Phonology and Phonetics of Rhythm:

States and Times or Cycles and Frequencies?

Dafydd Gibbon

Bielefeld University
Jinan University

13th Chinese Phonetics Conference
Guangzhou, November 2018

Theses: YARD (Yet Another Discussion of Rhythm)

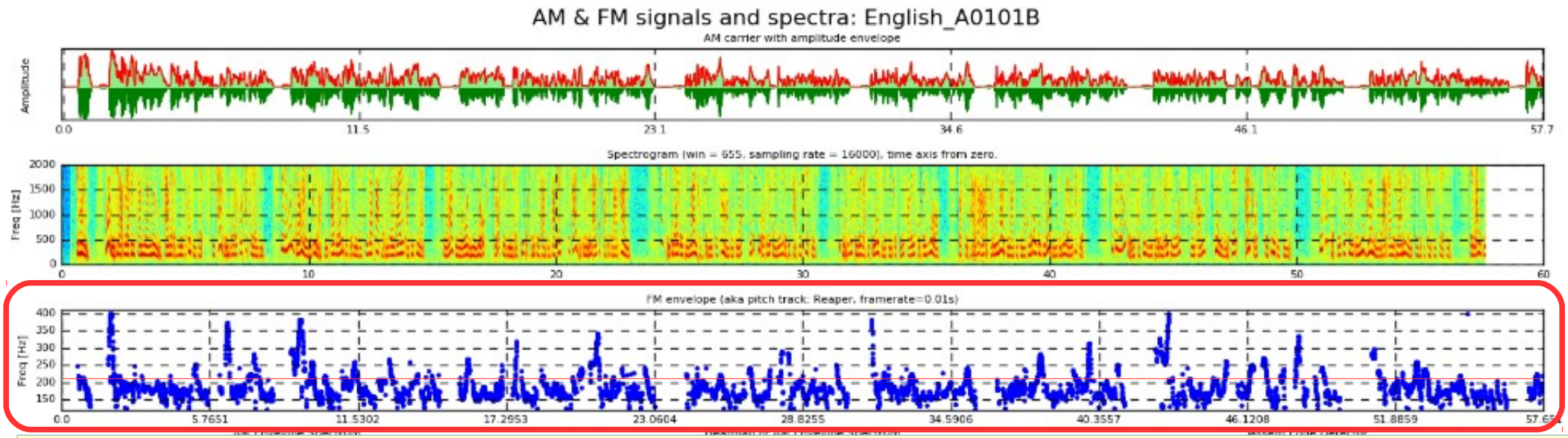
In speech (and in music):

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

So rhythms appear to be hierarchical, but the hierarchy

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

Linear Layered Discourse Rhythms



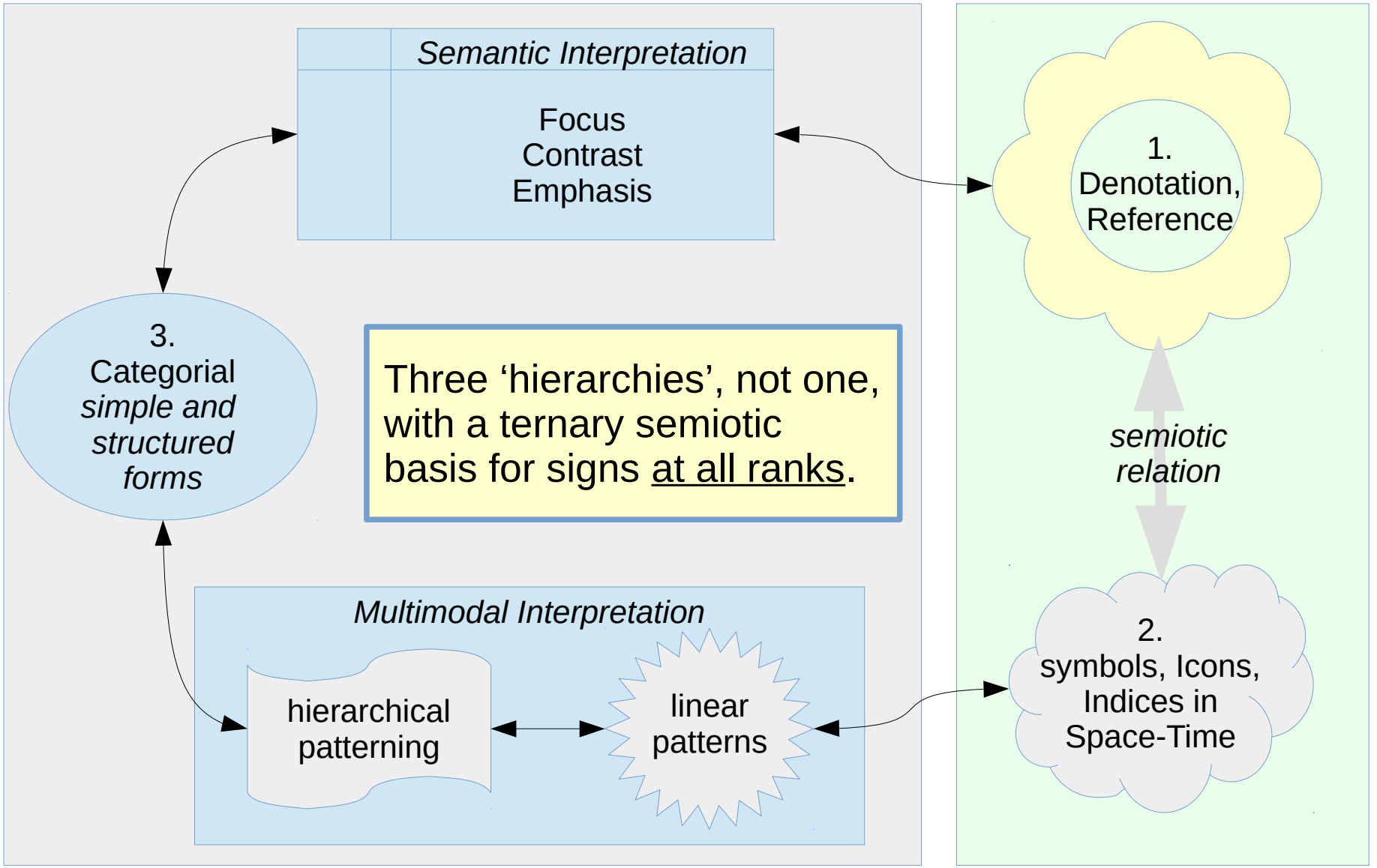
57 seconds of a BBC news broadcast, from Aix-MARSEC corpus (A0101B)

