# Computational Methods in Phonetics

## Visualisation of Waveforms and Frequencies

2019-07-17, 14:30-16:30

Dafydd Gibbon

Bielefeld University
Jinan University, Guangzhou

# *Types of Computing in Phonetics*

## Consumer computation

- Praat, Audacity
- Calc, Excel, SPSS

## Scripting and Programming

- Praat scripting
- R, Stata, Matlab
- Python

## Software development

- C
- Java
- Python

## *Theses: YARD (Yet Another Discussion of Rhythm)*

Domain of investigation: Prosody

- The rhythms and melodies of spoken language
- The focus here is on rhythm more than on melody

Scientific results depend on scientific methods:

- Cognitive or hermeneutic methods:
  - direct observation and analysis of speech based on understanding of speech
  - Classic methods in linguistics
- Physical methods:
  - Physical methods for analysis of speech production, transmission and perception
  - Visualisation

## *Theses: YARD (Yet Another Discussion of Rhythm)*

There are many rhythms in speech (and in music):

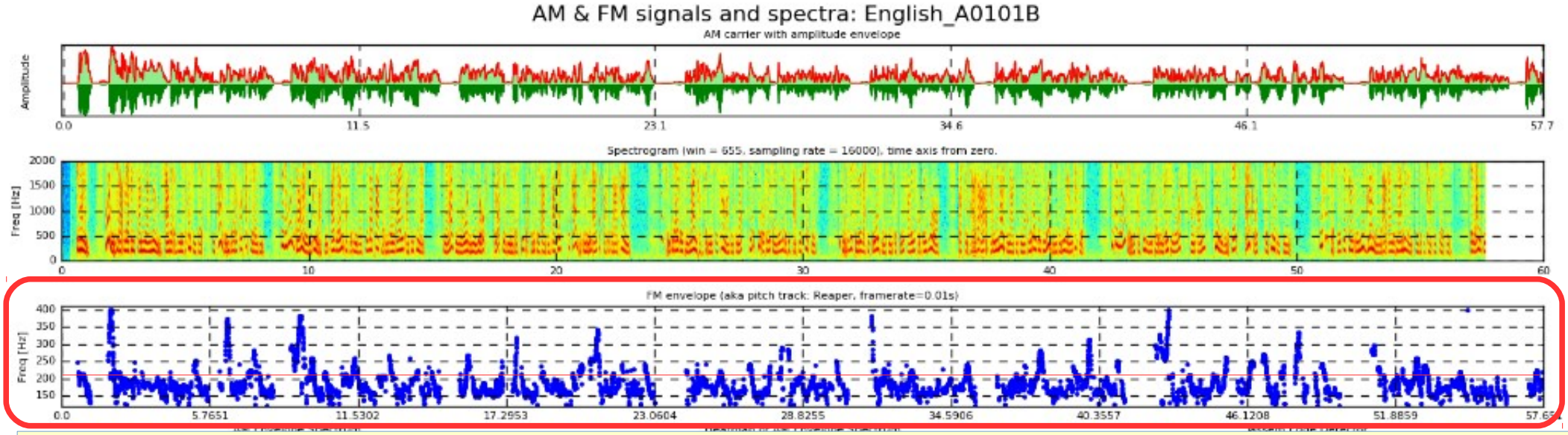
- Rhythms are oscillations, not sequences of isochronous states
- Rhythmic oscillations are essentially linear (‘finite state’)
- Rhythms occur in time domains of different sizes, each of them linear
- The time domains are associated with ranks of units of speech/language from discourse to phoneme

–

Rhythms appear to be hierarchical, but the hierarchy

- has limited depth
- has linear layers with iterative cycles
- is not a general recursive hierarchy

# *Linear Layered Discourse Rhythms*




57 seconds of a BBC news broadcast, from Aix-MARSEC corpus (A0101B)

Two Frequency zones:
1) Approximately 120 – 220 Hz
2) Approximately 300 – 400 Hz (on ‘paratone’ onsets)

<u>Finite depth linear 2-cycle iterative layered structure</u>
(FS, cf. Pierrhumbert, not a recursive hierarchy)

- Two independently motivated but structurally dependent linear layers
  (like hours & minutes on a clock)
- Each gradually declining and periodically resetting

# *Rhythm is oscillation*

## Rhythm is a marker of syntagmatic relations

- Rhythms mark syntagmatic relations in sequences of units which belong together

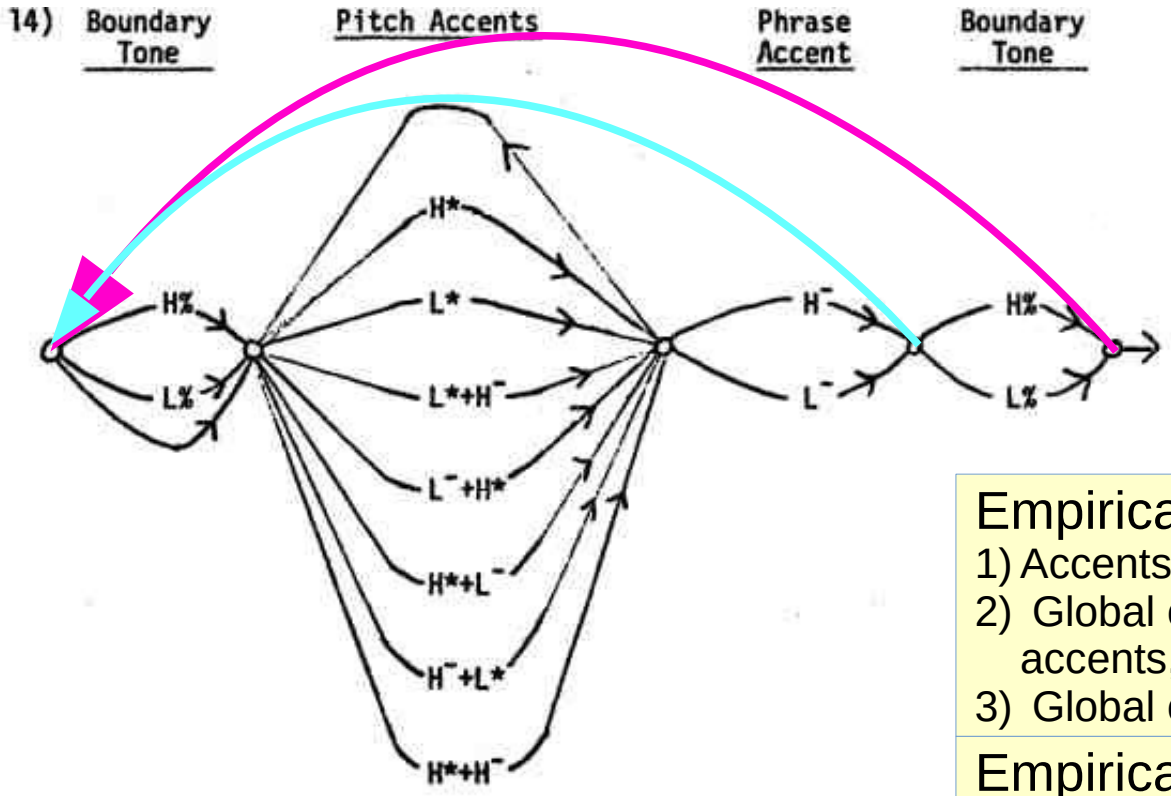
## Levels of abstraction:

1) Abstract oscillation: iterations, modelled by ‘loops’ in Finite State Machines
   - Intonation
     - Pierrehumbert’s model, with added loops
   - Tone
     - Gibbon’s model of Niger-Congo 2-level terraced tone
     - Jansche’s model of Tianjin tone
2) Physical oscillation of the amplitude of speech
   - Syllables
   - Phrases
   - ...

## *Abstract oscillation*

# *Phonological iteration as abstract oscillation*

**Pierrehumbert's regular grammar / finite state transition network**




Not the first (cf. Reich, 't Hart et al., Fujisaki, …)

But linguistically the most interesting.

Empirical overgeneration

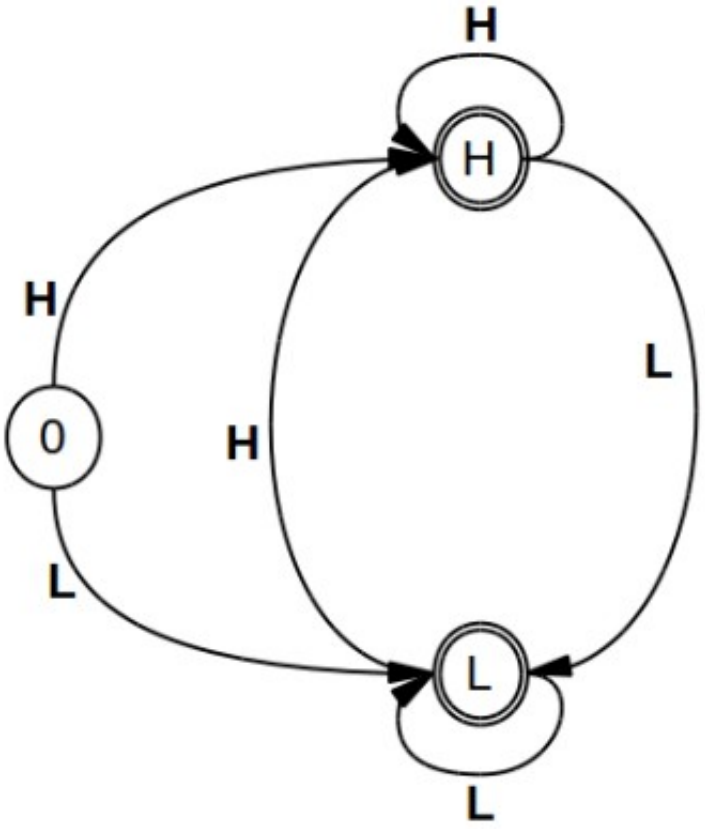
1) Accents in a sequence tend to be all H* or all L*
2) Global contours tend to be rising with L* accents, falling with H* accents
3) Global contours may span more than 1 turn

Empirical undergeneration

1) Paratone hierarchy not included
2) No time constraints

# *Phonological iteration as abstract oscillation*

**Niger-Congo Iterative Tonal Sandhi**
**(the most general case)**



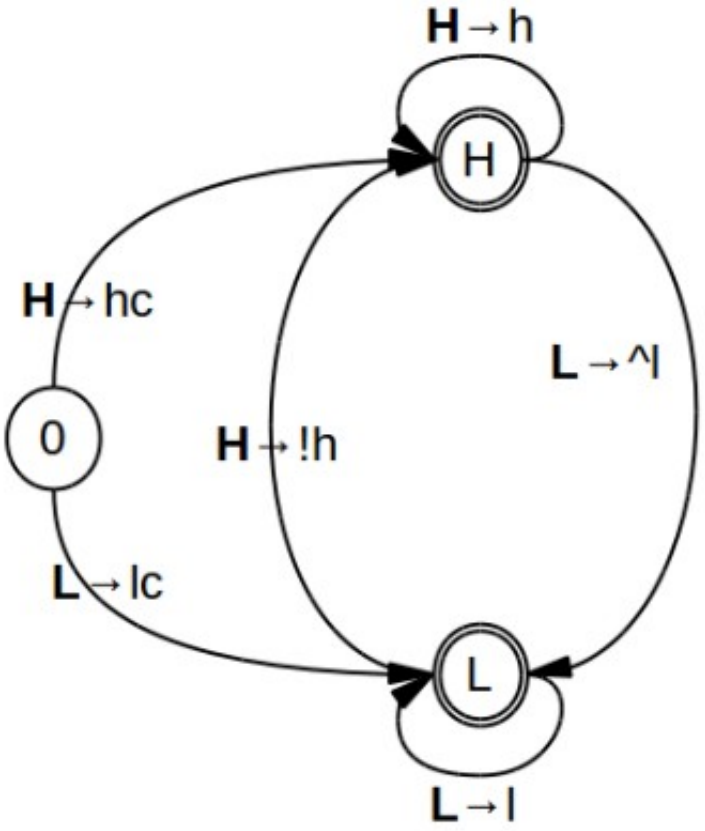

At the most abstract level, just one node with H and L cycling around it.

From an allotonic point of view:

- 3 cycles
- 1-tape (1-level) transition network

# *Phonological iteration as abstract oscillation*

**Niger-Congo Iterative Tonal Sandhi**
**(the most general case)**

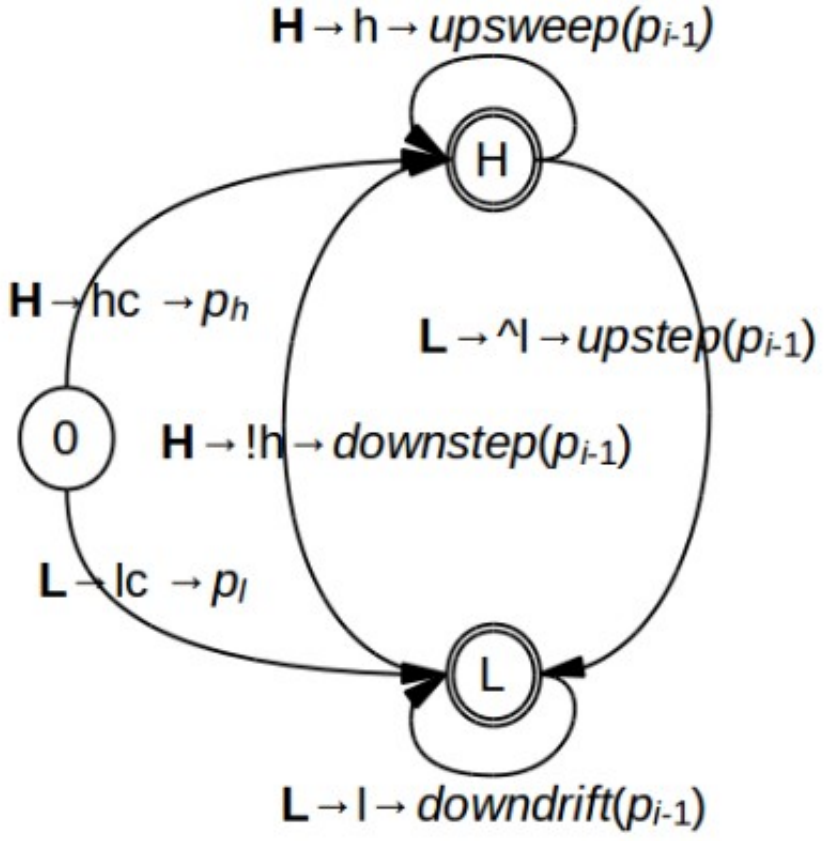



From an allotonic point of view:

- 3 cycles
- 2-tape (= 2-level) transition network

# *Phonological iteration as abstract oscillation*

**Niger-Congo Iterative Tonal Sandhi**
**(the most general case)**

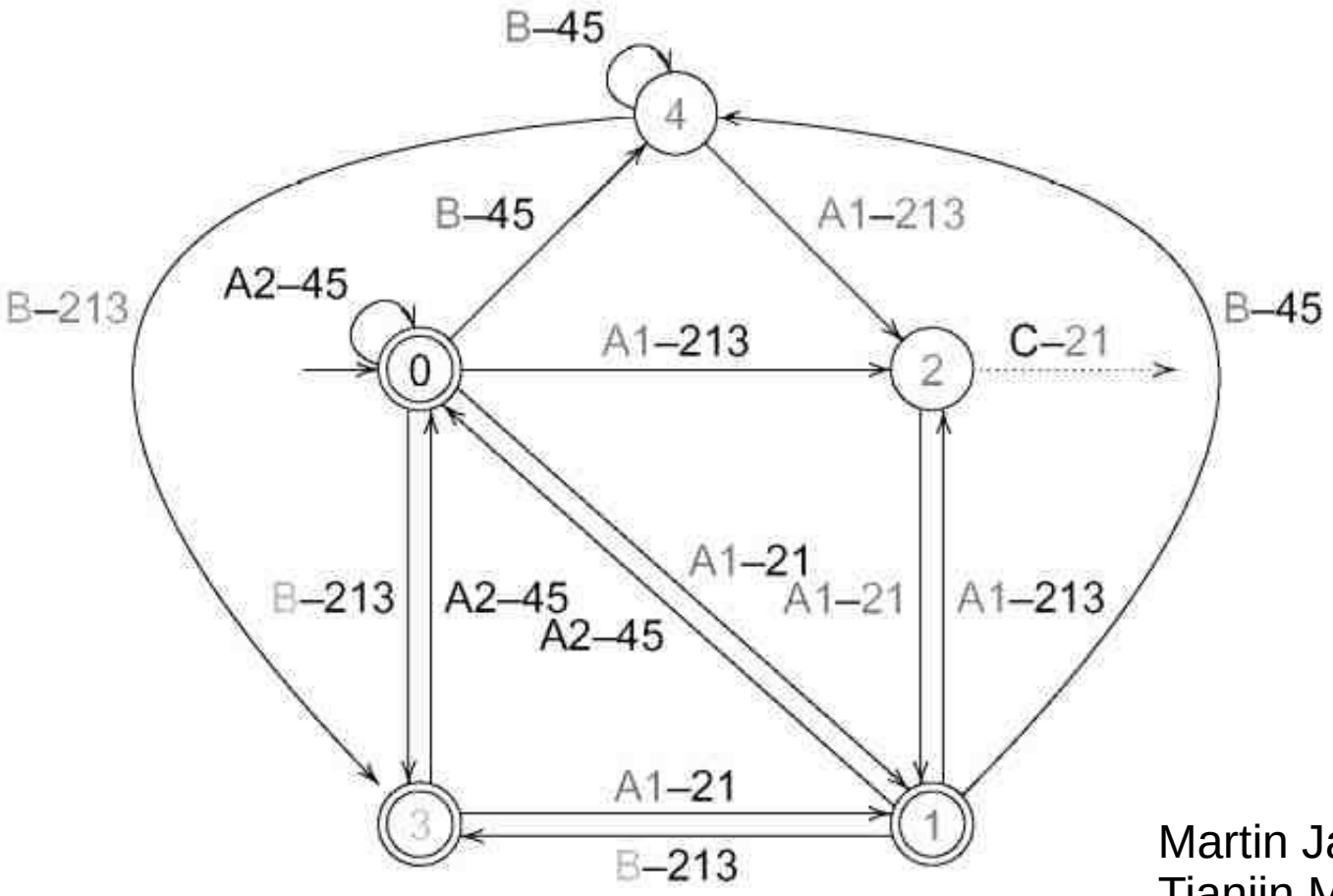



From phonetic signal processing point of view:

- 3 cycles
- 3-tape (= 3-level) transition network

# *Phonological iteration as abstract oscillation*

## Tianjin Dialect Iterative Tonal Sandhi




Martin Jansche 1998
Tianjin Mandarin tone sandhi

# *Physical oscillation*

## *Physical oscillation*

Oscillation means that intervals – cycles – tend to have the same length (duration).

But is it enough to focus only on duration of vowels, syllables, etc., using annotation mining?

What about the alternation feature of rhythm?

- 1-dimensional annotation mining of time-stamp durations
- 2-dimensional annotation mining of time-stamp durations
- 3-dimensional annotation mining of time-stamp durations

# *1-dimensional annotation mining of time-stamp durations*

One-dimensional because the result of the analysis is a single scale. The results are all comparable to variance or standard deviation, but differ in detail.

For example, with the *nPVI*, subtraction is between neighbouring data in a moving window (so a kind of AMDF, A*verage Magnitude Difference Function*), not between mean and data, thus factoring out tempo variations to some extent.

| | |
|---|---|
| $Variance(x_{1\ldots n}) = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$ | (or S*tandard Deviation*) |
| $PIM(x_{1\ldots n}) = \sum_{i \neq j} \left\lvert \log \frac{I_i}{I_j} \right\rvert$ | where $I_{i,j}$ are intervals in a given sequence |
| $PFD(d_{1\ldots n}) = \frac{\sum_{i=1}^n \lvert \bar{d} - d_i \rvert}{\sum_{j=1}^n d_j} \times 100$ | where $d$ is typically the duration of a *foot* |
| $nPVI(d_{1\ldots n}) = \frac{\sum_{k=1}^{k-1} \frac{\lvert d_k - d_{k+1} \rvert}{(d_k + d_{k+1})/2}}{n-1} \times 100$ | $d$ refers to duration of vocalic segment, syllable or foot, typically |

# *1-dimensional annotation mining of time-stamp durations*

One-dimensional because the result of the analysis is a single scale. The results are all comparable to variance or standard deviation, but differ in detail.

For example, with the *nPVI*, subtraction is between neighbouring data in a moving window (so a kind of AMDF, A*verage Magnitude Difference Function*), not between mean and data, thus factoring out tempo variations to some extent.

| | |
|---|---|
| $Variance(x_{1\ldots n}) = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$ | (or *Standard Deviation*) |
| $PIM(x_{1\ldots n}) = \sum_{i \neq j} \left\lvert \log \frac{I_i}{I_j} \right\rvert$ | where $I_{i,j}$ are intervals in a given sequence |
| $PFD(d_{1\ldots n}) = \frac{\sum_{i=1}^n \lvert \bar{d} - d_i \rvert}{\sum_{j=1}^n d_j} \times 100$ | where $d$ is typically the duration of a *foot* |
| $nPVI(d_{1\ldots n}) = \frac{\sum_{k=1}^{k-1} \frac{\lvert d_k - d_{k+1} \rvert}{(d_k + d_{k+1})/2}}{n-1} \times 100$ | $d$ refers to duration of vocalic segment, syllable or foot, typically |

# Isochrony as variance: Roach

Textual description hard to figure out, but maybe ...

$$\textit{Mean Foot Length (MFL)} = \frac{\Sigma_{i=1}^{n} |foot_i|}{n}$$

$$\textit{Percentage Foot Deviation (PFD)} = 100 \times \frac{\Sigma\ |MFL - len(foot_i)|}{n \times MFL}$$

Ignore syllables before initial and after final stresses

Calculate:

*average length* of interstress interval / foot (MFL)

*percentage deviation* of each interval from MFL, maybe ...

100 x (mean-interval$_i$) / mean

*variance* of percentage deviations (?)

This is a global measure:

ignores alternation and iteration criteria

# *1-dimensional annotation mining of time-stamp durations*

One-dimensional because the result of the analysis is a single scale. The results are all comparable to variance or standard deviation, but differ in detail.

For example, with the *nPVI*, subtraction is between neighbouring data in a moving window (so a kind of AMDF, A*verage Magnitude Difference Function*), not between mean and data, thus factoring out tempo variations to some extent.

| | |
|---|---|
| $Variance(x_{1\ldots n}) = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$ | (or *Standard Deviation*) |
| $PIM(x_{1\ldots n}) = \sum_{i \neq j} \left\lvert \log \frac{I_i}{I_j} \right\rvert$ | where $I_{i,j}$ are intervals in a given sequence |
| $PFD(d_{1\ldots n}) = \frac{\sum_{i=1}^n \lvert \bar{d} - d_i \rvert}{\sum_{j=1}^n d_j} \times 100$ | where $d$ is typically the duration of a *foot* |
| $nPVI(d_{1\ldots n}) = \frac{\sum_{k=1}^{k-1} \frac{\lvert d_k - d_{k+1} \rvert}{(d_k + d_{k+1})/2}}{n-1} \times 100$ | $d$ refers to duration of vocalic segment, syllable or foot, typically |

# Isochrony as ratio: Scott et al.

The Rh $Rhythmic\ Irregularity\ Measure\ (RIM) = \Sigma_{i \neq j} \left| \log \frac{I_i}{I_j} \right|$ dual
utter
ch
interval to each other interval.

Perfect isochrony: RIM = 0; non-isochrony is an open-ended log function.

RIM applies to utterances of the same length:

Scott & al. suggest generalising the RIM by dividing by *n* for interval sequences of length *n*.

This is incorrect: the RIM calculates a (triangular) matrix so a generalised RIM must be divided by $n^2$.

