CS 533
Natural Language Processing
Lecture 2 – February 3, 2003

Matthew Stone
(with input from Justine Cassell & Jennifer Venditti)

Department of Computer Science
Center for Cognitive Science
Rutgers University

Utterances – outline

Language in discourse
Intonation
Facial expression
Gesture and other action

Language in discourse

Take a conversation I recorded in a California drug store as I was buying a couple of items from a clerk I will call Stone.

- Clark, p. 31

Drugstore example

Clark walks up to a counter and places two items next to the cash register.

Stone is behind the counter marking off items on an inventory.

Clark, looking at Stone, catches her eye.

Stone, meeting Clark's eyes: "I'll be right there."

Clark: "OK."

Drugstore example, continued

Stone continues marking off items for fifteen seconds, puts the inventory aside, turns toward Clark, and manifestly begins to look for the items Clark is purchasing.

Clark, noting her search, points at the two items on the counter between them: "These two things over here."

Stone nodes, takes the items, examines the prices on them, and rings them up.

Drugstore example, continued

Stone: "Twelve seventy-seven."

Clark: "Twelve seventy-seven."

Clark takes out his wallet, extracts a twenty-dollar bill, hands it to Stone, then rummages in his coin purse for coins.

Clark: "Let's see that's two pennies I've got two pennies."

Clark hands Stone two pennies.

Drugstore example, continued

Stone: "Yeah."

Stone then enters $20.02 in the register, which computes the change.

Stone (handing change to Clark): "Seven twenty-five is your change."

Clark: "Right."

Clark puts the money in his wallet while Stone puts the items and receipt in a bag. She hands the bag to Clark, they break off…

Language in discourse

A discourse is simply a joint activity in which conventional language plays a prominent role.

Mostly linguistic	phone call, newspaper, radio conversation, tabloid, tv, science text transactions, plays, movies, coaching basketball, tennis, moving furniture
Mostly nonlinguistic	string quartet, waltz, catch

Meaning and participation draw on everything that goes on

"Transcript" from drugstore

Stone:	I'll be right there.
Clark:	OK.
Clark:	These two things over here.
Stone:	Twelve seventy-seven.
Clark:	Twelve seventy-seven.
Clark:	Let's see that's two pennies I've got two pennies.
Stone:	Yeah.
Stone:	Seven twenty-five is your change.
Clark:	Right.

Meaning and participation draw on everything that goes on

"Transcript" from drugstore

Stone: I'll be right there.

To know what this means, need to know that Clark had just caught Stone's eye and was waiting to be served.

Meaning and participation draw on everything that goes on

"Transcript" from drugstore

Stone: Twelve seventy-seven.

To know what this means, need to know that Stone has just rung up items on register.

Meaning and participation draw on everything that goes on

Conversely, *nonlinguistic action*s have a communicative role in taking things forward.

Clark's catching Stone's eye *was* a request for service.

Meaning and participation draw on everything that goes on

Conversely, *nonlinguistic actions* have a communicative role in taking things forward.

When Clark hands over the $20, there is a *joint action* of changing the $20 from Clark's possession to Stone's.

This is communicative since Clark could have been lending, asking for two tens, asking to check if it's counterfeit, etc…

Empirical motivation for discourse as joint activity

- The same channels carry *propositional* information (about the content of what is being said) and *interactional* information (about the process of conversation).
- Propositional and interactional information are carried by *verbal* (speech, intonation) and *visual* (facial expression, gesture, posture) means.

That is . .

- Propositional Layer
  - Verbal and visual behaviors that contribute to the intended meaning.
  - Verbal: content of speech & intonation
  - Visual / Non-verbal: deictic, iconic & metaphoric gestures
- Interactional Layer
  - Verbal and visual behaviors that regulate, coordinate and manage information flow.
  - Verbal: back-channels, "uh-huh"
  - Non-verbal / visual: gaze, nods, facial expressions, etc.