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

The Multilinear Rank Interpretation Model



The Multilinear Rank Interpretation Model

Multilinear Category Ranks

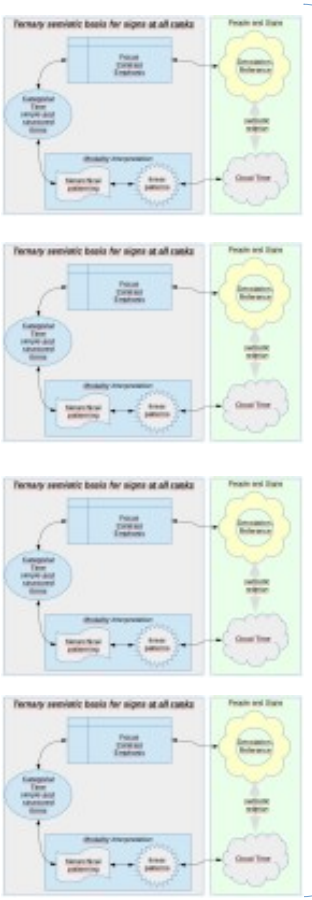
Ranks: Categories and their Interpretations

Discourse: Monologue, Dialogue

Utterance: turn, IPU, ...

Sentence, clause, phrase

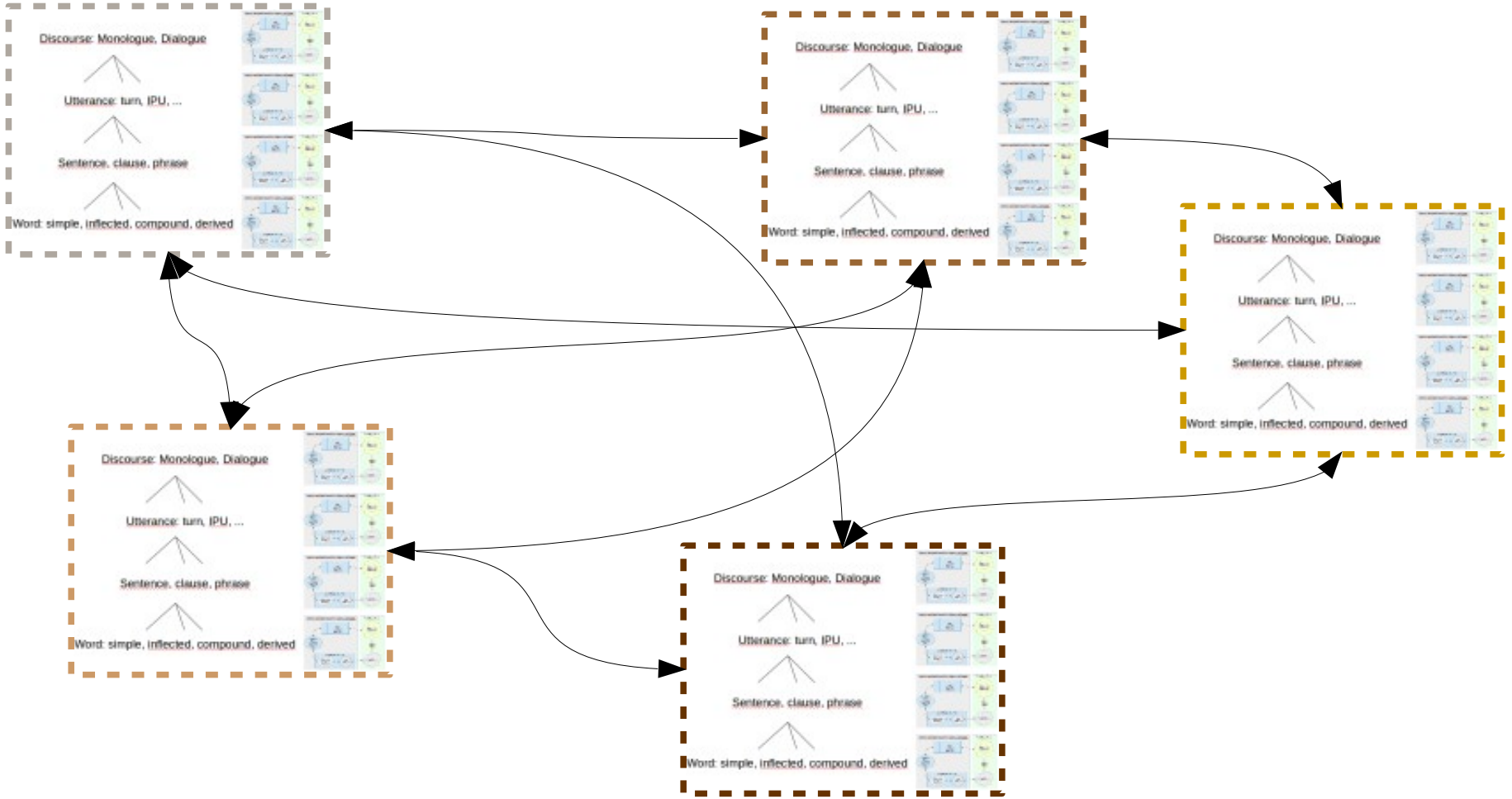
Word: simple, inflected, compound, derived



Multimodal Rank Interpretation Architecture

The Multilinear Rank Interpretation Model

The discourse dimension: people communicating with their multimodal rank interpretation architectures



States and Times or Cycles and Frequencies? - Overview

1. Rhythm Communities
2. Phonological cycles: abstract, iterative rhythm
3. Phonetic states: sequences of isochronous states as rhythm
4. Phonetic cycles: rhythm as oscillation:
   - Production as amplitude and frequency modulation
   - Perception as amplitude and frequency demodulation
5. Pitch cycles: the role of F0 in rhythm
   - AM and FM similarities
   - Some problems with F0 estimators (aka ‘pitch trackers’)
6. More roles of F0 in discourse rhythms

Rhythm Communities – a Selection

Conversation Analysis	Phonetic Annotation Analysis
Linguistic Analysis	Phonetic Signal Analysis

Rhythm Communities – a Selection

Conversation Analysis

‘Authentic’ everyday dialogue data
