The Music of Speech

Time and Tune

Rhythms and Melodies

Dafydd Gibbon

Bielefeld University, Germany

Mannheim, 16 November 2020

The Music of Speech – Speech Prosody

Accompaniment to song

Accompaniment to locutions

Prosody

Ancient Greek **προσῳδία** (prosōidía)
πρός (prós, “to”) + ᾠδή (ōidḗ, “song”).
song sung to music
pronunciation of syllable

Speech prosody

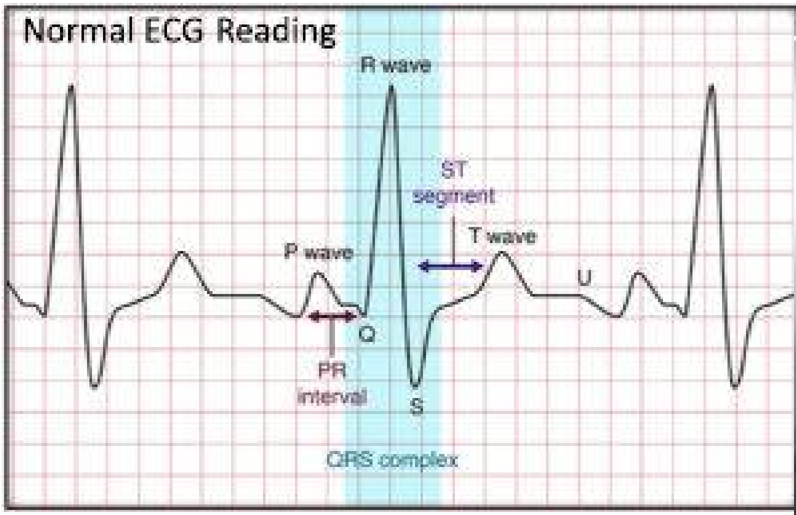
Rhythm – Time
tempo, rhythm
durations of sounds, syllables, words, phrases

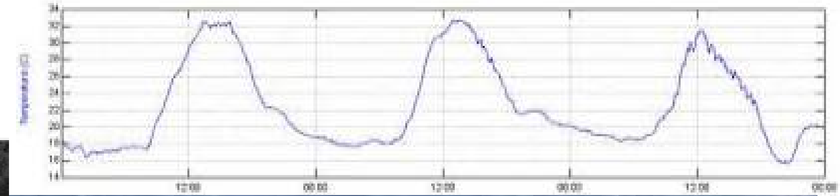
Melody – Tune
lexical prosody: pitch accents, tones in tone languages
phrasal prosody: marks grammatical structure
discourse prosody: marks argumentation, turn-taking, ...

Overview

- General questions:
  - Structure – syllables, words, phrases, discourse:
    - How are the accompaniment and song / locution aligned?
  - Meaning:
    - semantics: how does the accompaniment affect the meanings of words and phrases?
    - pragmatics: how does the accompaniment convey attitudes, meanings, emotions?
- Rhythms
  - Production / perception of rhythms
  - Synthesis / analysis of rhythms
- Melodies
  - Production / perception of melodies?
  - Synthesis / analysis of melodies

Rhythms


Normal ECG Reading
R wave
ST segment
T wave
P wave
U
Q
PR interval
S
QRS complex



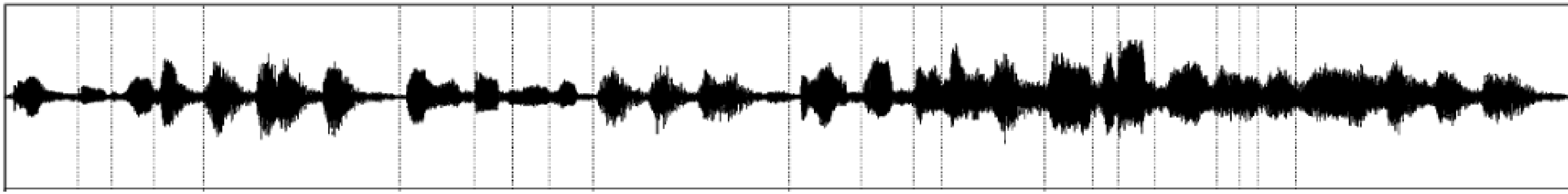

RHYTHM OF THE DANCE
20 YEARS


Music and Lyrics by George Gershwin and Ira Gershwin
Brightly
C Dm G7 C
I got rhy - thm, I got
I got dai - sies in green
Dm G7 C F Fm
mu - sic. I got my man. Who could
pas - tures. I got my man. Who could
C G7 C E7
ask for an - y - thing more?
ask for an - y - thing more? Old man

The rhythm of a song

Rhythms of speech + music

Music: tempo = allegro, time signature = 4/4, style = ‘brightly’



8 bars

8 x 4 = 32 notes

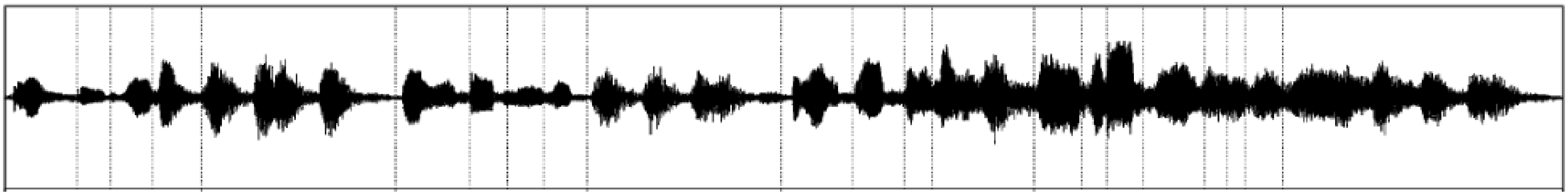
13.32 / 32 = 0.416 seconds per note

***rhythm frequency* = 1 / 0.416 = 2.4 Hz**

Tempo = 2.4 x 60 = 144 beats per minute: allegro

Rhythms of speech + music

Music: tempo = allegro, time signature = 4/4, style = ‘brightly’



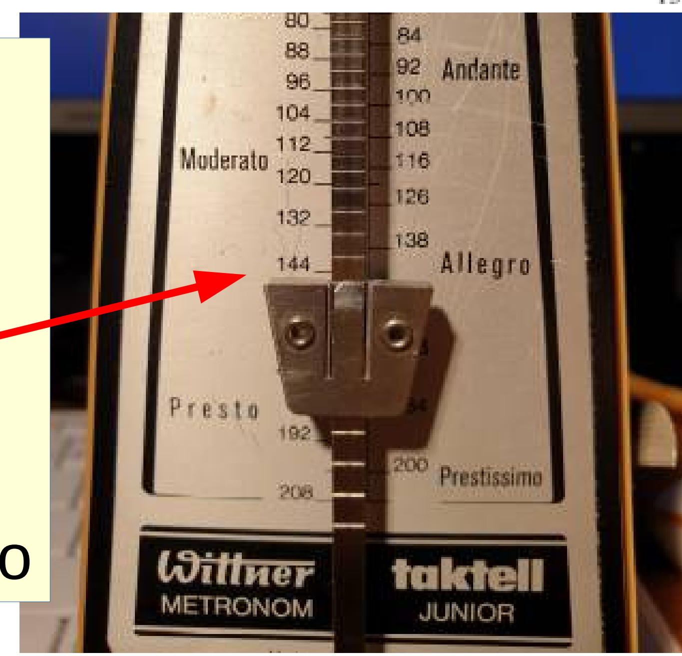
8 bars

8 x 4 = 32 notes

13.32 / 32 = 0.416 seconds per note

rhythm frequency = 1 / 0.416 = 2.4 Hz

Tempo = 2.4 x 60 = 144 beats per minute: allegro



We can perceive rhythm.

What is rhythm, physically?

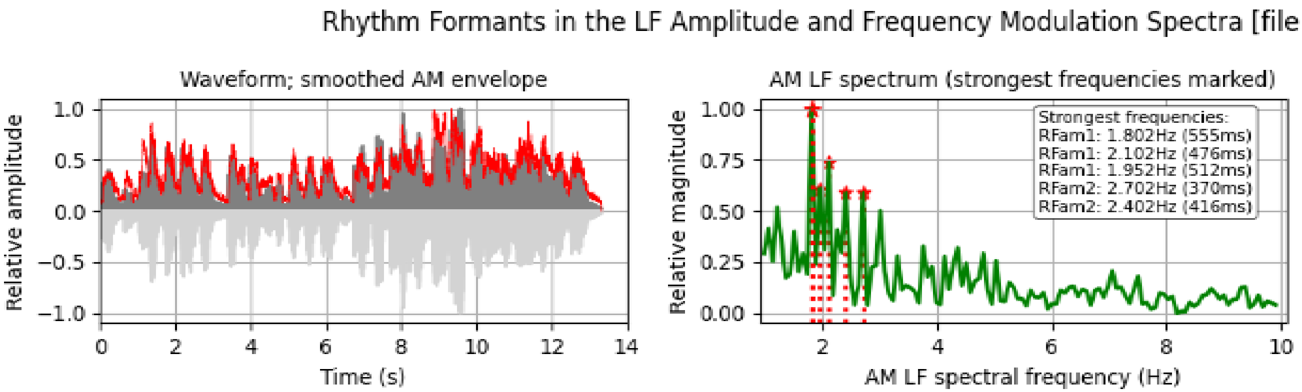
How can we detect rhythm?