Some conversational behaviors

- Speech
- Intonation
- Filled pauses ("umm" & other noises)
- Eye gaze towards & away from interlocutor
- Raising eyebrows
- Nods & head shakes
- Hand gestures

Some functions filled by conversational behaviors

- Conversation initiation
- Giving & taking turns
- Giving and requesting feedback
- Breaking away

Interactional (about process)

- Conveying information

Propositional (about content)

Summary and leadin

"If we take language use to include such communicative acts as eye gaze, iconic gestures, pointing, smiles and head nods – and we must – then all joint activities rely on language use. Chess may appear nonlinguistic, but every chess move is really a communicative act, and every chess game a discourse."

- Clark, p. 58

Utterances in detail: Intonation

An important ingredient in believable conversational utterances.

An "easy" case study for techniques that allow you to describe something that happens in the world in a way that shows it to be meaningful.

Intonation and meaning

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.

A: I'd like to fly to Davenport, Iowa on TWA.
B: TWA doesn't fly there ...
B1: They fly to Des Moines.
B2: They fly to Des Moines.

A1: I met Mary and Elena's mother at the mall yesterday.
A2: I met Mary and Elena's mother at the mall yesterday.

Speech production

oral & nasal cavities

larynx

lungs

air

Speech production

oral & nasal cavities

larynx

lungs

The vocal folds may be held wide open, or may vibrate.

Speech production

oral & nasal cavities

larynx

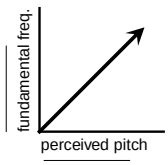
lungs

Positioning of the tongue, lips, etc. acoustically 'shapes' the air.

Vocal fold vibration

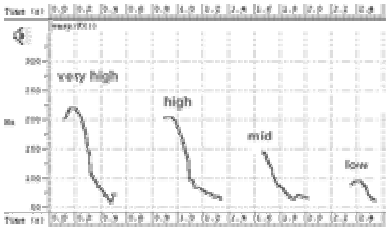
Physical: **Fundamental frequency (F0)**
⇒ rate of vibration of the vocal folds

Perceptual: **Pitch**



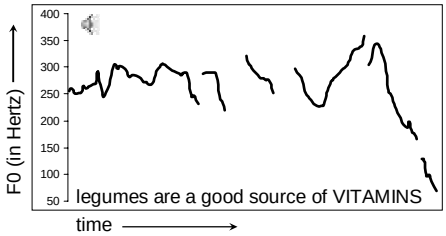
Pitch range

- Differences can be due to physical size, gender, social identity, excitement level, linguistic, etc ...

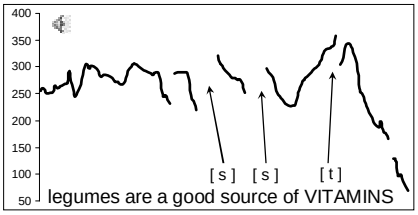


[from *Prosody on the Web* tutorial on pitch]

Graphic representation of F0

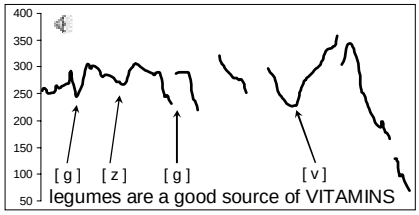


The 'ripples'



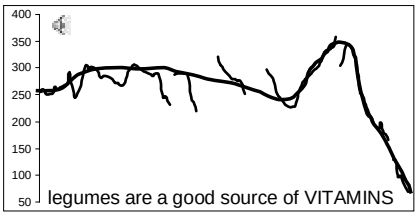
F0 is not defined for consonants without vocal fold vibration.

The 'ripples'



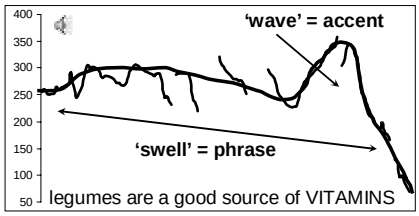
... and F0 can be perturbed by consonants with an extreme constriction in the vocal tract.