RIM is designed to be “symmetric”:

RIM therefore just measures isochrony, not rhythm, as it ignores rhythm alternation and iteration.

# *1-dimensional annotation mining of time-stamp durations*

One-dimensional because the result of the analysis is a single scale. The results are all comparable to variance or standard deviation, but differ in detail.

For example, with the *nPVI*, subtraction is between neighbouring data in a moving window (so a kind of AMDF, A*verage Magnitude Difference Function*), not between mean and data, thus factoring out tempo variations to some extent.

| | |
|---|---|
| $Variance(x_{1\ldots n}) = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$ | (or S*tandard Deviation*) |
| $PIM(x_{1\ldots n}) = \sum_{i \neq j} \left\lvert \log \frac{I_i}{I_j} \right\rvert$ | where $I_{i,j}$ are intervals in a given sequence |
| $PFD(d_{1\ldots n}) = \frac{\sum_{i=1}^n \lvert \bar{d} - d_i \rvert}{\sum_{j=1}^n d_j} \times 100$ | where $d$ is typically the duration of a *foot* |
| $nPVI(d_{1\ldots n}) = \frac{\sum_{k=1}^{k-1} \frac{\lvert d_k - d_{k+1} \rvert}{(d_k + d_{k+1})/2}}{n-1} \times 100$ | $d$ refers to duration of vocalic segment, syllable or foot, typically |

# Isochrony as local distance: Grabe & al.

$$nPVI = 100 \times \Sigma_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| /(m-1)$$

Normalises locally between neighbouring intervals for speech rate, using a distance measure:

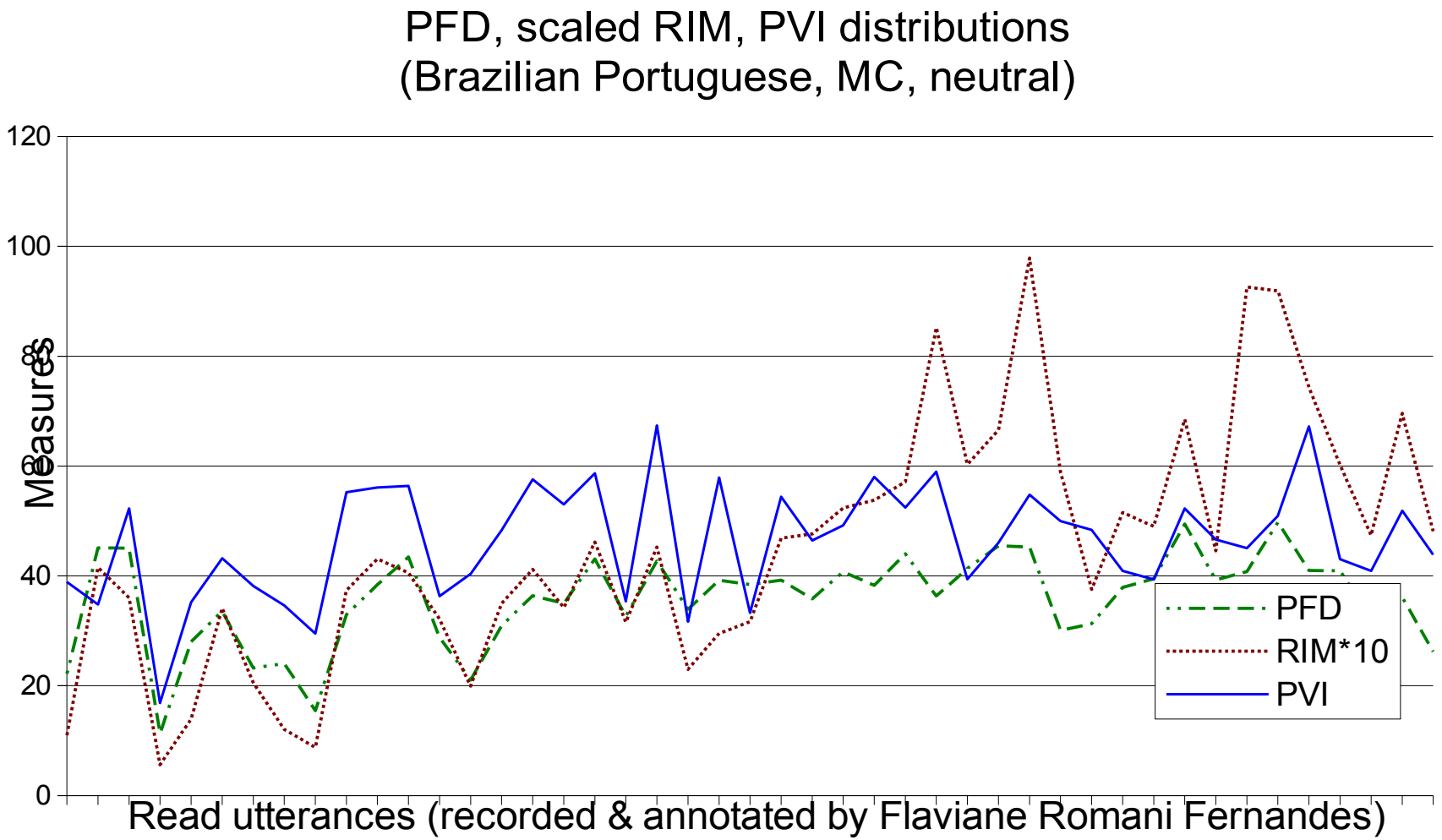
$DISTANCE_i = | INT_i – INT_i+1 | / AVG(INT_i, INT_i)$

PVI = 100 * AVG(DISTANCE) (range 0...200, asymptote)

Problems:

- Magnitude operation:
  - If PVI = 0, then isochrony holds – this is ok.
  - But if PVI ≠ 0, then intervals are somehow irregular, use of the absolute value means many sequences (increasing, decreasing, mixed, non-binary, ...) may have the same PVI
- Binary comparison (supposes iambs/trochees?), but
  - Spondaic: *That big black bear swam fast past Jane's boat.*
  - Dactylic: *Jonathan Appleby trundled along with a tune on his lips.*

# Empirical comparison of PFD, RIM, PVI




The models should at least correlate...

... but they don’t correlate too well

# Interval duration approaches and typology

Ramus:

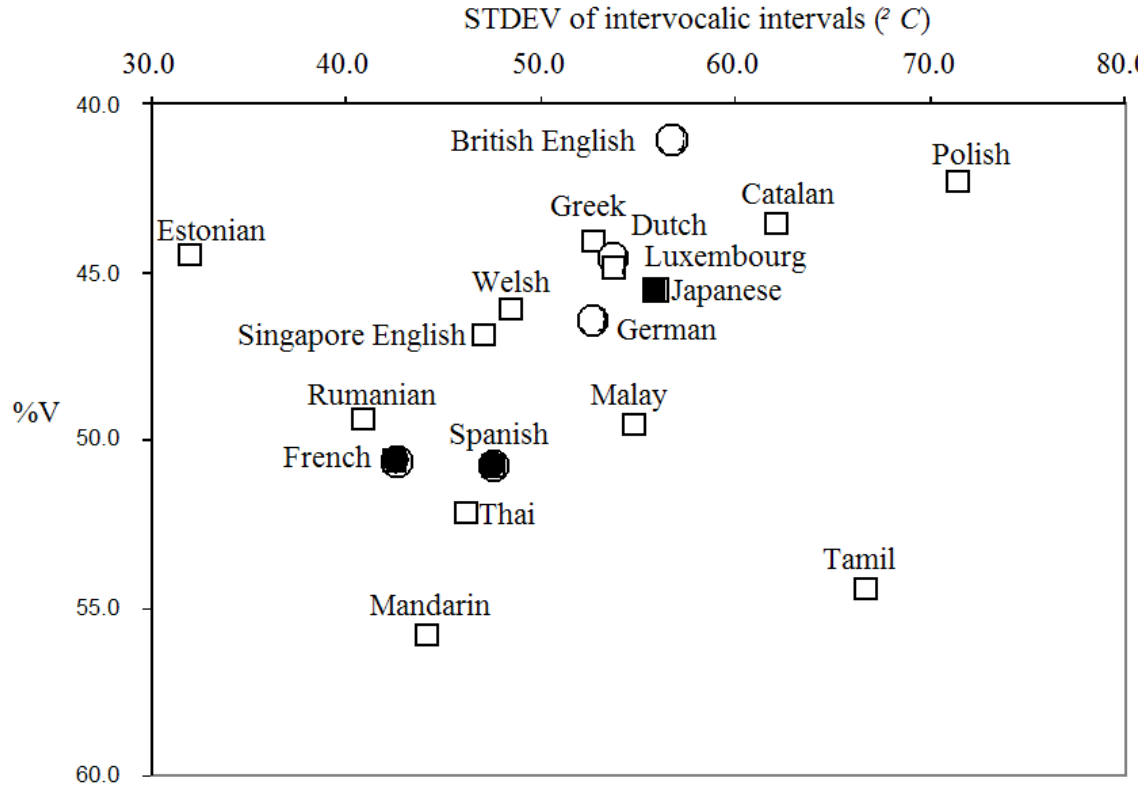
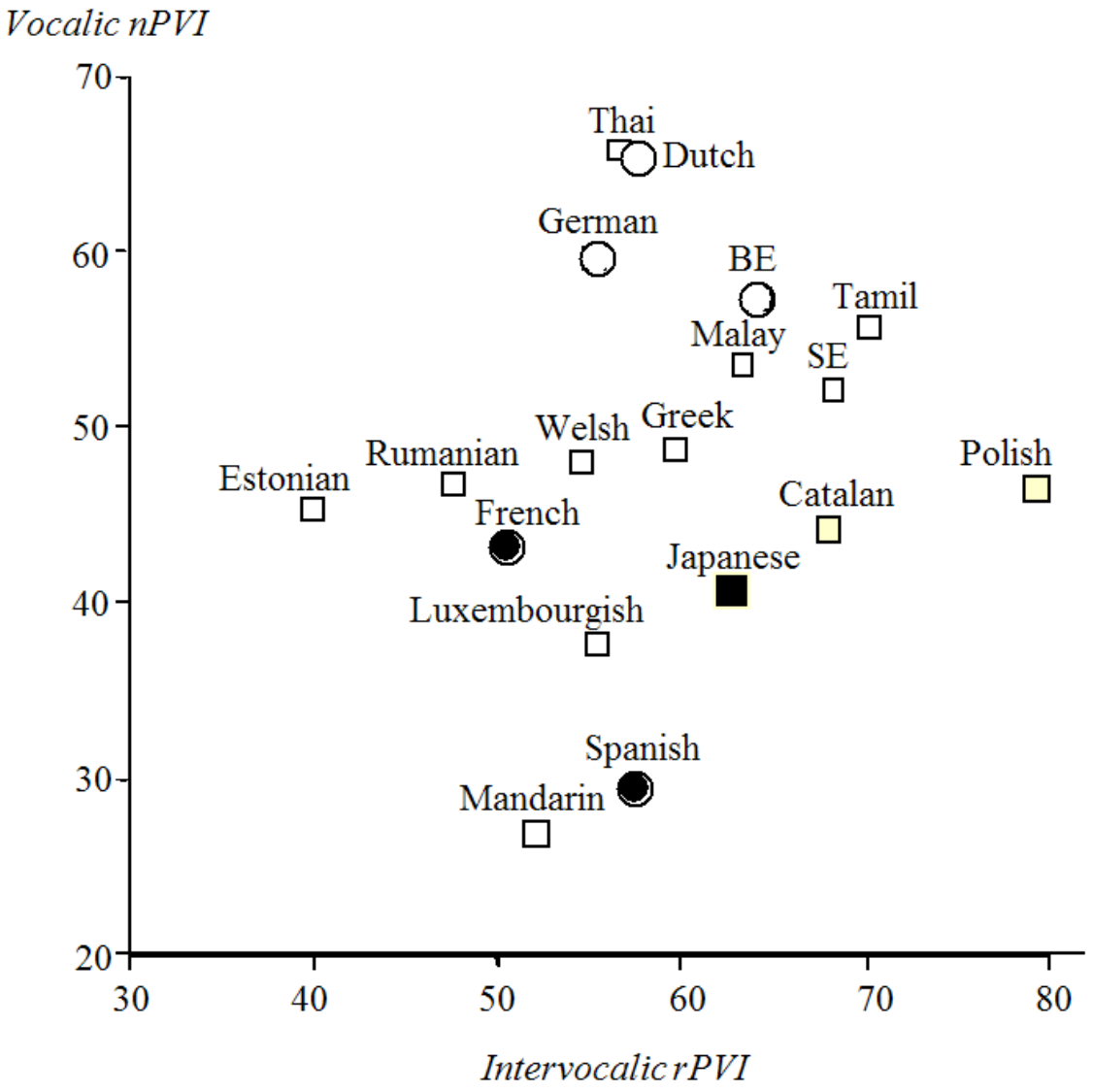



Figure 3. The measure *%V* is plotted on the y-axis, in reverse order. The standard deviation of intervocalic intervals $\Delta C$, is given on the x-axis.

# Interval duration approaches and typology

Grabe & al.:



# Interval duration approaches and typology

Ramus:

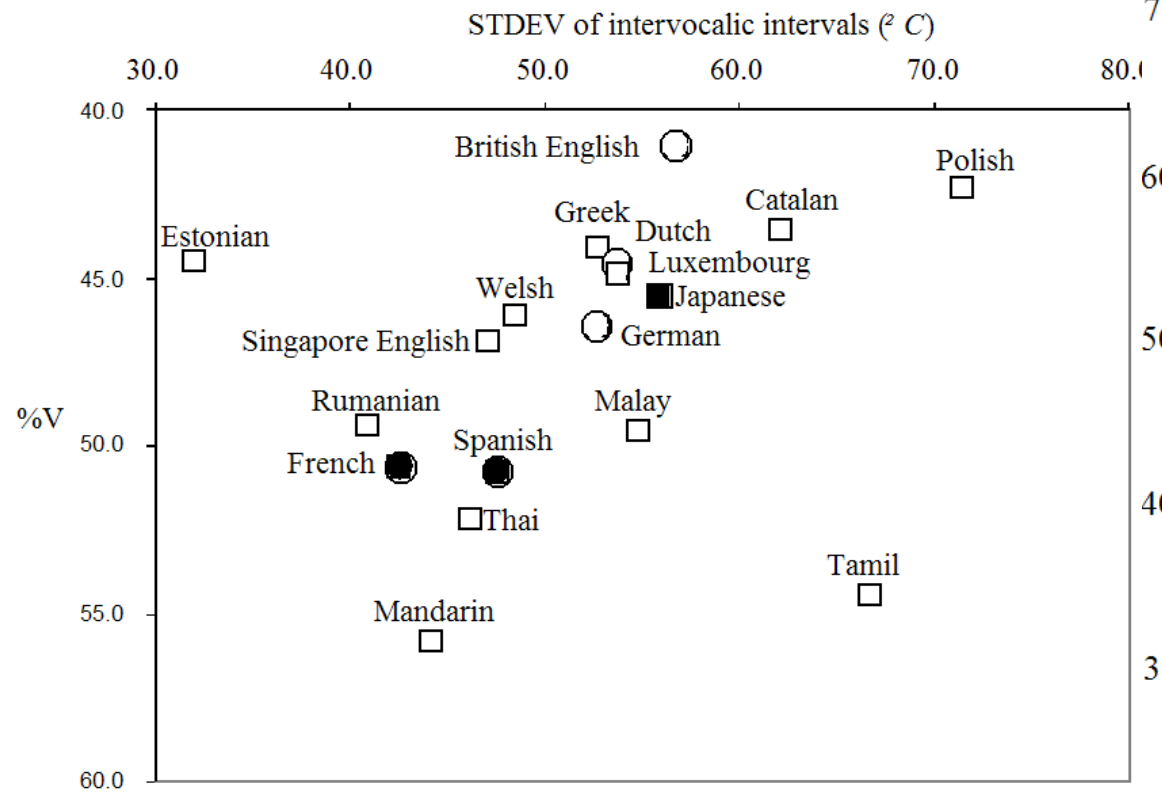
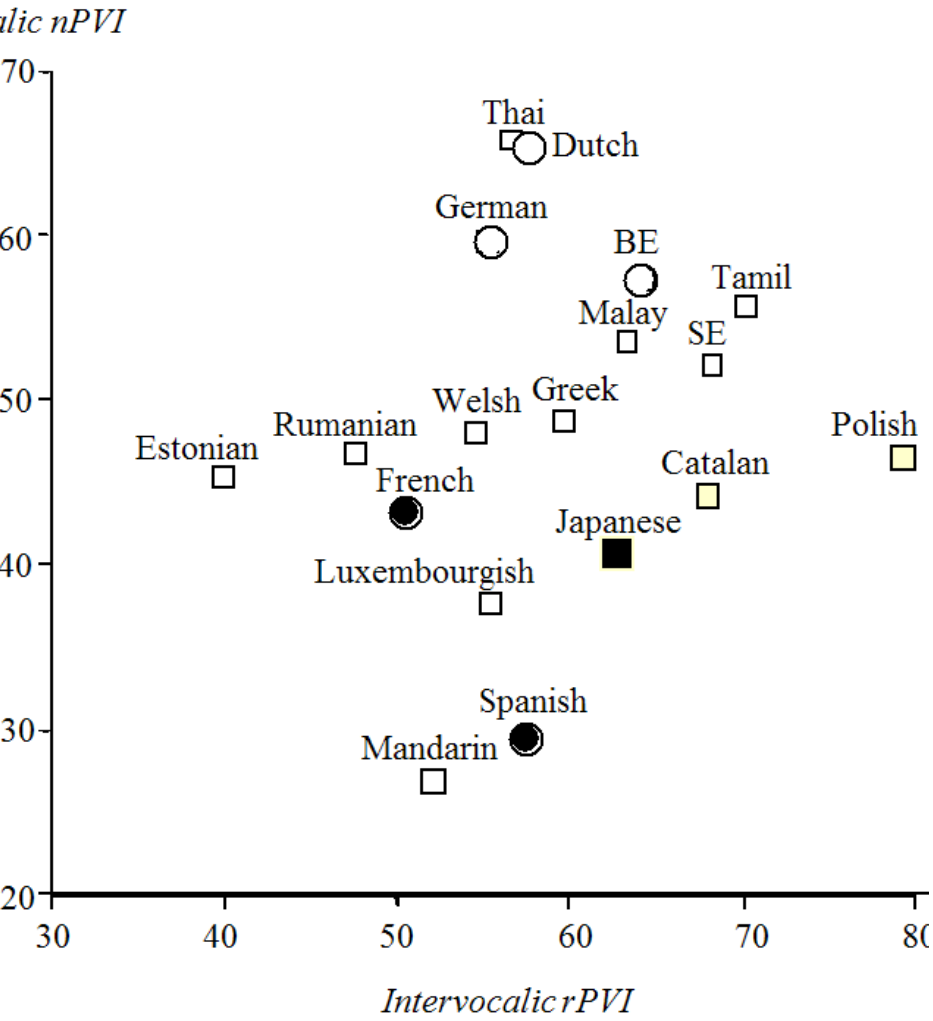



Figure 3. The measure *%V* is plotted on the y-axis, in reverse order. The standard deviation of intervocalic intervals $\Delta C$, is given on the x-axis.

Grabe et al.:




Daniel Hirst pointed out that these relations can be obtained phonologically, by comparing the phonotactics of the languages, rather than phonetically, with measurements.

# Summary: 1-dimensional interval duration approaches

There are many other interval duration measures

- perhaps most prominently in the past 5 years the non-isochronous Ramus model: ΔC x %V

Isochrony/irregularity is not a sufficient condition:

- cf. Cummins (2002) on Ramus:
  - *Where is the bom-di-bom-bom in %V?*
- Interval duration isochrony approaches ignore the *ordering and directionality*, of rhythm, *alternation* within Rhythm Units and *iteration* of Rhythm Units.

And

- The interval duration approaches assume the relevant event is duration of segmental constructs
  - which it may or may not be

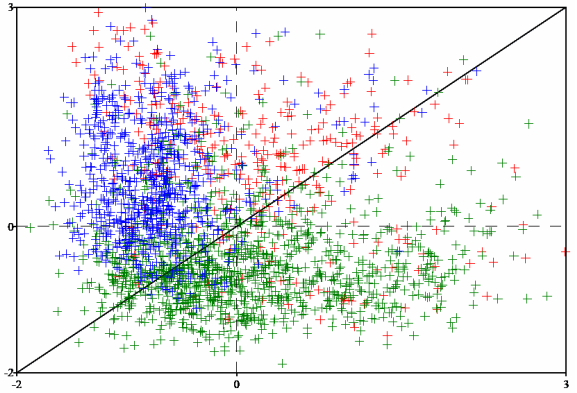
# A 2-dimensional interval duration approach

# 2-dimensional annotation mining of time-stamp durations

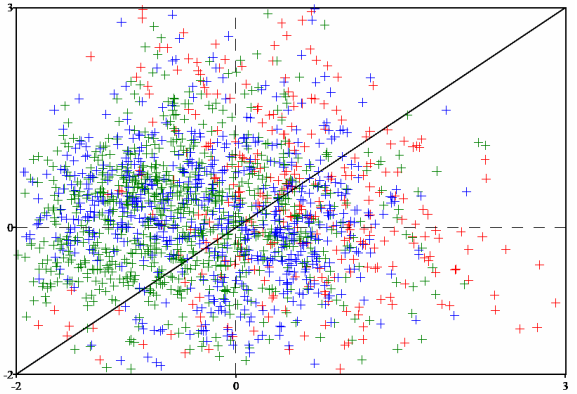
Wagner:

- addresses the absolute magnitude problem
- plots $duration_i$ x $duration_{i+1}$
- creates typologically interpretable clusters

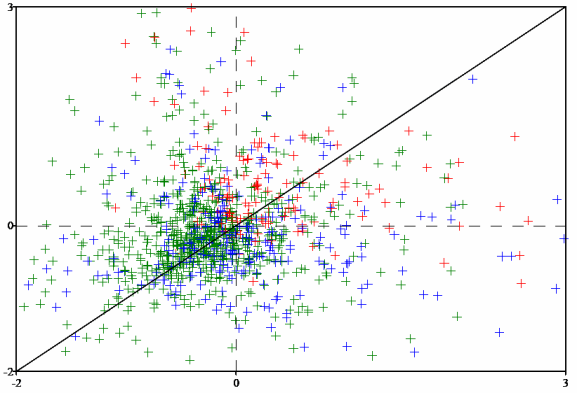
English



French



Polish



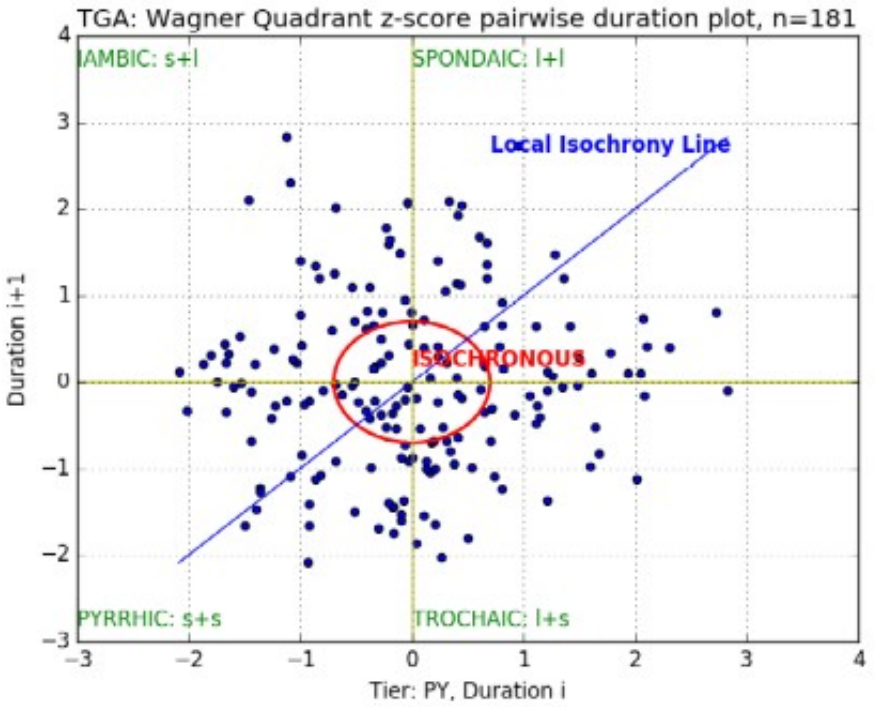
green: stressed x unstressed
blue: unstressed x stressed
red: phrase-final

# *2-dimensional annotation mining of time-stamp durations*

Following Wagner, two-dimensional duration relations are represented in a z-scored scatter plot, not as a single scale.
Result, visualising the scale in two dimensions:

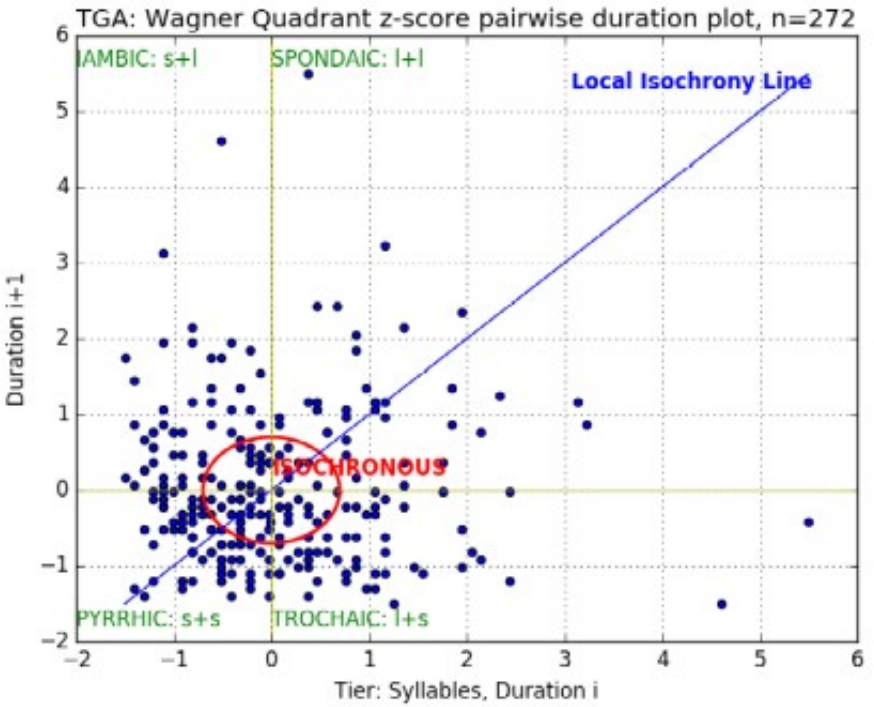
Mandarin: means are scattered relatively evenly around the centre
English: e.g. *count*(*short-short*) **>** *count*(*long-long*), not binary!

