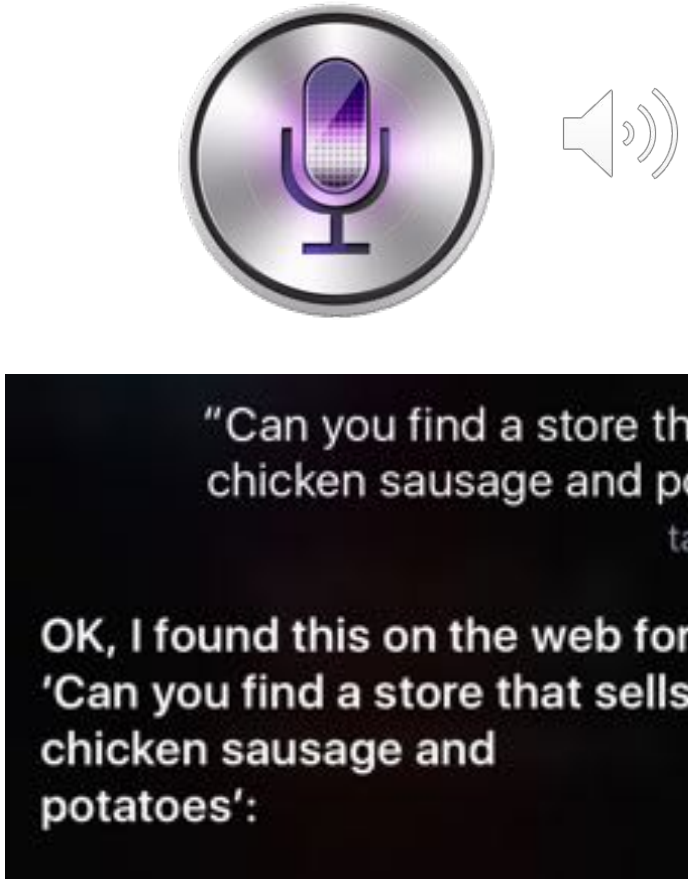
# Intonation for Conversational AI

By Joseph Tyler

# Siri example:




How can we improve?

# Siri example:

Joseph:

What’s the difference?

# Siri example:

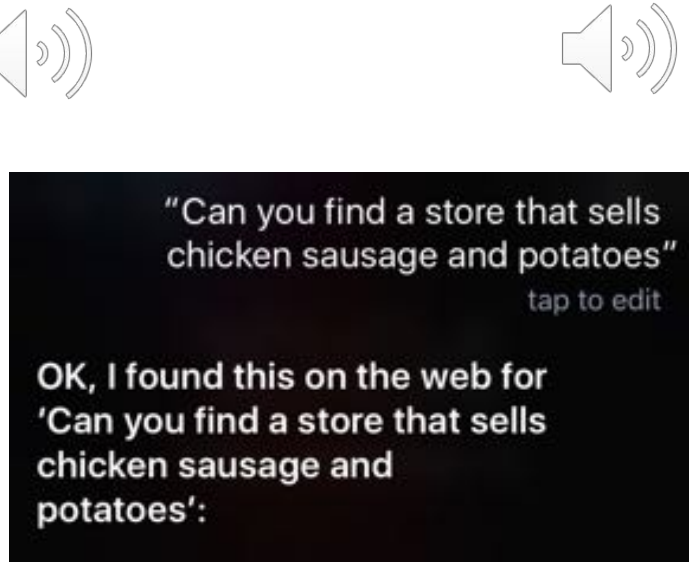
Joseph:

What's the difference?

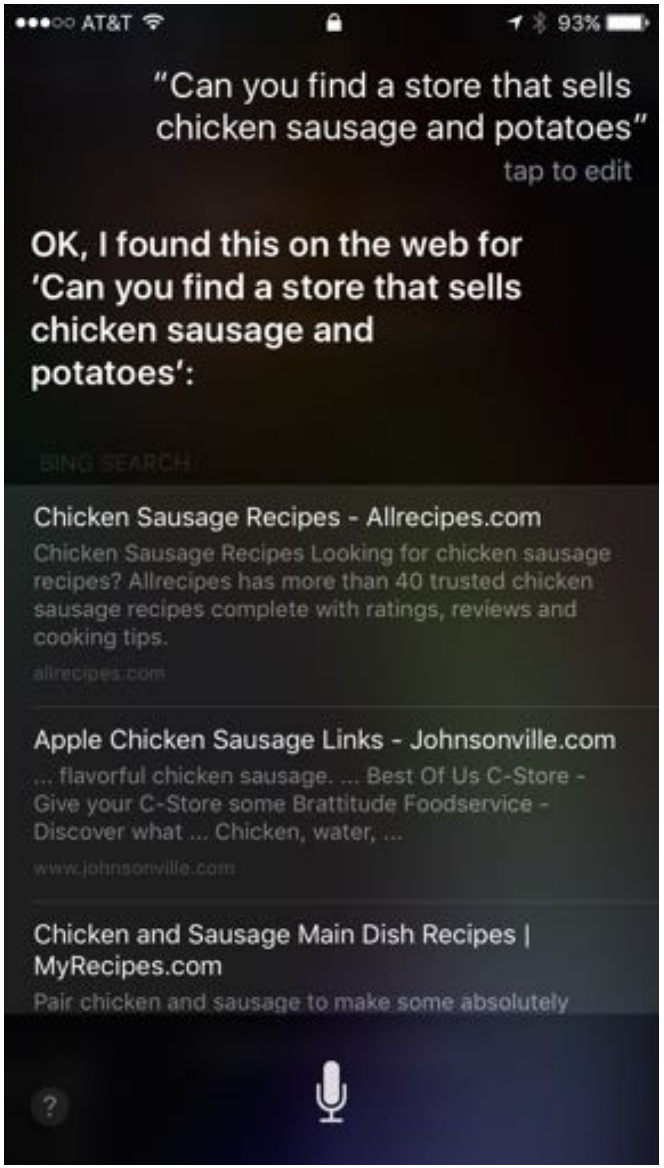

# Siri example:

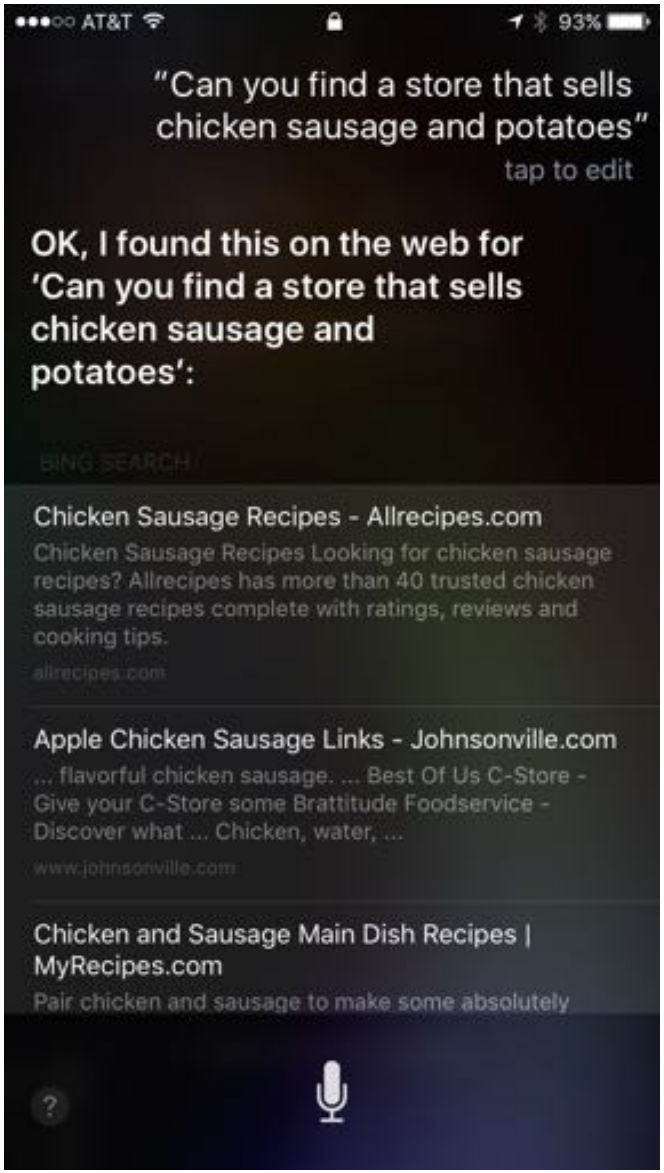
Joseph:

Siri:



Same Results!