Abstraction of the F0 contour



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

The 'waves' and the 'swells'



Symbolic description

Phrases: group words together
- use parentheses informally
- formally: mark phrasal tone and maybe boundary

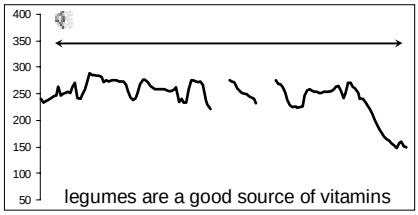
Accents: mark words as prominent
- use capital letters informally
- formally: mark kind of pitch movement at accent

(legumes are a good source of VITAMINS)

or

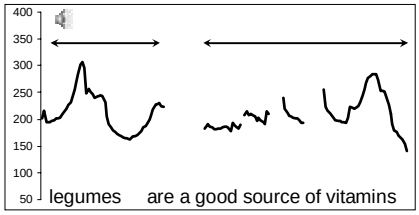
legumes are a good source of vitamins H* L-L%

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.

Stress vs. accent

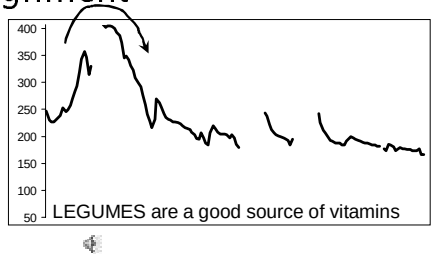
- ***Stress*** is a structural property of a word — it marks a potential (arbitrary) location for an accent to occur, if there is one.
- ***Accent*** is a property of a word in context — it is a way to mark intonational prominence in order to 'highlight' important words in the discourse.

(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	

Which word receives an accent?

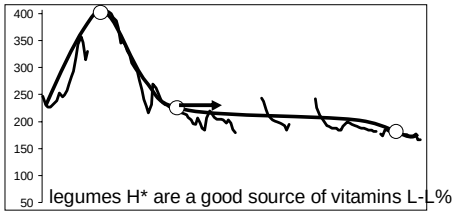
- It depends on the context. For example, the 'new' information in the answer to a question is often accented, while the 'old' information usually is not.
  - Q1: What types of foods are a good source of vitamins?
  - A1: LEGUMES are a good source of vitamins.
  - Q2: Are legumes a source of vitamins?
  - 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.

Same 'tune', different alignment



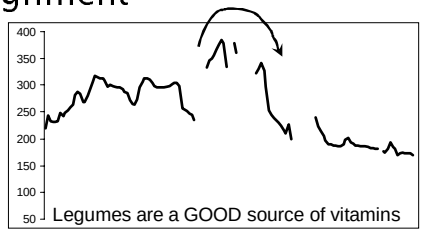
The main **rise-fall** accent (= "I assert this") shifts locations.

Same ‘tune’, different alignment



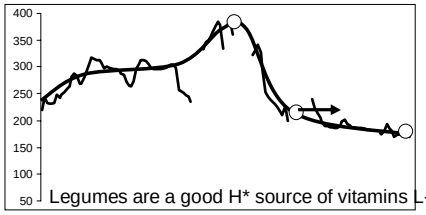
The main **rise-fall** accent (= “I assert this”) shifts locations.

Same ‘tune’, different alignment



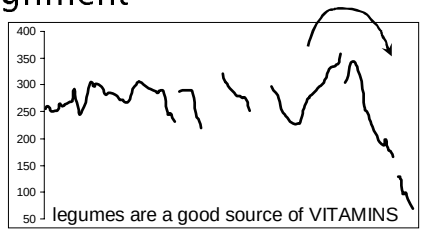
The main **rise-fall** accent (= “I assert this”) shifts locations.

Same ‘tune’, different alignment



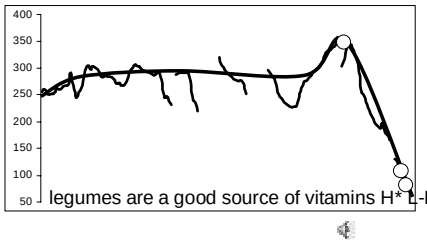
The main **rise-fall** accent (= “I assert this”) shifts locations.