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

# A 3-dimensional interval duration approach

# 3-dimensional annotation mining of time-stamp durations

General strategy:

- take the local difference/distance measure from the PVI
- do not throw directionality away by taking absolute values of differences
- but use directionality (polarity) to determine grouping

Specific procedure:

- using annotation time-stamps, recursively build tree structures (Time Trees):
  - iambic parametrisation:
    if right neighbour is stronger,
    then group
    else stack and wait for a stronger right neighbour
  - trochaic parametrisation:
    if right neighbour is stronger,
    then group
    else stack and wait for a weaker right neigbour

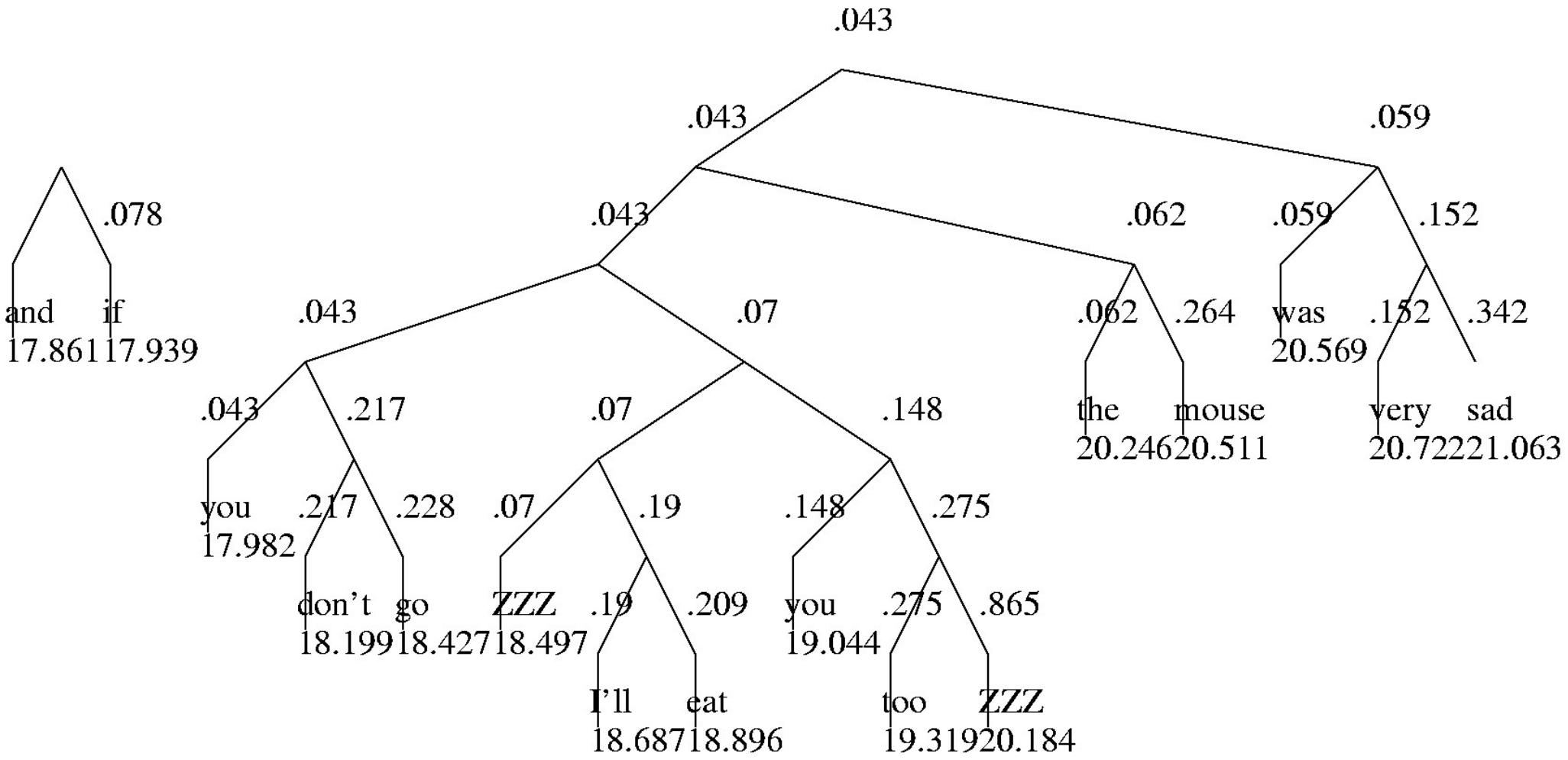
# Data – reading style presumed optimal

A tiger and a mouse were walking in a field when they saw a big lump of cheese lying on the ground. The mouse said: "Please, tiger, let me have it. You don't even like cheese. Be kind and find something else to eat." But the tiger put his paw on the cheese and said: "It's mine! And if you don't go I'll eat you too." The mouse was very sad and went away.

The tiger tried to swallow all of the cheese at once but it got stuck in his throat and whatever he tried to do he could not move it. After a while, a dog came along and the tiger asked it for help. "There is nothing I can do." said the dog and continued on his way. Then, a frog hopped along and the tiger asked it for help. "There is nothing I can do." said the frog and hopped away.

Finally, the tiger went to where the mouse lived. She lay in her bed in a hole which she had dug in the ground. "Please help me," said the tiger. "The cheese is stuck in my throat and I cannot remove it." "You are a very bad animal," said the mouse. "You wouldn't let me have the cheese, but I'll help you nonetheless. Open your mouth and let me jump in. I'll nibble at the cheese until it is small enough to fall down your throat." The tiger opened his mouth, the mouse jumped in and began nibbling at the cheese. The tiger thought: "I really am very hungry.."

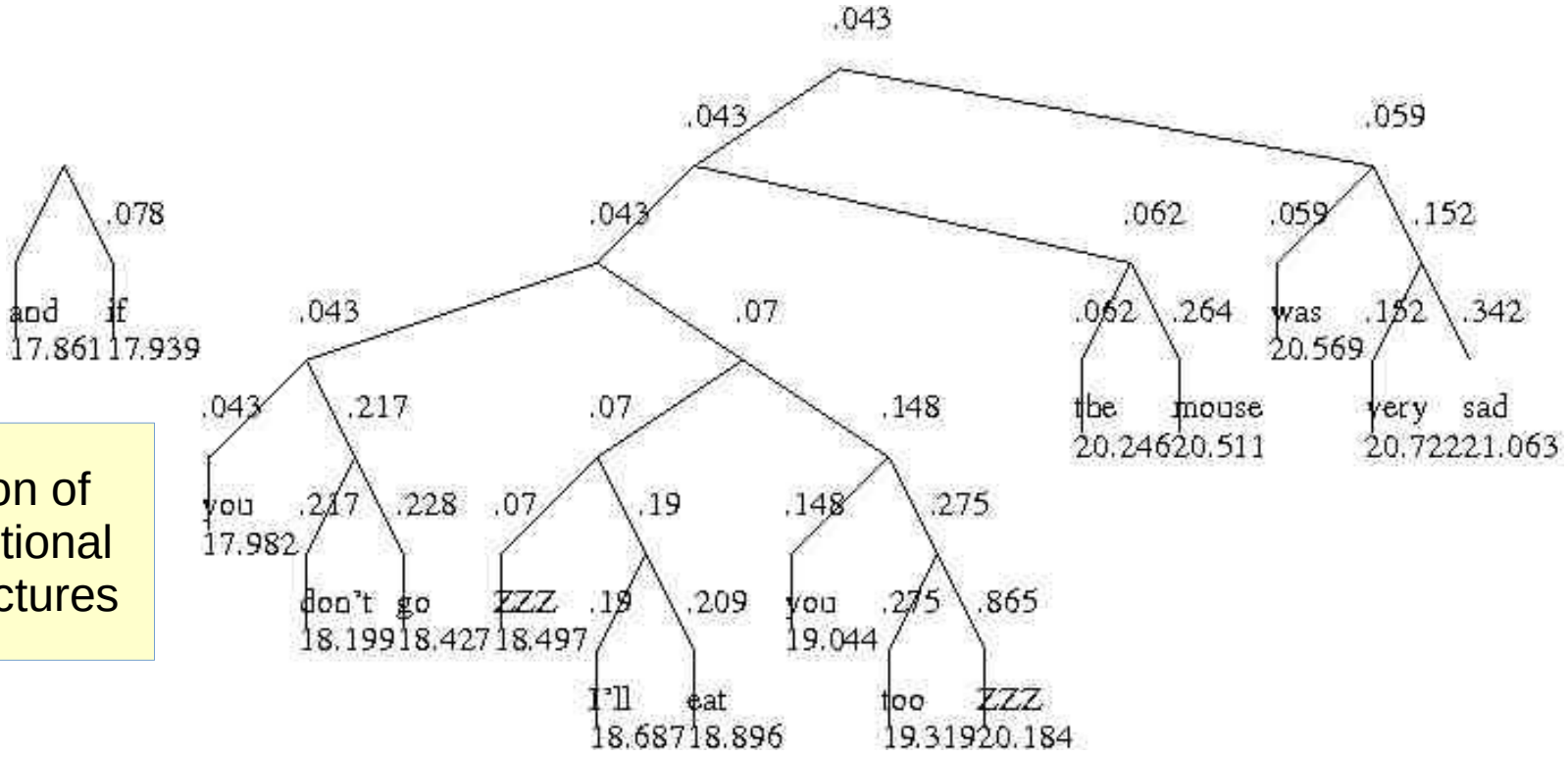
# Interpreting Time Trees



# *Two-dimensional annotation mining of time-stamp durations*

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.

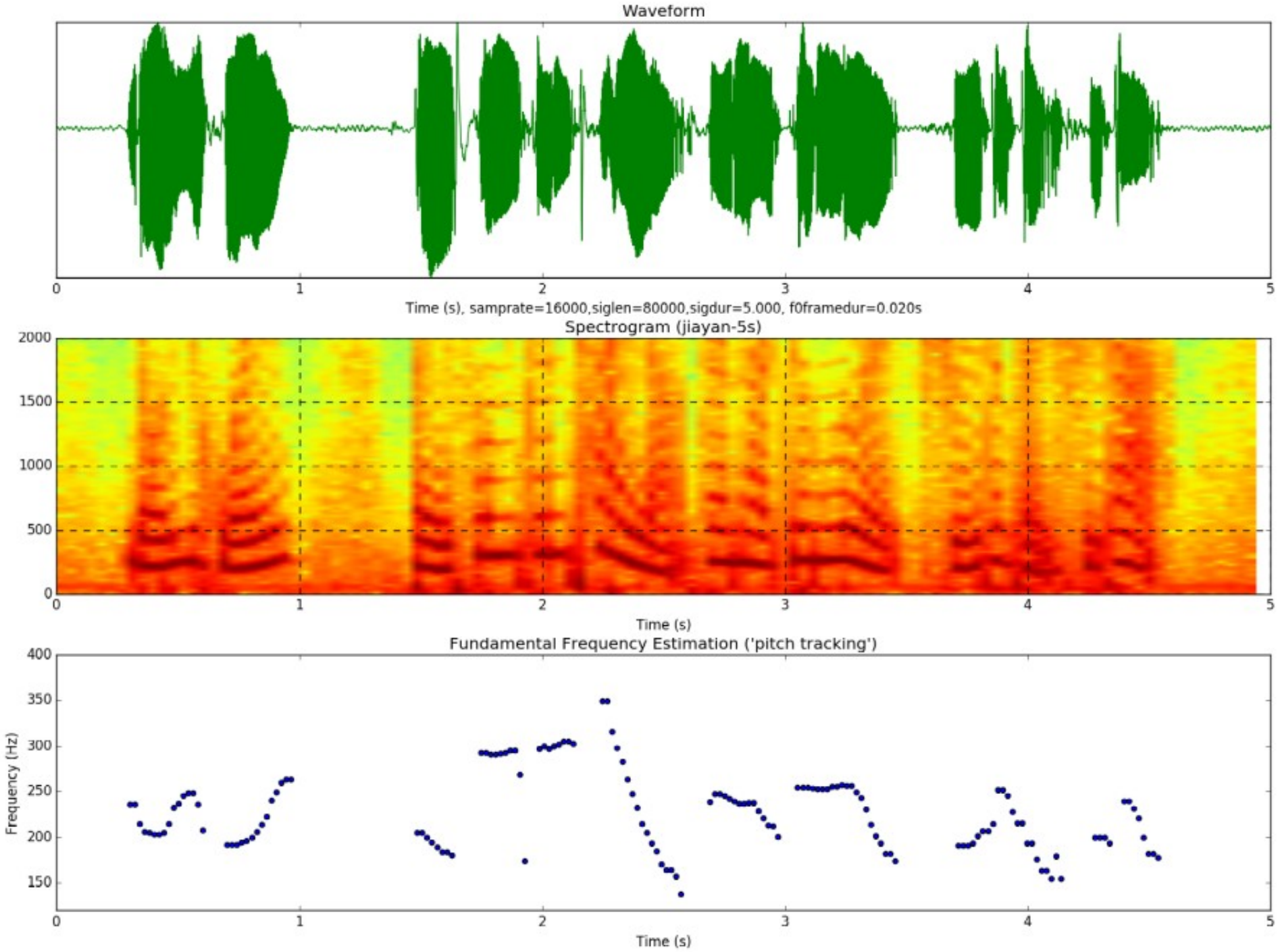
**But the annotation mining approaches ...**

- are based on cognitive filtering by linguists
  (for example with annotated time-stamps, Praat etc.)
- model static durations
- use pairwise relations
- focus on *isochrony*, equal timing of durations

- ignore the essential property of *rhythm*, which is *alternation of units with approximately equal duration*
- in other words: *oscillation*

**There is an alternative: speech signal analysis**

# *Rhythms in Amplitude AND Frequency Modulation*



# *Rhythm as Physical Oscillation*

## *Amplitude Modulation and Frequency Modulation*

INFORMATION:
SENDER

SIGNAL

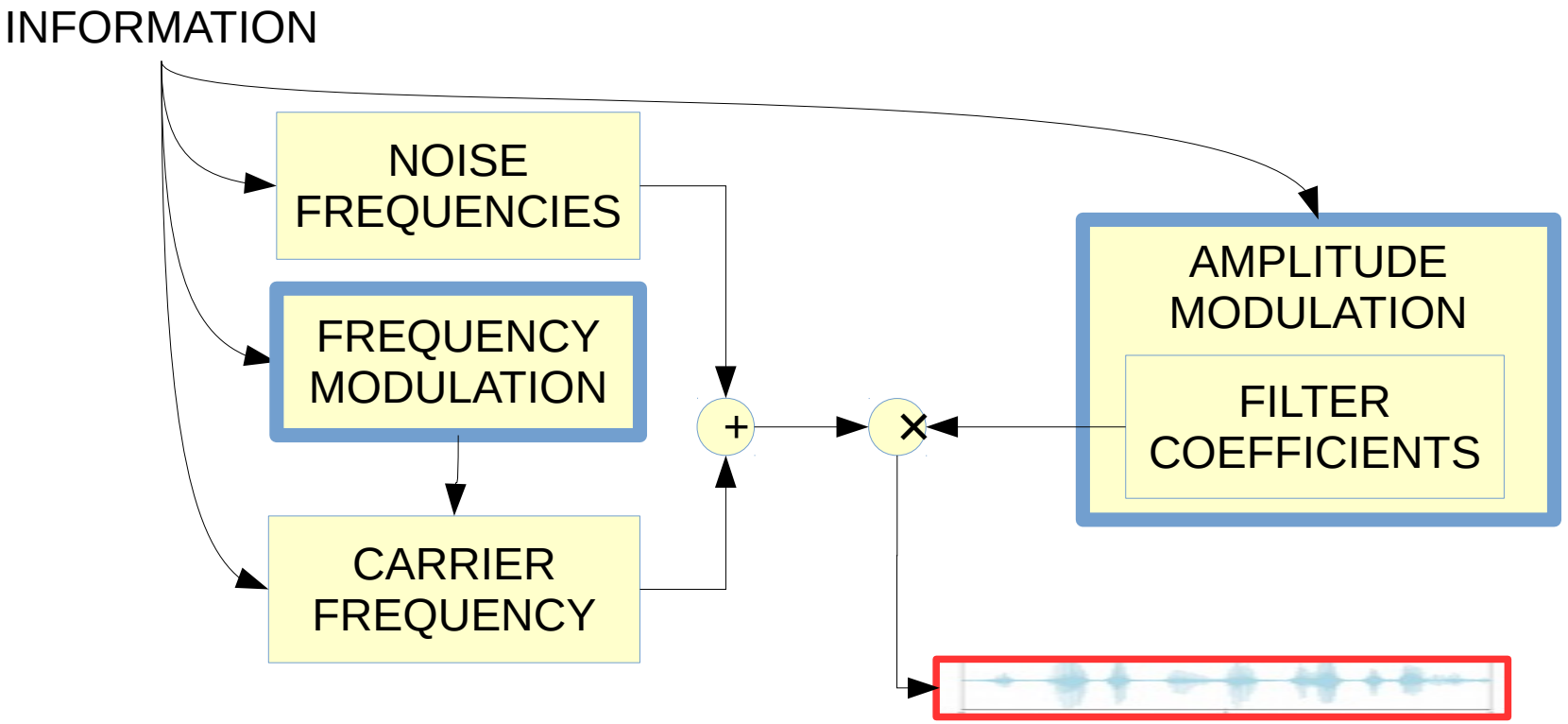
INFORMATION:
RECEIVER

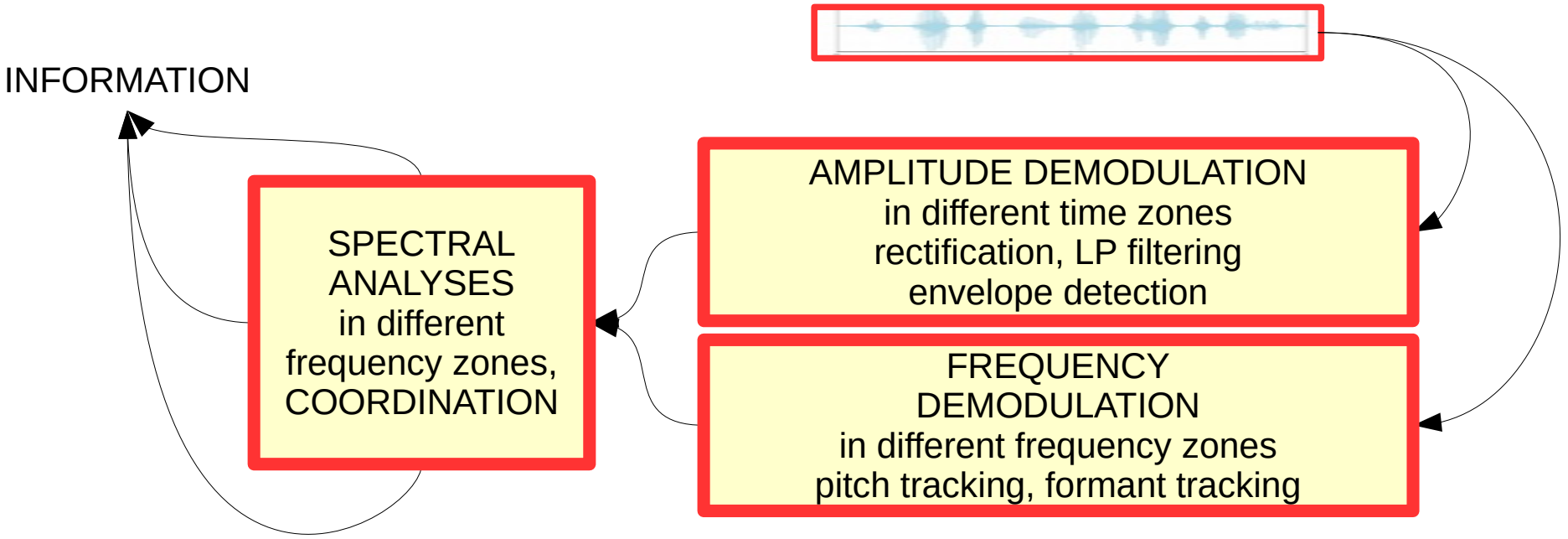
INFORMATION:
SENDER

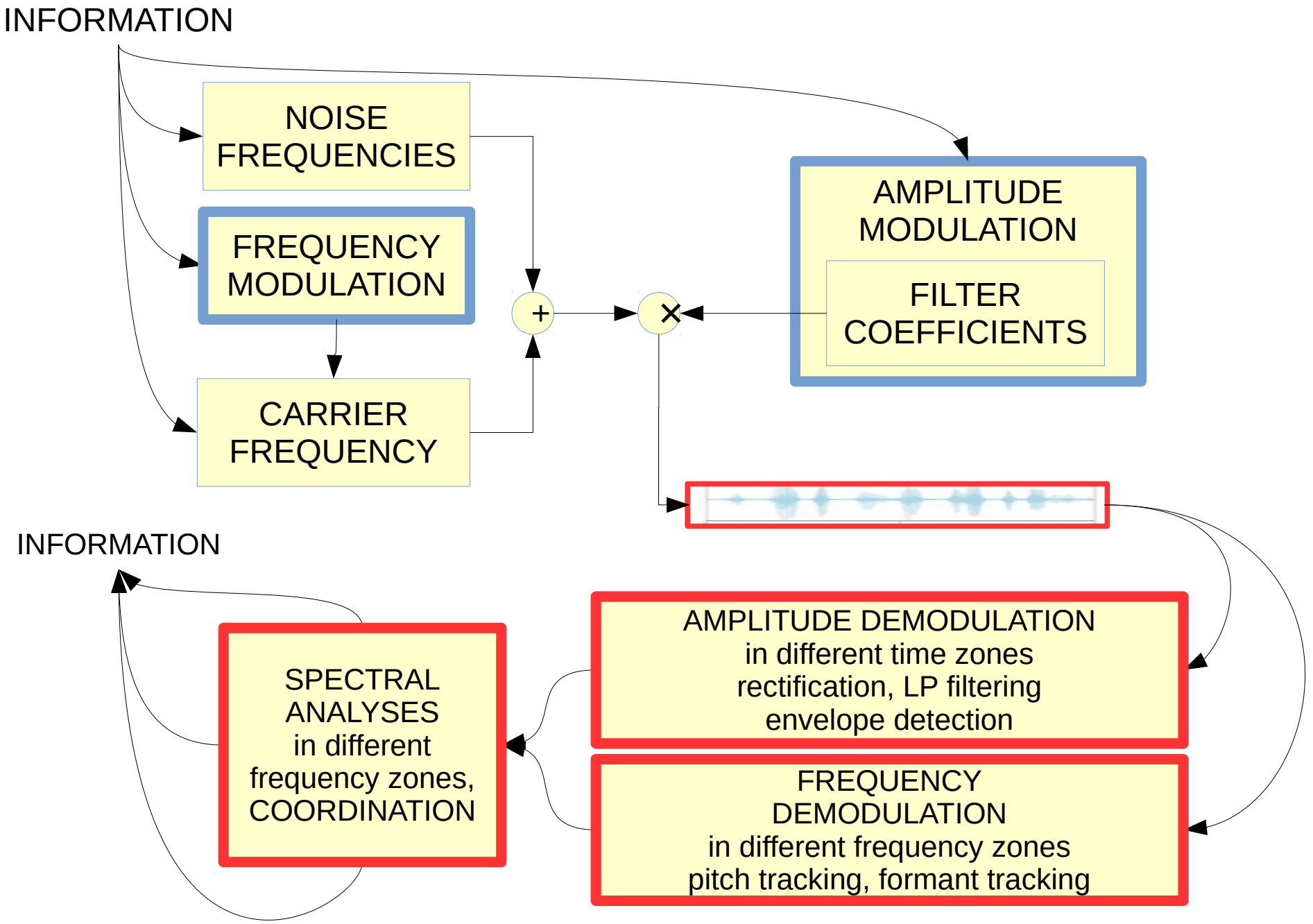
Which information? Semantics
Which structure? Grammar
Which code? Modulations