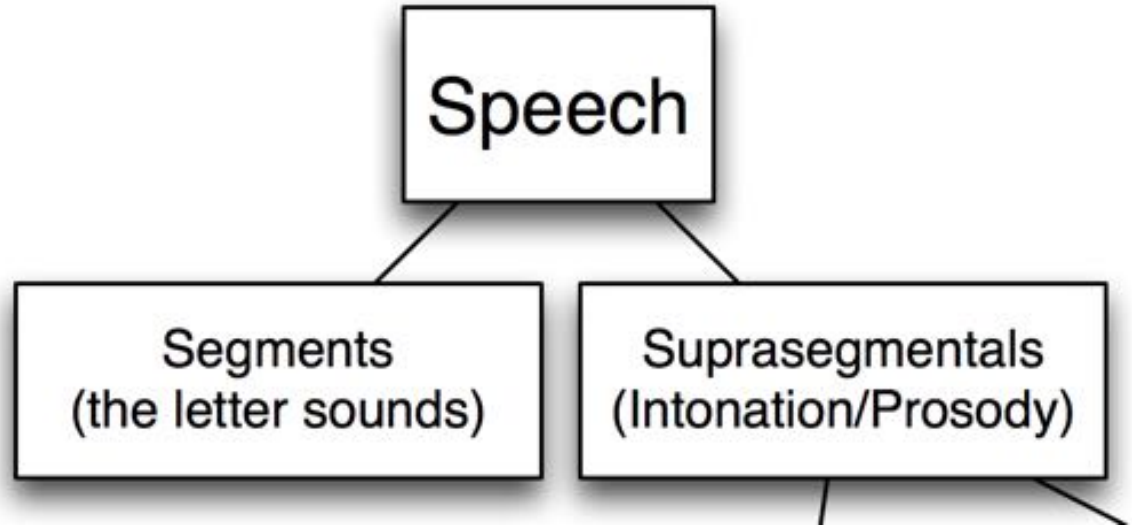



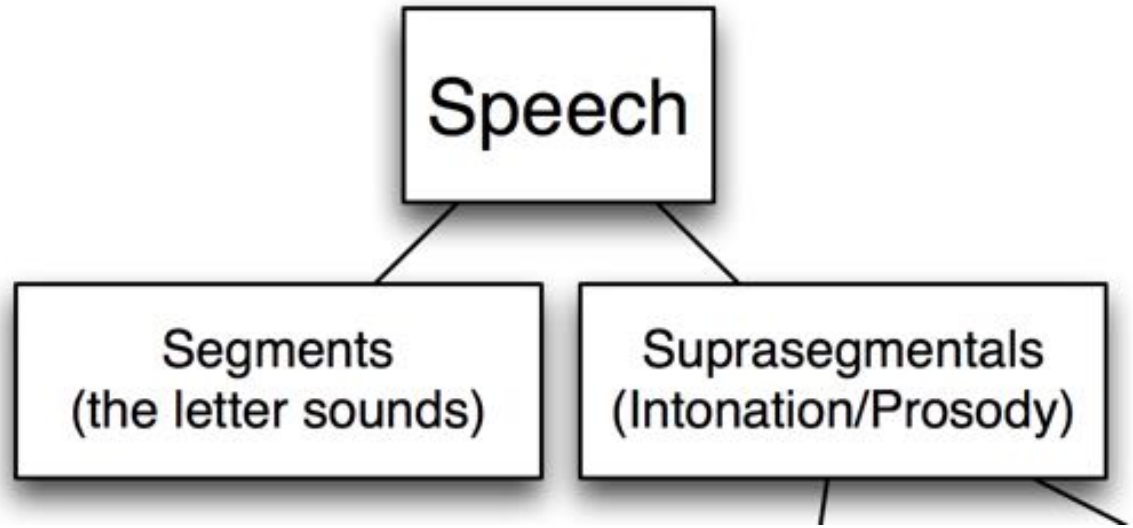
# Brainstorm Siri (Alexa/OkGoogle) frustrations

- What are your biggest frustrations with Siri/Alexa/etc…?
  - Where do they currently fail?
  - What do you wish theydid better?
  - What seems unnatural?

- Introduce yourself to people around you and share.

# Defining speech, intonation, and prosody

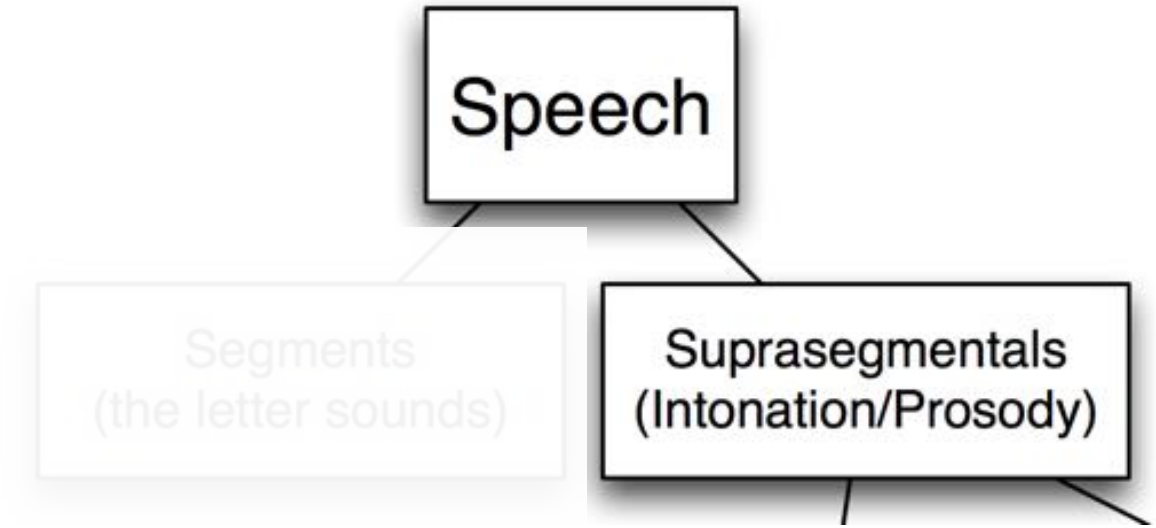



Speech
Segments
(the letter sounds)
Suprasegmentals
(Intonation/Prosody)

E.g. [p] vs [b]

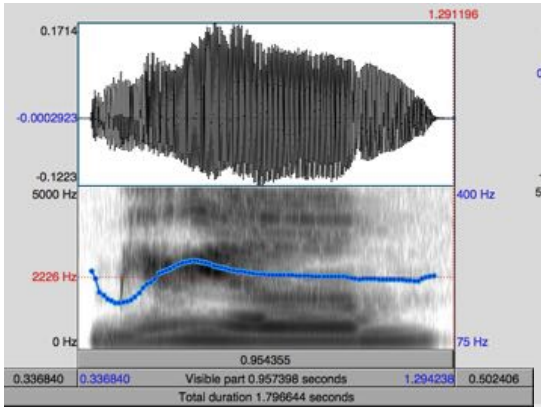
vs.


PAT
PAT

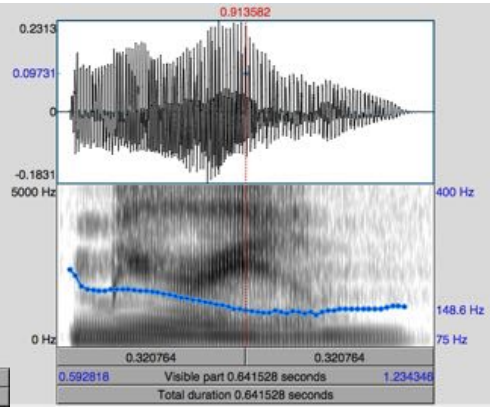



Spanning multiple segments!