First: from intuition to definition

Speech rhythms are …

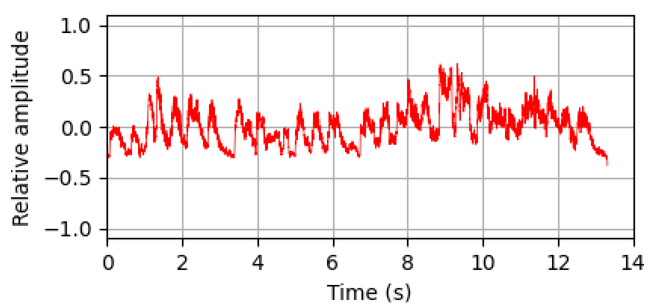
- fairly regular **oscillations** below about 10 Hz
  - which **modulate** the speech source **carrier signal**
  - and are detectable in **spectral analysis**
    - as **magnitude peaks** in the **low frequency spectrum** of
      - both the **amplitude modulation** (AM) of the speech signal, related to the syllable, word, phrase outline of the waveform
      - and the **frequency modulation** (FM) of the signal, related to fundamental frequency (F0) or perceived pitch contours of the carrier signal, related to tones, pitch accent and intonation.

Second: how do we detect speech rhythm?

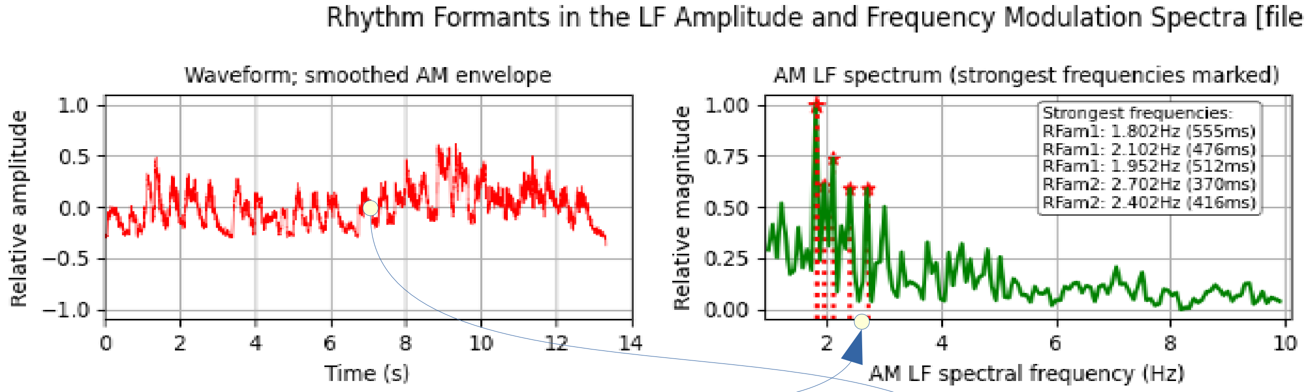


Detect and extract amplitude envelope

Convert envelope to low frequency spectrum



Second: how detect speech rhythm?



specfreqs,specmags = fft(signal,fs)

Procedure (from red to green):

- read waveform file (grey)
- separate positive amplitudes (dark grey)
- detect amplitude envelope (red)
- apply spectral analysis to envelope (green)
- and ...

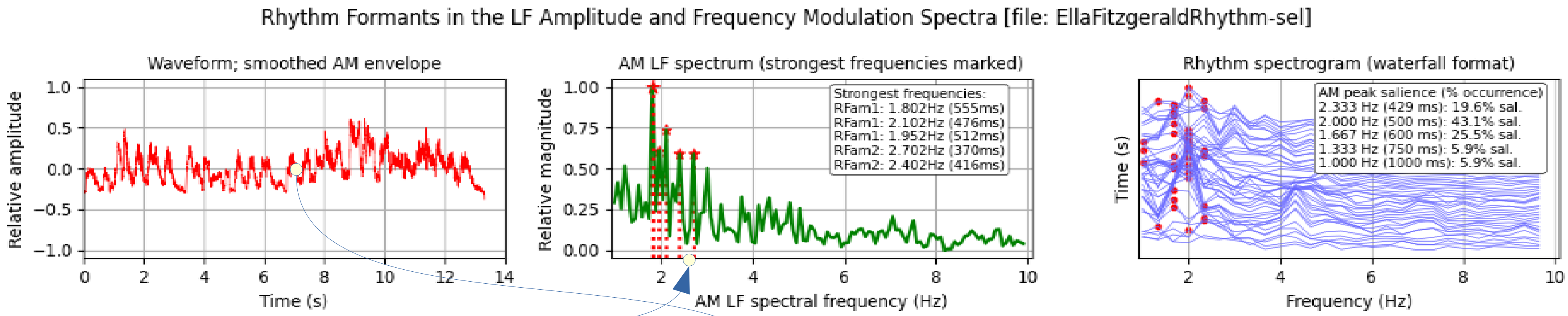
Note the difference:

frequency convention (2.4 Hz)
≠
measured frequencies
(average of top five: 2.2 Hz)

Why?

time signature ≠ beat

Third: how detect speech rhythm variation?



specfreqs,specmags = fft(signal,fs)

Procedure (from red to green):

- read waveform file (grey)
- separate positive amplitudes (dark grey)
- detect amplitude envelope (red)
- apply spectral analysis to envelope (green)
- track the rhythm changes

Conventional, abstract, static time signature in the score

vs.

Instance of physical, variable beat in the performance

Jakobson’s distinction (1960): *design* vs. *performance*

Similarly with poetry

UNSTOPPABLE

unstoppable
my words race
forward
while I'm still
dragging my feet

almost
faster than light
sound jumps
the space
between us

faster yet
your recognition
then
your smile

R. T.

Abstract <u>design</u> and physical <u>performance</u>

Roman Jakobson's distinction (*Linguistics and Poetics,* 1960):

*<u>design</u> – 'versification', foot, **metre**, line, verse, poem*

*<u>performance</u> – stress clash, enjambement, **rhythm**, ...*

Different performances of the same metre

may have a different rhythm

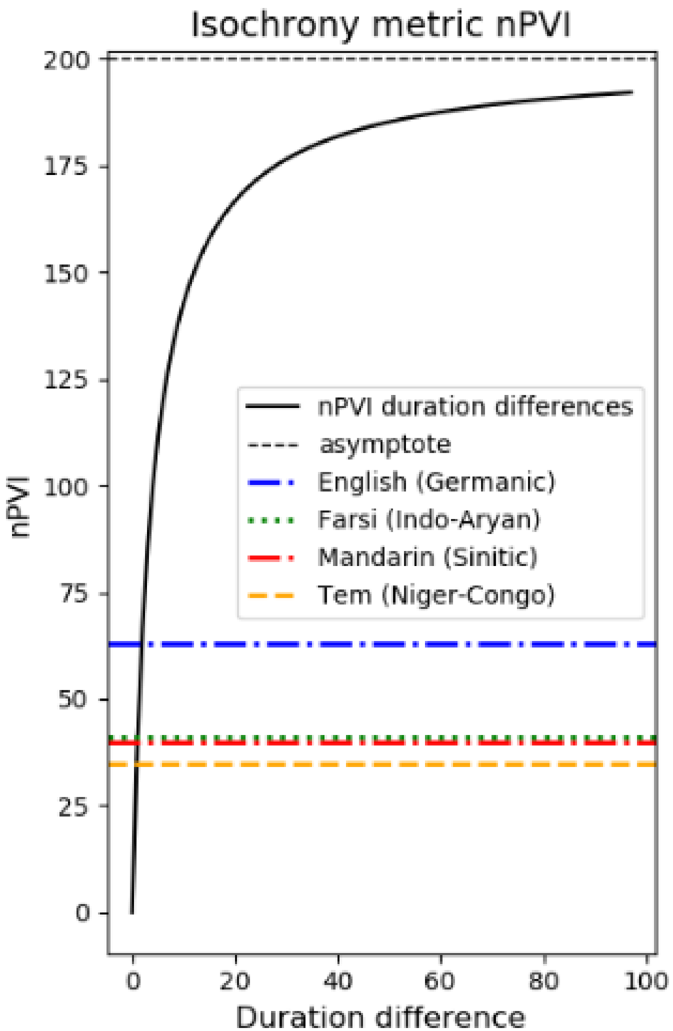
(and of course different intonation)

Partial rhythm analysis: relative duration

Annotation mining of time-stamp durations

One-dimensional becaus … single scale. The results are all comparabl … ition, but differ in detail.

For example, with the *nF* … ghbouring data in a moving window (so a kin … *Difference Function*), not between mean and d … ariations to some extent.