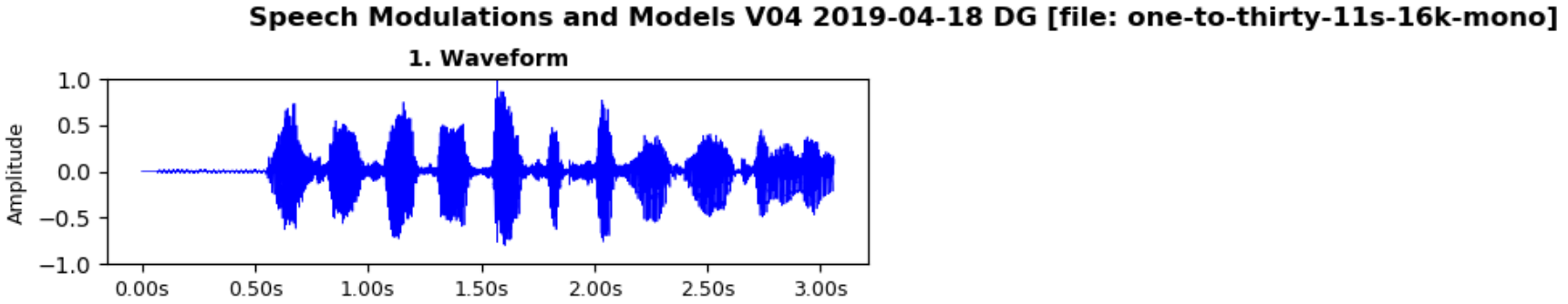
SIGNAL

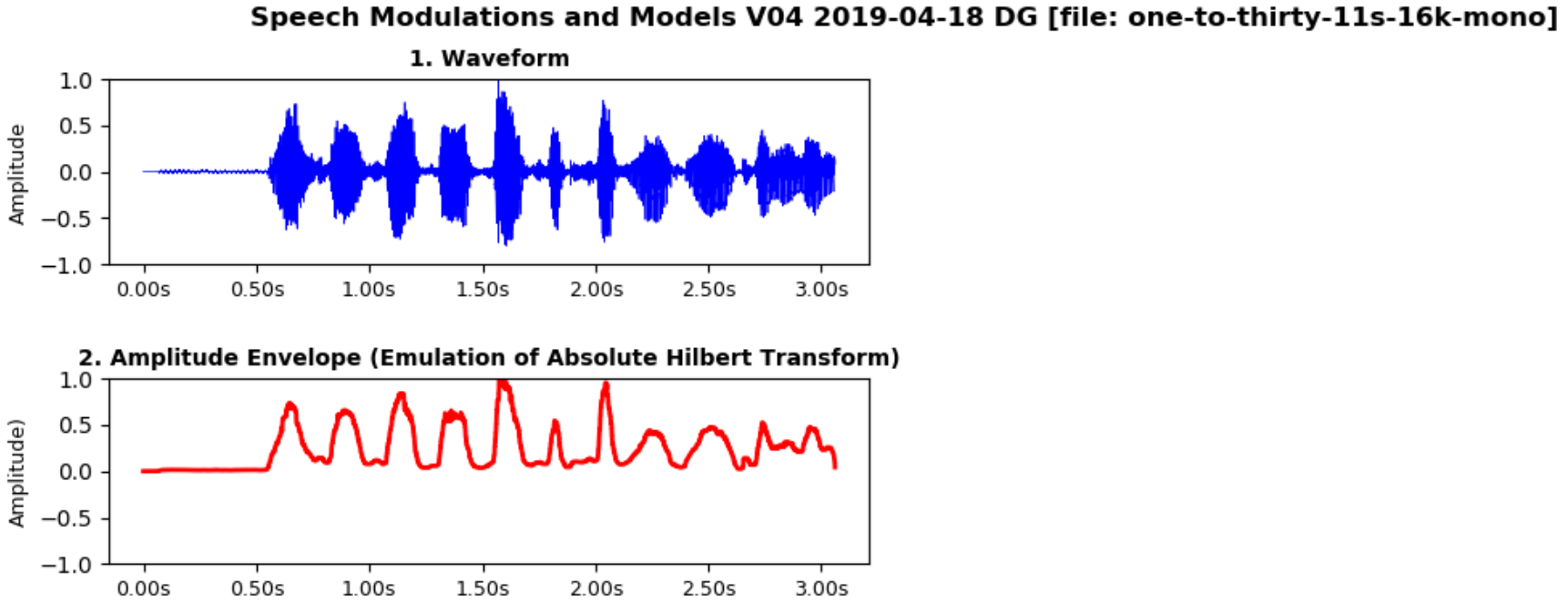
INFORMATION:
RECEIVER

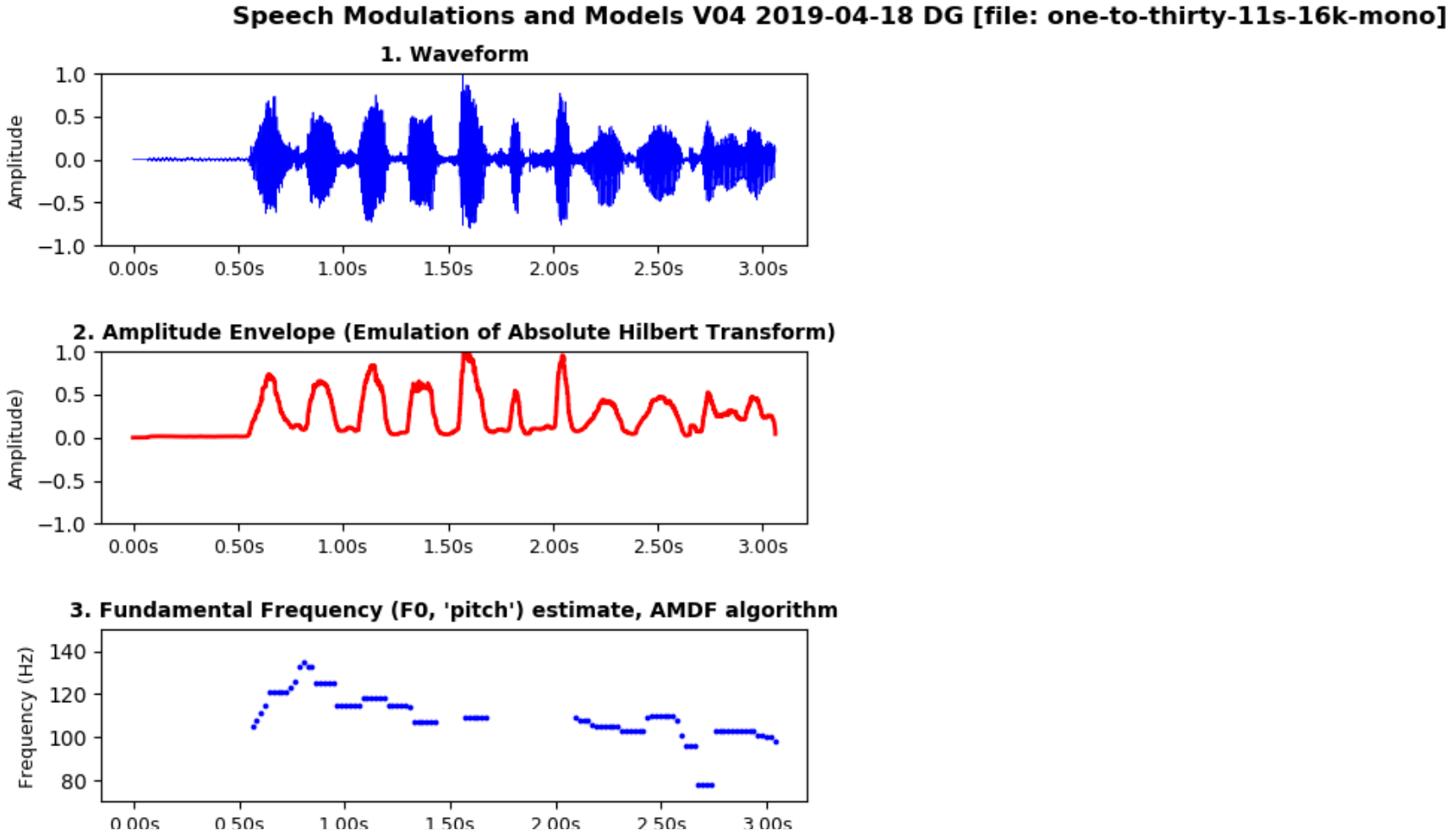

INFORMATION
NOISE FREQUENCIES
FREQUENCY MODULATION
CARRIER FREQUENCY
+
×
AMPLITUDE MODULATION
FILTER COEFFICIENTS

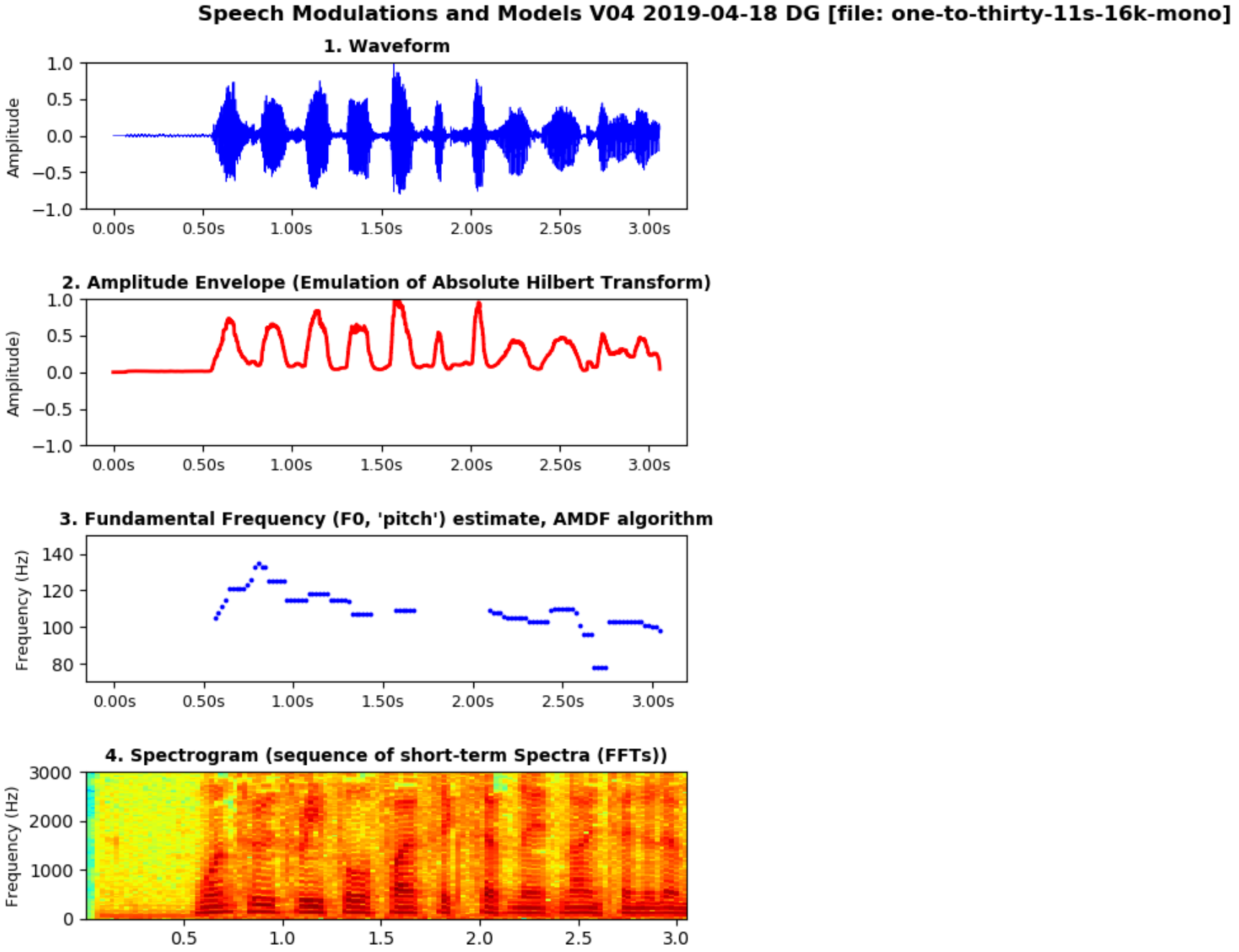

INFORMATION
SPECTRAL ANALYSES in different frequency zones, COORDINATION
AMPLITUDE DEMODULATION in different time zones rectification, LP filtering envelope detection
FREQUENCY DEMODULATION in different frequency zones pitch tracking, formant tracking


INFORMATION
NOISE FREQUENCIES
FREQUENCY MODULATION
CARRIER FREQUENCY
+
×
AMPLITUDE MODULATION
FILTER COEFFICIENTS
INFORMATION
SPECTRAL ANALYSES in different frequency zones, COORDINATION
AMPLITUDE DEMODULATION in different time zones rectification, LP filtering envelope detection
FREQUENCY DEMODULATION in different frequency zones pitch tracking, formant tracking


Speech Modulations and Models V04 2019-04-18 DG [file: one-to-thirty-11s-16k-mono]
1. Waveform
Amplitude
1.0
0.5
0.0
−0.5
−1.0
0.00s
0.50s
1.00s
1.50s
2.00s
2.50s
3.00s

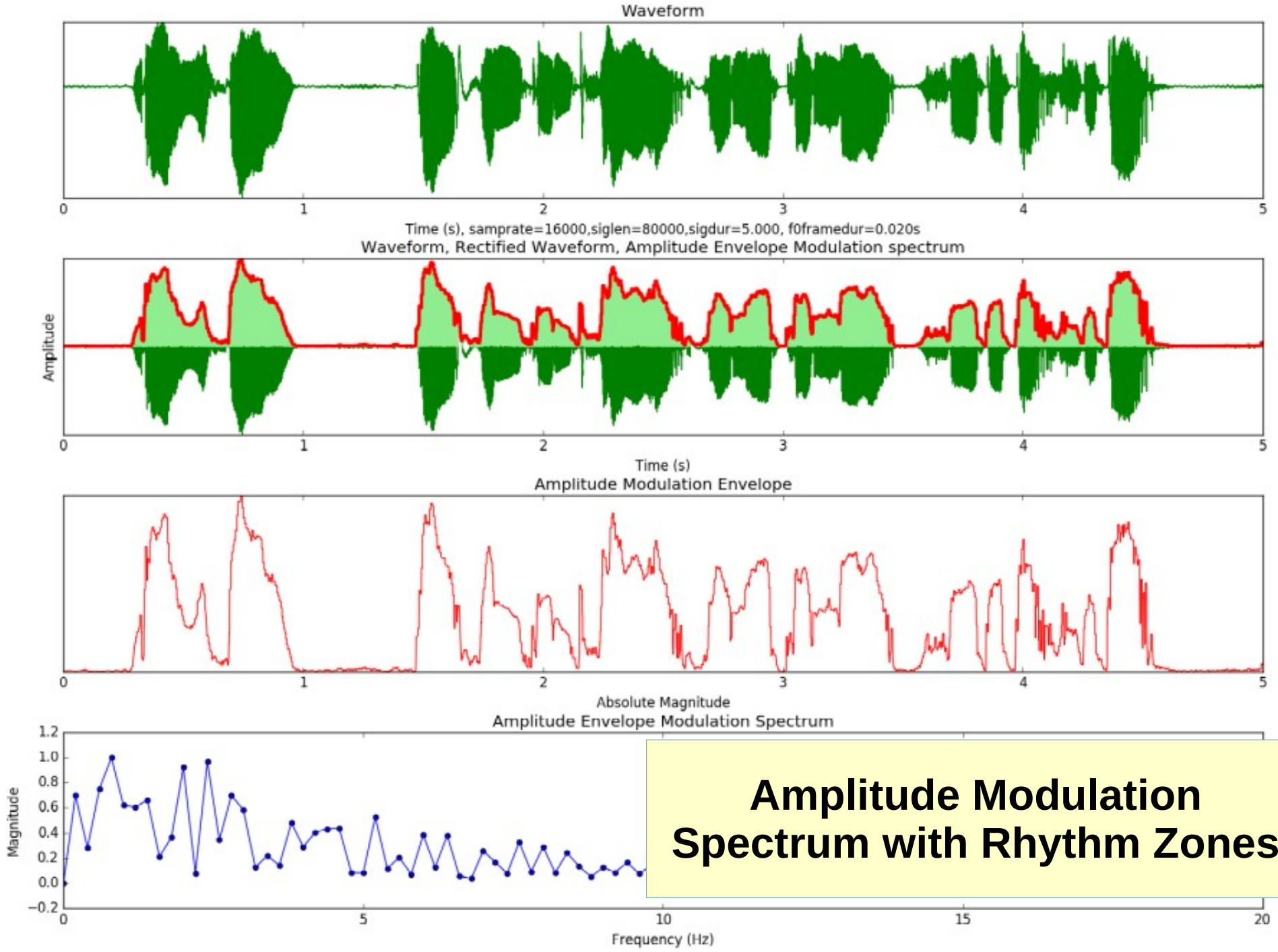

Speech Modulations and Models V04 2019-04-18 DG [file: one-to-thirty-11s-16k-mono]
1. Waveform
Amplitude
1.0
0.5
0.0
−0.5
−1.0
0.00s
0.50s
1.00s
1.50s
2.00s
2.50s
3.00s
2. Amplitude Envelope (Emulation of Absolute Hilbert Transform)
Amplitude)
1.0
0.5
0.0
−0.5
−1.0
0.00s
0.50s
1.00s
1.50s
2.00s
2.50s
3.00s


Speech Modulations and Models V04 2019-04-18 DG [file: one-to-thirty-11s-16k-mono]
1. Waveform
Amplitude
1.0
0.5
0.0
−0.5
−1.0
0.00s
0.50s
1.00s
1.50s
2.00s
2.50s
3.00s
2. Amplitude Envelope (Emulation of Absolute Hilbert Transform)
Amplitude)
3. Fundamental Frequency (F0, 'pitch') estimate, AMDF algorithm
Frequency (Hz)
140
120
100
80


Speech Modulations and Models V04 2019-04-18 DG [file: one-to-thirty-11s-16k-mono]
1. Waveform
Amplitude
1.0
0.5
0.0
−0.5
−1.0
0.00s
0.50s
1.00s
1.50s
2.00s
2.50s
3.00s
2. Amplitude Envelope (Emulation of Absolute Hilbert Transform)
Amplitude)
3. Fundamental Frequency (F0, 'pitch') estimate, AMDF algorithm
Frequency (Hz)
140
120
100
80
4. Spectrogram (sequence of short-term Spectra (FFTs))
Frequency (Hz)
3000
2000
1000
0
0.5
1.0
1.5
2.0
2.5
3.0

# *The Reality of Rhythm!*

# *How to measure physical rhythm*




**Amplitude Modulation Spectrum with Rhythm Zones**

# The Role of F0

If a spectrum can be derived from the **AM envelope**,

why not derive a spectrum from the **FM track**

and see whether they correlate?

**Preliminary answer:**

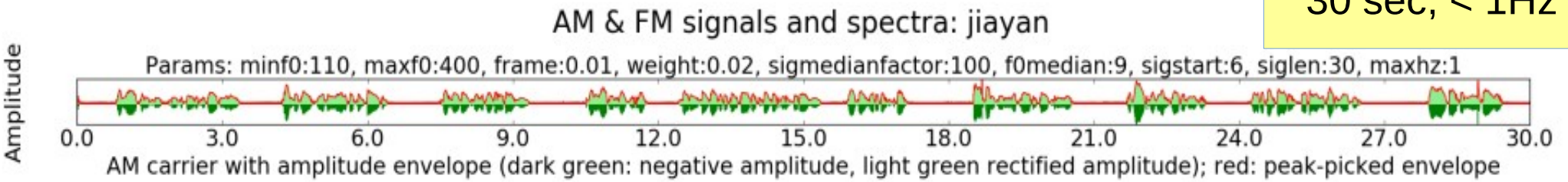
Yes, they do correlate to some extent, but not overwhelmingly strongly!

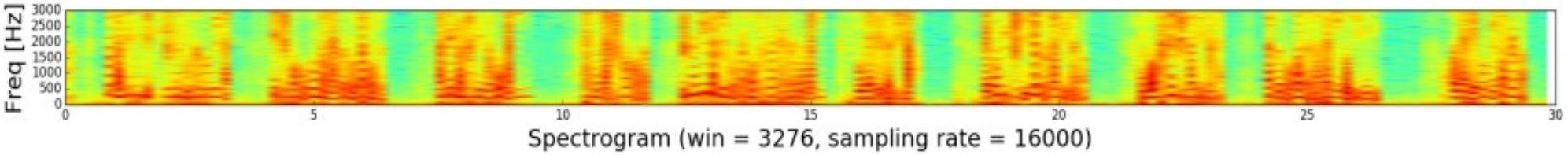
This is not very surprising, of course, since they are partly co-extensive locally and globally, though locally not too similar.