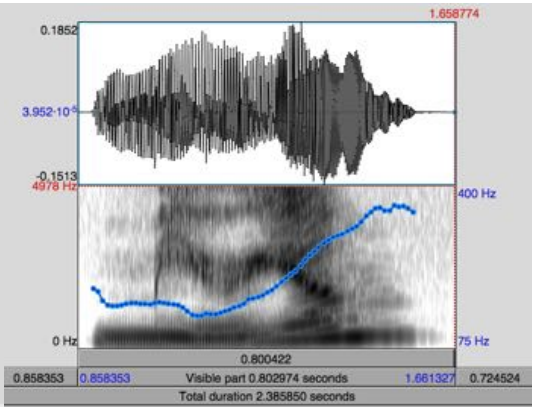
1

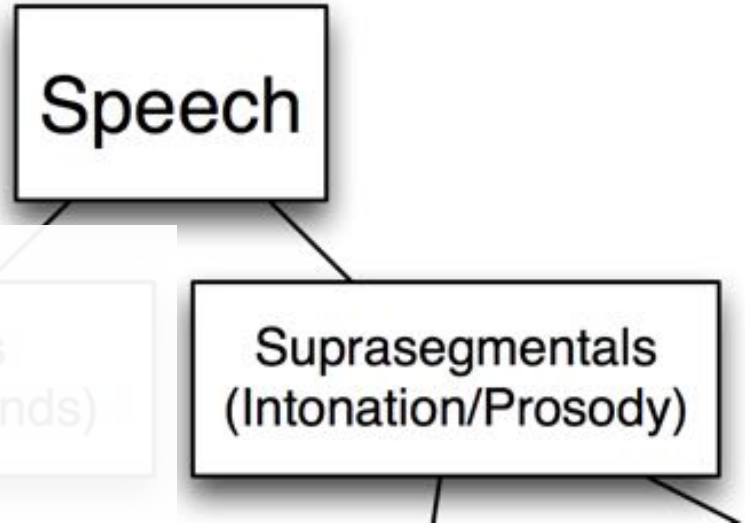



2



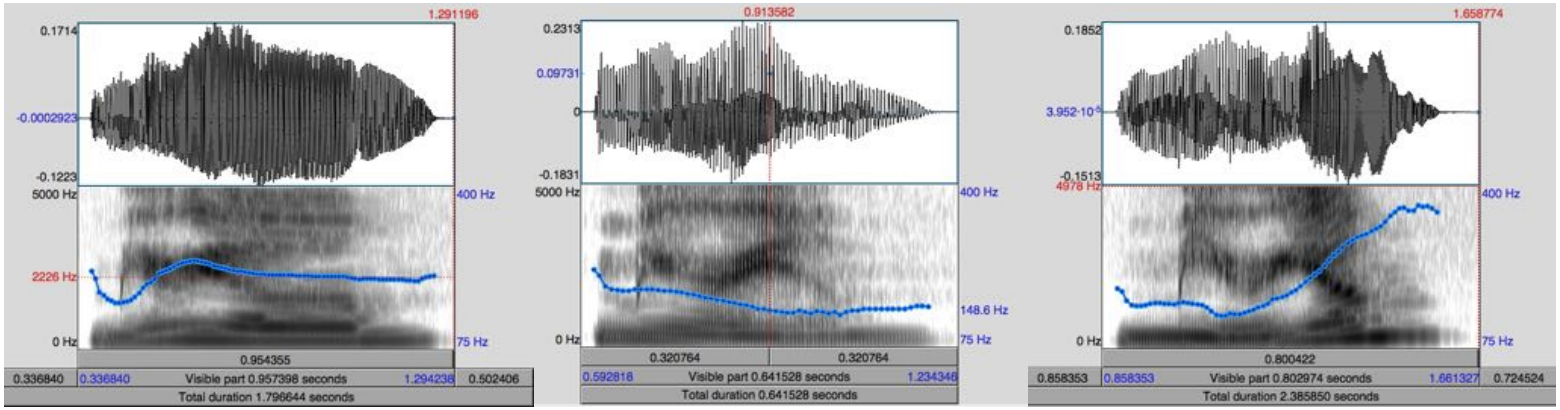

3



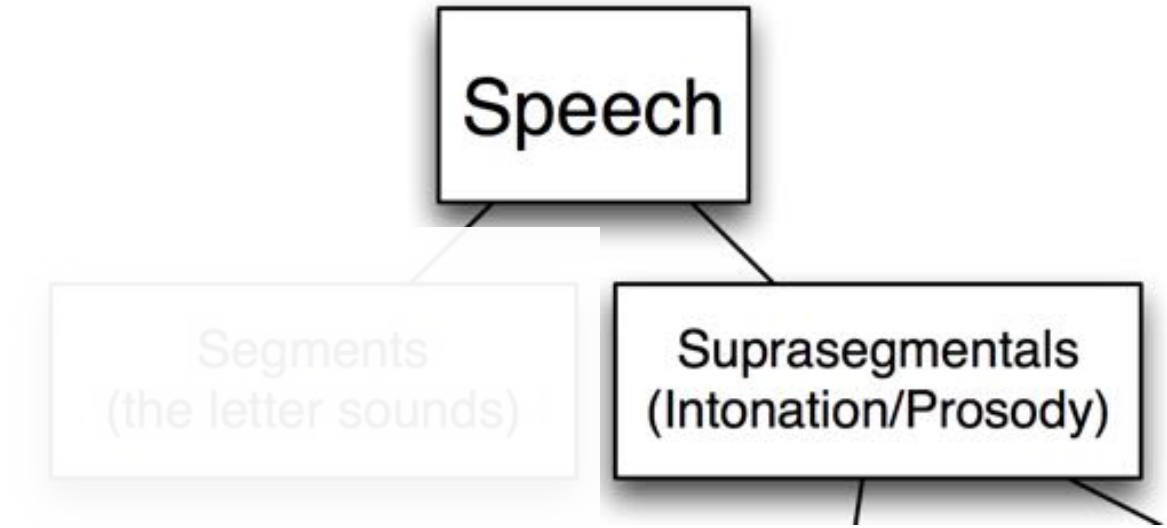



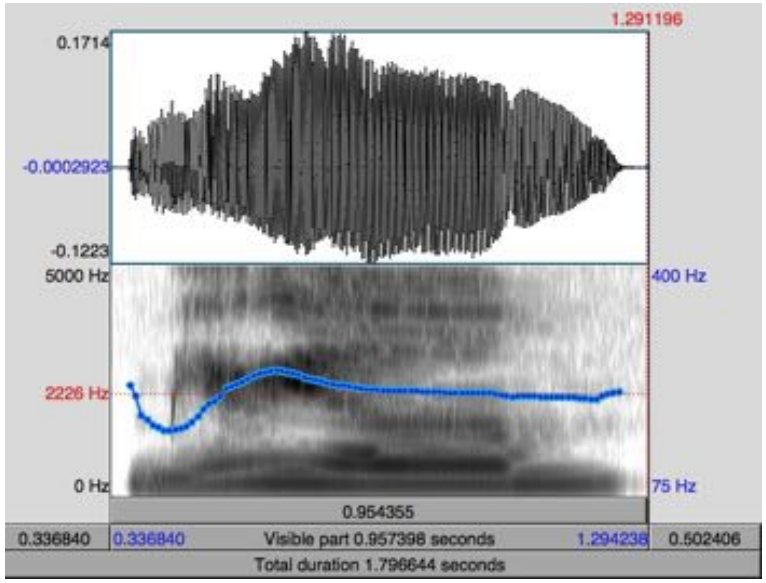

Spanning multiple segments!