$Variance(x_{1...n}) = \sum$ … *ndard Deviation*)

$PIM(x_{1...n}) = \sum_{i \neq j}$ … $I_{i,j}$ are intervals in a … equence

$PFD(d_{1...n}) = \frac{\sum_{i=1}^{n}}{\sum_{j=}^{n}}$ … d is typically the … n of a *foot*

$$nPVI(d_{1...n}) = \frac{\sum_{k=1}^{k-1} \frac{|d_k - d_{k+1}|}{(d_k + d_{k+1})/2}}{n-1} \times 100$$

d refers to duration of vocalic segment, syllable or foot, typically

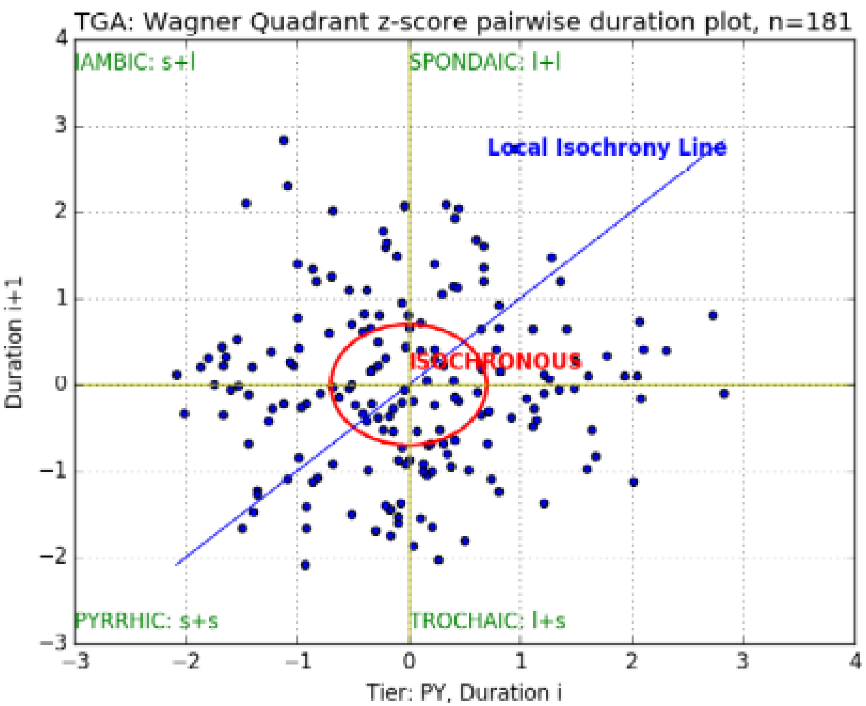
Two-dimensional annotation mining

Two-dimensional because duration relations are represented in a z-scored scatter plot, not as a single scale.

Result, visualising the scale in two dimensions:

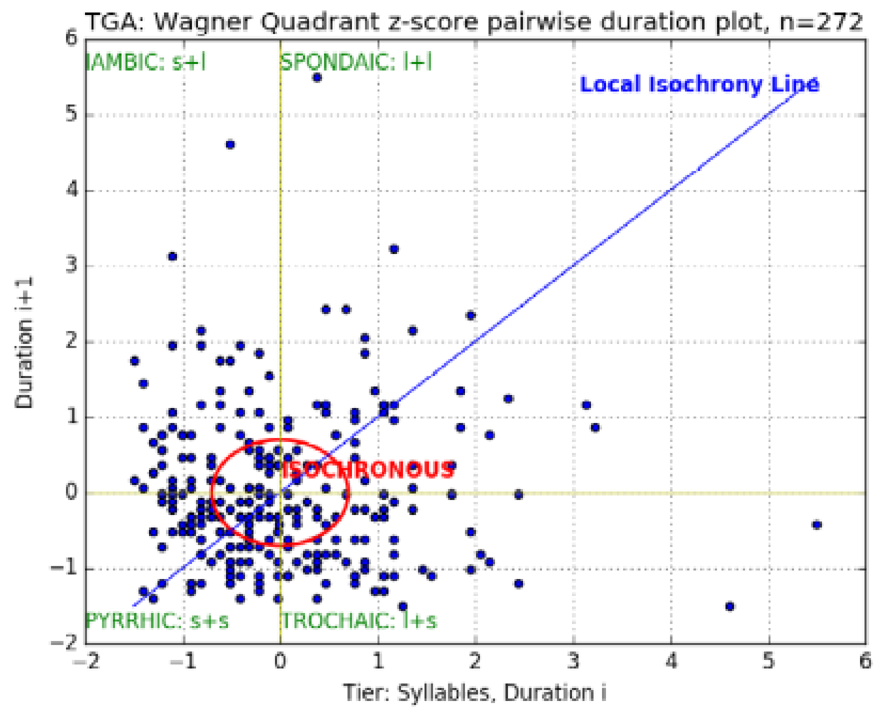
- Mandarin: means are scattered relatively evenly around the centre
- English: e.g. *count*(*short-short*) > *count*(*long-long*), not binary!

SHORT-LONG LONG-LONG



SHORT-SHORT LONG-SHORT

SHORT-LONG LONG-LONG



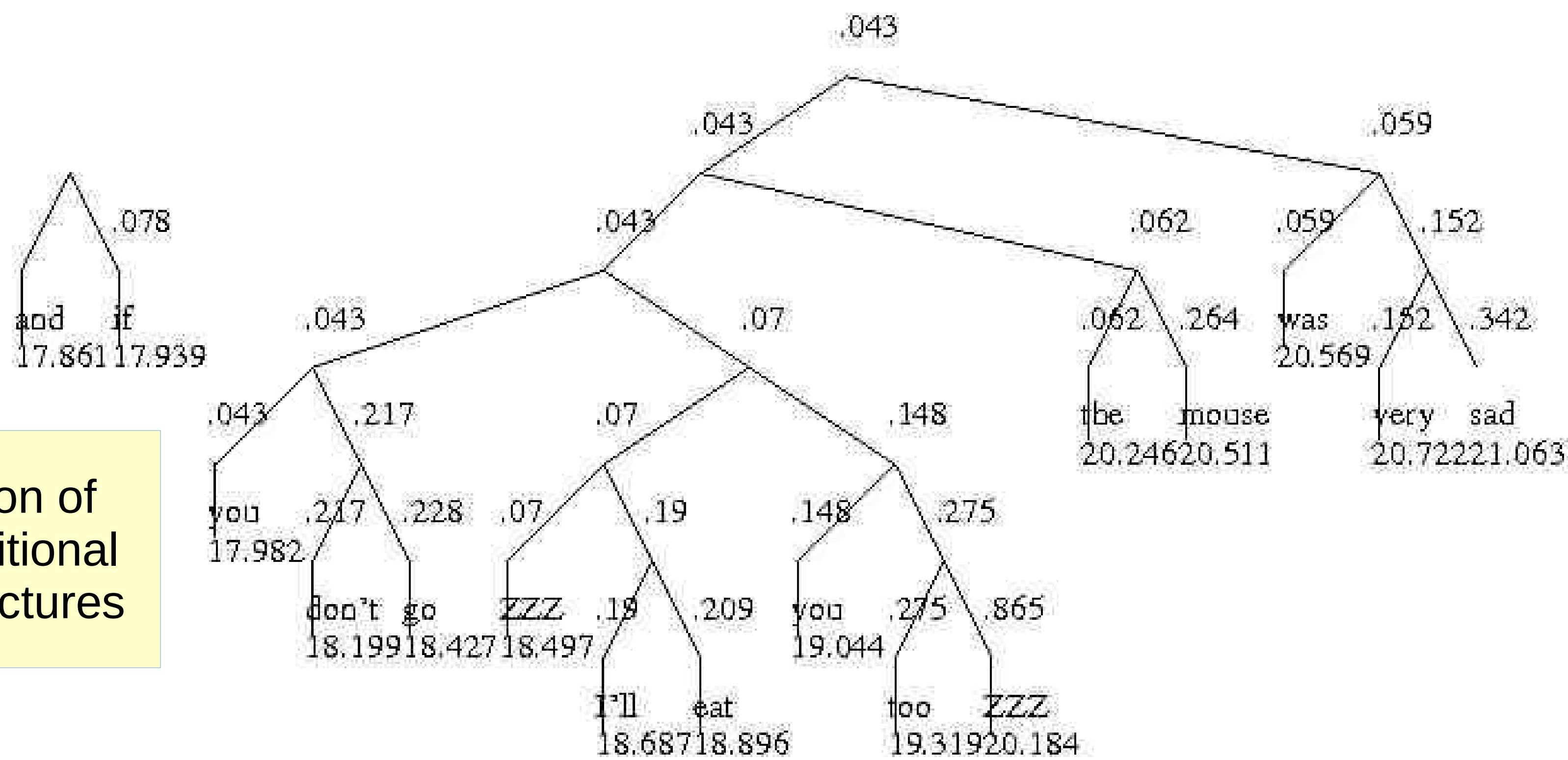
SHORT-SHORT LONG-SHORT

Wagner, Petra (2007). "Visualizing levels of rhythmic organisation." *Proc. International Congress of Phonetic Sciences, Saarbrücken 2007*, pp. 1113-1116, 2007

Three-dimensional annotation mining

Three-dimensional because alternative trees are possible, depending on the algorithm settings:

- binary/nonbinary, lower/higher percolated
- related to phrasal and discourse patterns



Induction of compositional tree structures

Gibbon, Dafydd. 2006. “Time types and time trees: Prosodic mining and alignment of temporally annotated data”. In: Stefan Sudhoff, et al., eds. *Methods in Empirical Prosody Research*. Berlin: Walter de Gruyter, pp. 281–209, 2006.