Same ‘tune’, different alignment



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

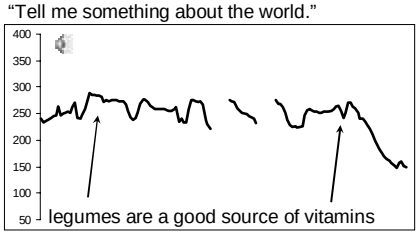
Same ‘tune’, different alignment



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

Broad focus

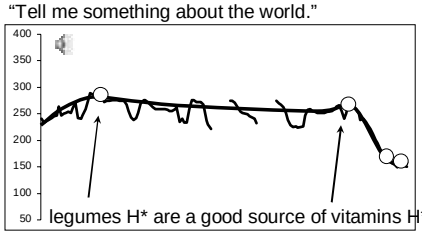
“Tell me something about the world.”



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

Broad focus

"Tell me something about the world."



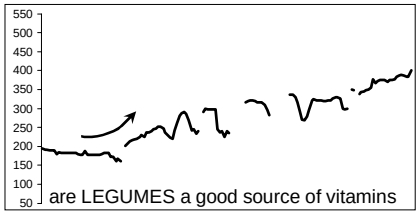
In the absence of narrow focus, English tends to mark the first and last 'content' words with perceptually prominent accents.

Key point

Symbolic description

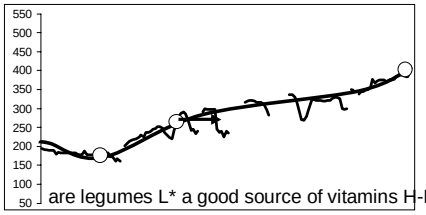
- Abstracts away from continuous event into descrete categories.
- Highlights meaningful differences among utterances.

Yes-No question tune



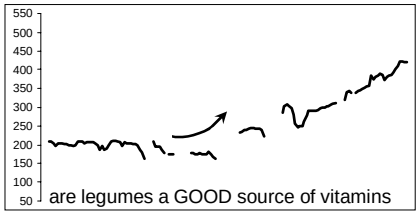
Rise from the main accent to the end of the sentence.

Yes-No question tune



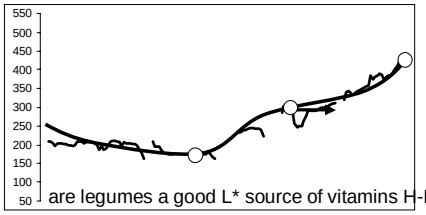
Rise from the main accent to the end of the sentence.

Yes-No question tune



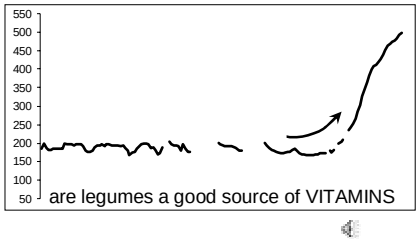
Rise from the main accent to the end of the sentence.

Yes-No question tune



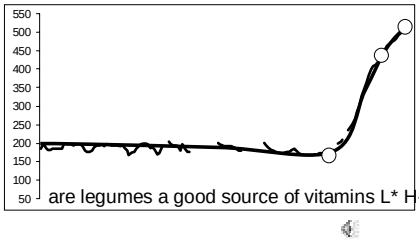
Rise from the main accent to the end of the sentence.

Yes-No question tune



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

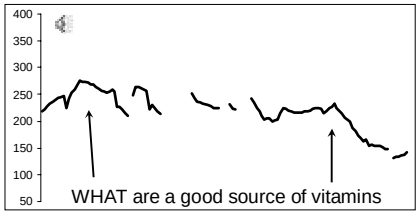
Yes-No question tune



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

WH-questions

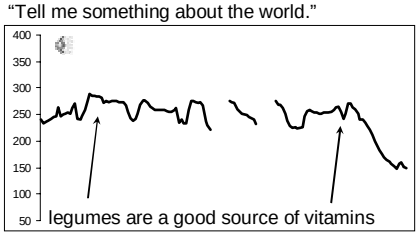
[I know that many natural foods are healthy, but ...]