1 2 3



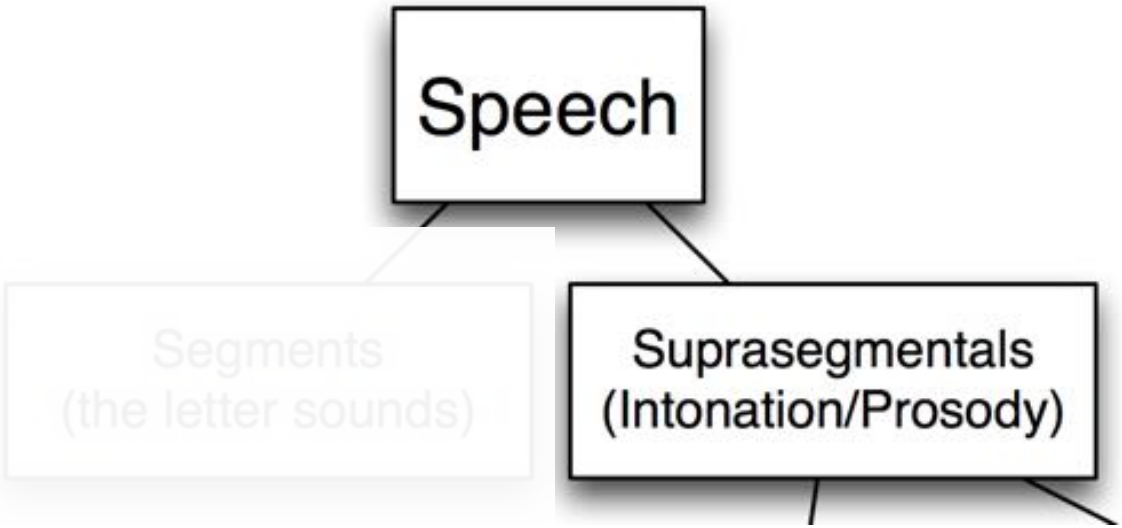

3 2 1



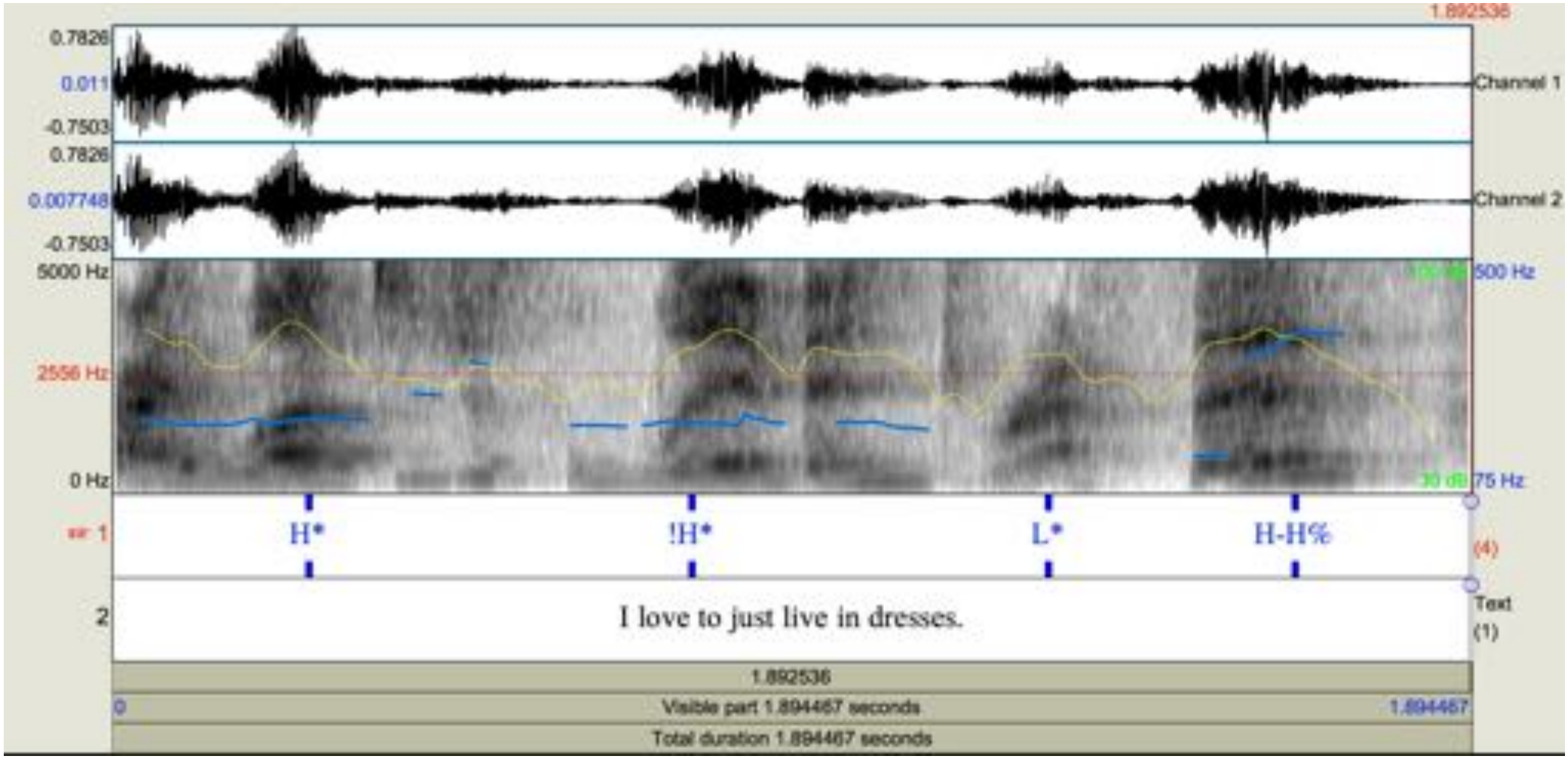

## Acoustic features of prosody:

- Intonation and tone (f0, measured in Hz, perceived as pitch)
- Intensity/amplitude (perceived as loudness)
- Pausing
- Speech rate
- Voice quality (voicing, whispering, breathy voice, creaky/vocal fry, more)
- Many more…




Structural categories of prosody:

- Boundary tones (H and L)
- Pitch accents (H and L)


1.892536
0.7826
0.011
-0.7503
Channel 1
0.7826
0.007748
-0.7503
Channel 2
5000 Hz
500 Hz
2556 Hz
0 Hz
75 Hz
H*
!H*
L*
H-H%
(4)
I love to just live in dresses.
Text
(1)
1.892536
Visible part 1.894467 seconds
1.894467
Total duration 1.894467 seconds

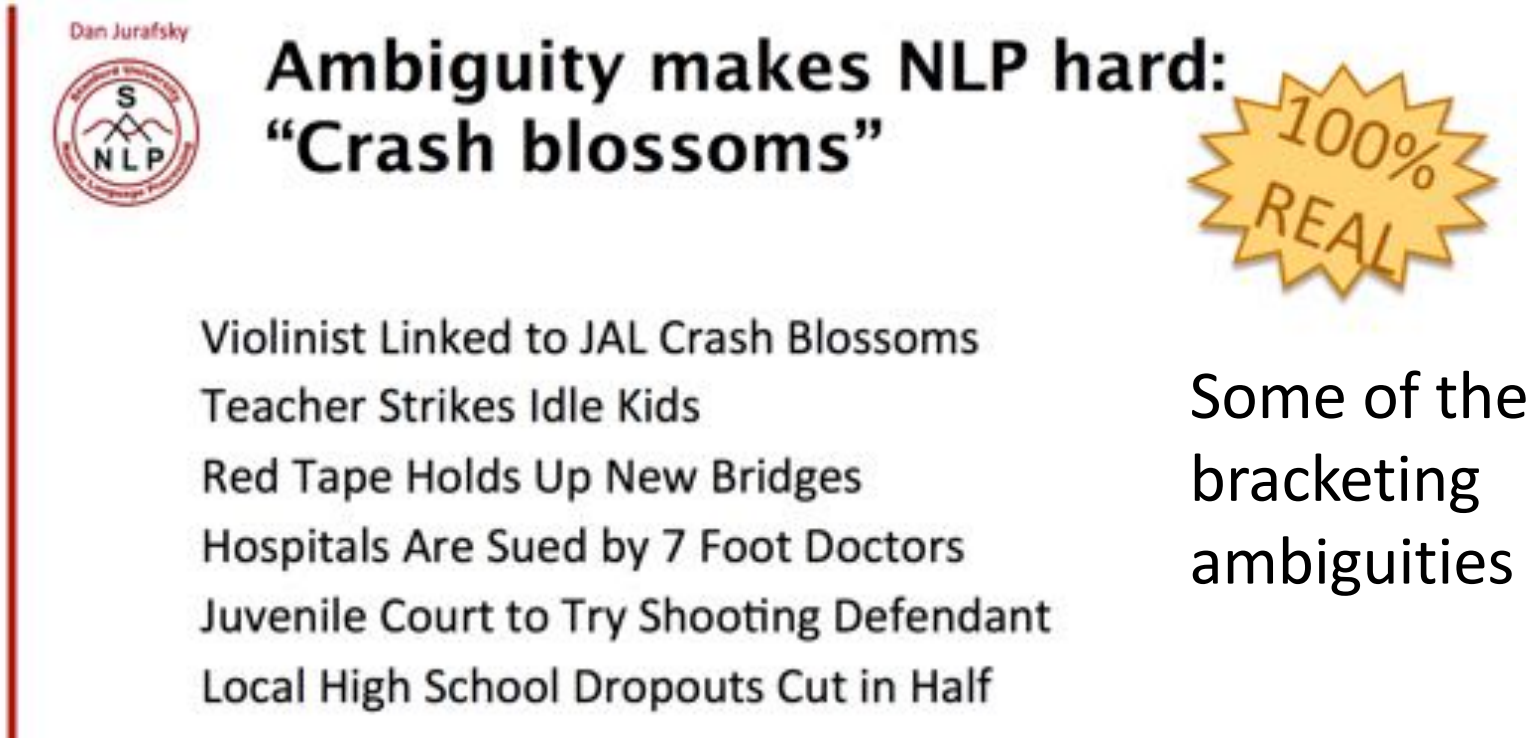
# What can prosody do (in English)?