From physics to function

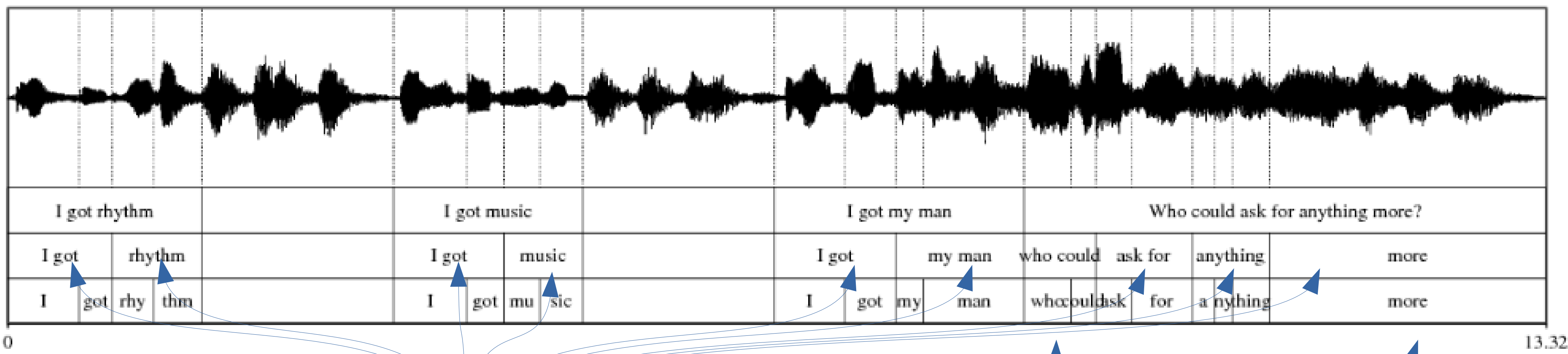
From physics to function

Rhetorical and structural functions of rhythm:

- Dialogue: coordination and alignment of interlocutors
  A: WHO saw JACK? - B: JIM saw Jack.
  cf. shadowing
- Utterance: structuring of narrative, arguments, …
  All Greeks are democrats. Socrates was a Greek. Socrates is a democrat.
- Sentence: ‘metadeixis’, pointing to focussed lexical items
  JOHN and RICHARD married SUSAN and KATE respectively.
- Word: ‘metadeixis’, pointing to the stress positions of modifying morphemes
  - The BLACKbird landed on the black BOARD.
  - This whisky wasn’t EXported, it was DEported.

Back to the song – music and speech

Music: tempo = allegro, time signature = 4/4, style = ‘brightly’



Abstract, structural rhythm:
foot / word timing

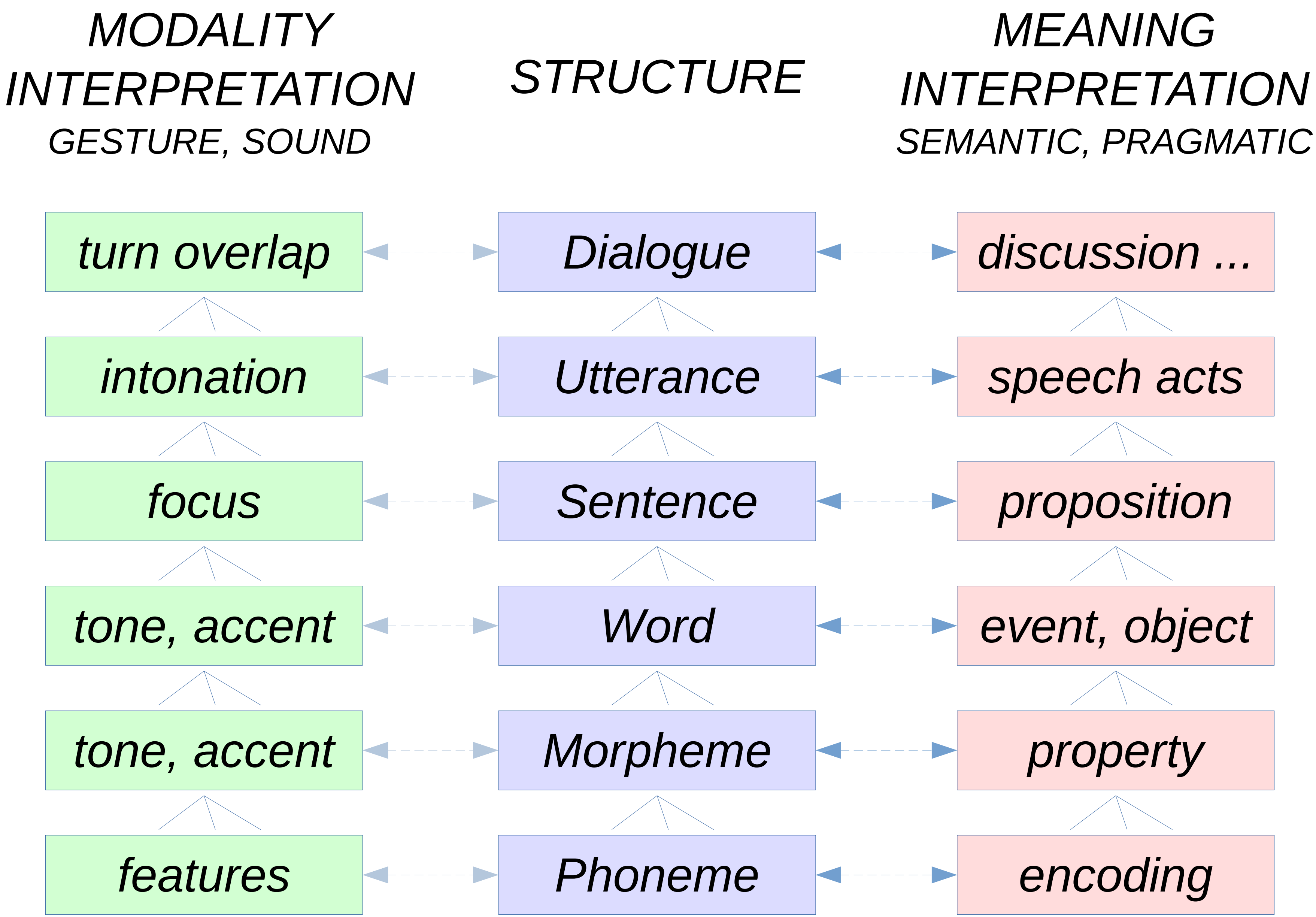
Special case:
Final lengthening

Special case:
only unstressed syllables

Ella Fitzgerald
“I got rhythm”