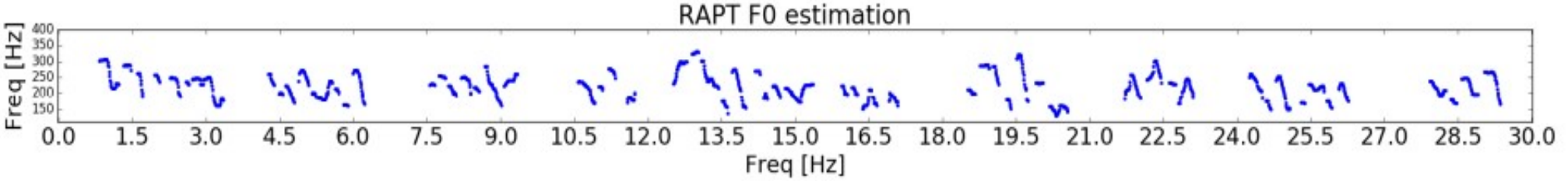
I will look at both AM and FM spectra.

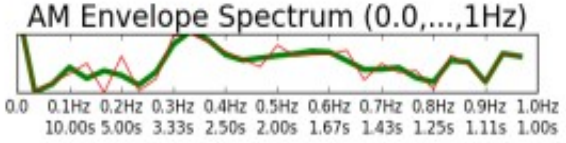
# *AM and FM Demodulation*

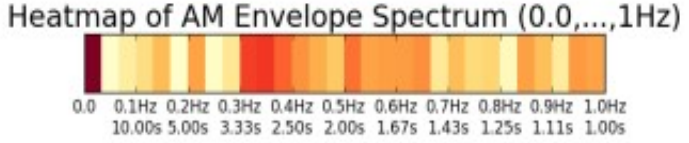
Mandarin, female
30 sec, < 1Hz

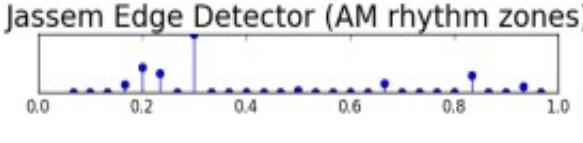


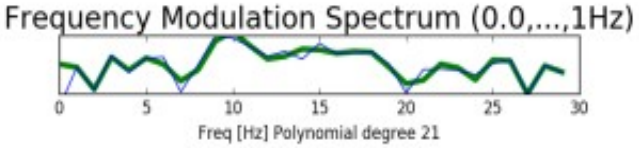



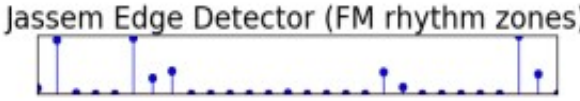



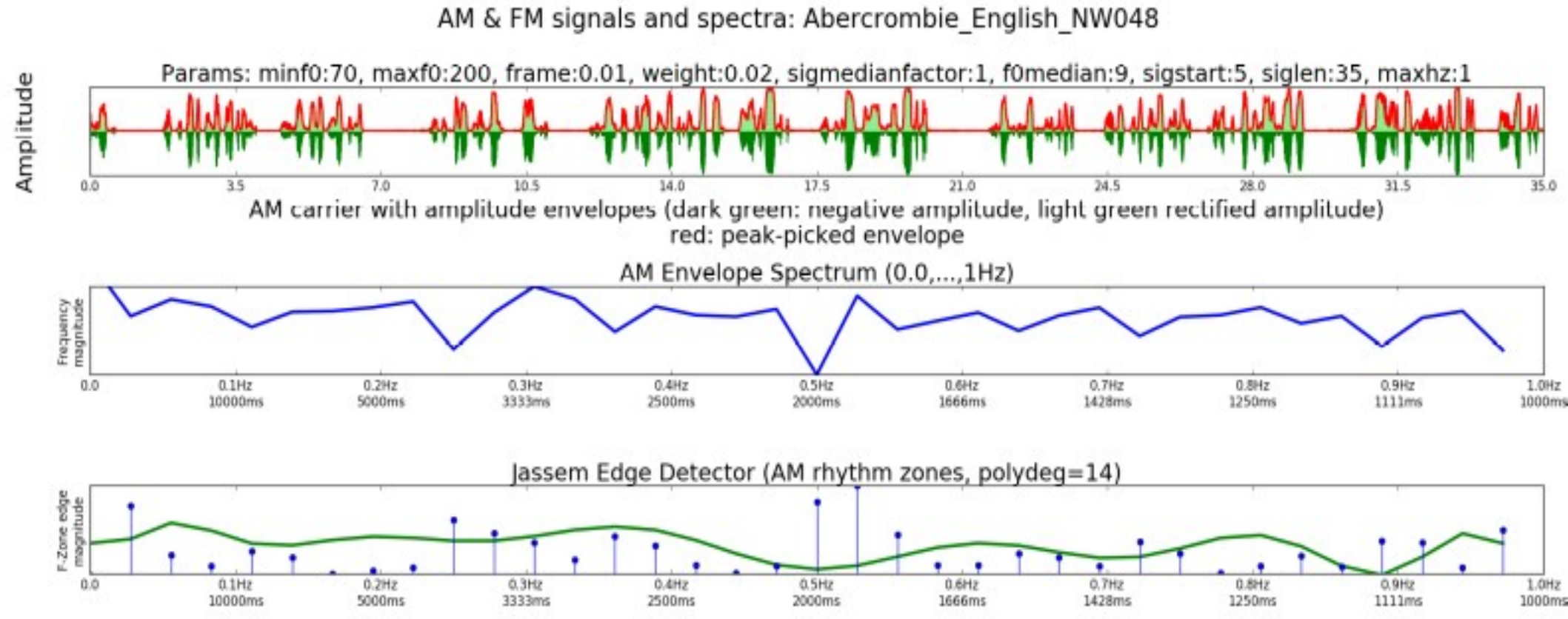





Correlation AME:FME=0.74
Correlation AMS:FMS=0.29

# *Envelope Demodulation: Extending to Discourse Spectra*

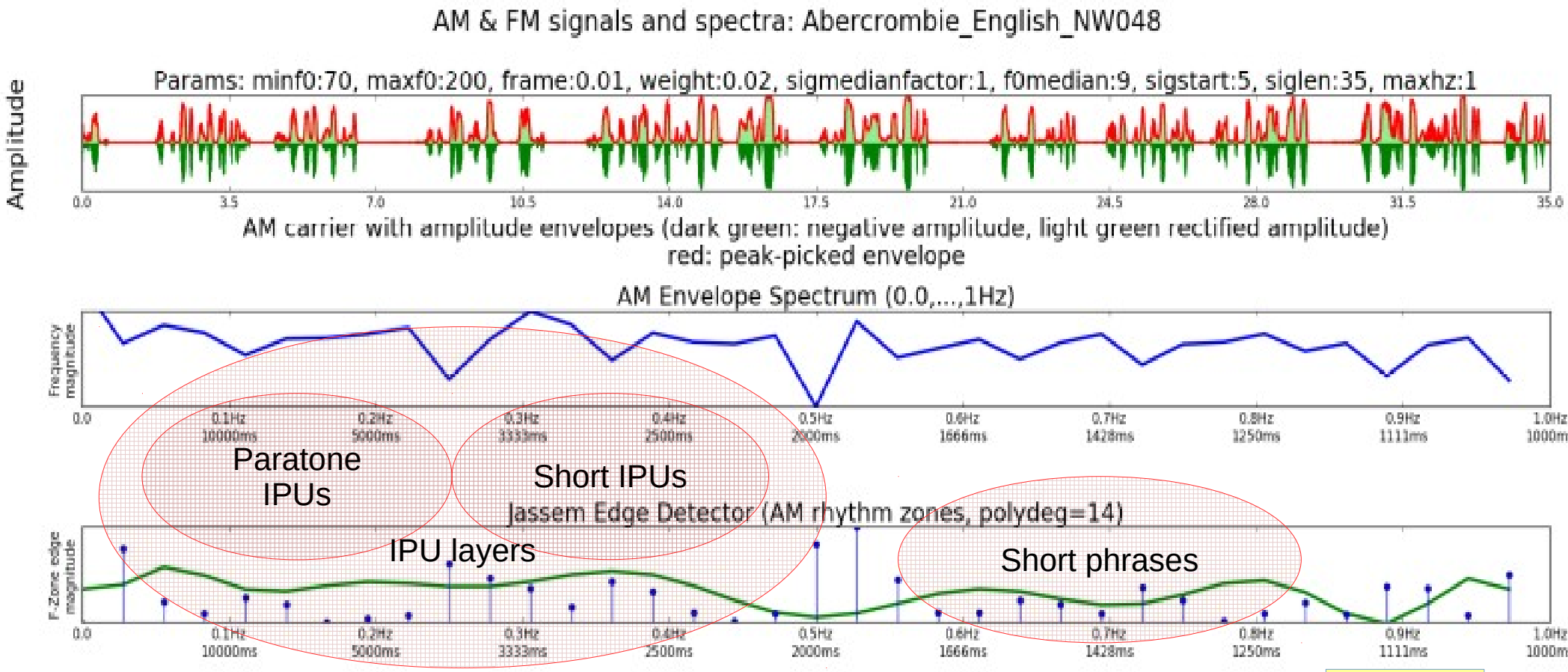
## English (RP) Edinburgh corpus *"The North Wind and the Sun"*




1 Hz

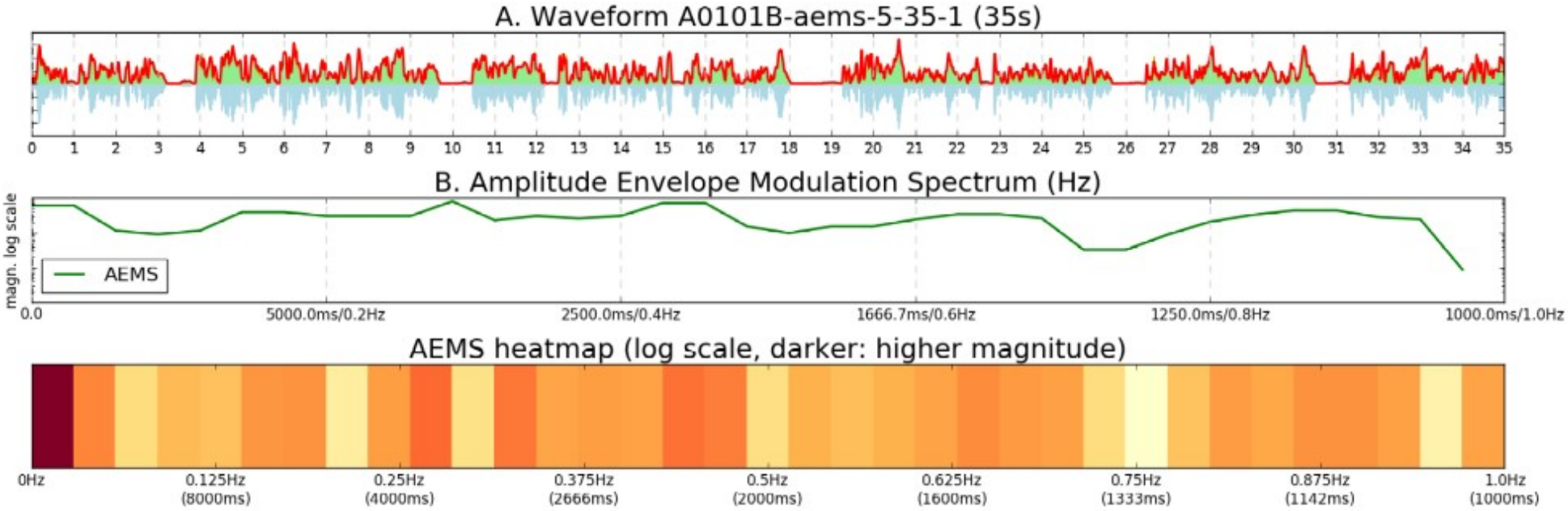
# *Envelope Demodulation: Extending to Discourse Spectra*

## English (RP) Edinburgh corpus *"The North Wind and the Sun"*



# *Extending to Discourse Spectra: English Genres*

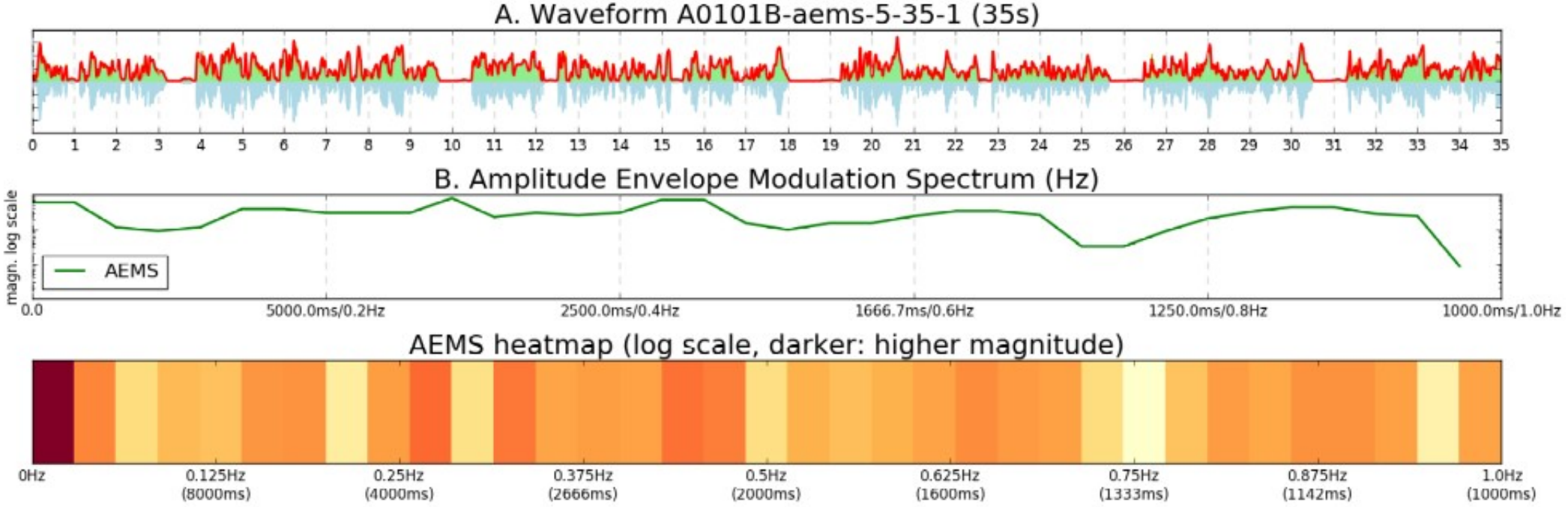
**English Newsreading**

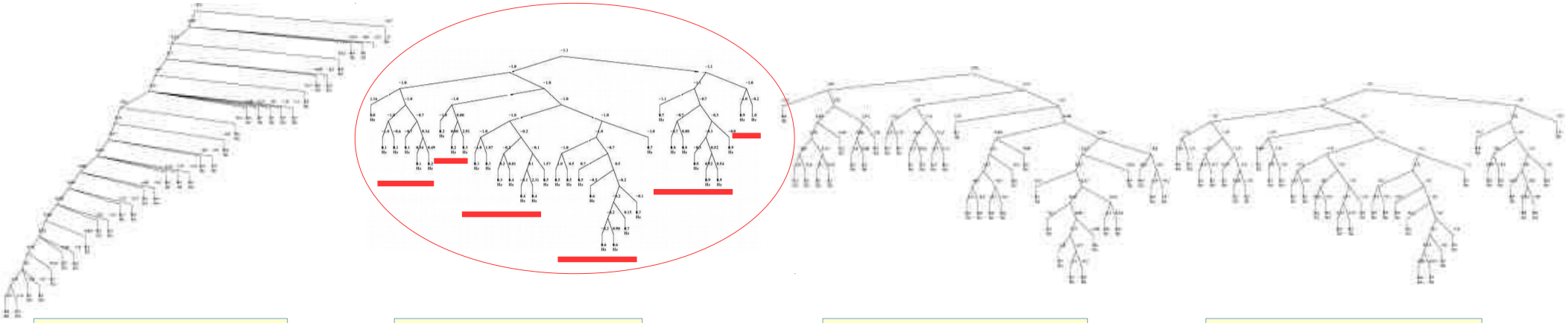



**Spectral Rhythm Zone Boundaries (lighter colours)**

# *AM and FM Demodulation and Spectral Tree Induction*

## English Newsreading

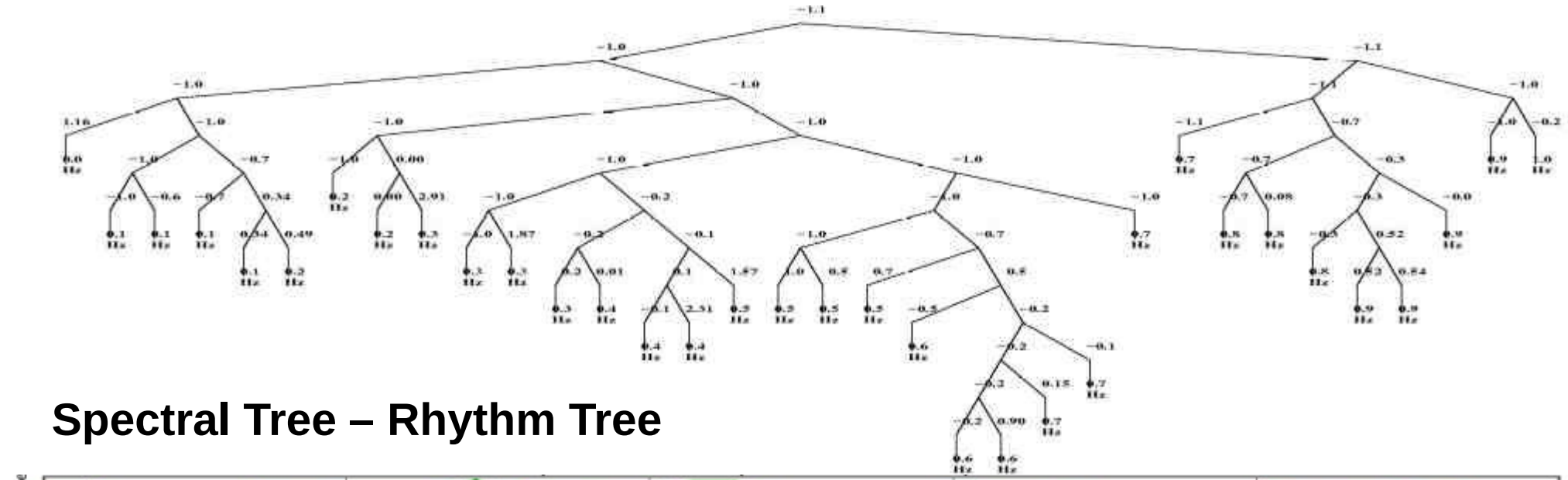


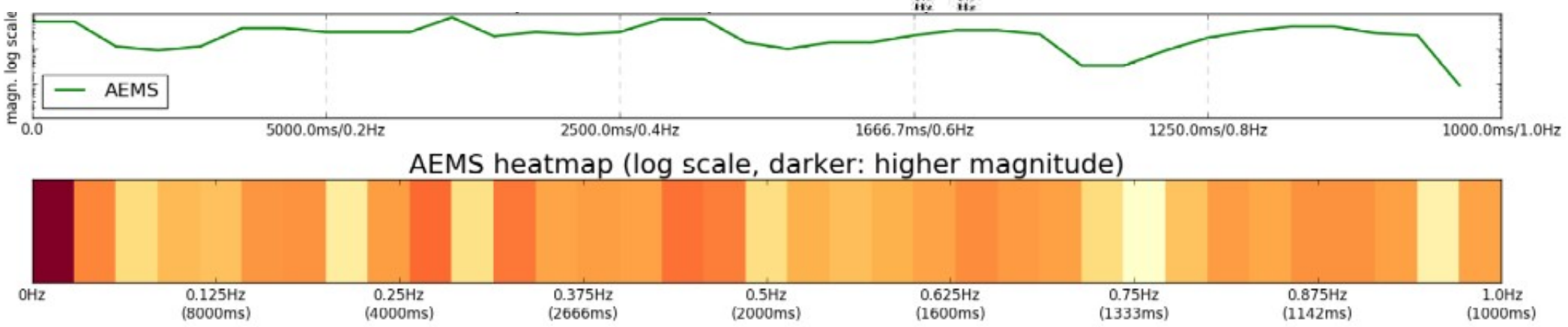
L-strong, > | L-strong, < | R-strong, > | R-strong, <