WH-questions typically have falling contours, like statements.

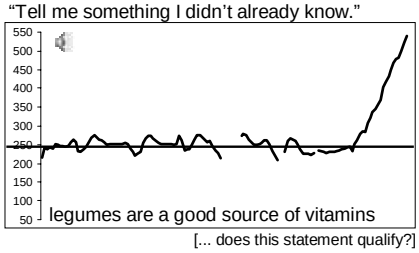
Broad focus

"Tell me something about the world."



Rising statements

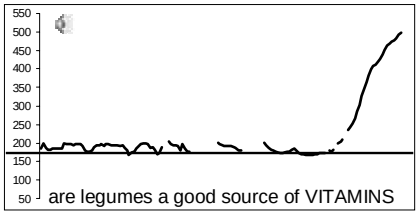
"Tell me something I didn't already know."



[... does this statement qualify?]

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

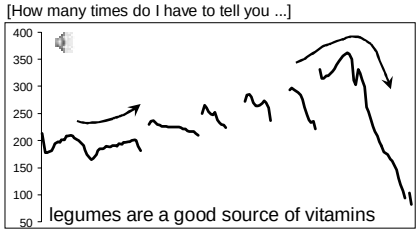
Yes-No question



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

‘Surprise-redundancy’ tune

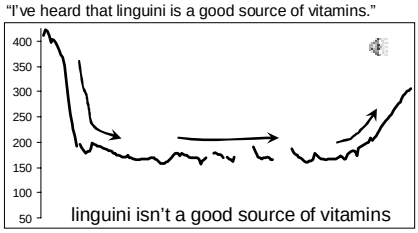
[How many times do I have to tell you ...]



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.

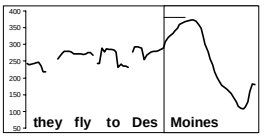
Intonation makes the difference

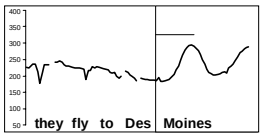
A: What types of foods are a good source of vitamins?
B1: Legumes are a good source of vitamins.
B2: Legumes are a good source of vitamins.

A: I’d like to fly to Davenport, Iowa on TWA.
B: TWA doesn’t fly there ...
B1: They fly to Des Moines.
B2: They fly to Des Moines.

Alignment differences cue “assertion” vs. “suggestion”

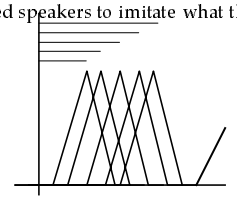
A: I’d like to fly to Davenport, Iowa on TWA.
B: TWA doesn’t fly there ...





Two distinct alignment categories

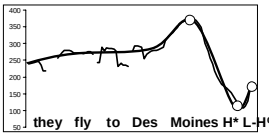
- Pierrehumbert & Steele (1989) synthesized many intonation contours with varying degrees of peak delay, and asked speakers to imitate what they heard.

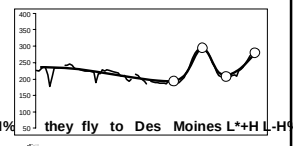


- Peak delay of speakers’ responses patterned in two categories: early (‘assertion’) and late (‘suggestion’).

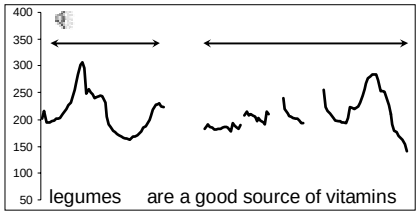
Alignment differences cue “assertion” vs. “suggestion”

A: I’d like to fly to Davenport, Iowa on TWA.
B: TWA doesn’t fly there ...