- Prosody can help with:
  - Disambiguation (within and between sentences)
  - Making language easier to understand
  - Improving naturalness

- Bracketing (indicate word groupings)
- Speech acts (e.g. Q vs A, informing vs. reminding)
- Focus and sets of alternatives (John/JOHN didn't cheat on the test).
- Social meaning (stereotypes, social groups, dialects)

# Prosody and Syntactic Bracketing

From Jurafsky's NLP Course:



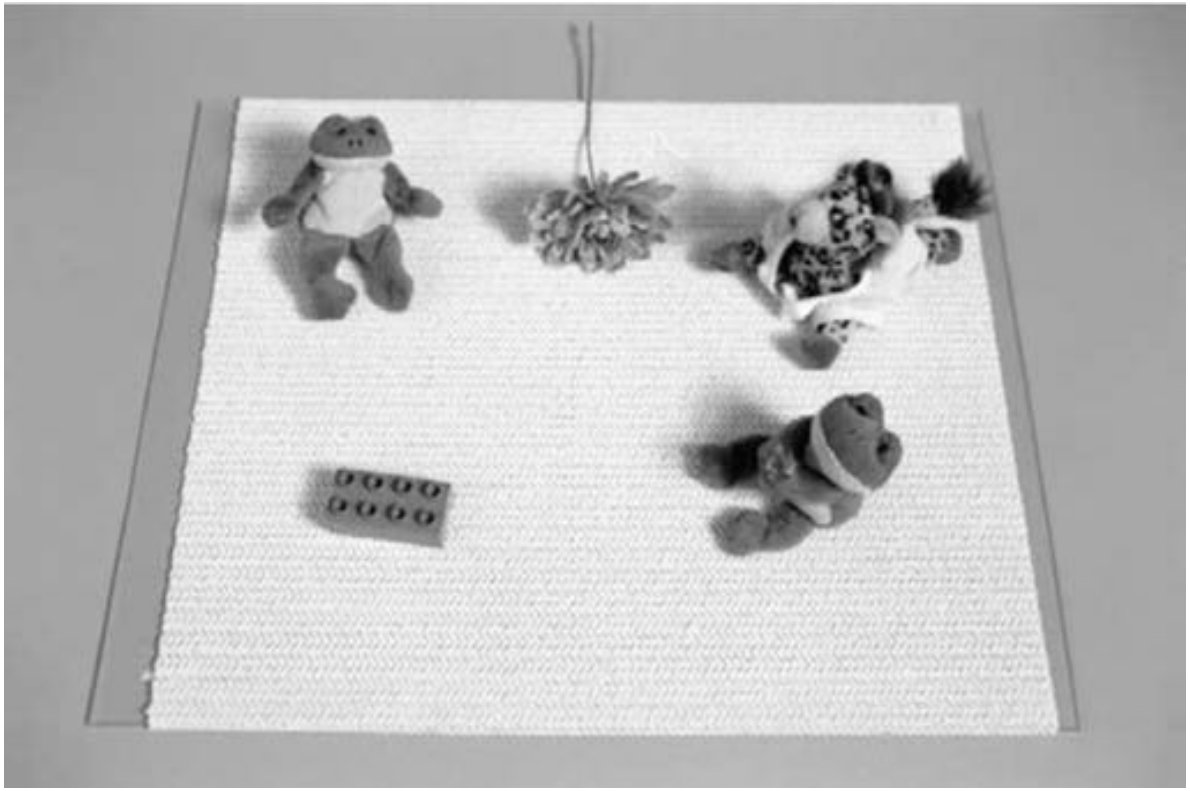

Some of these are bracketing ambiguities

# Prosody and Syntactic Bracketing

- Linguistic ambiguity: one set of words, 2+ meanings
  - Prosodic boundaries can disambiguate
  - Great place to see prosody's work
- "Old men and women"
  - Old men, and women
  - Old (men and women)
- "Paula phoned a friend from Alabama."
  - Paula phoned (a friend from Alabama).
  - Paula phoned a friend (from Alabama).

# Prosody and Syntactic Bracketing

- Tap the frog with a flower.
  - Tap (the frog with a flower).
  - Tap the frog (with a flower).



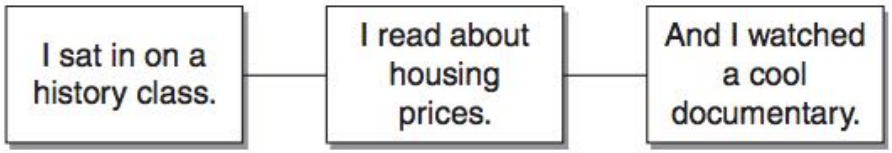
# Prosody and Discourse Bracketing

- Sally says: "I sat in on a history class. I read about housing prices. And I watched a cool documentary."

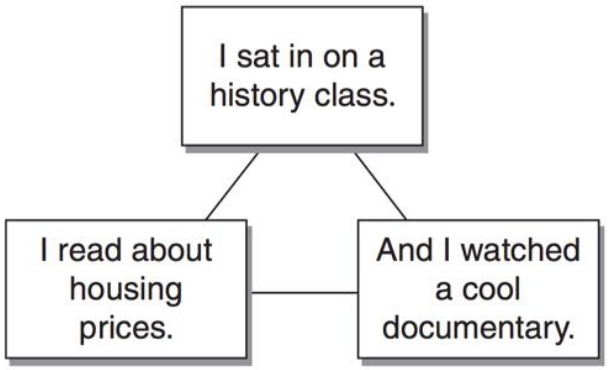
Did Sally mean that she read about housing prices and watched a cool documentary in history class?

# Prosody and Discourse Bracketing

- Sally says: "I sat in on a history class. I read about housing prices. And I watched a cool documentary."




Coord Interpretation




Subord Interpretation

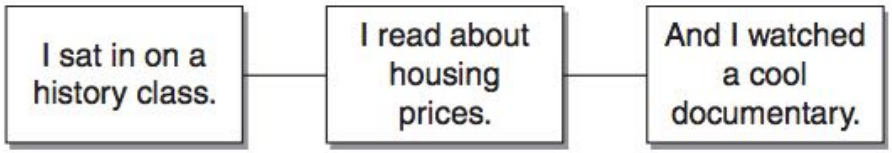
(Tyler 2014)

# Prosody and Discourse Bracketing

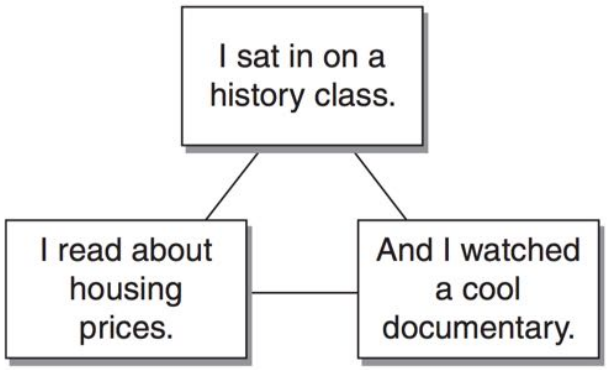
- Sally says: "I sat in on a history class. I read about housing prices. And I watched a cool documentary."