Ethnomethodology: task scenario
Systematic but subjective judgments
Focus on interpretation:
- discourse grammar: framing
- semantics: topic development
- pragmatics: speaker, addressee

Phonetic Annotation Analysis

Standardised reading data
- sentences
- stories (*The North Wind…*)

Annotation:
- manual, semi-automatic
- syllables, feet, C or V chunks

Focus on isochrony

Linguistic Analysis

Constructed data paradigms
Rules for verbal categories
Abstract stress position patterns:
- linear structures, grids
- hierarchical structures, trees
  - words
  - sentences

Phonetic Signal Analysis

Smoothing models
- splines, polynomials

Oscillation models
- production oriented
- perception oriented
  - Amplitude Modulation Spectrum
  - Frequency Modulation Spectrum

Rhythm Communities – a Selection

Tree Structures

Linguistics: deductive, top-down
- paradigmatic (classificatory)
- syntagmatic (compositional)

Phonetics: inductive, bottom-up
- classificatory (CART etc.)
- compositional

Phonetic Annotation Analysis

Standardised reading data
- sentences
- stories (*The North Wind…*)

Annotation:
- manual, semi-automatic
- syllables, feet, C or V chunks

Focus on isochrony

Linguistic Analysis

Constructed data paradigms
Rules for verbal categories
Abstract stress position patterns:
- linear structures, grids
- hierarchical structures, trees
  - words
  - sentences