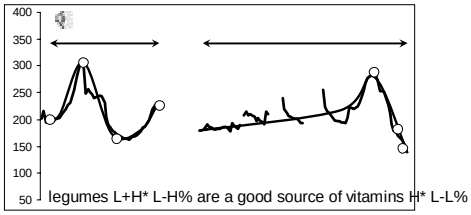


Multiple phrases



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

Multiple phrases



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

Facial conversational signals

Speakers can use non-verbal signals to highlight the interpretation of their utterances

Judy Fortin on CNN

"…far greater than any similar object ever discovered."

Goal: High-level input

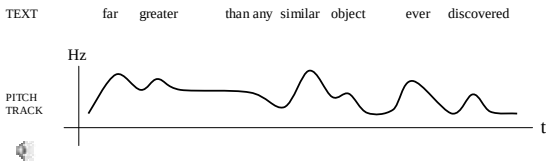
- Makes it possible to choose among a set of actions when generating utterances
- Allows facial conversational signals to combine with other behaviors

Strategy: coding

- Coding systems describe the meaningful elements of human action, i.e.
  - FACS (*Facial Action Coding System*) [Ekman & Friesen 77]
  - ToBI (*Tones and Break Indices*) [Silverman et al 92]
- We make a coding system for facial conversational displays
  - motivated by both FACS and ToBI
  - specifies format of input for animation

Intonation coding: ToBI for American English

- ToBI is designed to analyze pitch



- ToBI describes changes in pitch *qualitatively*

Intonation coding: ToBI for American English

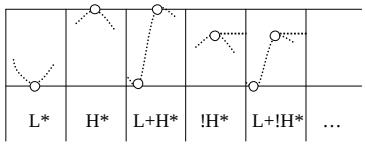
- Utterances are broken into *phrases*

  far greater than any similar object ever discovered

- Each phrase is described separately in terms of:
  - Pitch accents (*)
  - Phrase accents (-)
  - Boundary tones (%)
- When described this way, the various accents and tones have compositional meanings
  [Pierrehumbert & Hirschberg 90]

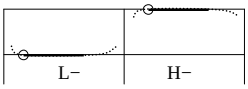
Pitch accents

Describe maxima (H) and minima (L) aligned with the stressed syllable



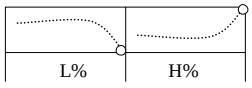
Phrase accents

- Describe pitch in phrase after last accent



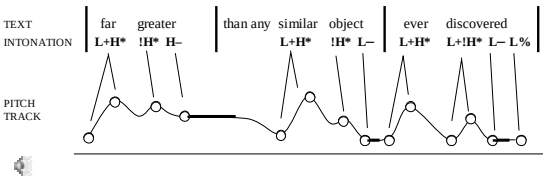
Boundary tones

- Describe pitch at end of phrase



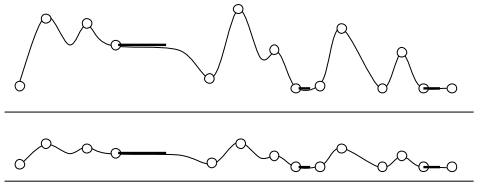
ToBI coding example

- By stringing together accents and tones, you can describe the significant changes in pitch



ToBI doesn't code quantitative information

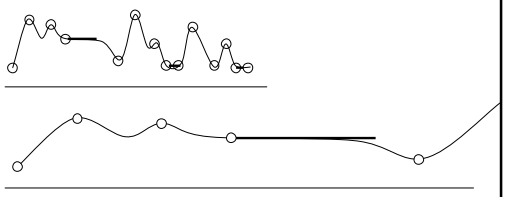
Changes in pitch range don't change ToBI coding



- listeners might perceive different levels of excitement

ToBI doesn't code quantitative information

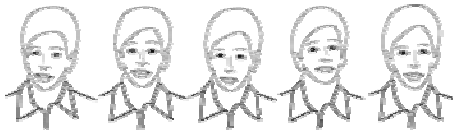
Changes in speaking rate don't change ToBI coding