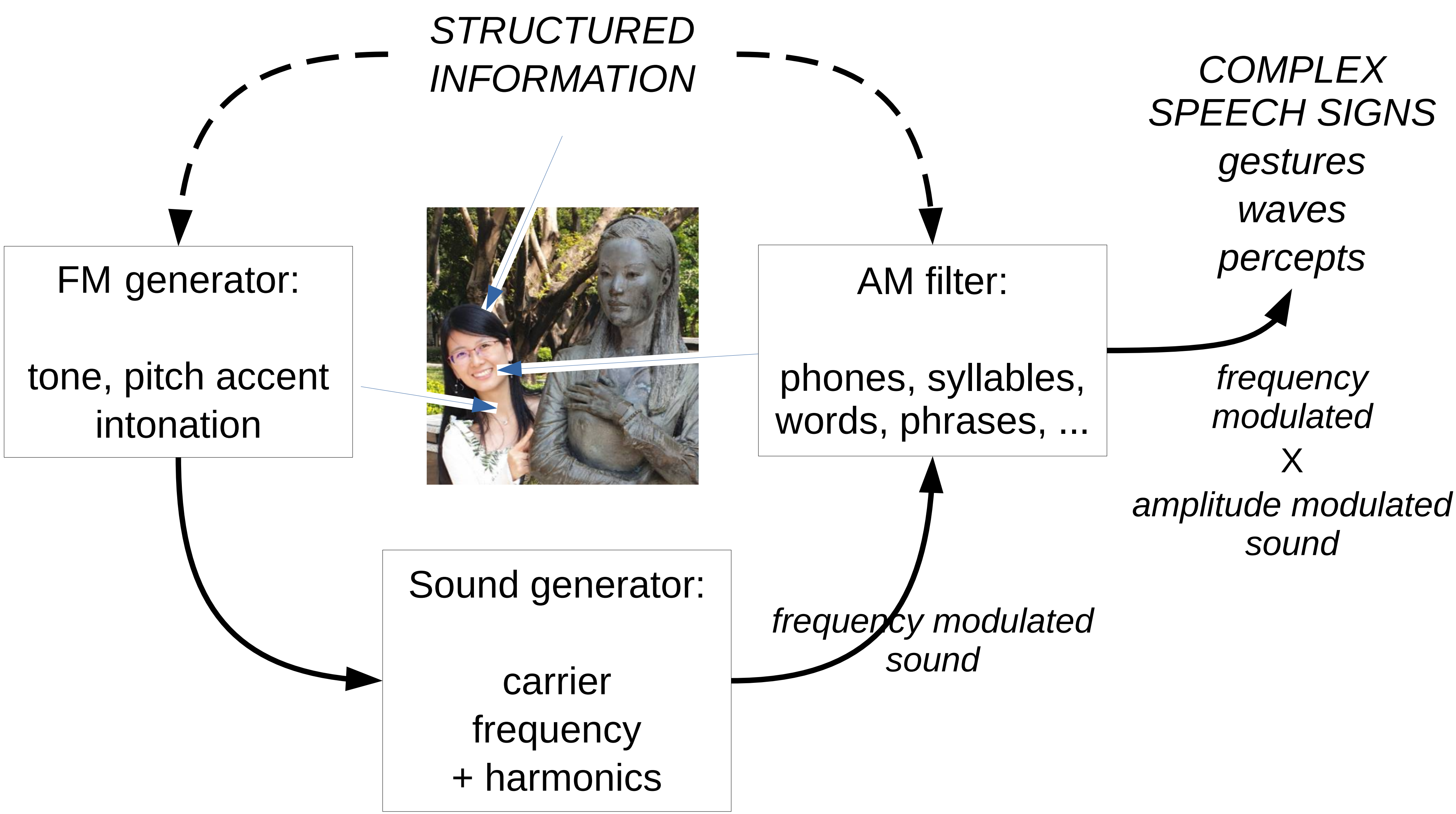
So where does rhythm align with language?

Summary: sources of rhythm (and melody)



Information conveyed by modulation of a sound

Amplitude modulation (AM) and Frequency Modulation (FM)



Frequency modulation

tones
pitch accents
intonation

Lexical tone: Mandarin Chinese

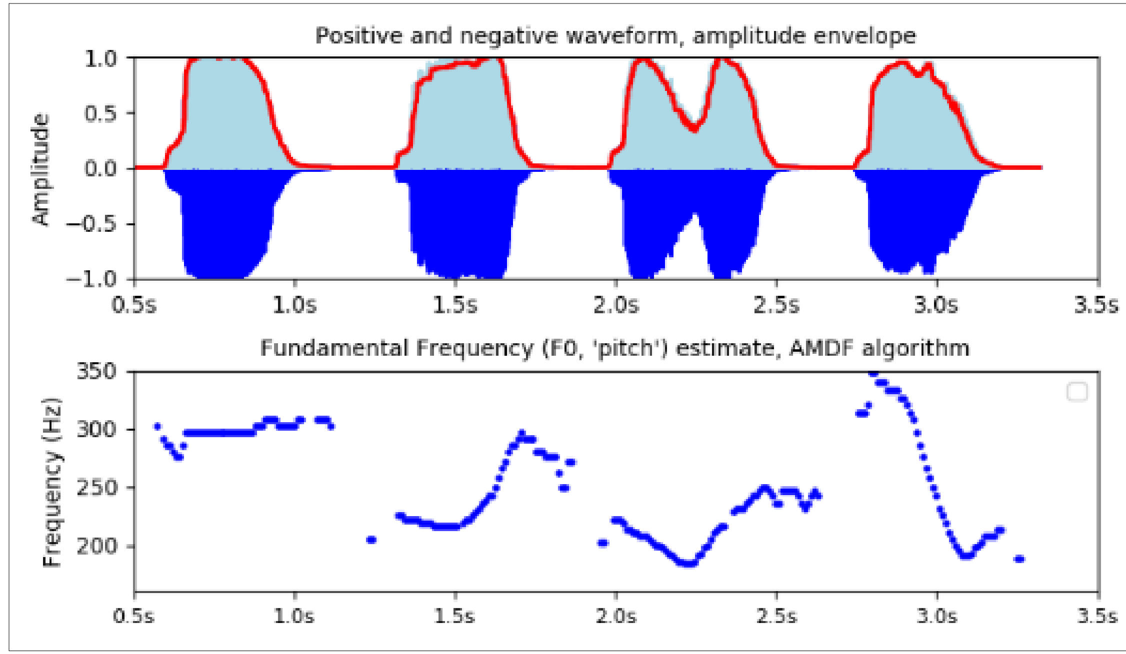
Mandarin Chinese:

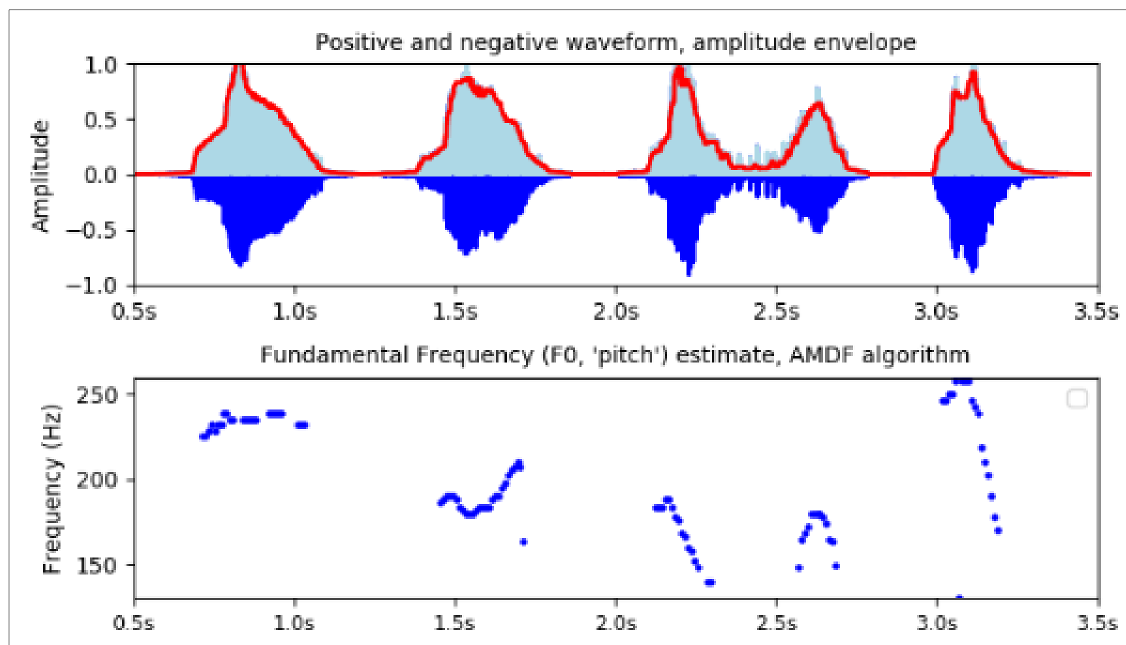
First tone:	ma1 mā	*mother*
Second tone:	ma2 má	*hemp*
Third tone:	ma3 mǎ	*horse*
Fourth tone:	ma4 mà	*scold*

Two female speakers of Beijing Mandarin.

Note the gap in Tone 3 of the second speaker, due to creaky voice on low pitch.

Approximately 50% of speakers in my classes in China have the creaky voice version.