Phonetic Signal Analysis

Smoothing models
- splines, polynomials

Oscillation models
- production oriented
- perception oriented
  - Amplitude Modulation Spectrum
  - Frequency Modulation Spectrum

Phonological Iteration as Abstract Oscillation:

Iterative intonation

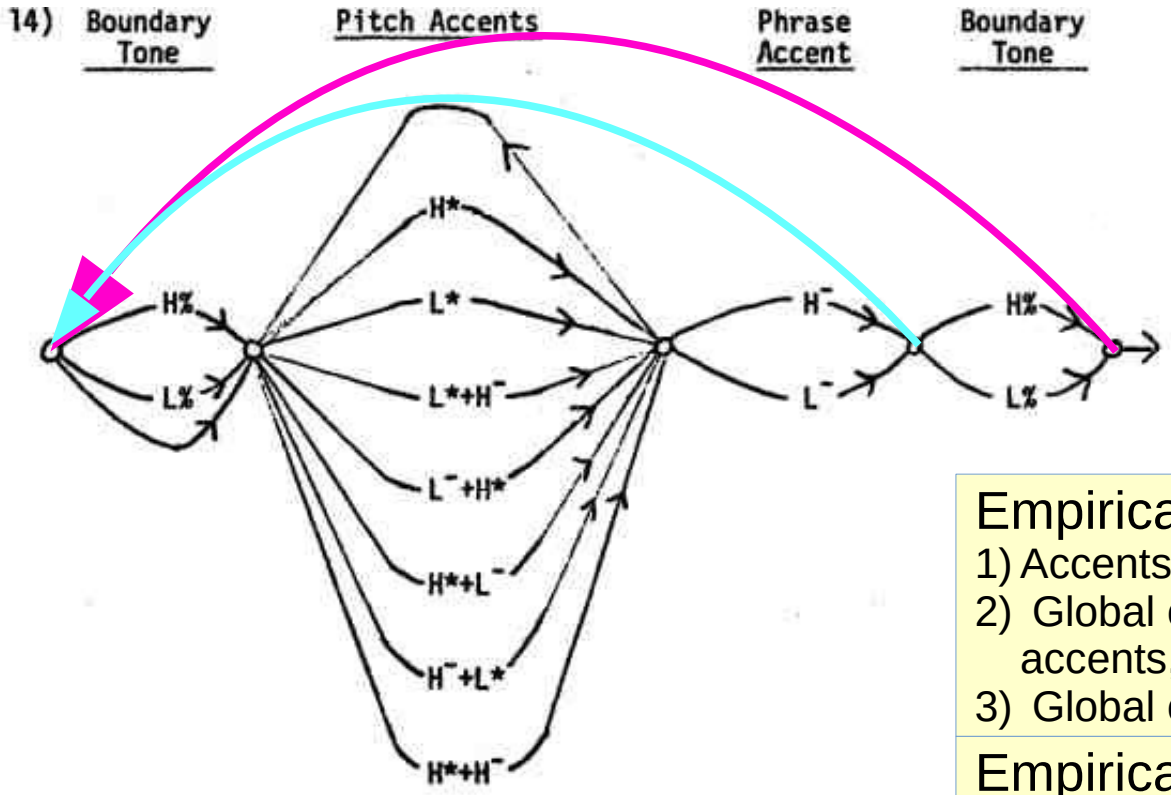
Iterative tonal sandhi (Niger Congo)

Iterative tonal sandhi (Tianjin dialect)

Note: by 'cycle' I do not mean tone cycles in the paradigmatic , classificatory sense, but syntagmatic, compositional iterations

Phonological iteration as a layered hierarchy of linear abstract oscillations

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

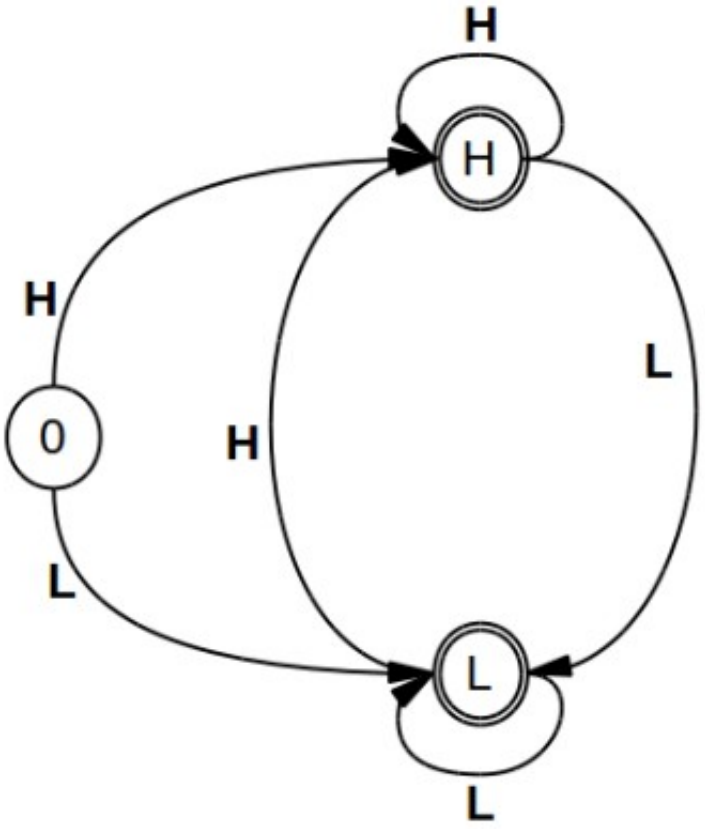
1) Accents in a sequence tend to be all H* or all L*
2) Global contours tend to be rising with L* accents, falling with H* accents
3) Global contours may span more than 1 turn