# *AM and FM Demodulation and Spectral Tree Induction*

**English Newsreading**



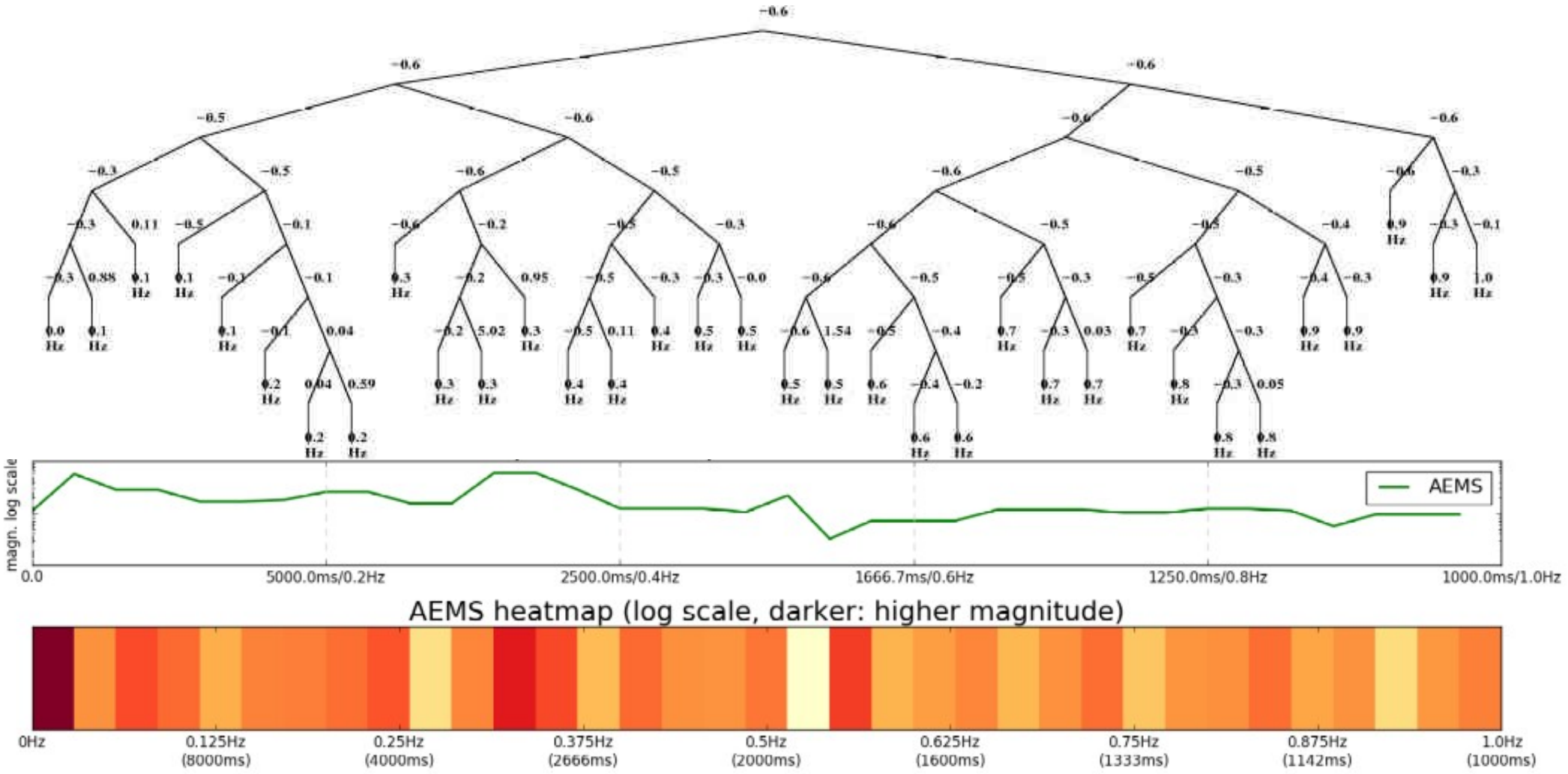
## Spectral Tree – Rhythm Tree




L-strong, <

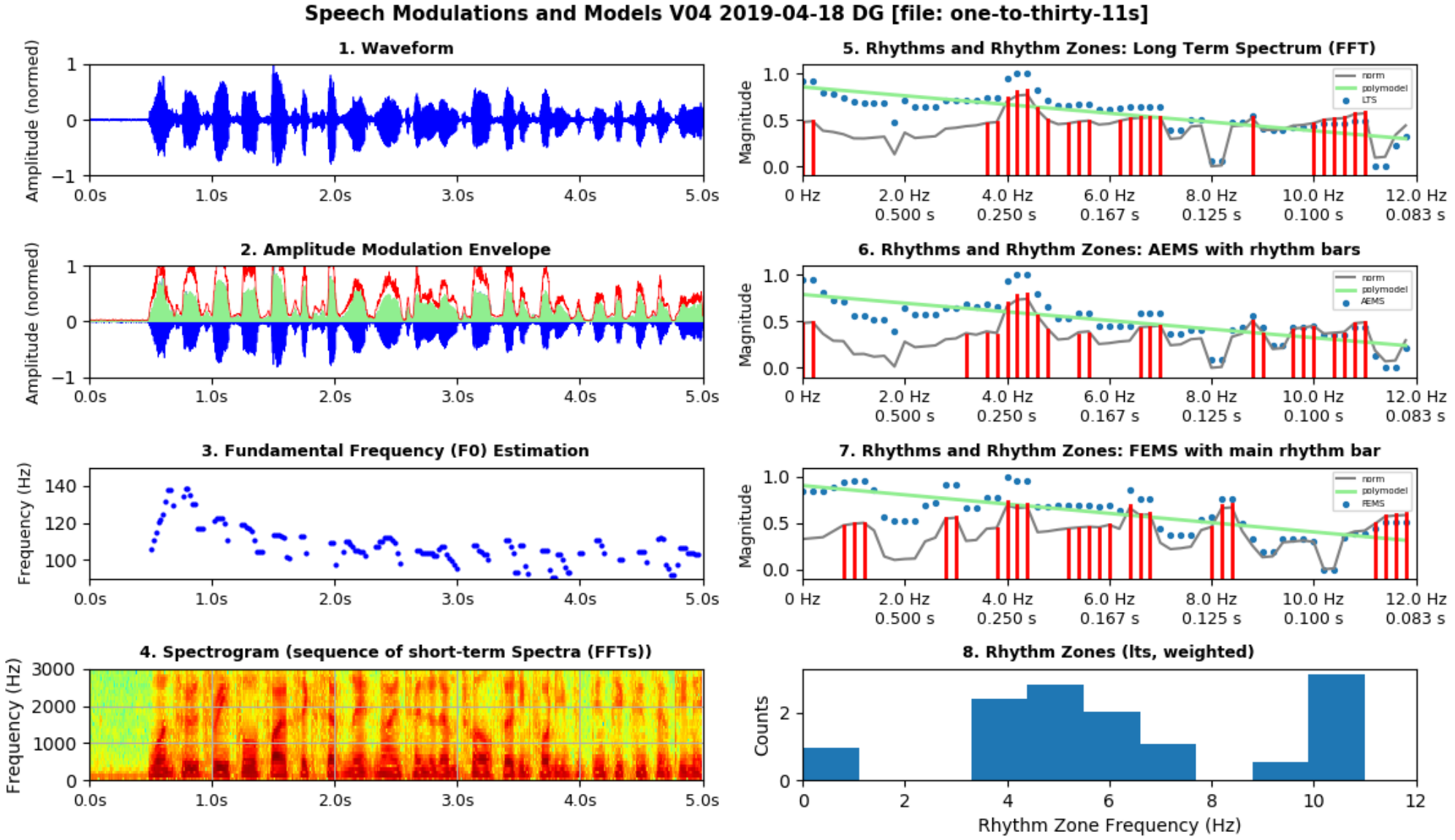
# *AM and FM Demodulation and Spectral Tree Induction*

## English (RP) Edinburgh corpus *"The North Wind and the Sun"*

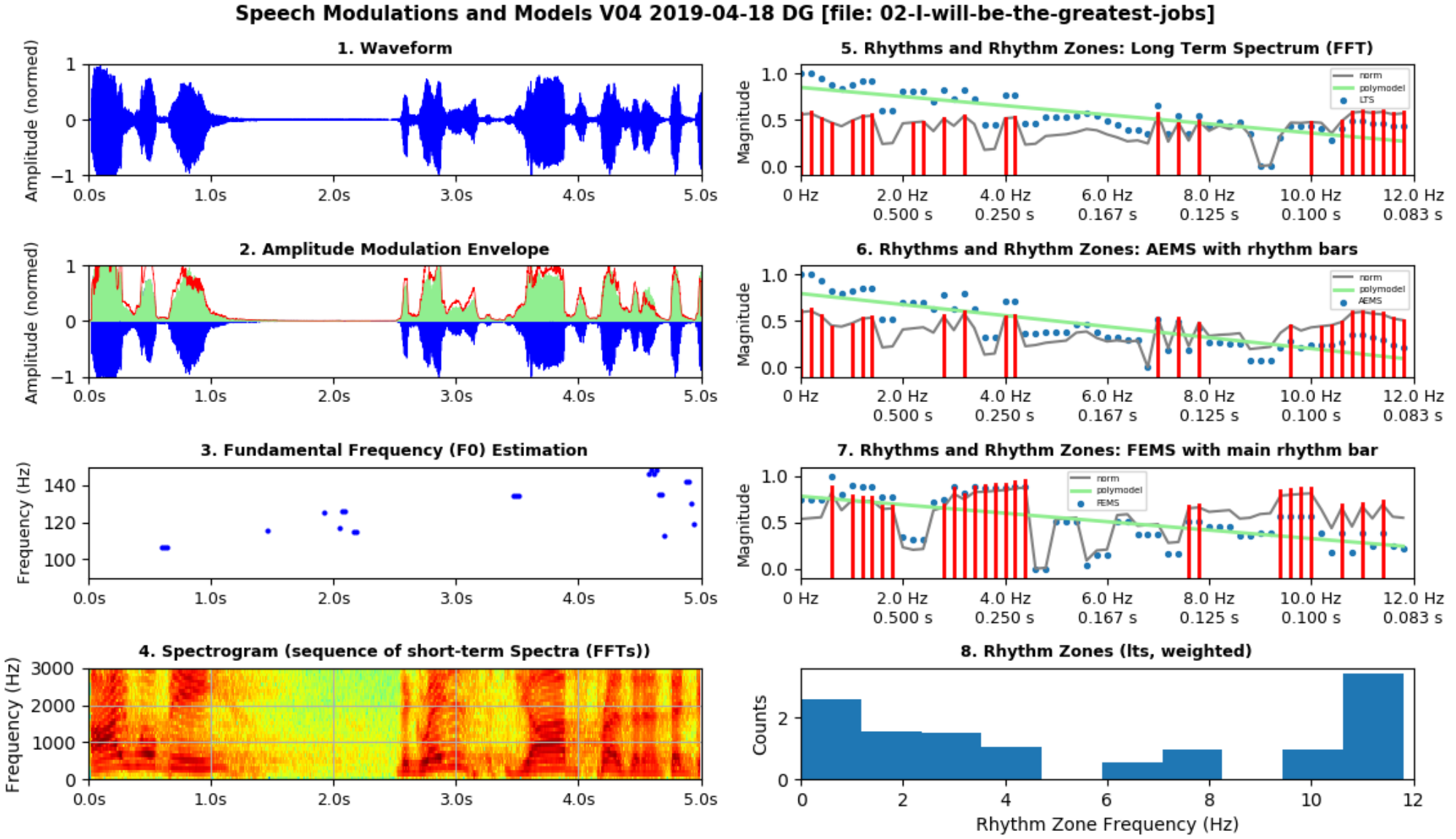



L-strong, <

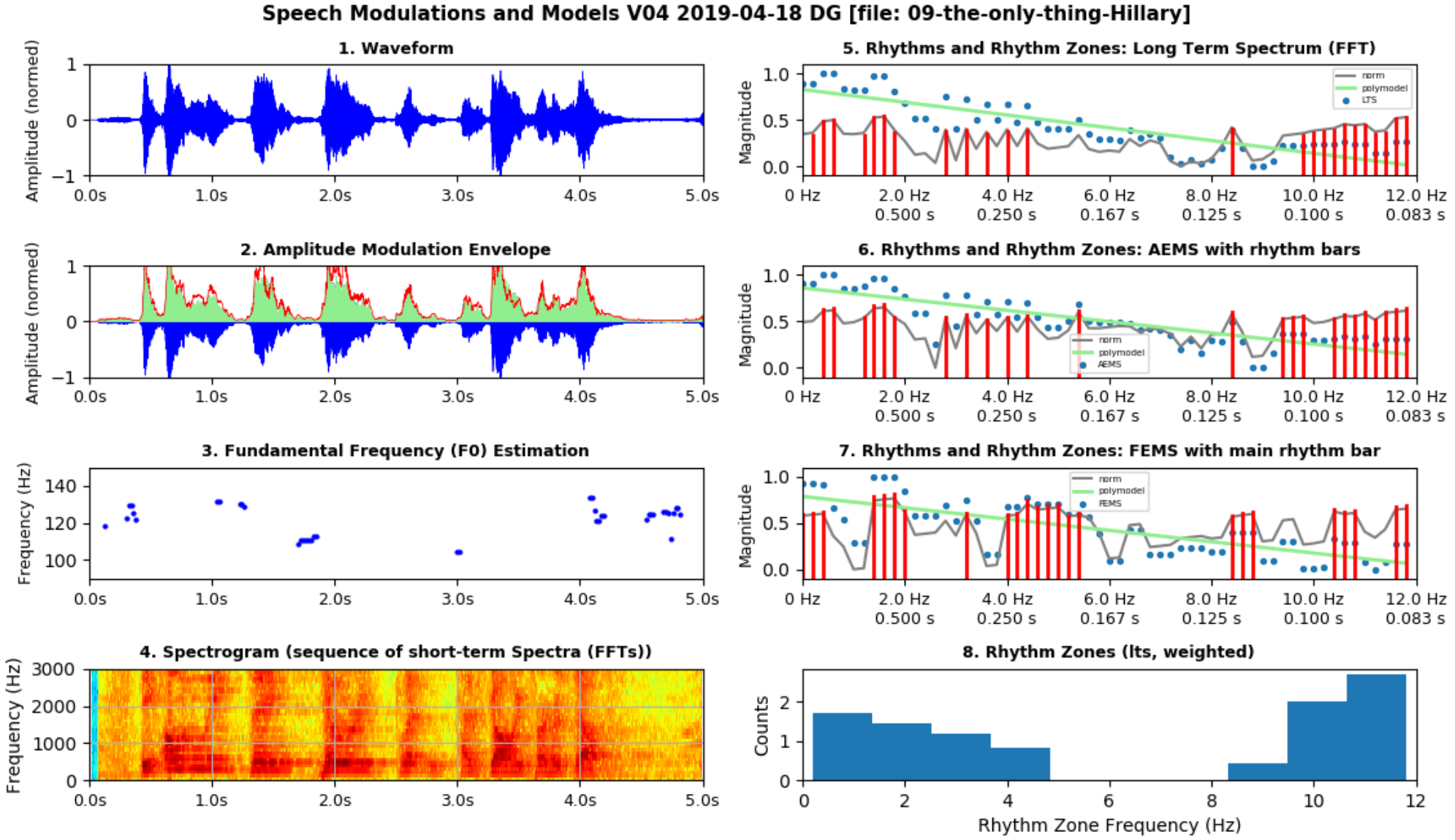
# *Rhythm Zones*



# *Rhythm Zones*



# *Rhythm Zones*



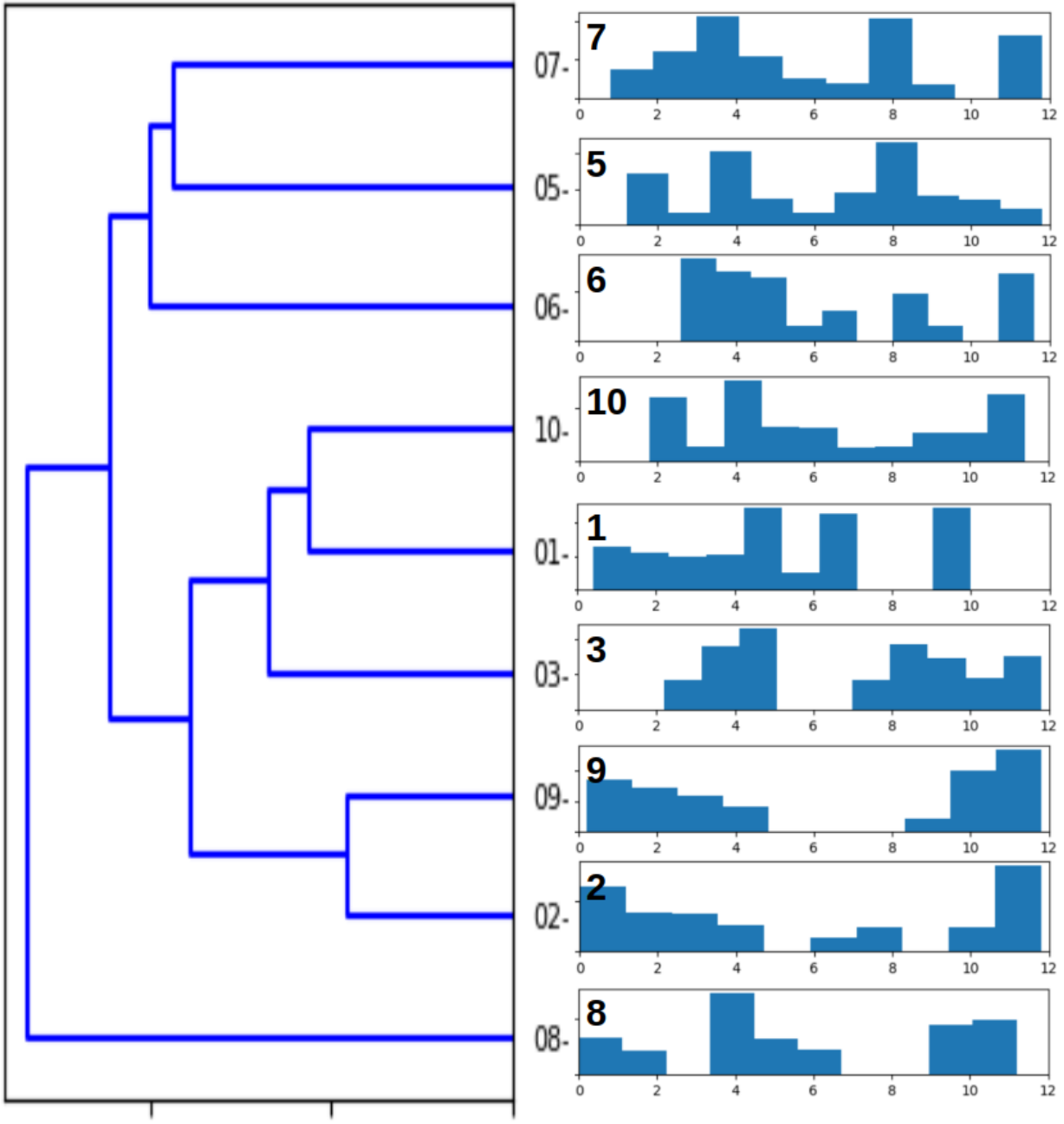
# *Classification of Utterances by Rhythm Zone Similarity*

- Data: rhythm zone vectors
- Calculate Manhattan Distance (differences)

$$d_i(p,q)=\sum_1^n |p_i-q_i|$$

- Group distances by Average Pair Group Method with Arithmetic Mean (UPGMA)

$$d_i(p,q)=\sum_{i,j}\frac{dist\,(p_i,q_j)}{|p|\times|q|}$$



# 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

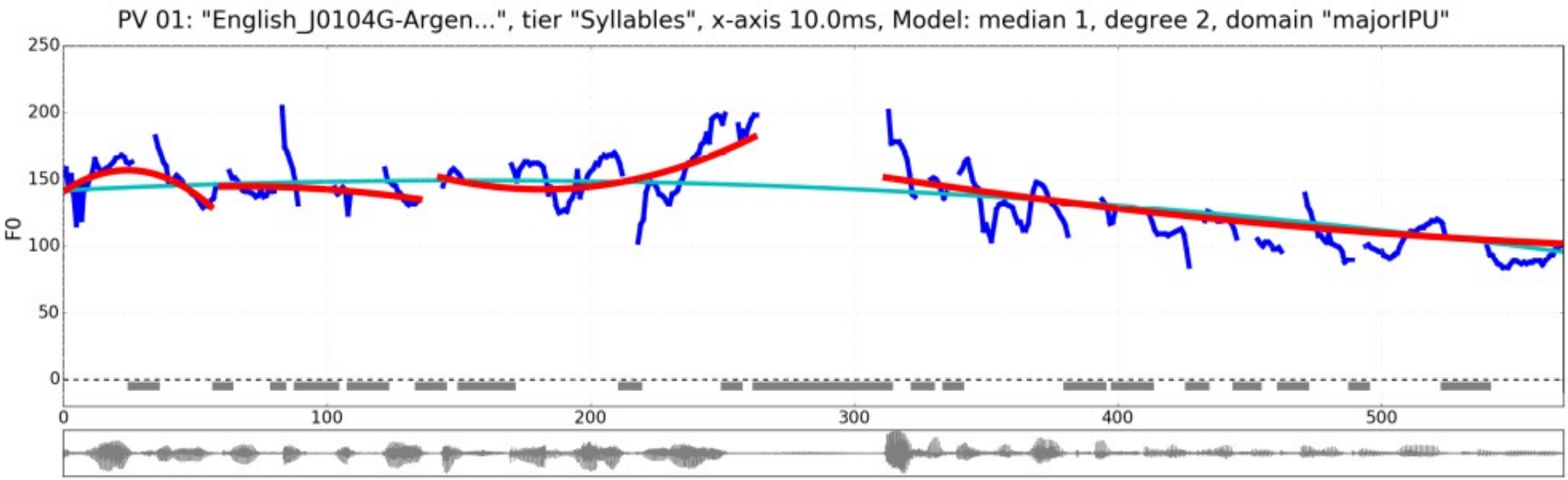
**Note: in infant speech,**

- **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 ☺

# *Discourse Rhythms: Long FM contours*

**Question:**
**rising utterance contour**

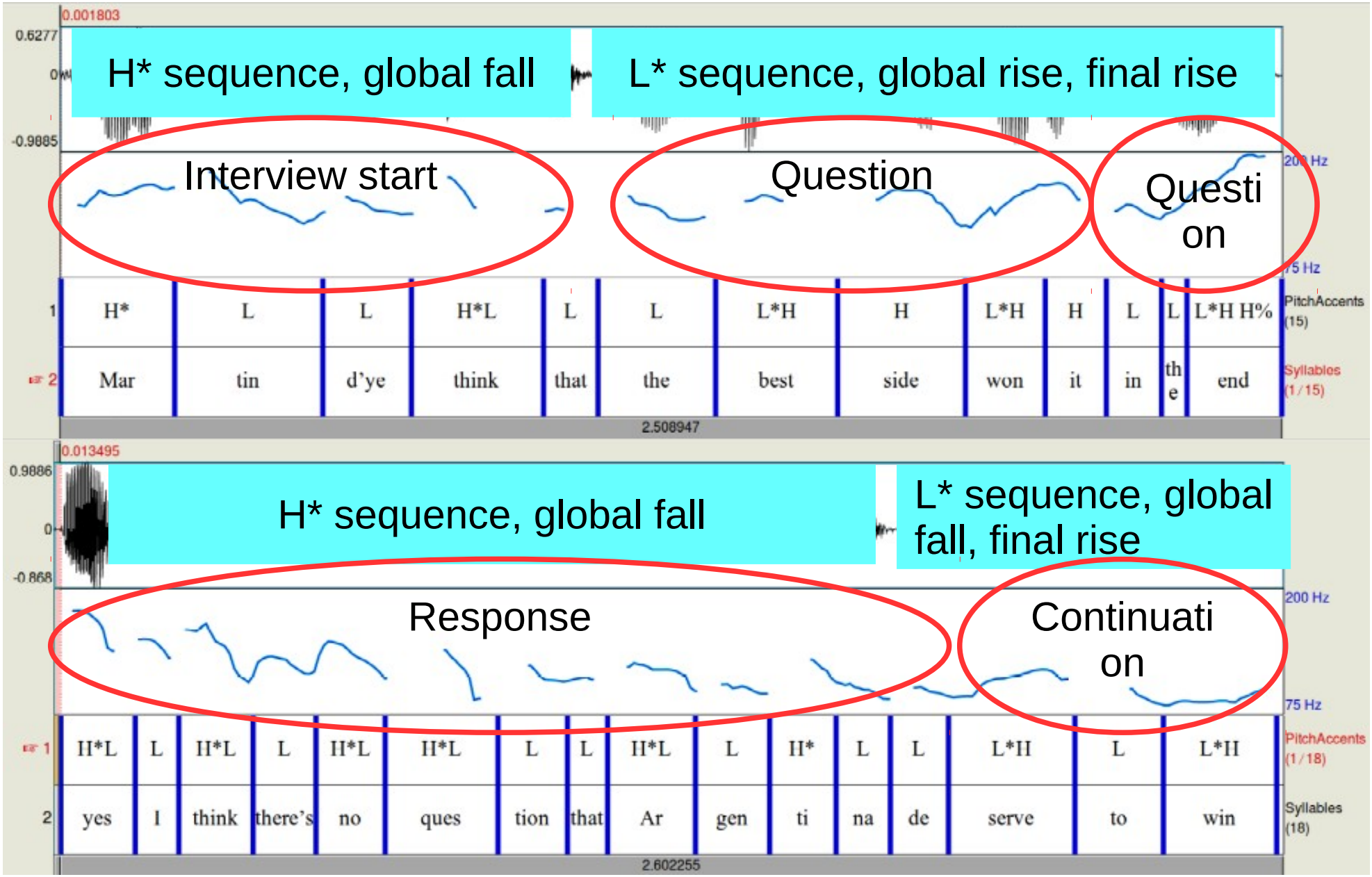
**Answer:**
**falling utterance contour**