- Rising pitch: 51.9% Coord Interp
- Falling pitch: 43.2% Coord Interp

8.7% effect




Coord Interpretation

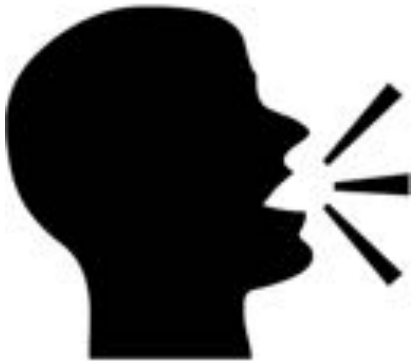



Subord Interpretation

(Tyler 2014)

# Prosodic correlates of discourse in production

# Prosodic correlates of discourse in production



BBC News segments tend to start with high onset pitch

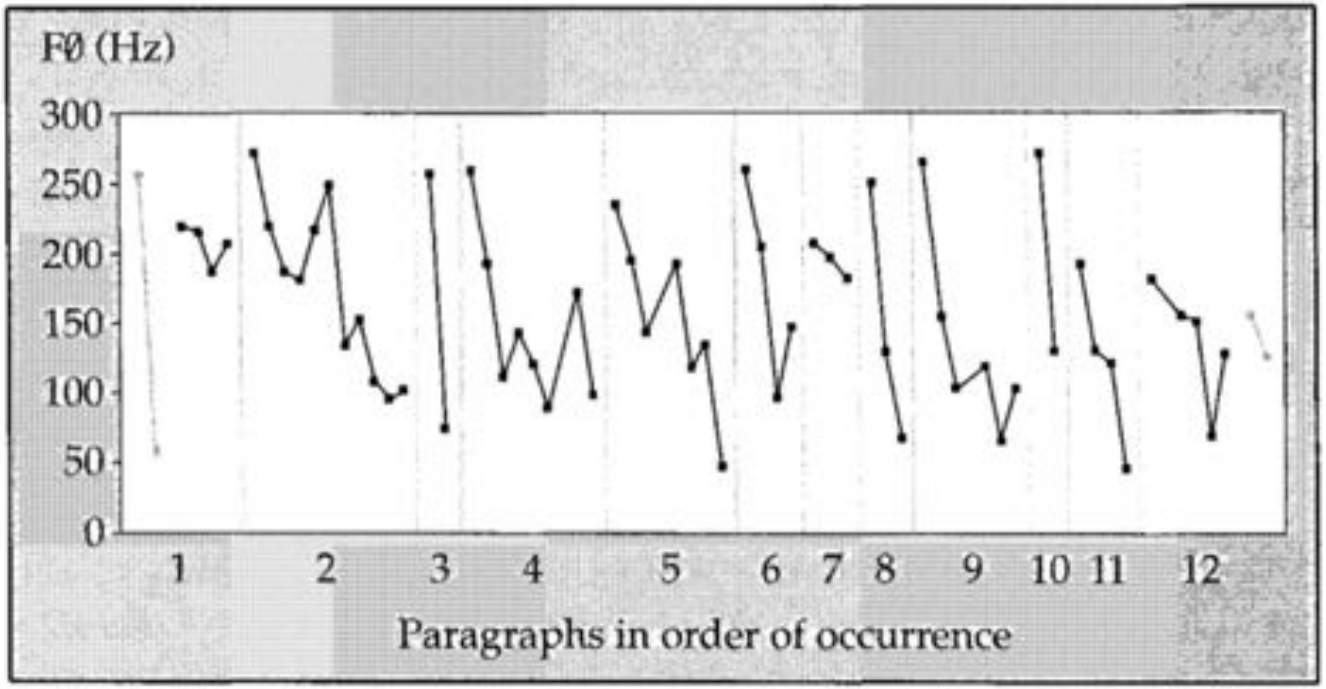



*Figure 2.3* The height of each 'sentence beginning' (major tone group onset) in a news report[4] (SEC B01). Sections 1 and 12 contain the opening and closing headlines, preceded and followed respectively by metatextual links.

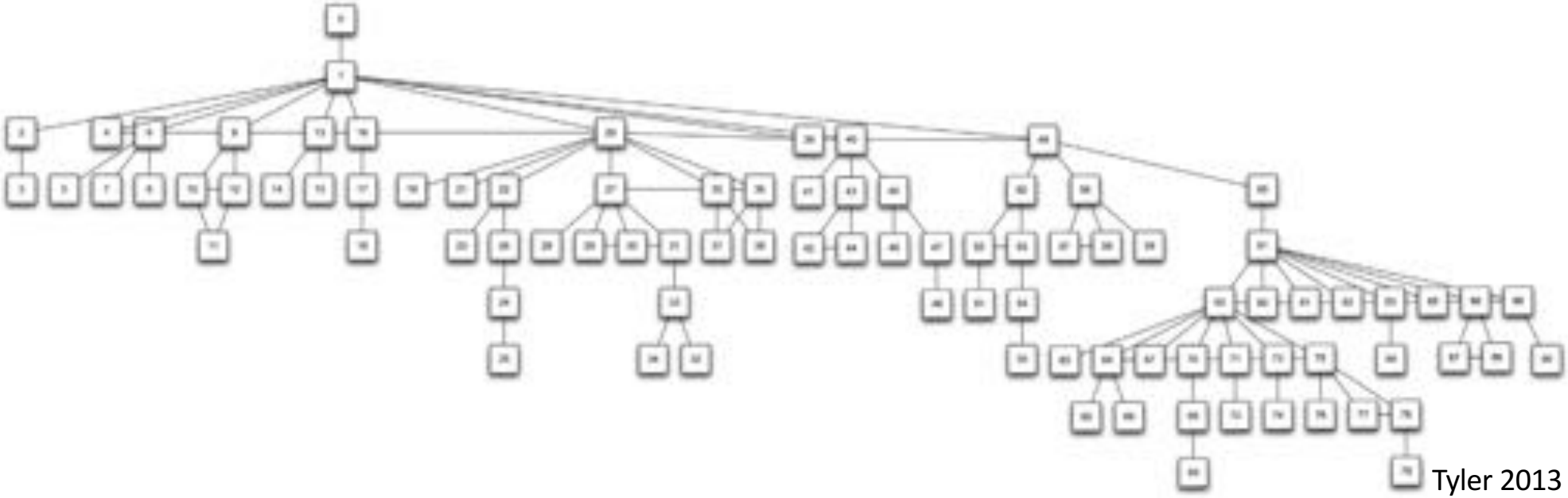
# Prosodic correlates of discourse in production

- Monologues have structure beyond the sentence. Aspects of this structure is visible in prosody.
- Reading newspaper article

Tyler 2013

# Prosodic correlates of discourse in production

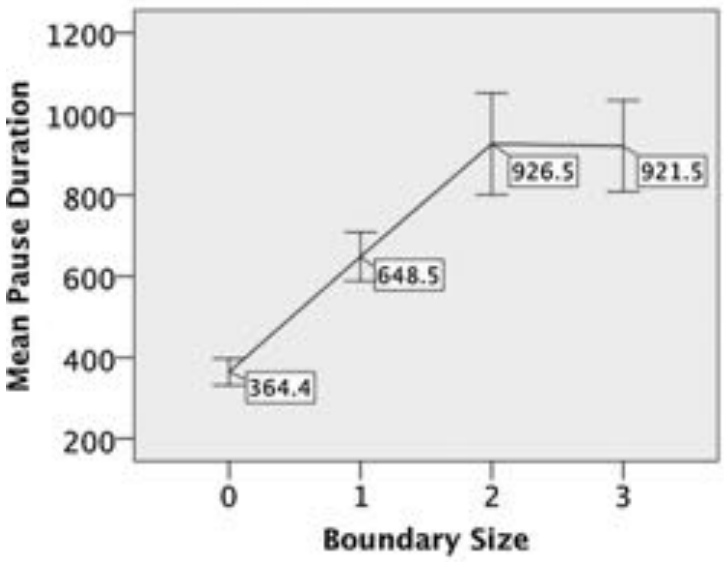
- Monologues have structure beyond the sentence. Aspects of this structure is visible in prosody.
- Reading newspaper article: segmented, relations build full structure