Empirical undergeneration

1) Paratone hierarchy not included
2) No time constraints

Phonological iteration as a layered hierarchy of linear abstract oscillations

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



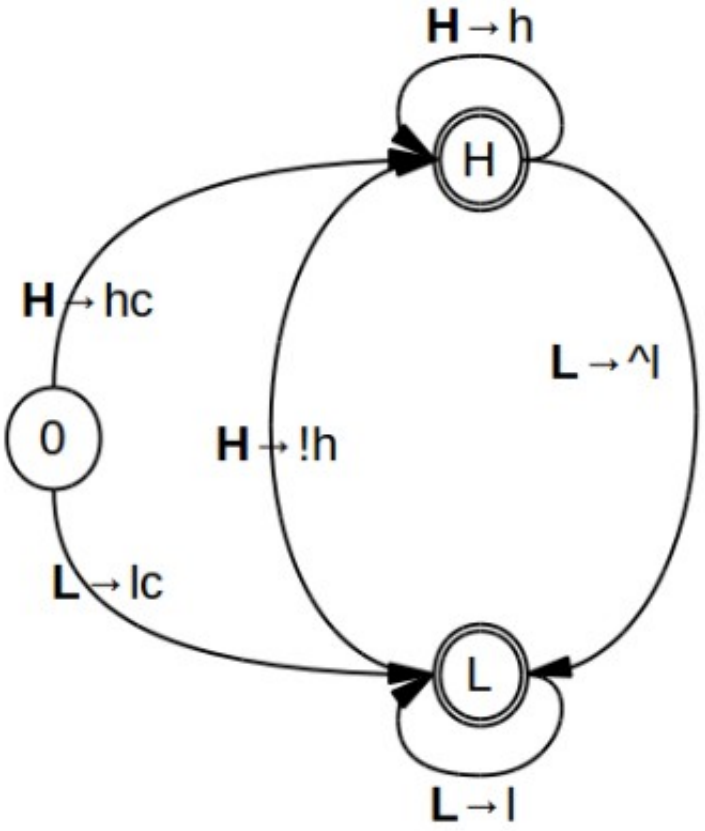
At the most abstract level, just one node with H and L cycling around it.

From an allotonic point of view:

