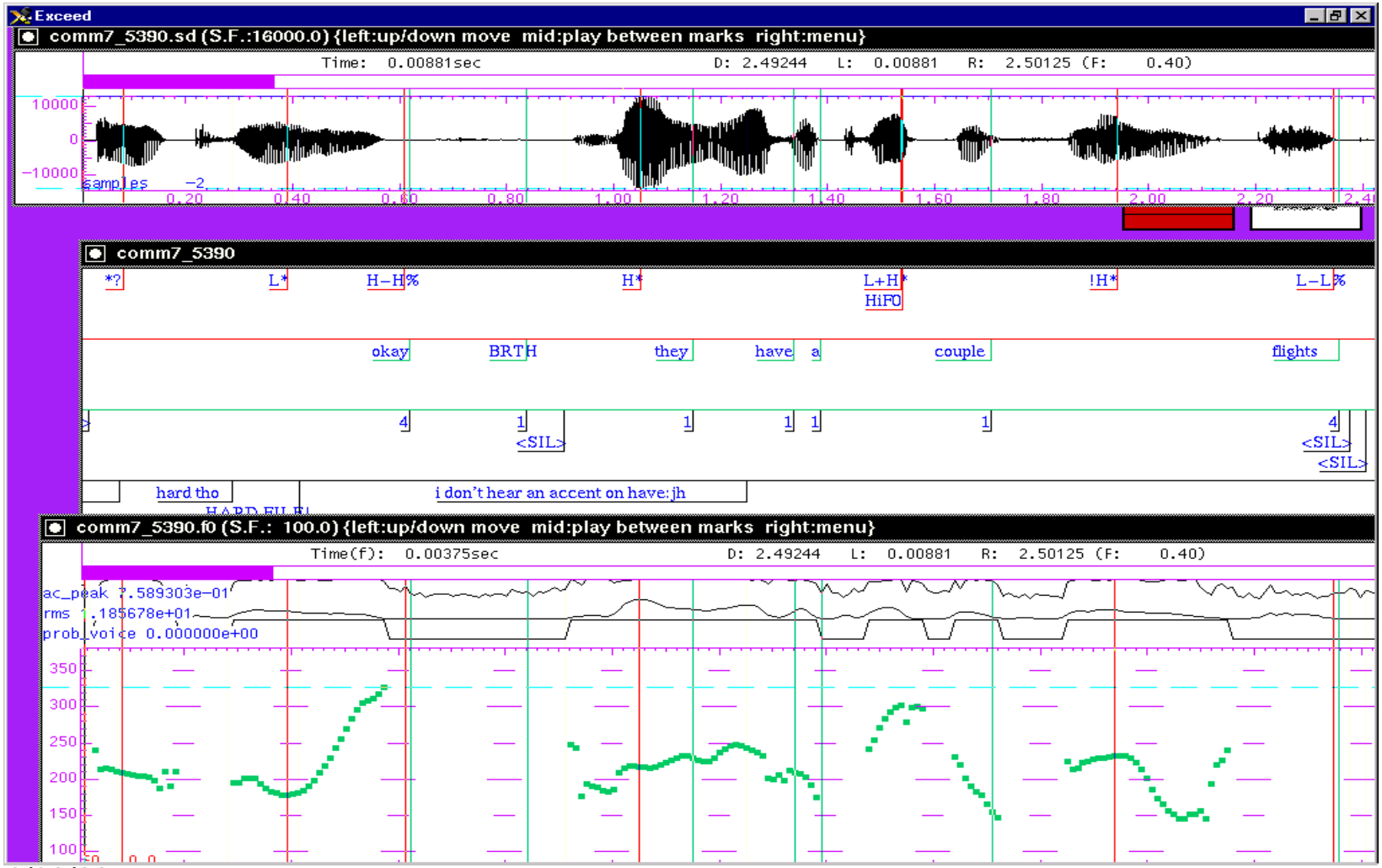
- Break indices: degree of juncture between words
- 0 → 8 (none to ‘a lot’)
  - *What I’d like is a nice roast beef sandwich*.