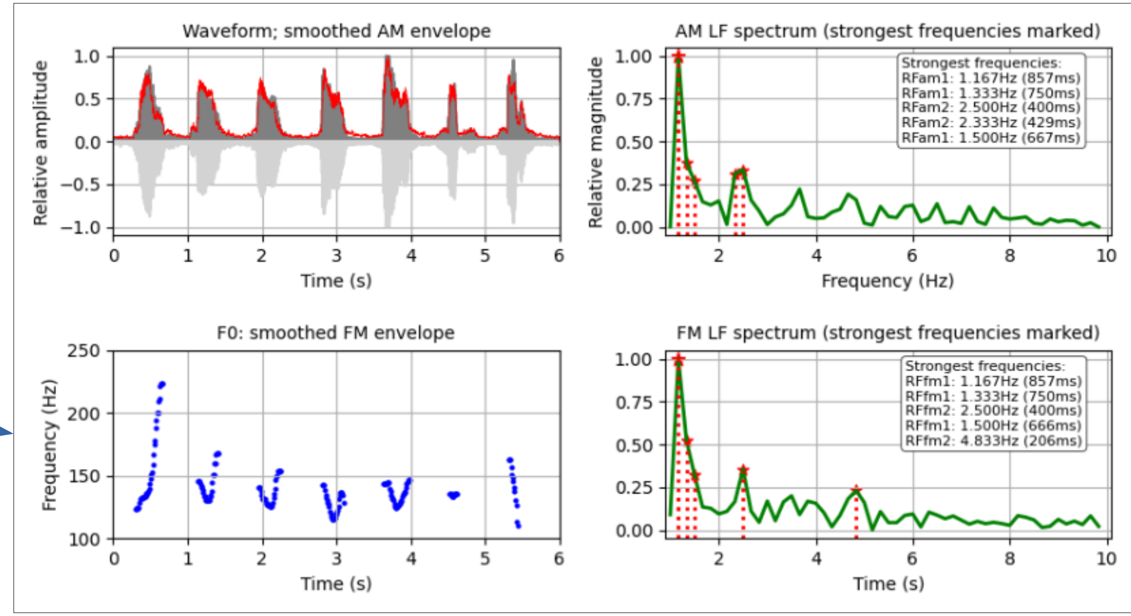


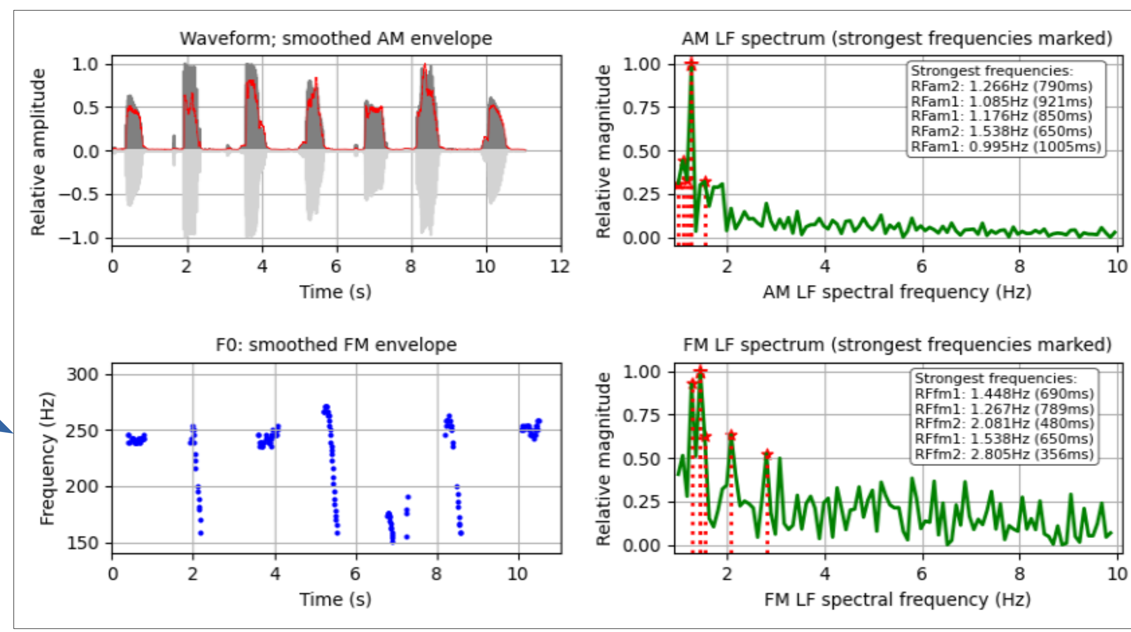
Frequency Modulation: pitch accent, intonation

Why does the 'music' of Chinese speech sound different?

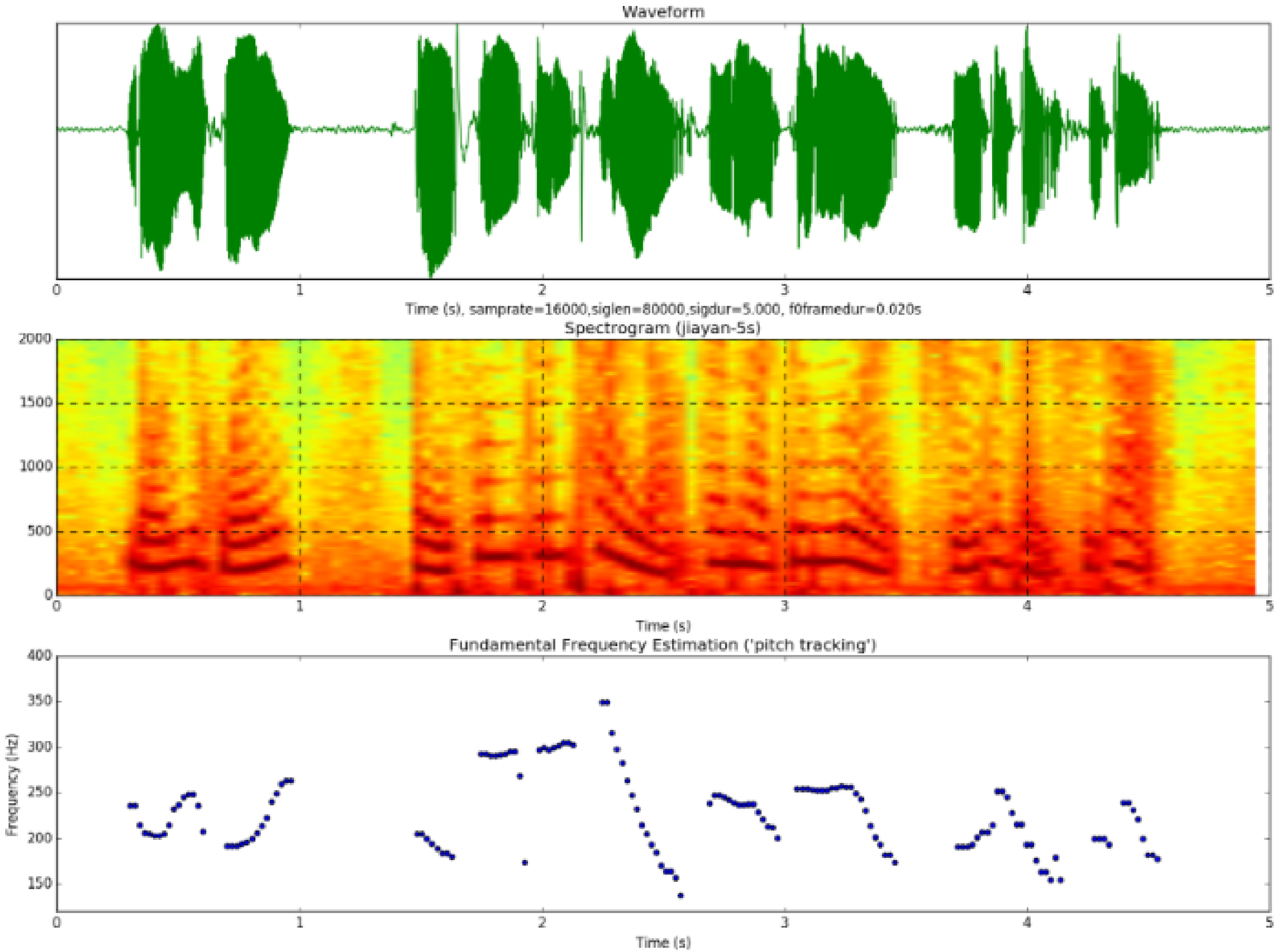
Because ...

- English has <u>pitch accents</u> which tend to remain the same in a tone group
- Chinese has <u>lexical tones</u> which tend to differ from word to word.