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

Phonological iteration as a layered hierarchy of linear abstract oscillations

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

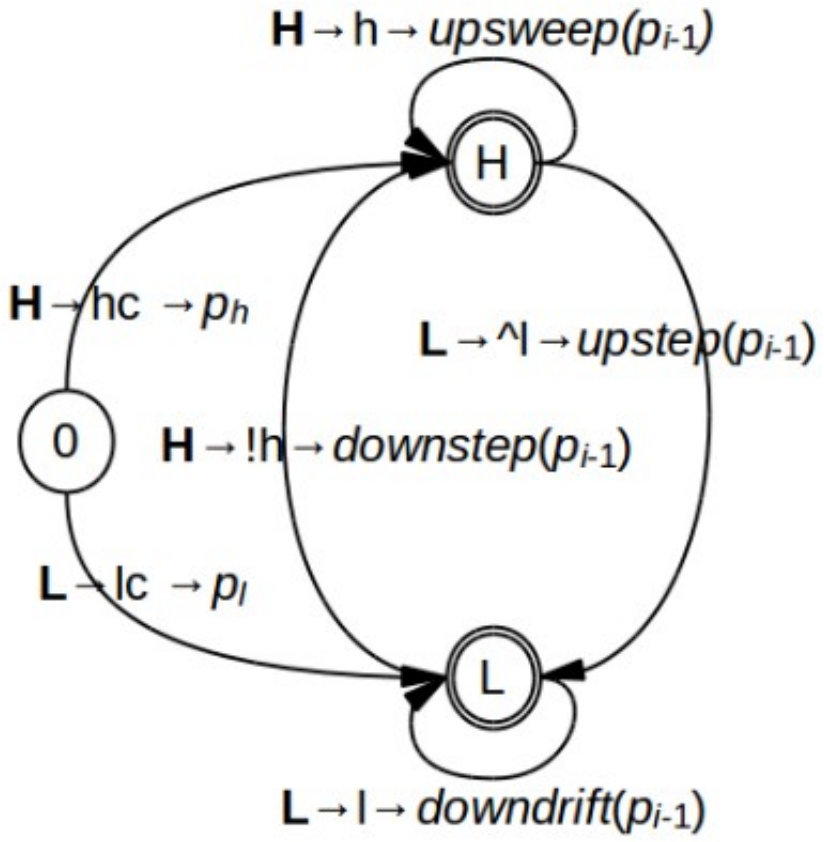


From an allotonic point of view:

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

Phonological iteration as a layered hierarchy of linear abstract oscillations

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

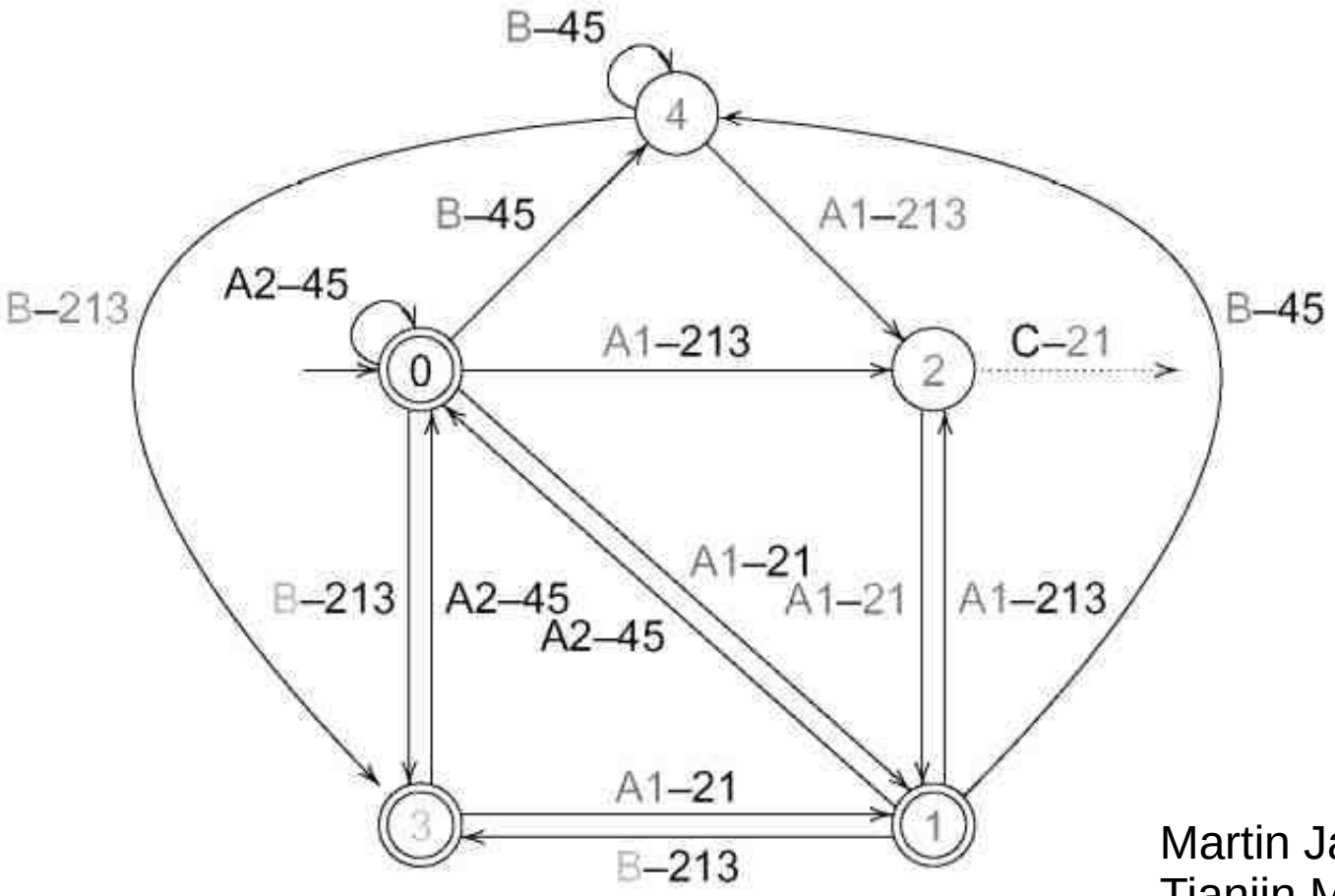


From phonetic signal processing point of view:

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

Phonological iteration as a layered hierarchy of linear abstract oscillations

Tianjin Dialect Iterative Tonal Sandhi



Martin Jansche 1998
Tianjin Mandarin tone sandhi

Phonetics, the Search for Rhythm I

States and Times: the Isochrony Approach

1-dimensional
2-dimensional
3-dimensional

One-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.