Question+Answer: rising-falling adjacency pair contour

***syntagmatic entrainment***

# *Discourse Rhythms: Long FM contours*

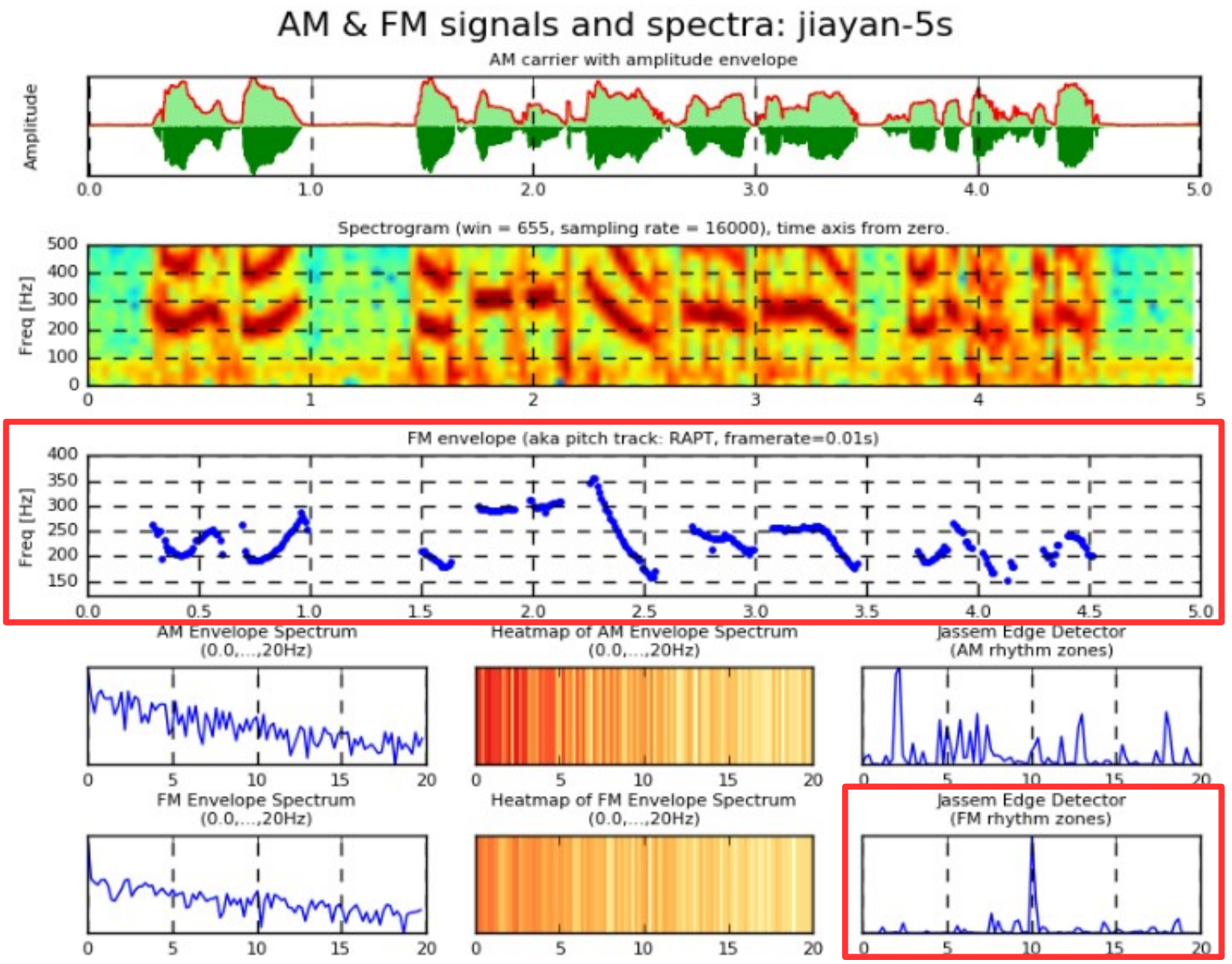


## *But there are Methodological Problems in F0 / Pitch estimation*

1. Terminology:
   - *articulation rate* (production)
   - *F0* (acoustic transmission)
   - *pitch* (perception)
2. Measurement:
   - F0 estimation implementations yield slightly different results
     - Autocorrelation
     - Normalised Cross-correlation
     - Average Magnitude Difference Function (AMDF)
     - FFT peak detection
     - Cepstrum
   - Environment differences
     - Preprocessing: low-pass filter; centre-clipping
     - Postprocessing: moving median

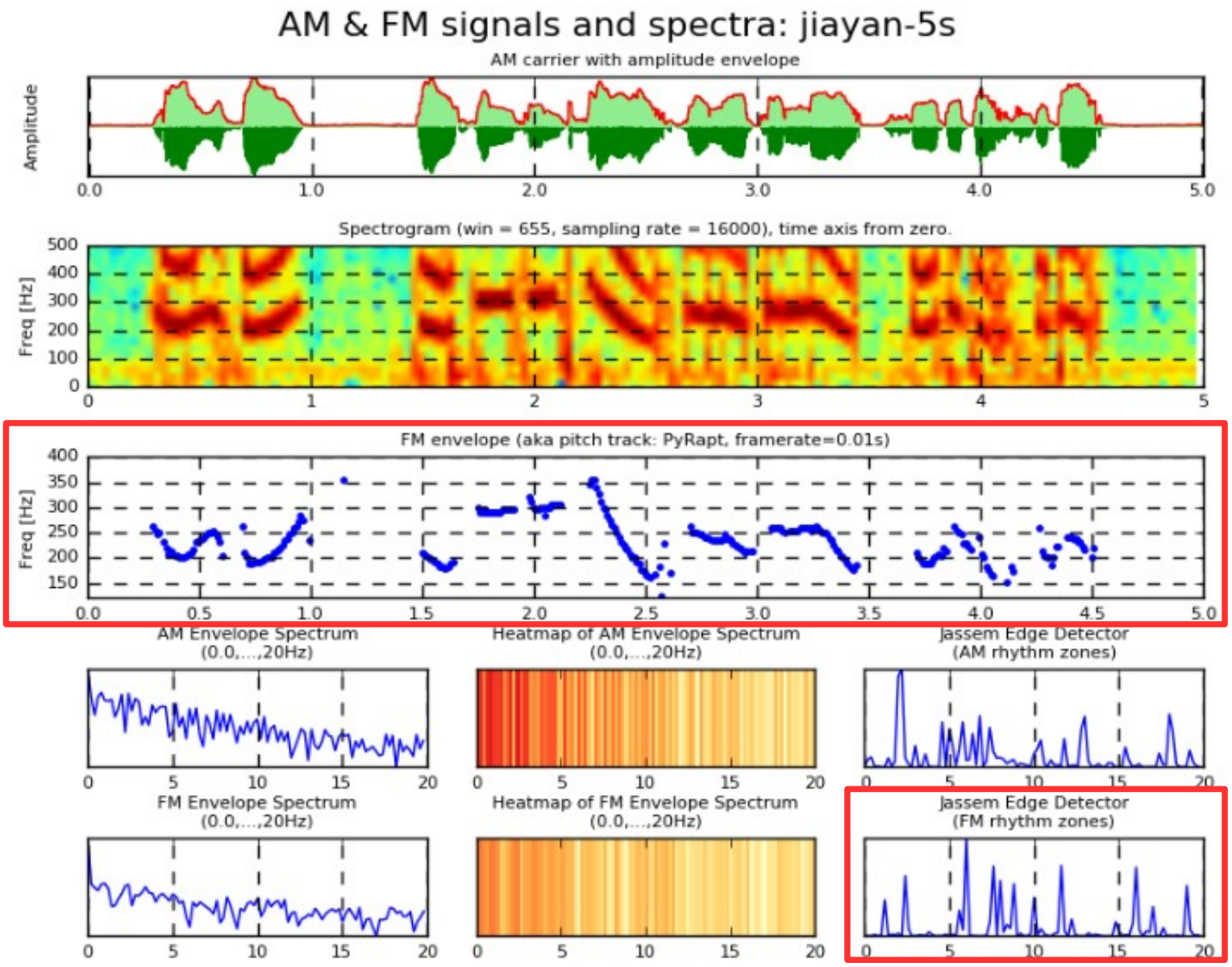
# *RAPT (Robust Algorithm for Pitch Tracking)*

David Talkin



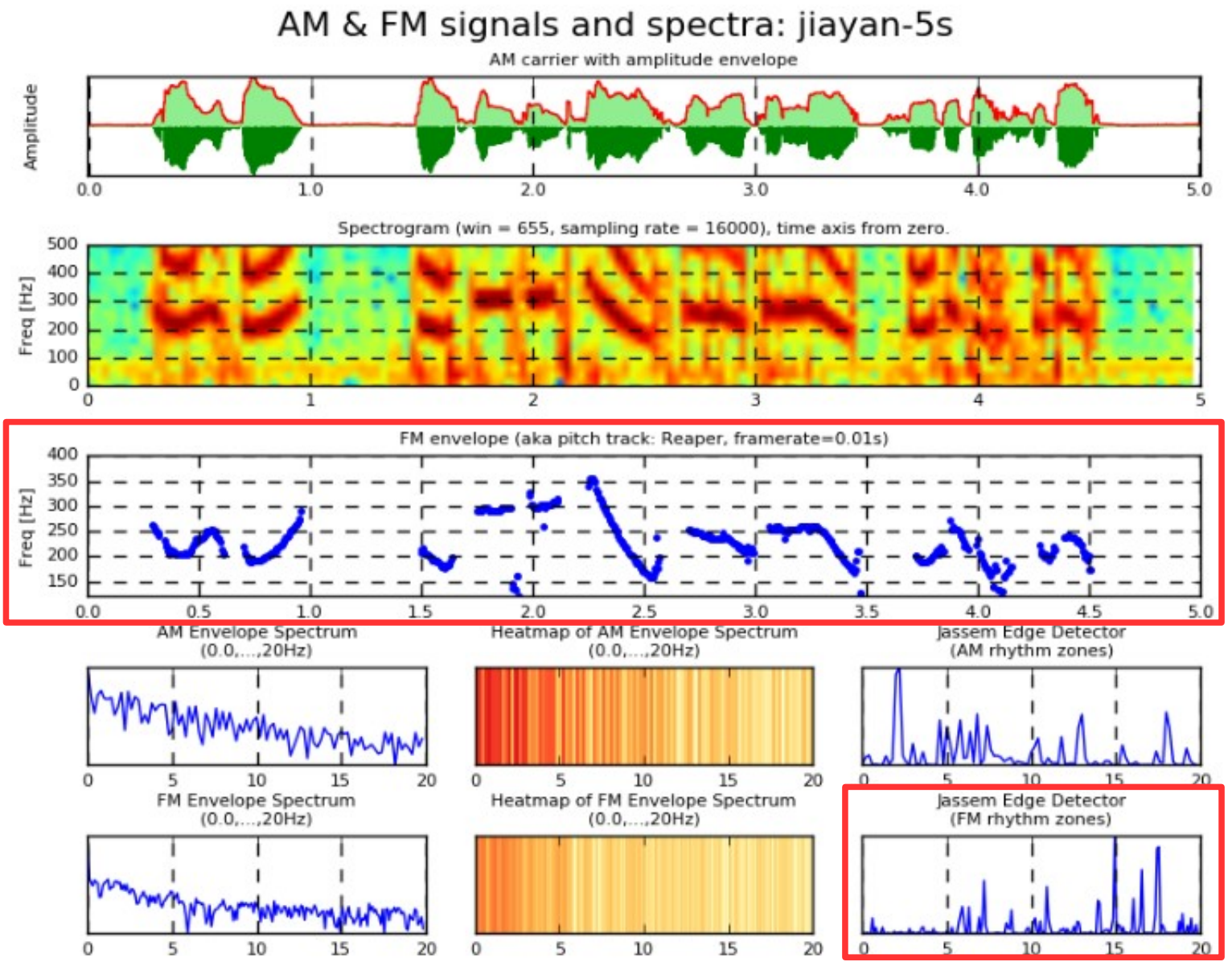
# *RAPT (Python emulation)*

Daniel Gaspari



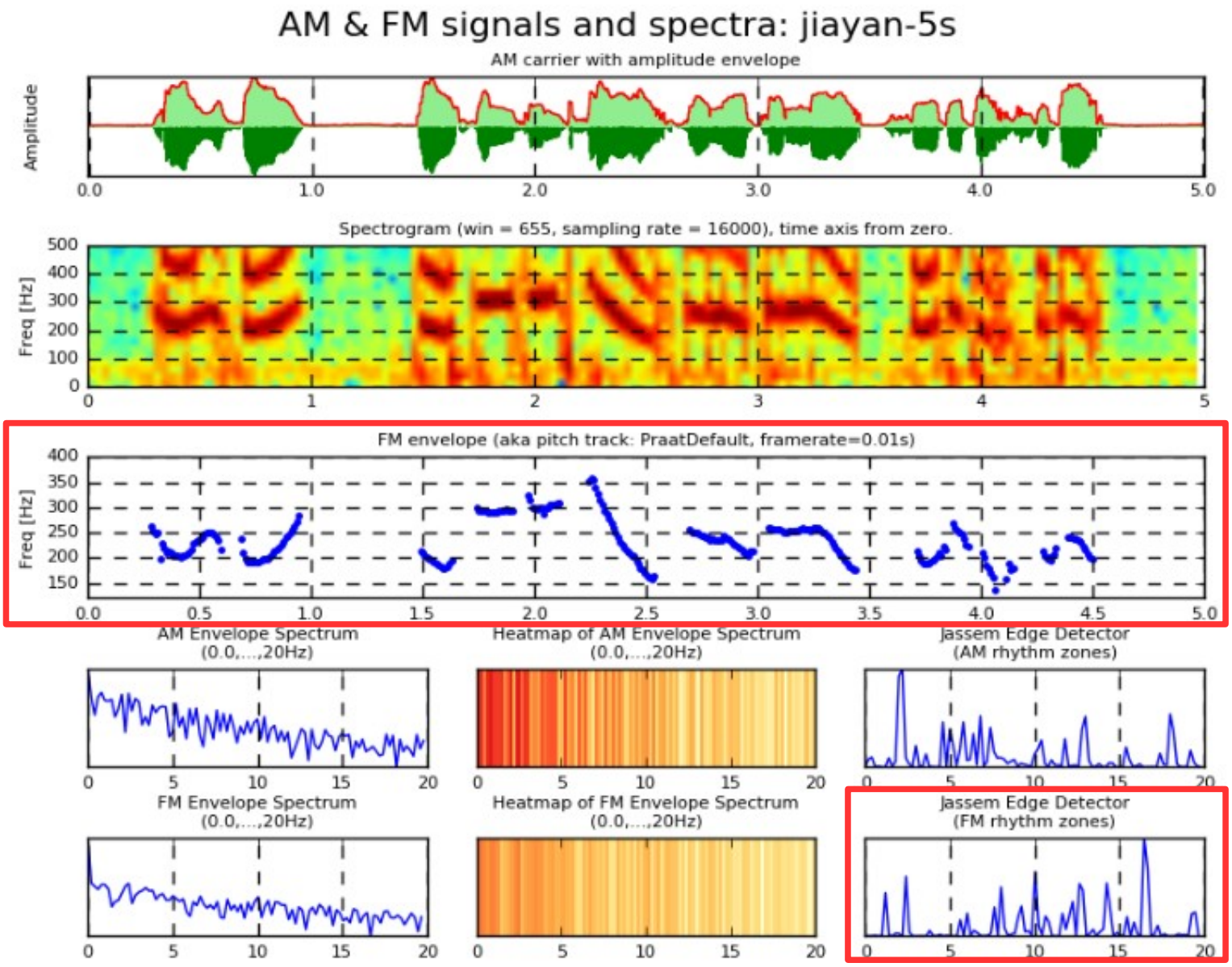
# *Reaper (Robust Epoch And Pitch EstimatoR)*

David Talkin



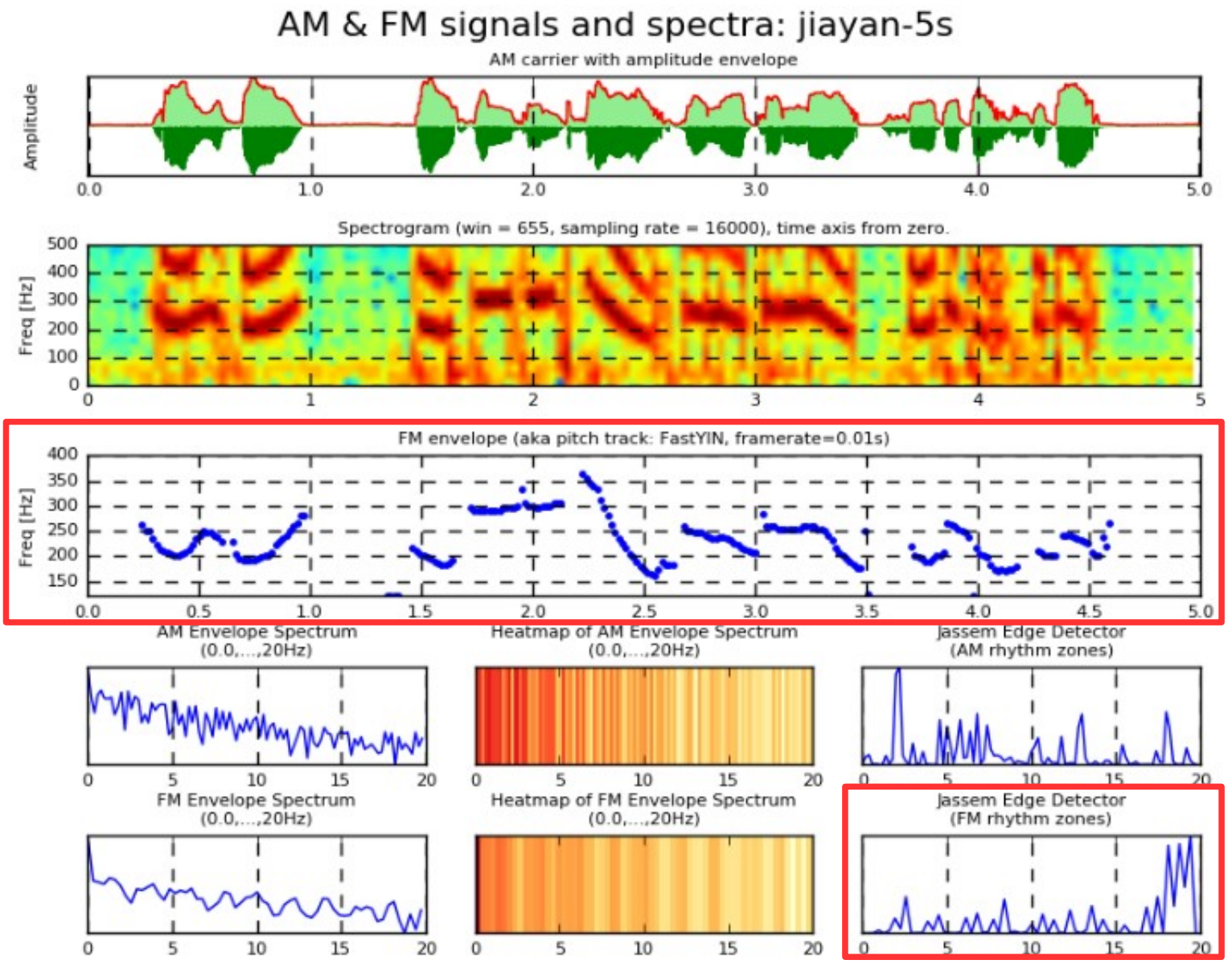
# *Praat*

Paul Boersma



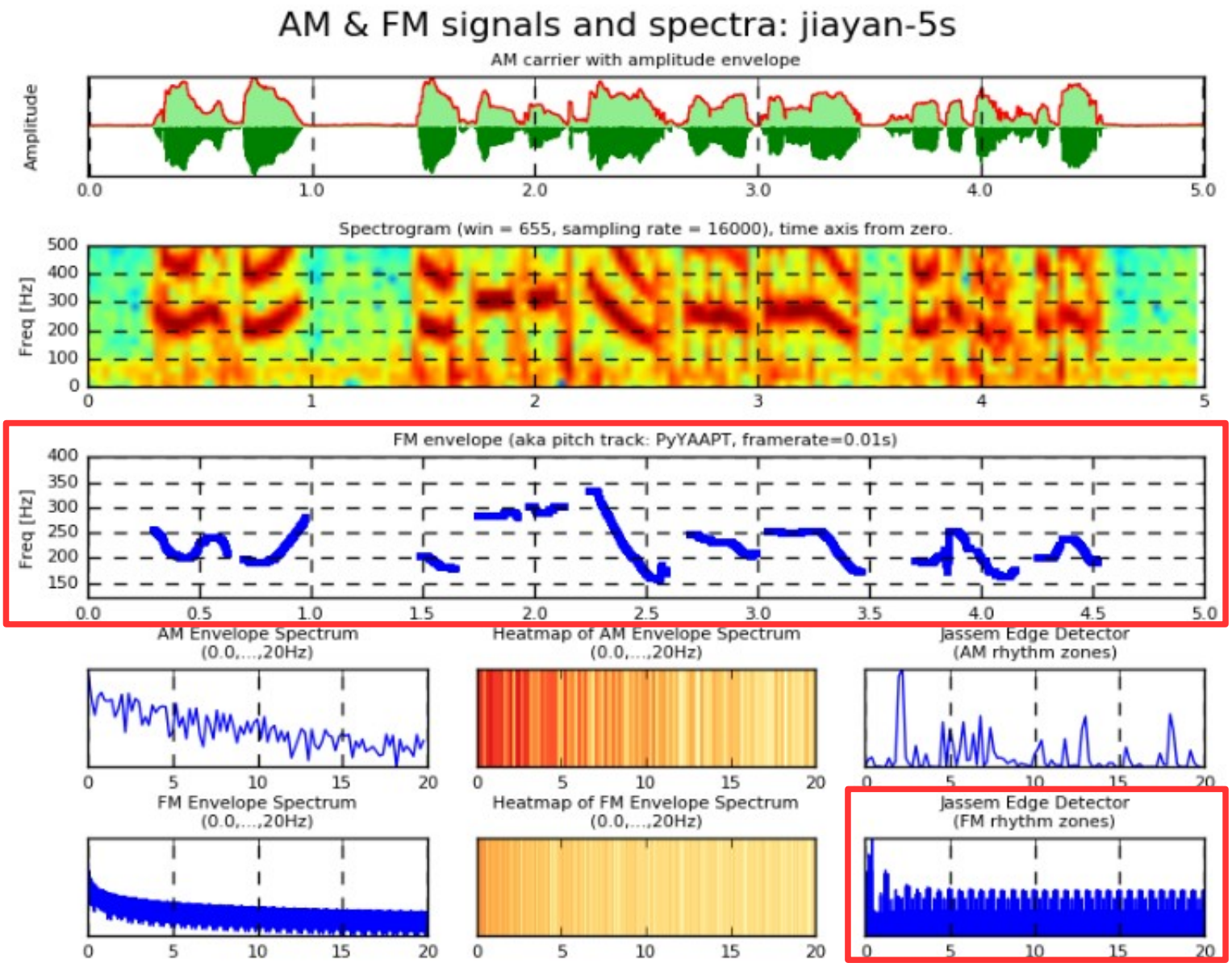
# *YIN (as opposed to YANG, Python emulation)*