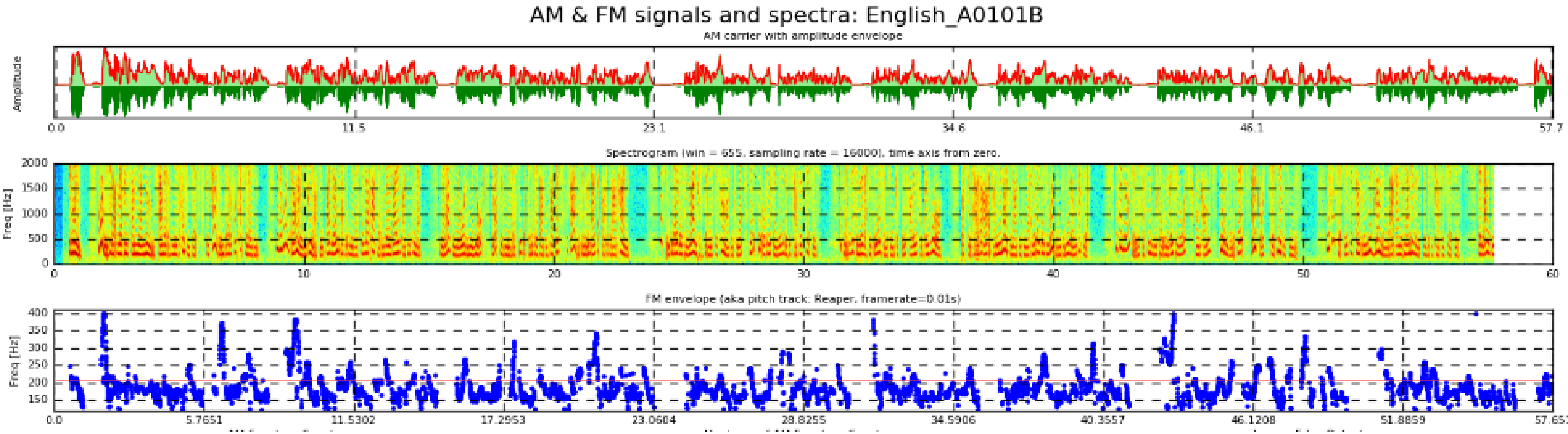




Narrative prosody: Mandarin Chinese



Narrative prosody: English



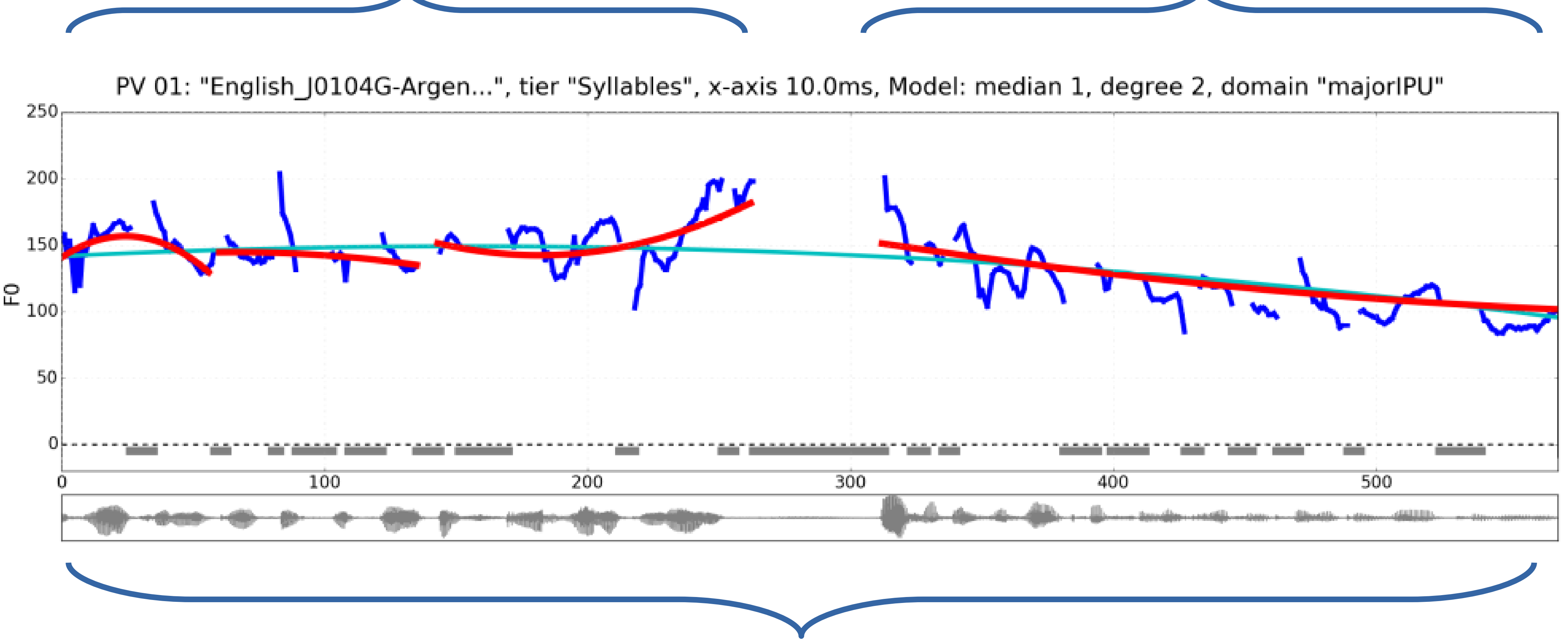
Hierarchy of pitch patterns (female voice):

- ‘paratone’ with very high initial pitch
- local ‘tone groups’ with mid to pitch