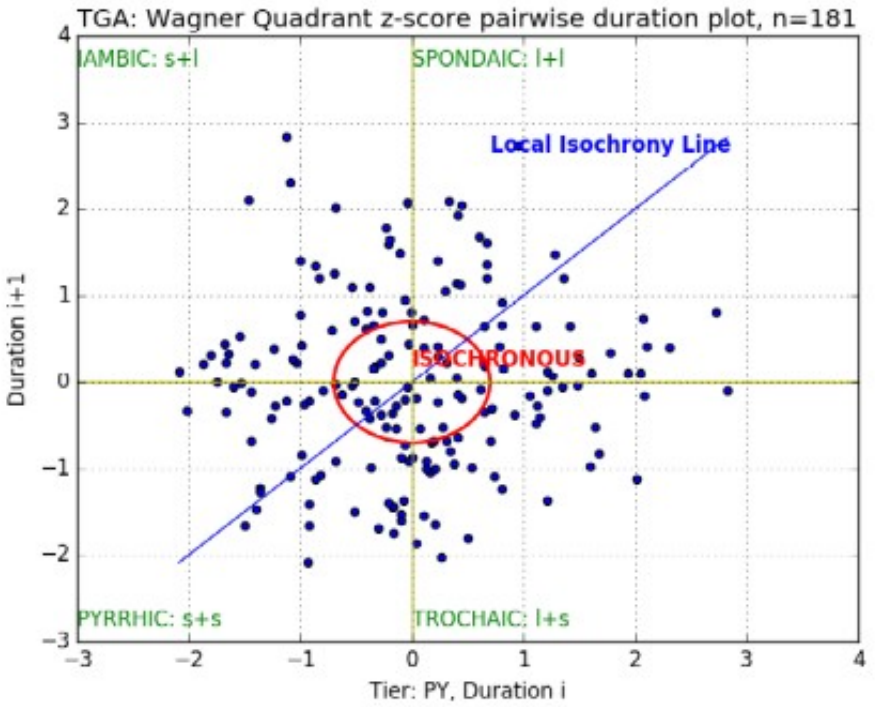
$Variance(x_{1\ldots n}) = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}$	(or S*tandard Deviation*)
$PIM(x_{1\ldots n}) = \sum_{i \neq j} \left\lvert \log \frac{I_i}{I_j} \right\rvert$	where $I_{i,j}$ are intervals in a given sequence
$PFD(d_{1\ldots n}) = \frac{\sum_{i=1}^n \lvert \bar{d} - d_i \rvert}{\sum_{j=1}^n d_j} \times 100$	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

Two-dimensional annotation mining of time-stamp durations

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:

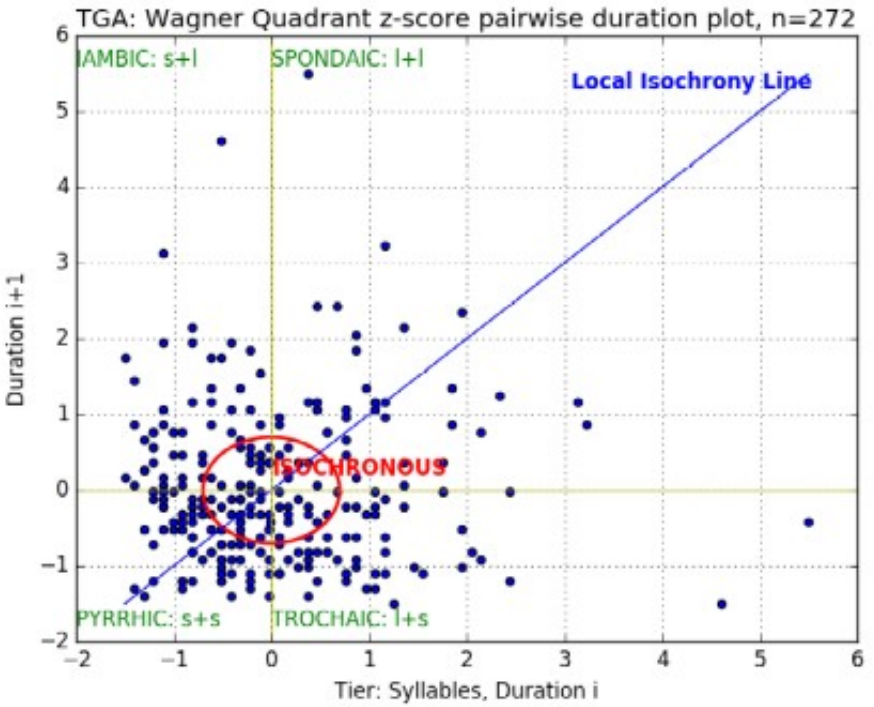
Mandarin: means are scattered relatively evenly around the centre
English: e.g. *count*(*short-short*) **>** *count*(*long-long*), not binary!