To(nes and)B(reak)I(ndices)

- Developed by prosody researchers in four meetings over 1991-94
- Putting Pierrehumbert ’80 and Price, Ostendorf, et al together
- Goals:
  - devise common labeling scheme for Standard American English that is robust and reliable
  - promote collection of large, prosodically labeled, shareable corpora

- ToBI standards also proposed for Japanese, German, Italian, Spanish, British and Australian English,....
- Minimal ToBI transcription:
  - Recording of speech
  - F0 contour
  - ToBI tiers:
    - orthographic tier: words
    - break-index tier: degrees of junction (Price et al ‘89)
    - tonal tier: pitch accents, phrase accents, boundary tones (Pierrehumbert ‘80)
    - miscellaneous tier: disfluencies, non-speech sounds, etc.

Sample ToBI Labeling



- Online training material,available at: http://anita.simmons.edu/~tobi/index.html
- Evaluation
  - Good inter-labeler reliability for expert and naive labelers: 88% agreement on presence/absence of tonal category, 81% agreement on category label, 91% agreement on break indices to within 1 level (Silverman et al. ‘92,Pitrelli et al ‘94)

Pitch Accent/Prominence in ToBI

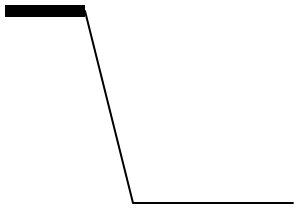
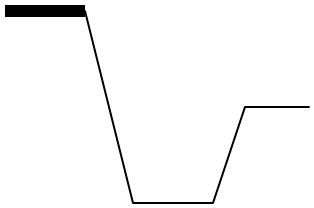
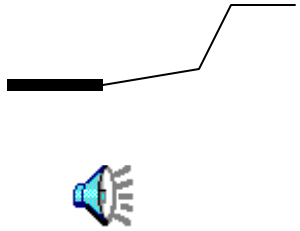
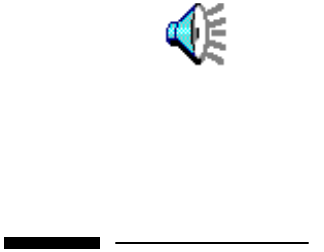
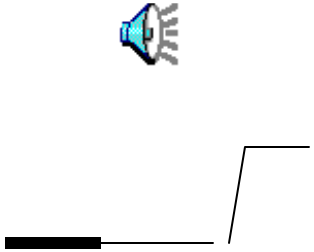
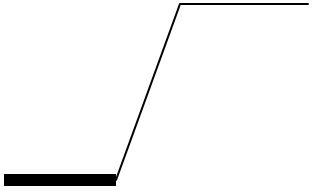
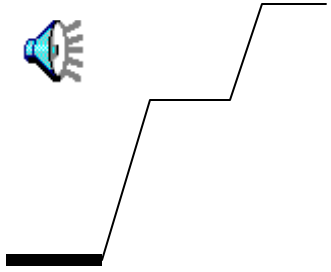
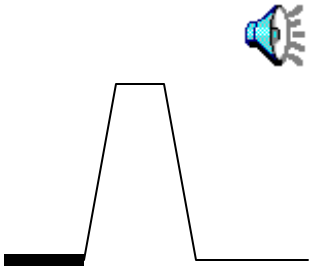
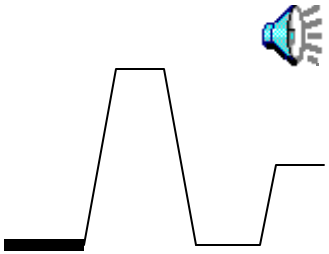
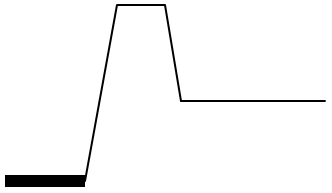
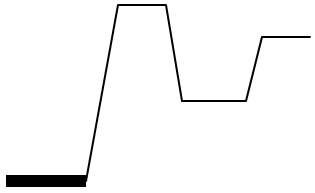
- Which items are made intonationally prominent and how: tonal targets/levels not movement
- Accent type:
  - H* simple high (declarative)
  - L* simple low (ynq)
  - L*+H scooped, late rise (uncertainty/ incredulity)
  - L+H* early rise to stress (contrastive focus)
  - H+!H* fall onto stress (implied familiarity)