Patrice Guyot



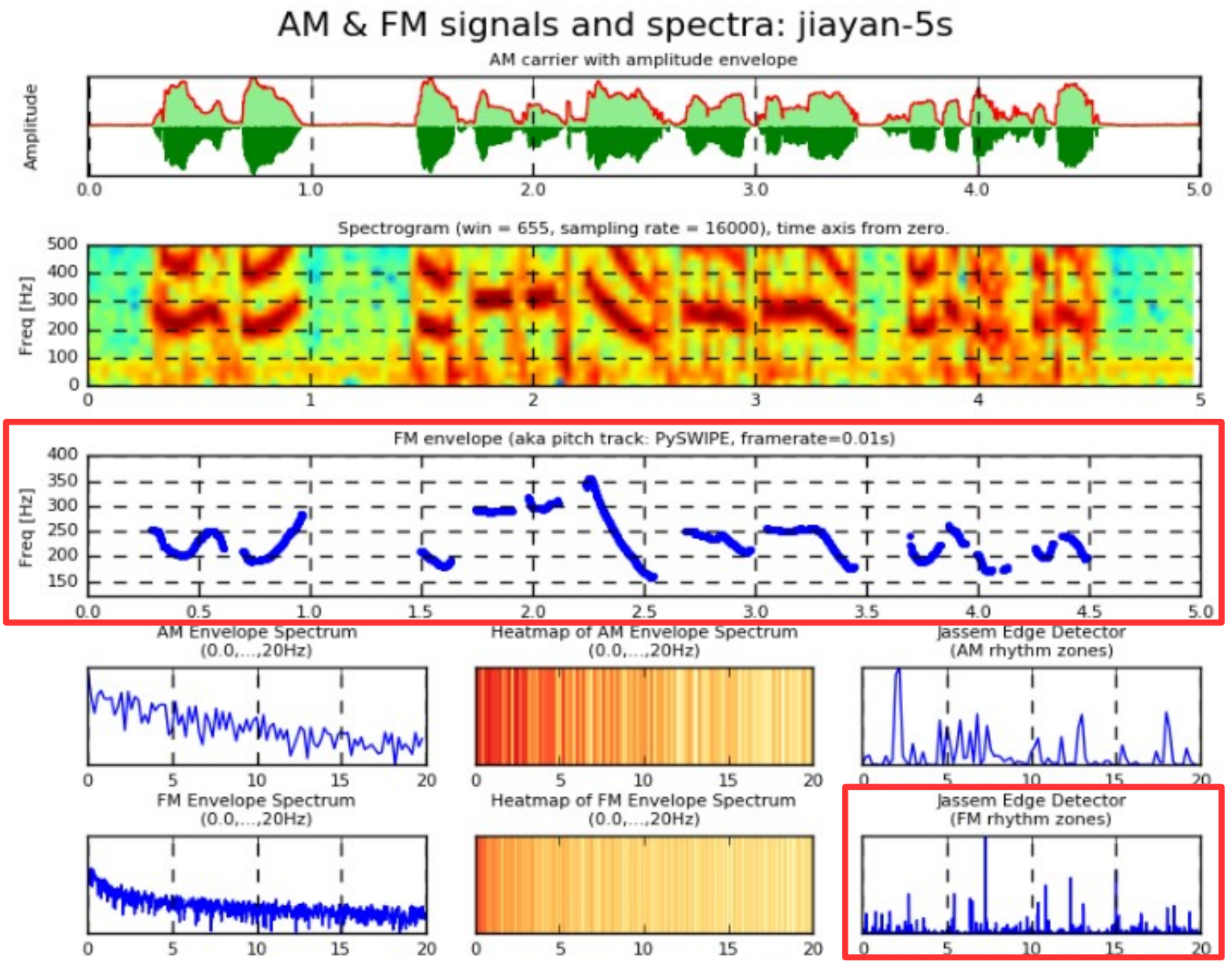
# *YAAPT (Yet Another Algorithm for Pitch Tracking, Python emulation)*

Bernardo J. B. Schmitt



# *SWIPE (Square Wave Inspired Pitch Estimator, Python emulation)*

Disha Garg



# *F0 – Pitch: A Constructive Do-It-Yourself Strategy*

## Time domain:

## AMDF

A kind of auto-correlation, but with subtraction minima not correlation maxima

Preprocessing:
- centre-clipper
- low-pass filter

Postprocessing:
- moving median

Simple: no
- voice detection
- candidate weighting

(code on GitHub)

## Frequency domain:

## FFT+spectrum peak-picking

Finding the lowest frequency spectral peak in the Fourier transform

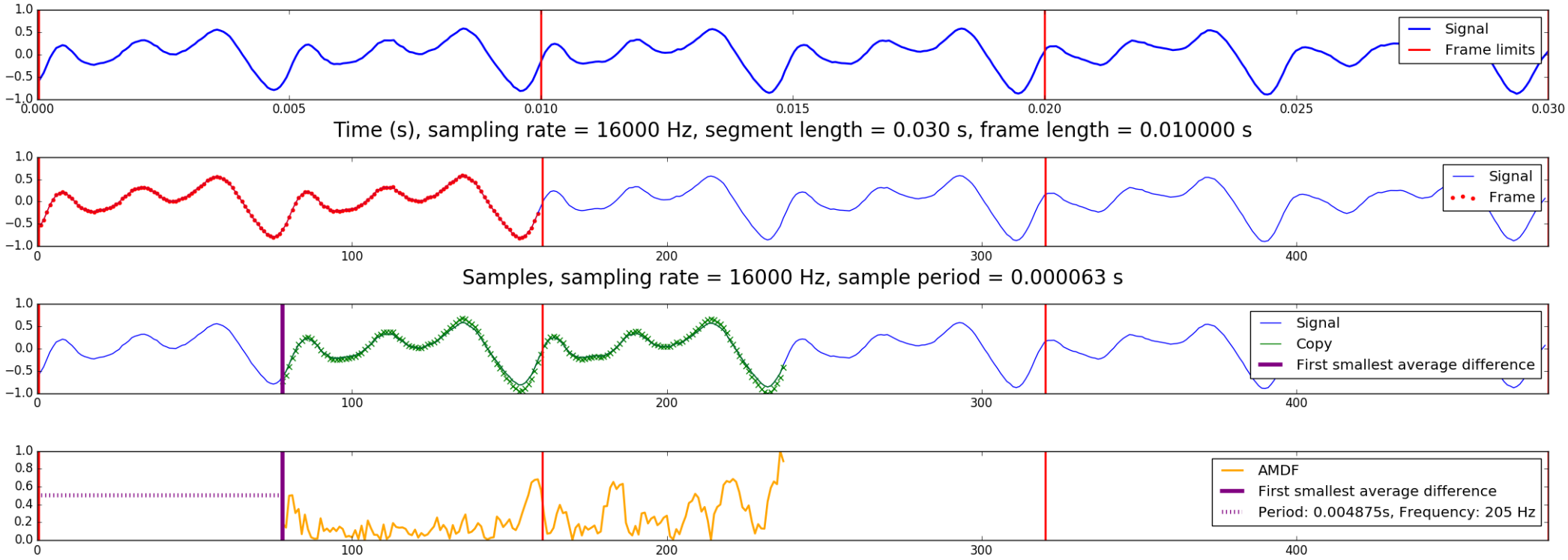
Preprocessing:
- centre-clipper
- low-pass filter

Postprocessing:
- moving median

Simple - no
- voice detection
- candidate weighting

(code on GitHub)

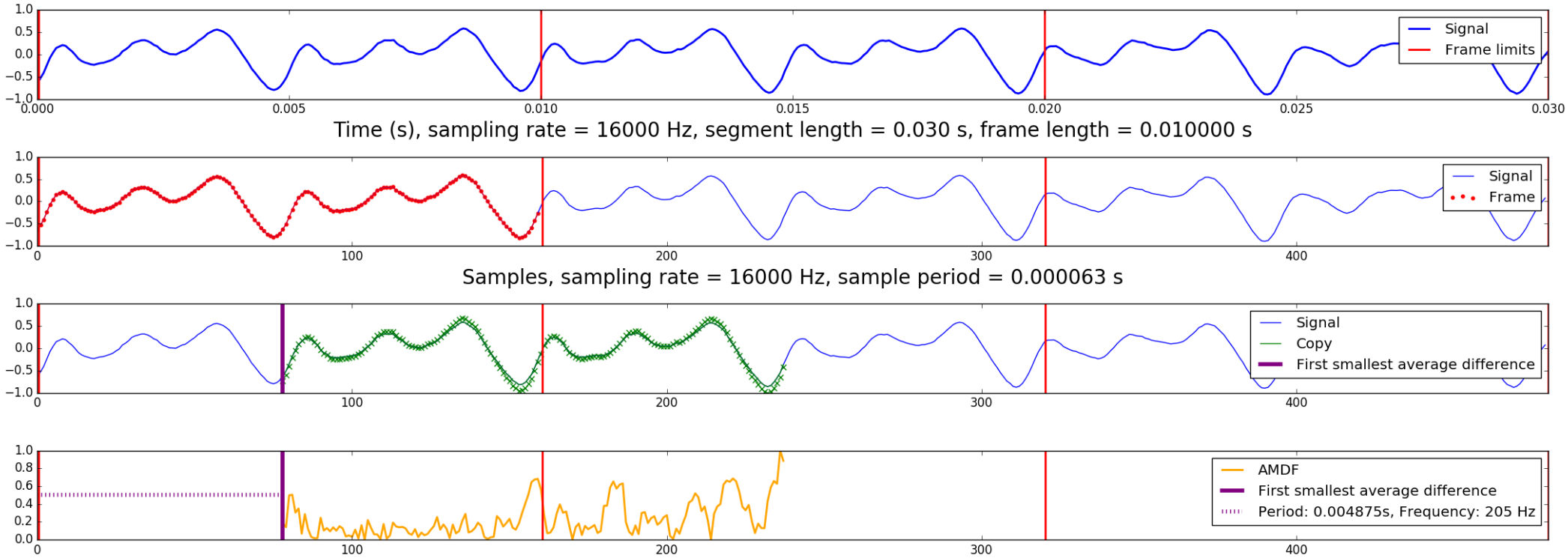
# Frequency Demodulation – F0 estimation - ‘pitch’ extraction




**The Average Magnitude Difference Function** works by subtracting segments of the signal from each other and picking the smallest difference, giving the *fundamental period*.

Then: *fundamental frequency* = 1 / *fundamental period*

# Frequency Demodulation: the AMDF method




**Divide the signal into frames into frames, and for each frame (current frame in red):**

**Define a frame-sized copy of the frame, and move the copy along the signal for the length of the frame (moving window). For each sample in the frame:**

**At each step collect average differences between frame and moving window**

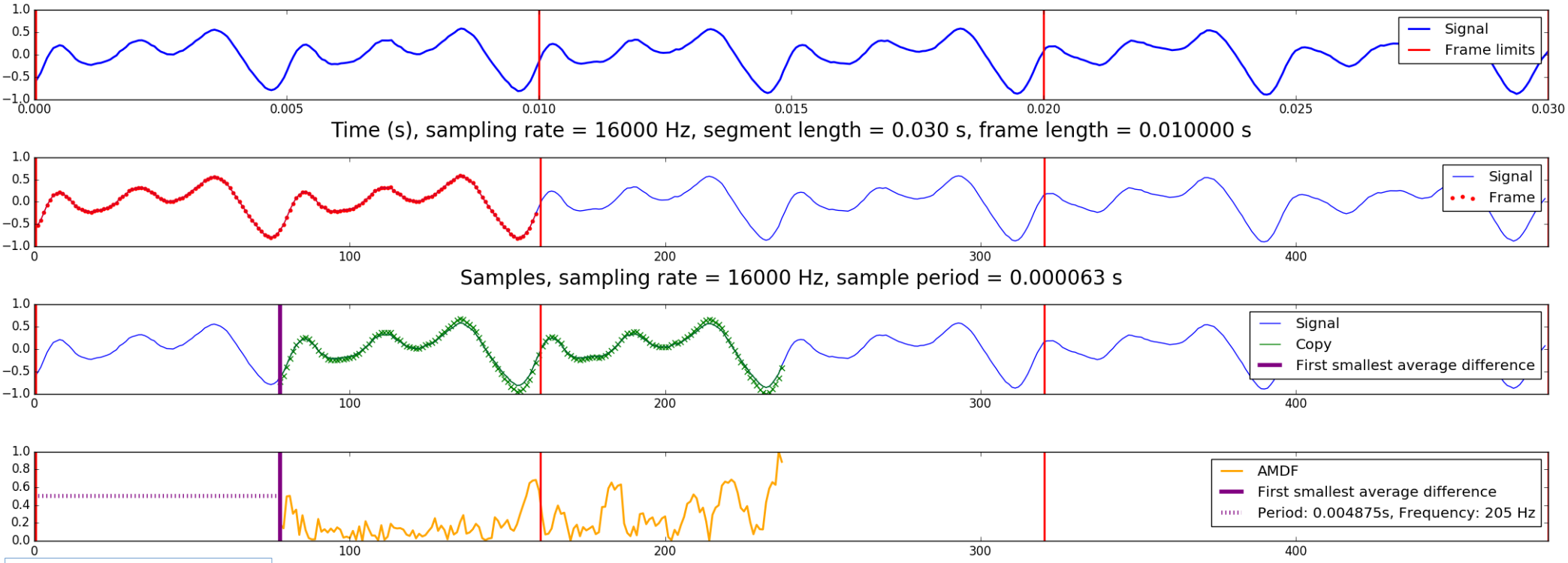
**Find the smallest average difference for all average differences in the frame**

**Frame start subtracted from time of difference = duration of *fundamental period***

***fundamental frequency* = 1 / *fundamental period***

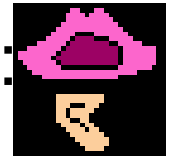
***fundamental frequency* = 1 / *fundamental period***

# Frequency Demodulation: the AMDF method




AMDF-demo

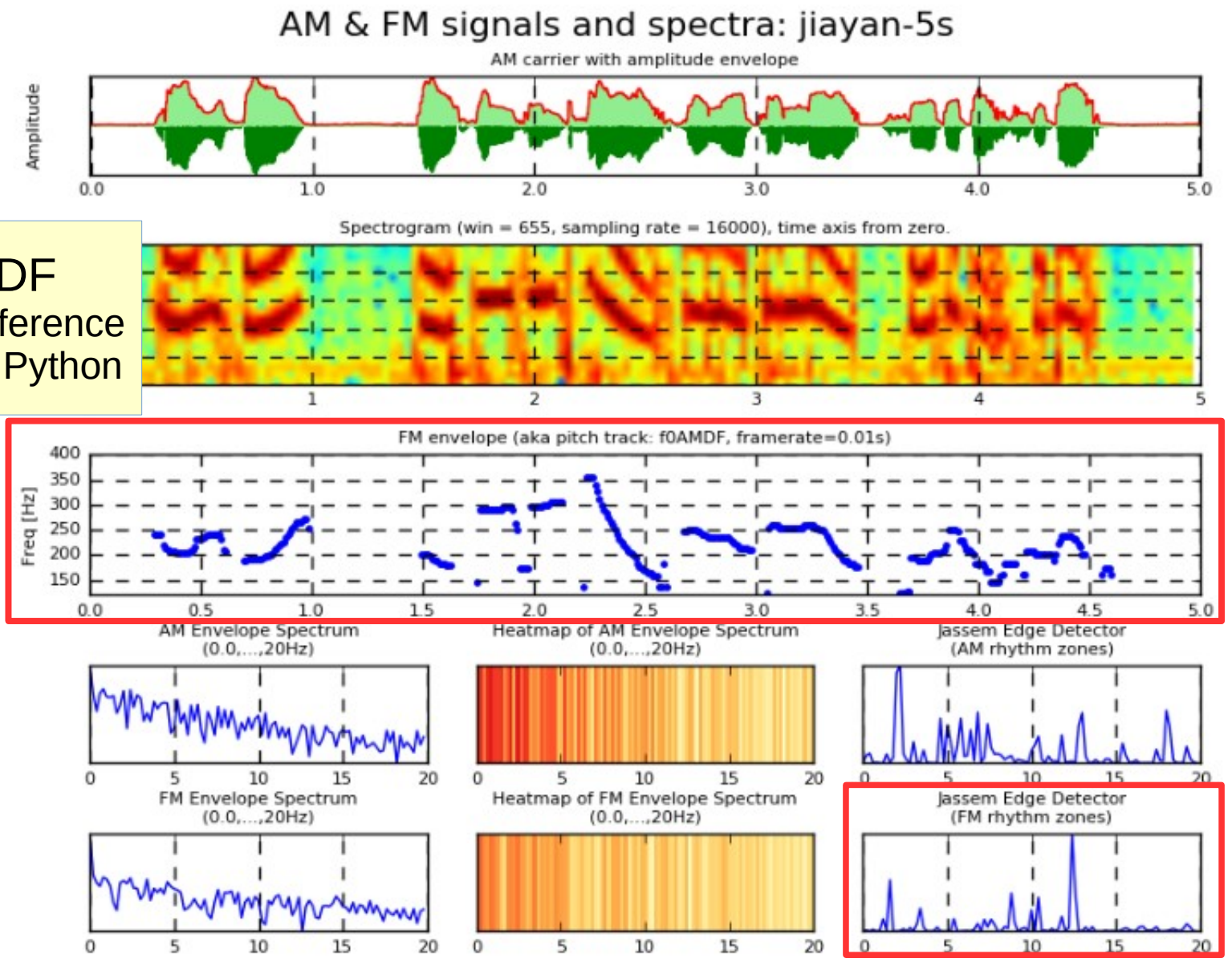
**Autocorrelation** functions on the same principle, except that copies are multiplied, not subtracted, and the largest product is taken.

This is how Praat works: 

# *How about AMDF (Average Magnitude Difference Function, Python)*

Dafydd Gibbon

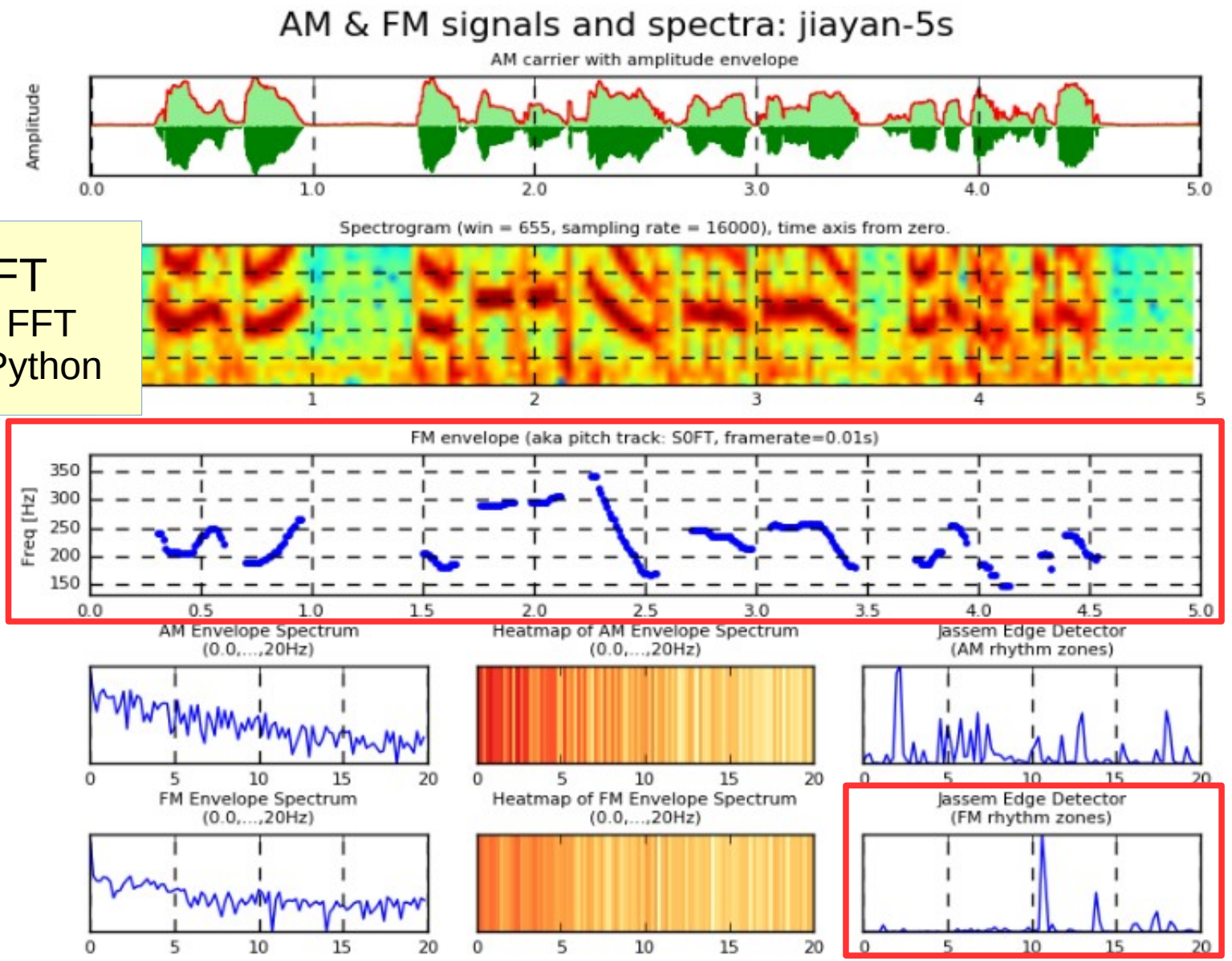
AMDF
naive difference minima, Python



# *FFTpeak (Simple F0 Tracker, Python)*

Dafydd Gibbon*

S0FT
naive FFT
peak, Python




* inspired by a snippet from ‘Jonathan’, gist.github.com/endolith/255291

# *Comparing F0 estimators with RAPT as 'gold standard'*

| F0 estimator pair | Correlation |
|---|---|
| | |
| RAPT:PyRAPT | 0,8902 |
| RAPT:Praat | 0,8657 |
| **RAPT:FFTpeak** | **0,9096** |
| RAPT:f0AMDF | 0,8605 |
| | |
| **FFTpeak:RAPT** | **0,9096** |
| FFTpeak:Praat | 0,8409 |
| FFTpeak:PyRAPT | 0,8352 |
| FFTpeak:f0AMDF | 0,8016 |

*Benchmark against standard F0 estimators*

Encouraging for FFTpeak (problems remain, of course):

- correlation ignores some relevant properties such as overall difference in pitch height
- (slightly positively biased) idea of the relationship
- not enough test data
- not as robust as RAPT
- but suggests that RAPT is fit for purpose

# Frequency Modulations – Emotive Rhythms

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!***

# *Selected Work on Amplitude Envelope Demodulation Spectra*


☛ **[1]** **Cummins**, Fred, Felix **Gers** and Jürgen **Schmidhuber**. “Language identification from prosody without explicit features.” *Proc. Eurospeech*. 1999.

**[2]** **He**, Lei and Volker **Dellwo**. “A Praat-Based Algorithm to Extract the Amplitude Envelope and Temporal Fine Structure Using the Hilbert Transform.” In: *Proc. Interspeech* 2016, San Francisco, pp. 530-534, 2016.

**[3]** **Hermansky**, Hynek. “History of modulation spectrum in ASR.” *Proc. ICASSP* 2010.

**[4]** **Leong**, Victoria and Usha **Goswami**. “Acoustic-Emergent Phonology in the Amplitude Envelope of Child-Directed Speech.” *PLoS One* 10(12), 2015.

**[5]** **Leong**, Victoria, Michael A. **Stone**, Richard E. **Turner**, and Usha **Goswami**. “A role for amplitude modulation phase relationships in speech rhythm perception.” *JAcSocAm*, 2014.

**[6]** **Liss**, Julie M., Sue **LeGendre**, and Andrew J. **Lotto**. “Discriminating Dysarthria Type From Envelope Modulation Spectra.” *Journal of Speech, Language and Hearing Research* 53(5):1246–1255, 2010.

**[7]** **Ludusan**, Bogdan Antonio **Origlia**, Francesco **Cutugno**. “On the use of the rhythmogram for automatic syllabic prominence detection.” *Proc. Interspeech*, pp. 2413-2416, 2011.

**[8]** **Ojeda**, Ariana, Ratree **Wayland**, and Andrew **Lotto**. “Speech rhythm classification using modulation spectra (EMS).” Poster presentation at the 3rd Annual Florida Psycholinguistics Meeting, 21.10.2017, U Florida. 2017.

☛ **[9]** **Tilsen** Samuel and Keith **Johnson**. “Low-frequency Fourier analysis of speech rhythm.” *Journal of the Acoustical Society of America*. 2008; 124(2):EL34–EL39. [PubMed: 18681499]

☛ **[10]** **Tilsen**, Samuel and Amalia **Arvaniti**. “Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages.” *The Journal of the Acoustical Society of America* 134, p. 628 .2013.

☛ **[11]** **Todd**, Neil P. McAngus and Guy J. Brown. “A computational model of prosody perception.” *Proc. ICSLP* 94, pp. 127-130, 1994.

**[12]** **Varnet**, Léo, Maria Clemencia **Ortiz-Barajas**, Ramón Guevara **Erra**, Judit **Gervain**, and Christian **Lorenzi**. “A cross-linguistic study of speech modulation spectra.” *JAcSocAm* 142 (4), 1976–1989, 2017.

Thank you