- listeners might perceive differences in affect

Coding conversational facial displays

- We need to describe eyebrow and head movements (from video + audio)



- Again, we start from a continuous signal

Coding conversational facial displays

We think head and eyebrow movements work the same way as prosody

- they are synchronized with prosodic units (phrases and accents)
- they have a qualitative specification

Head and eyebrow movement coding

- We rely on the same prosodic structure
  far greater **than any similar object** **ever discovered**
- We describe movements in synchrony with
  - accents (compare Ekman's *batons*)
  - phrases (compare Ekman's *underliners*) [Ekman 79]
- Brow actions are in terms of FACS AUs **1**, **2** and **4**
  - **1+2** (neutral raise), …
- Head movements are in qualitative directions
  - **U** (up), **D** (down), **TR** (tilt right), …

Batons

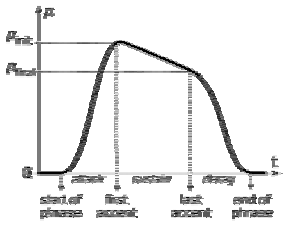
It turns out there are many *brief* motions of head or brows that *peak* on pitch accents

- like the head nods on "similar" and "ever"

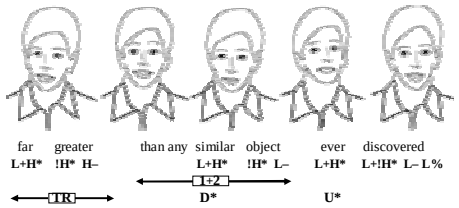
Underliners

It turns out there are also longer and more gradual motions that extend over an entire prosodic phrase

- like TR on "far greater" or **1+2** on "than any similar object"



Full coding example



Full coding example

Machine-readable form:

```
((far ((register "HL") (accent "L+H*") (jog "TR")))
 (greater ((accent "!H*") (tone "H-") (blink) (jog)))
 (than ((register "HL-H") (brow "1+2")))
 (any ())
 (similar ((accent "L+H*") (jog "D*")))
 (object ((pos nn) (tone "L-") (blink) (brow)))
 (ever ((register "L") (accent "H*") (jog "U*")))
 (discovered ((accent "L+!H*") (tone "L-L%") (blink))))
```

Capable of being rendered back

Coding results from pilot study

Working from newscaster video (Judy Fortin on CNN) and a text transcription and ToBI coding, four analysts regularly observed:

- **D*** downward nod (baton)
- **TR** tilting nod (underliner)
- **1+2** neutral brow raise (underliner)

Other speakers make different motions

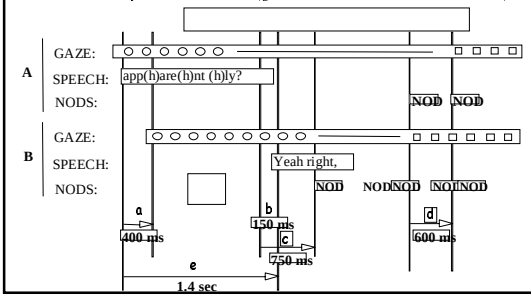
"Asteroid" example

Eye & Head Movement:

- Gaze & Head turns also mark
  - status of turn-taking
  - attention to task
  - cognitive activity

An example

- A conversation becomes increasingly synchronized (entrainment)
- Time scales vary widely -- 400 ms to 1.4 sec (multi-threadedness)
- Multi-modality follows function(gesture is there because we need it).



What is 'Gesture'?

- Pen gestures
- Command language gestures
- Articulatory gestures
- (Sign languages)
- Emblems
- Propositional gestures
- *Co-verbal gestures*

Co-verbal Gestures

- Iconics:
  - represent feature of accompanying speech.
- Metaphorics:
  - depict metaphorically a feature of speech.
- Deictics:
  - indicate places in space
- Beats:
  - Occur for emphasis, with turntaking, etc.