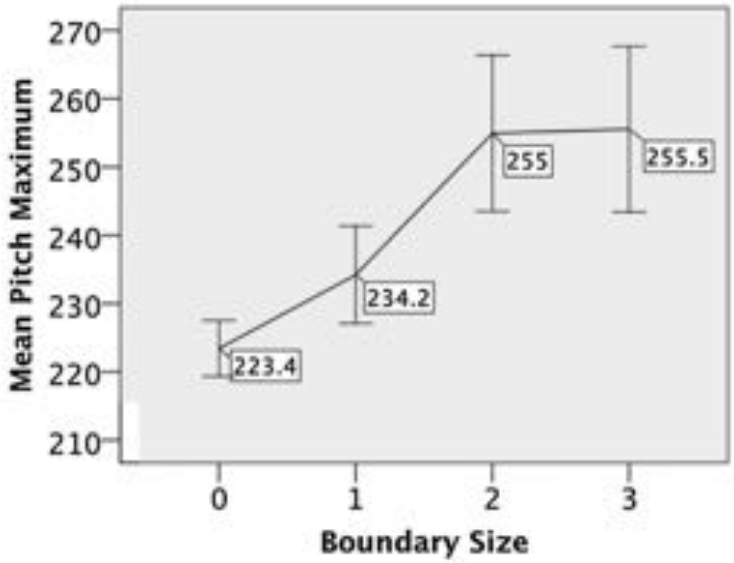


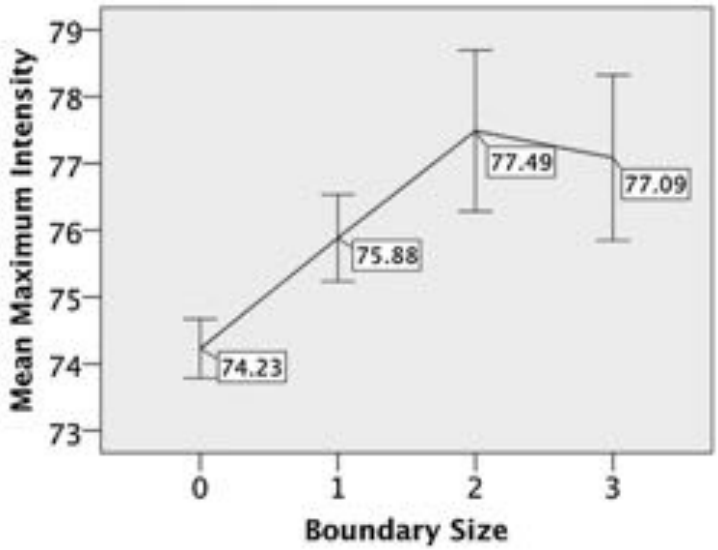
Tyler 2013

Sentences after big breaks have:

1. Higher pitch
2. Higher intensity
3. Longer preceding pauses

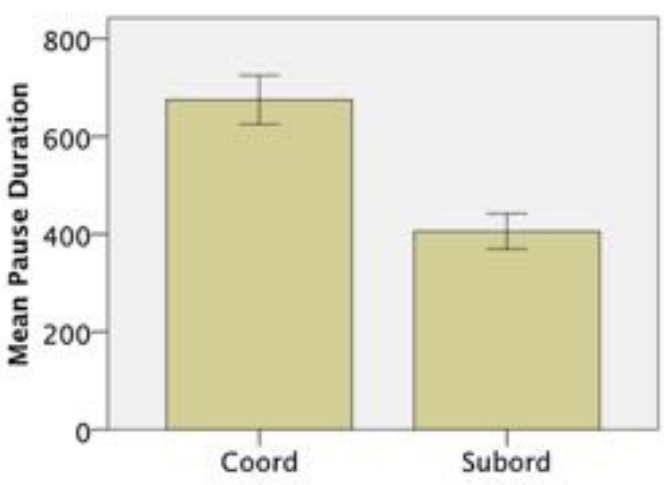


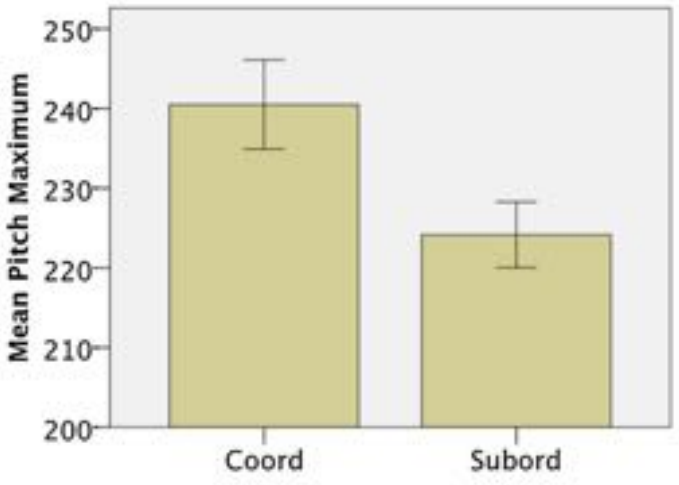



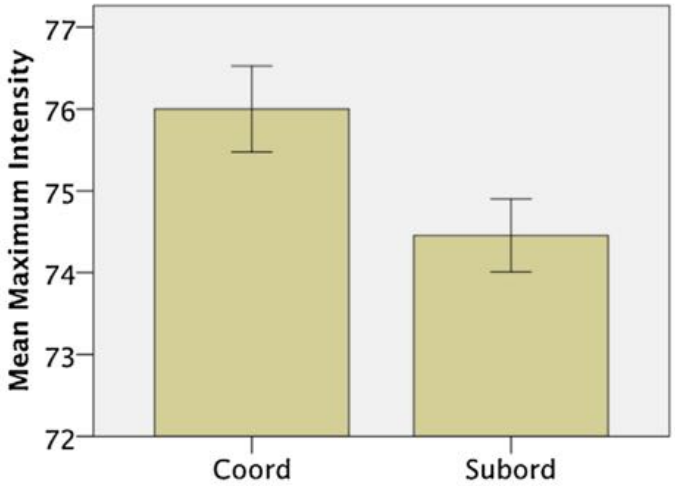




Tyler 2013

# Prosodic correlates of discourse in production



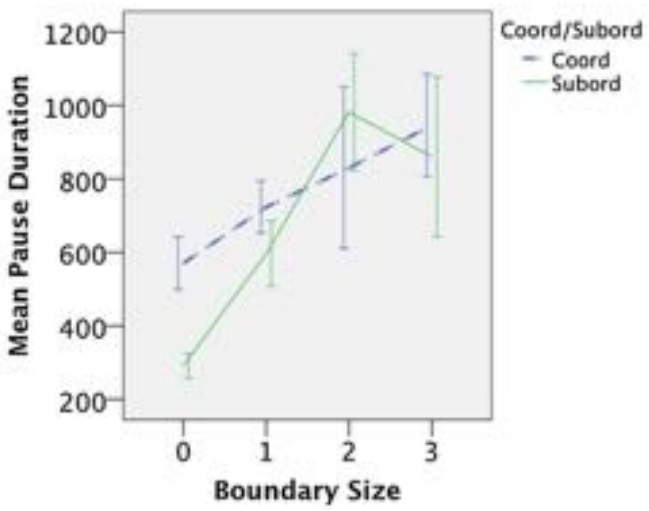



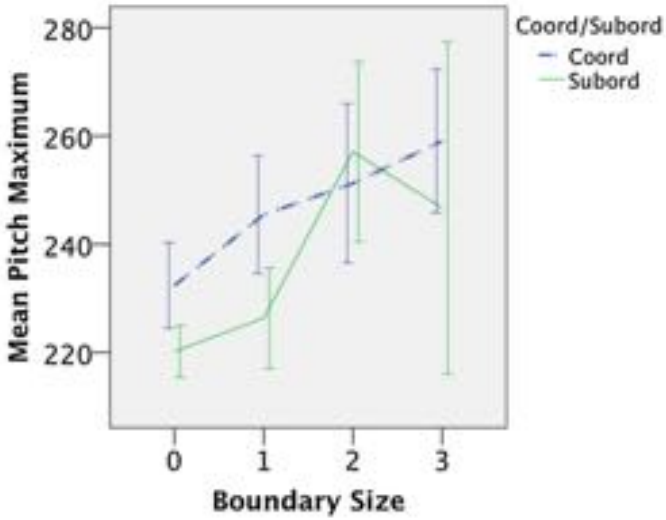

Sentences at same level of detail (not lower) have:

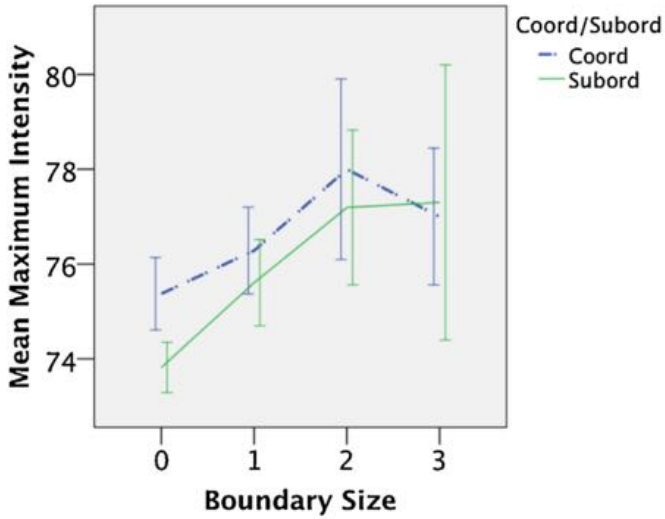
1. Higher pitch
2. Higher intensity
3. Longer preceding pauses

Tyler 2013

# Prosodic correlates of discourse in production




Local relations (Coord/Subord) only matter after small boundaries. Otherwise, boundary size washes out effect.