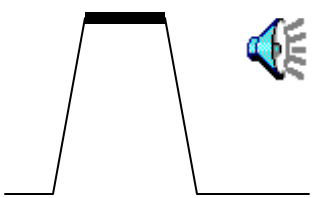
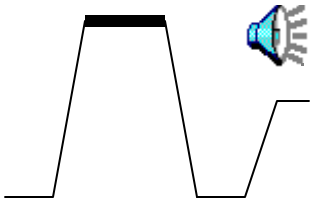
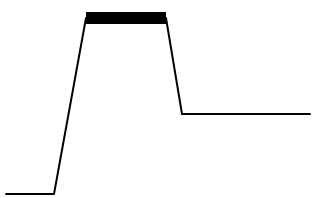
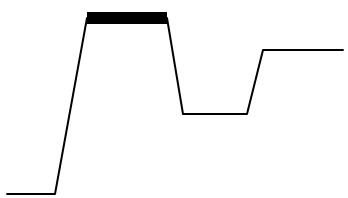
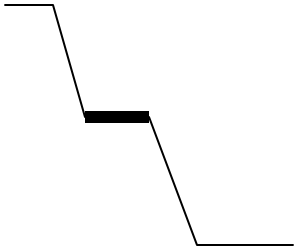
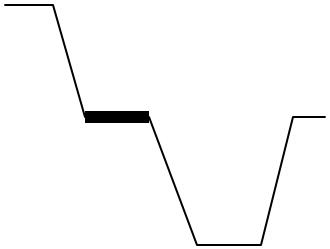
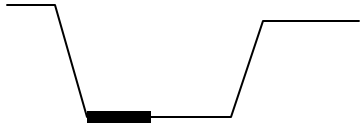
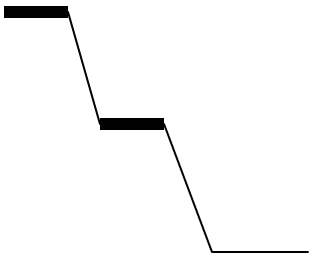
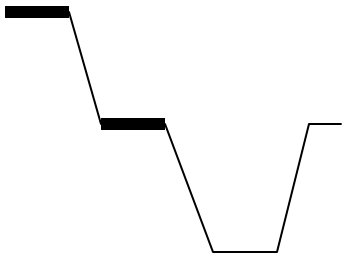
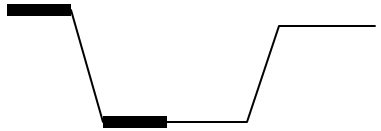
- **Downstepped accents:**
  - **!H*,**
  - **L+!H*,**
  - **L*+!H**
- **Degree of prominence:**
  - **within a phrase: HiF0 (~nuclear accent)**
  - **across phrases ??**

Prosodic Phrasing in ToBI

- 'Levels' of phrasing:
  - intermediate phrase: one or more pitch accents plus a phrase accent, H- or L-
  - intonational phrase: 1 or more intermediate phrases + boundary tone, H% or L%
- ToBI break-index tier
  - 0 no word boundary
  - 1 word boundary

- 2 strong juncture with no tonal markings
- 3 intermediate phrase boundary
- 4 intonational phrase boundary

	L-L%	L-H%	H-L%	H-H%
H*				
L*				
L*+H				

	L-L%	L-H%	H-L%	H-H%
L+H*				
H+!H*				
H* !H*				

- ToBI exercises

Next Class

- Predicting prosodic assignments from text