Gesture in Human Conversation is

- Integrated into production of discourse at temporal, semantic, pragmatic/discourse level
- Used in understanding to build representation of communicative intent semantically and pragmatically

Temporal Integration

- Iconics & Metaphorics consist of 3 phases: preparation, stroke, retraction.
- Deictics & Beats collapse prep & stroke phases.
- Most effortful part of gesture (the stroke) co-occurs with stressed part of speech (pitch accent).
- Gestural phrase co-occurs with semantically parallel unit.
- Holds ensure that gesture is synchronized.

Semantic Integration

- Gestures convey complementary information to speech
- Gestures are sometimes redundant -- for the purposes of the discourse
- When gestures are non-redundant, semantic features are distributed across speech and gesture.

Discourse Integration

- Gestures mark
  - information as new and otherwise important
- add
  - point of view
  - perspective or spatialization of people & events
  - speaker's beliefs about discourse

Examples

- "The road runner [zips] over him"
  + redundant speed (fast gst)
- "The road runner [zips over him]"
  + path/manner (zig-zag gst)
- "The road runner goes [pschew]"
  + non-redundant manner & path (fly up gst)

Example

"A kid can make a device that will have real behavior (...) that two of them [will interact] in a - to - to do a [dance together]"

Does Conversational Embodiment Matter?

2 user studies: communicative task & collaborative task

- When Gandalf exhibited conversational smarts (and did **not** exhibit emotions), he was judged to be
  - more credible
  - more helpful
  - more collaborative

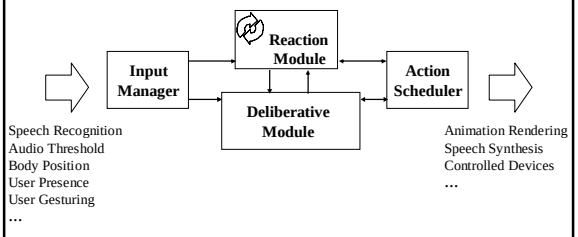
3rd Embodied Conversational Agent: Case Study, in detail

- Support Multi-Modal Input and Graphical Output
- Operate in Real-Time
- Process Propositional and Interactional Information
- Use Conversational Functions (over modalities)
- Be Modular and Extensible
- Actually generate verbal and non-verbal output

REA, Experiment in Virtual Realty

- Shows clients through houses
- Engages in small talk
- Answers questions about particular houses
- Obeys requests to show houses/rooms
- Asks questions about client's housing needs

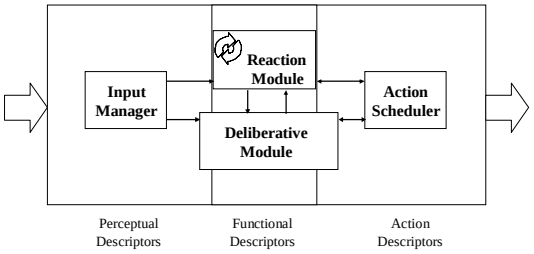
Rea Architecture



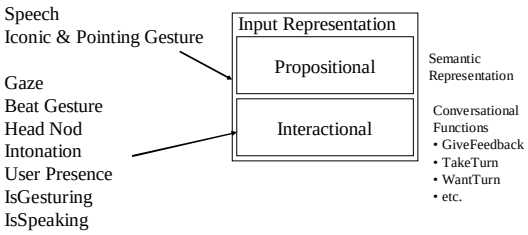
Rea I/O Components

- Inputs:
  - Stereo Vision: Stive
    - User present/absent, location, gesturing
    - Azarbayejani, A., Wren, C. and Pentland A. Real-time 3-D tracking of the human body. In *Proceedings of IMAGE'COM 96*, (Bordeaux, France, May 1996).
  - Audio threshold (speaking/paused/idle)
  - ASR: IBM ViaVoice
- Outputs:
  - Animation: SGI OpenGL
  - TTS: Microsoft Whisper

Conversational Functions



Interactional and Propositional Information



REA

See Cassell, et al., 1999 (Proceedings of CH