Tyler 2013

# Rises vs. Rise-Plateaus

- Podcast example
- On NPR all the time


http://www.chronicle.com/blogs/linguafranca/2014/11/25/the-list-lilt/

# Rises vs. Rise-Plateaus

- Mark is lost on the east side of town! He calls Stacie and asks how to get home. She says:
- ["Well, what you need to do is get on Hudson and turn on 11th avenue. Then you get on 71, get on 670, take a right on 27…”]
- When Mark’s cell phone dies. Does Stacie think that Mark knows how to get home?

# Rises vs. Rise-Plateaus

- Rise are for lists that **inform**
  - Listener doesn't know the items in the list.
- Rise-plateaus are for lists that **remind**
  - Listener does know the items in the list, or at least can finish it based on the examples you gave.

# Prosody (pitch accents) and Focus Marking

- Pitch accents!
  - JOHN didn’t eat the cake.
  - John didn’t EAT the cake.
  - John didn’t eat the CAKE.

- The meaning difference?

# Example 1: Pitch Accent and Focus for Siri?

- Ask: “Are there any movies playing near me right now?”
- Siri: “I found quite a number of movies playing near Oakland.”

vs.

“I don’t see any HORROR movies playing near Oakland, but there are quite a few others.”

# Example 1: Pitch Accent and Focus for Siri?

- Ask: “Are there any movies playing near me right now?”
- Siri: “I found quite a number of movies playing near Oakland.”

vs.

“I don’t see any HORROR movies playing near Oakland, but there are quite a few others.”

(Assuming Siri knows you like horrow movies)

It’s more personal. It’s more relevant. This discourse should sound better with this intonation. The pitch accent indicates an awareness of context.

# Example 2: Pitch Accent and Focus for Siri?

- Ask: “What’s traffic look like for getting home?”
- Siri: “The traffic to home is about average, so it should take about 1 minute.”

“Traffic to HOME is about average, but traffic to Yosemite is terrible right now.”

# Example 2: Pitch Accent and Focus for Siri?

- Ask: “What’s traffic look like for getting home?”
- Siri: “The traffic to home is about average, so it should take about 1 minute.”

“Traffic to HOME is about average, but traffic to Yosemite is terrible right now.”

What if Siri knew that, while you were asking about traffic home, you really wondered about traffic to Yosemite (maybe it is on your calendar).

# Prosody and Social Meaning

Who do you visualize?

# Prosody and Social Meaning

- Uptalk and the uptalk stereotype

# The Uptalk Epidemic

Can you say something without turning it into a question?

Published on October 6, 2010 by Hank Davis in Caveman Logic

Psychology Today

Illustration: Athena Gubbe

I've done everything I can to stop it. Whatever modest sphere of influence I have, I've used. Teaching large undergraduate classes, writing newspaper articles, giving interviews - all to no avail. I'm fighting a steamroller here or, in the more colorful language of Evolutionary Psychology, a very powerful meme. This is the meme from Hell. The kind of cultural thing Richard Dawkins must have had in mind when he introduced the term in The Selfish Gene in 1976. This was, he argued, the way culture spreads - longitudinally as a virus spreads within a population. The meme is the basic unit of culture. As Dawkins argued, memes "travel horizontally, like viruses in an epidemic." They compete with other memes and the winners take up residence in our minds, defining what our culture looks and sounds like. When Susan Blackmore wrote The Meme Machine in 1999, she didn't have the topic of this column as an example to draw upon. That's unfortunate. This one is the equivalent of a viral video. About all you can do is stand back and watch it spread. In this case, of course, you'd have to listen to it spread, since it has become part of speech.

Lan

Language Log Home About Comments policy

## Uptalk anxiety

September 7, 2008 @ 6:31 am · Filed by Mark Liberman under Language and gender, Psychology of language

« previous post | next post »

For additional background, here are some of the earlier Language Log posts that deal with related questions:

This is, like, such total crap? (5/15/2005)
Uptalk uptick (12/15/2005)
Angry Rises (2/11/2006)
Further thoughts on "the Affect" (3/22/2006)
Uptalk is not HRT (3/28/2006)
Poem in the key of what (10/9/2006)
Satirical cartoon uptalk is not HRT either (11/14/2006)
Intonation contours and polonium poisoning (12/16/2006)
Uptalk anxiety (9/7/2008)
The phonetics of uptalk (9/13/2008)
Word (in)constancy (9/16/2008)

# Prosody and Social Meaning

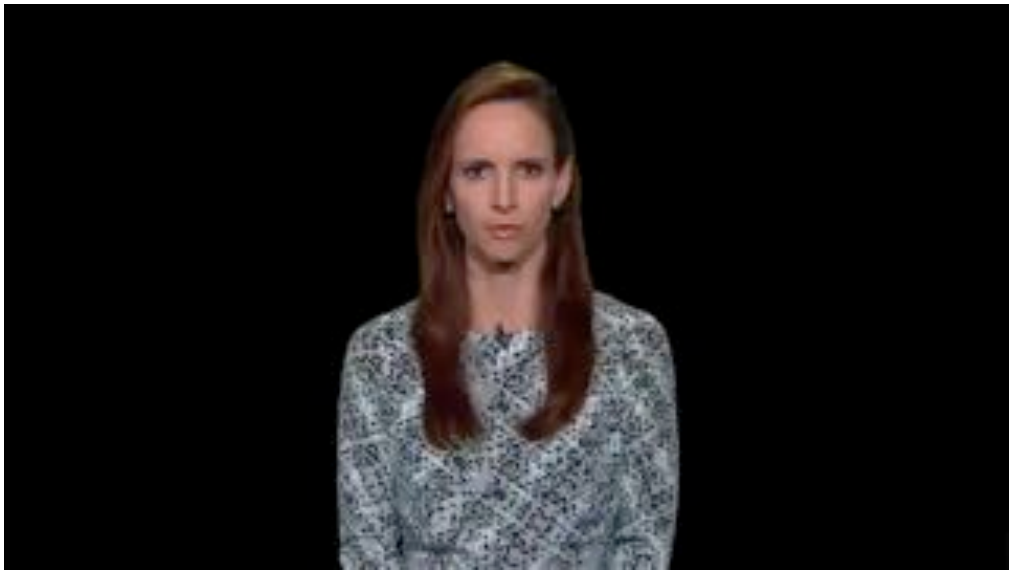
- The stereotype might not equal reality, meanings are more varied than the popular discussion of uptalk suggests.
  - Positive (emphasis, excitement, normal, happy)
  - Negative (dumb, like a tic)
  - Linguistic (unfinished, conditional)
  - Social (young, female, urban, California)

  - Correlation analysis of these perceptions shows two clusters:
    - Unfinished (“continuation rise”, holding the floor)
    - Everything else (social, emotional, regional)

(Tyler 2015)

# Prosody and Social Meaning

- Vocal fry (aka creaky voice)
- https://www.youtube.com/watch?v=YEqVgtLQ7qM



# Prosody and Social Meaning

- Vocal fry (aka creaky voice)
- Men do it too, but they tend not to be criticized for it:

# Prosody and Social Meaning

- Dialect differences
- Black English tends to have more frequent pitch accents

**Sandra Bland: Talking While Black**

August 15, 2015 @ 5:41 pm · Filed by Mark Liberman under Language and politics, Language and society

« previous post | next post »

Below is a guest post by Nicole Holliday, Rachel Burdin, and Joseph Tyler:

---

Sandra Bland's traffic stop and the tragic series of events that occurred afterwards have been the subject of many recent think pieces, but few authors have examined why the initial traffic stop went wrong in the first place. The most obvious explanation might be simple racial profiling, which almost certainly played a role, but the dash cam video of the event also shows an interaction that escalated at an alarmingly rapid pace. The

# Prosody and Social Meaning

- Dialect differences
- Jewish English has lots of low-high pitch accents (L+H*), distinct listing intonation

Why Linguists are Fascinated by the American Jewish Accent

The linguistic field of prosody, the story of melody, pitch, and other hard-to-study verbal traits, is suddenly hot.

(Burdin 2016)

# Summary:


- Prosodic Boundaries == linguistic bracketing
- Speech acts (Q vs. A, Informing vs. Reminding)
- Pitch accents == sets of alternatives
- Social meaning
  - Construct characters, convey attitudes, define groups
- Lots more I didn't have time for!

- Questions?
- Where is the low hanging fruit for Siri/Alexa?