Discourse prosody: English

Question:
rising utterance contour

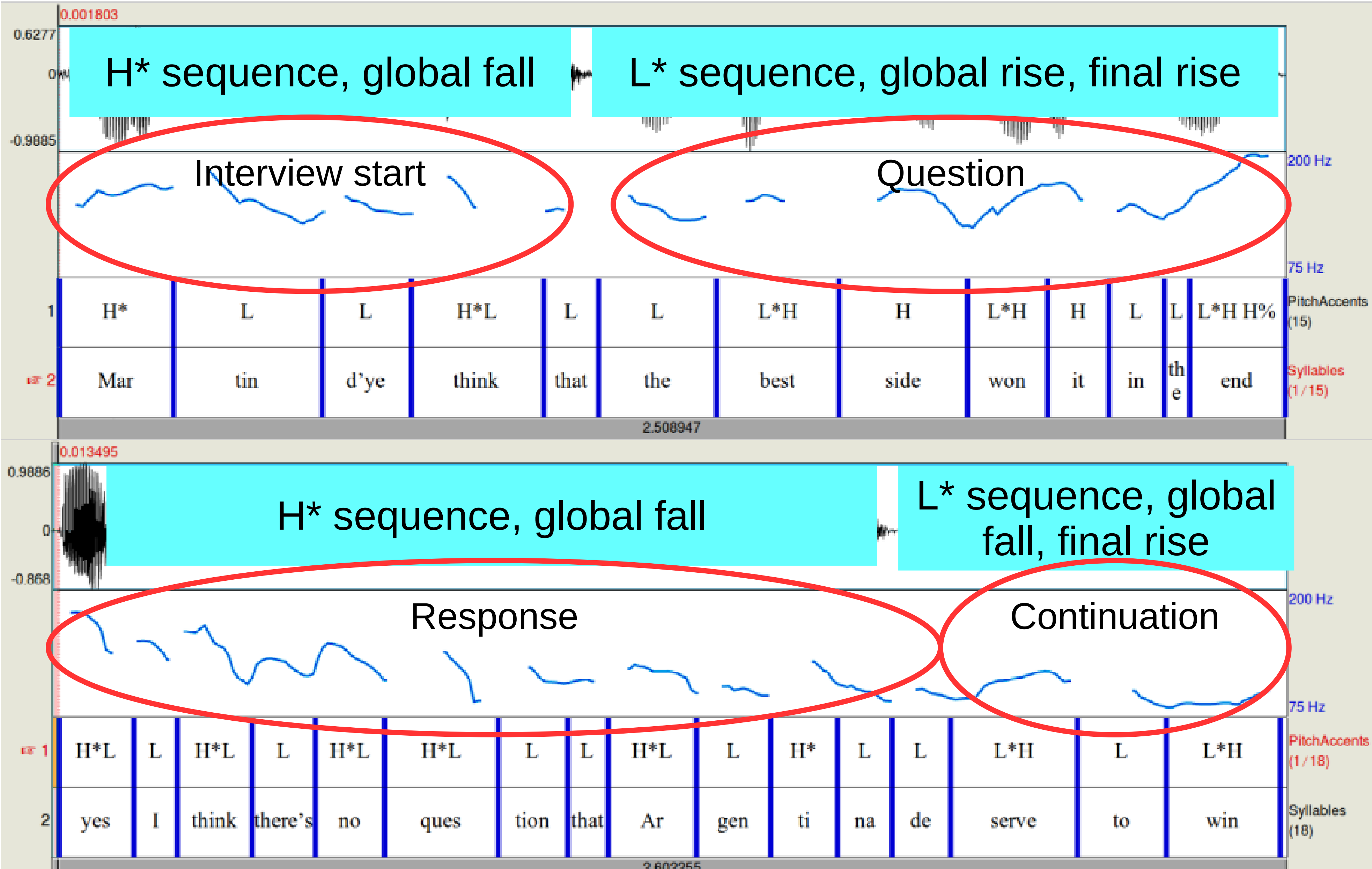
Answer:
falling utterance contour



Question+Answer: rising-falling adjacency pair contour

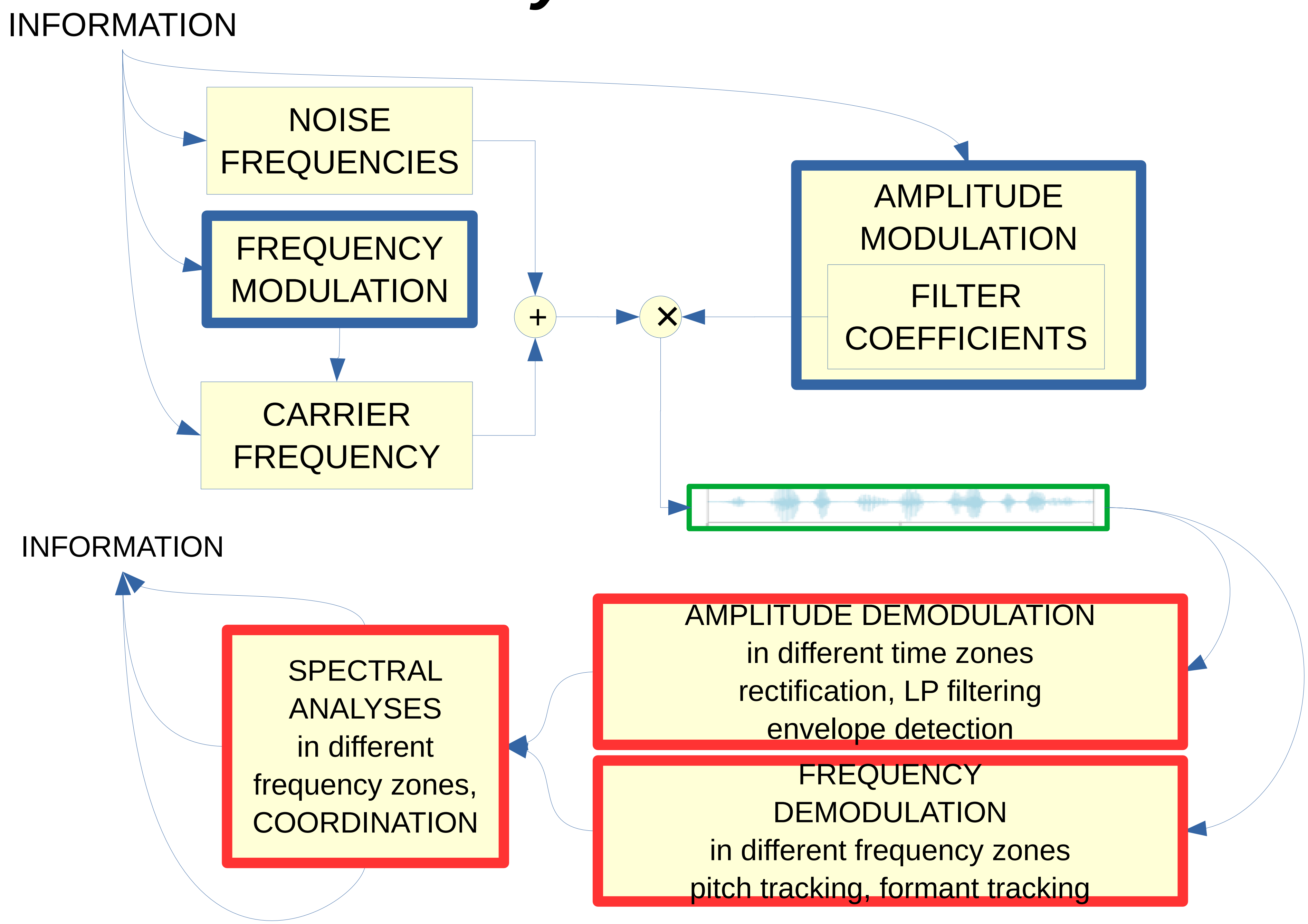
syntagmatic entrainment

Discourse prosody: English



Summary

Summary for Techies ...



Summary of Rank-Interpretation Architecture

Multilinear Category Ranks

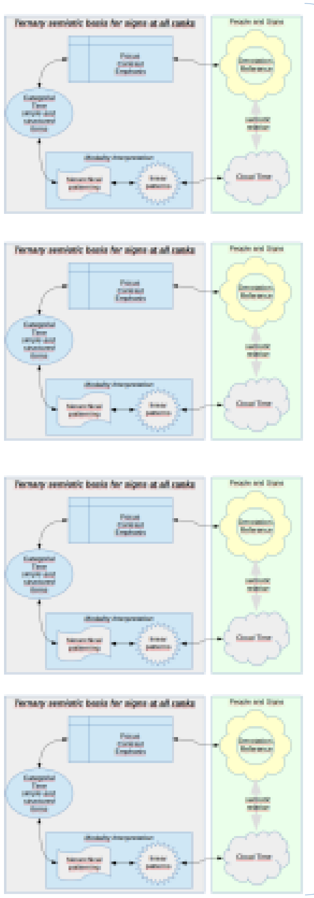
Ranks: Categories and their Interpretations

Discourse: Monologue, Dialogue

Utterance: turn, IPU, ...

Sentence, clause, phrase

Word: simple, inflected, compound, derived



Multimodal Rank Interpretation Architecture

Envoi

Speculations

on the evolution of speech prosody

Discourse Rhythms: Long FM contours

Thesis: in evolution,

- **frequency modulation and rhythm came first**
  - **emotional cries**
  - **turn-taking came before grammar,**

    Levinson, “Turn-taking in Human Communication – Origins and Implications for Language Processing”, 2015

Note: in infant speech,

- **frequency modulation and rhythm also come first**
  - **emotional cries**

    Wermke, Sebastian-Galles
  - **turn-taking**

    cf. the ‘bootstrapping’ literature

    the infant ‘twin-talk’ videos on YouTube ☺

Prosodic modulations – emotive speech

Thesis 1:

In the evolutionary time domain: emotive 'animal' modulations came before structural modulations

Thesis 2:

In the beginning was "Wow!" (Or "Aaah!")

Thesis 3:

Or the wolf whistle (it's not simply 'cat-calling')

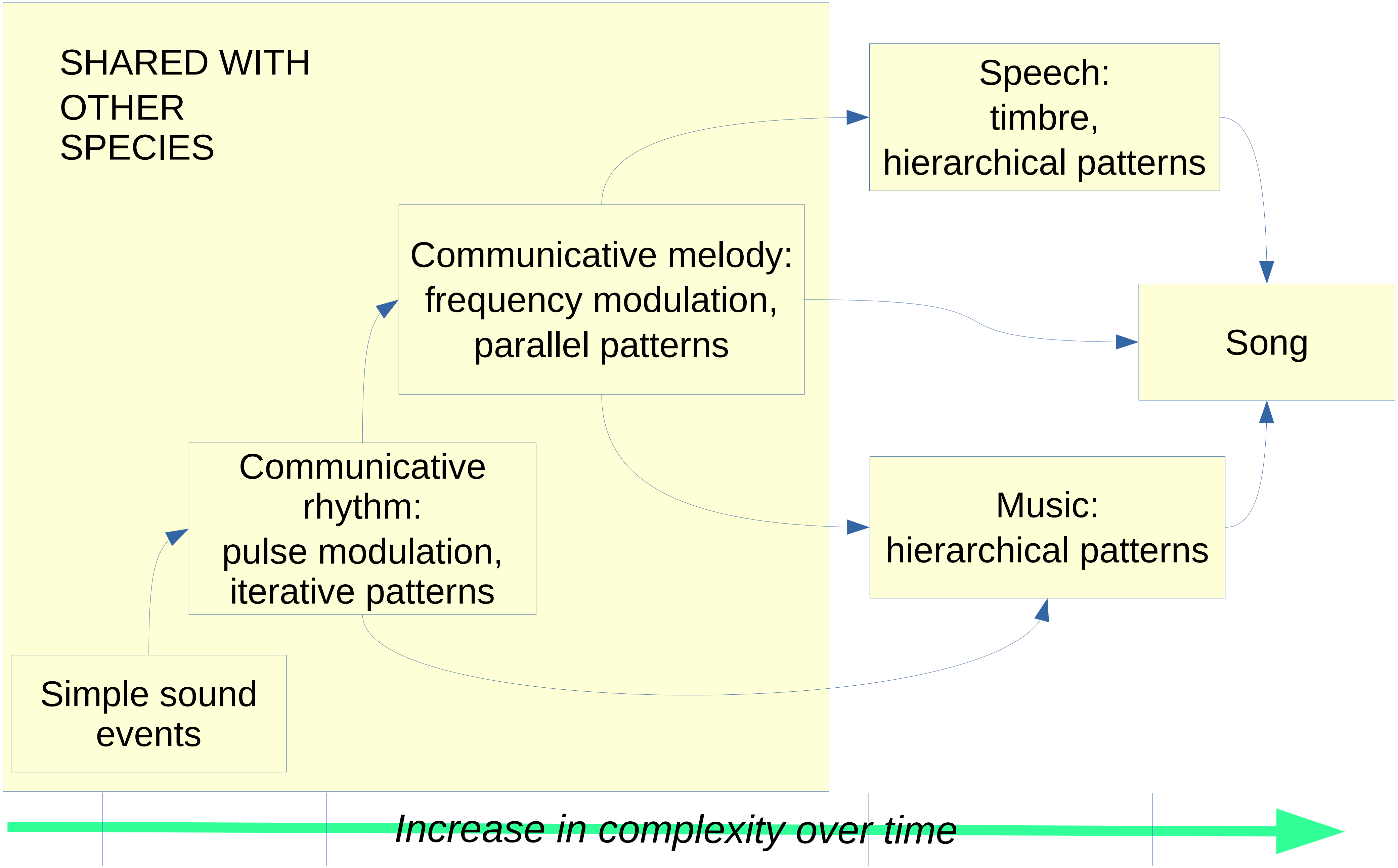
Thesis 4:

Other primates wowed, aahed and whistled first.

Humans continued the custom.

… I recommend these topics for future M.A. theses!

Speculation on prosody evolution



Thank you!