SHORT-LONG LONG-LONG



SHORT-SHORT LONG-SHORT

SHORT-LONG LONG-LONG



SHORT-SHORT LONG-SHORT

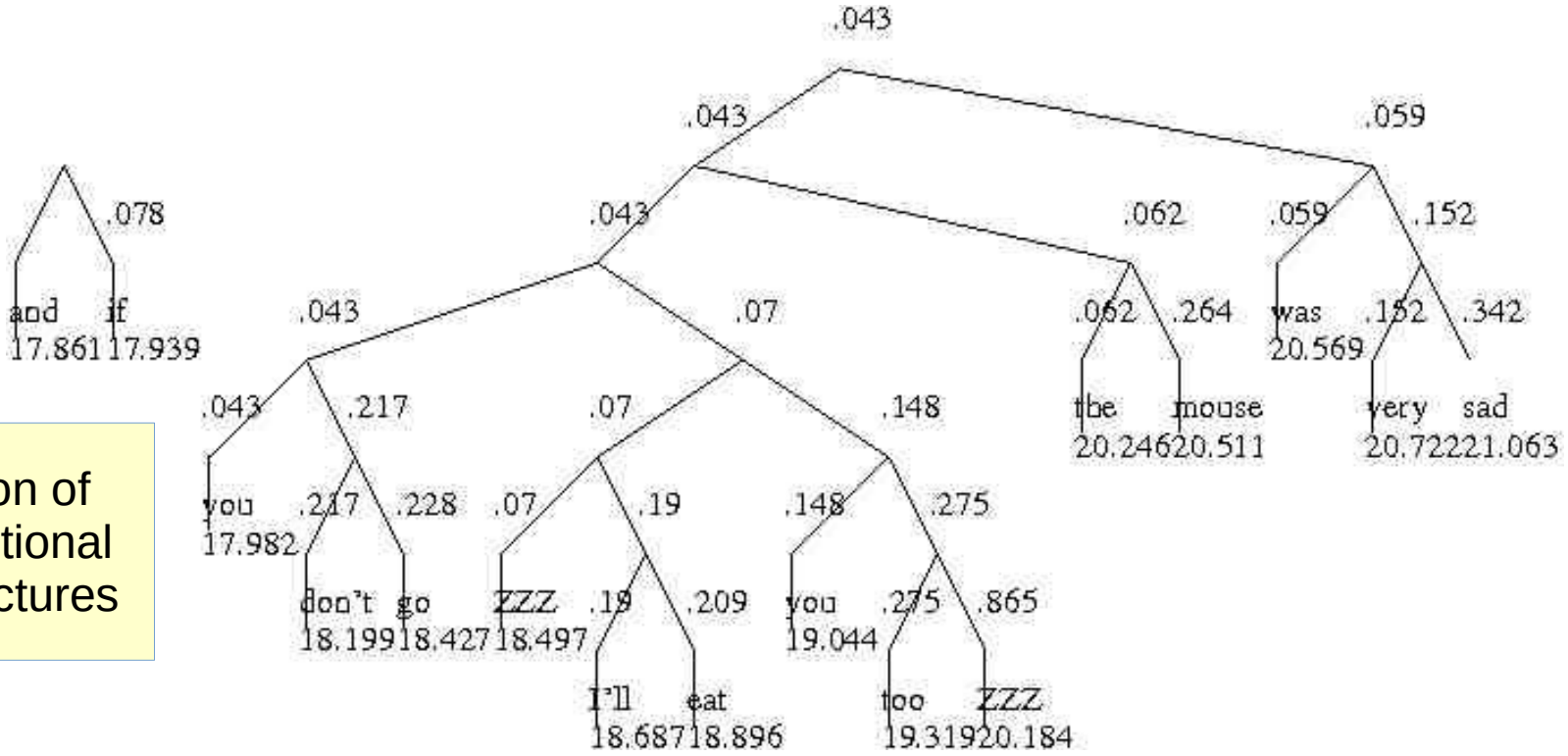
Wagner, Petra (2007). "Visualizing levels of rhythmic organisation." *Proc. International Congress of Phonetic Sciences, Saarbrücken 2007*, pp. 1113-1116, 2007

Two-dimensional annotation mining of time-stamp durations

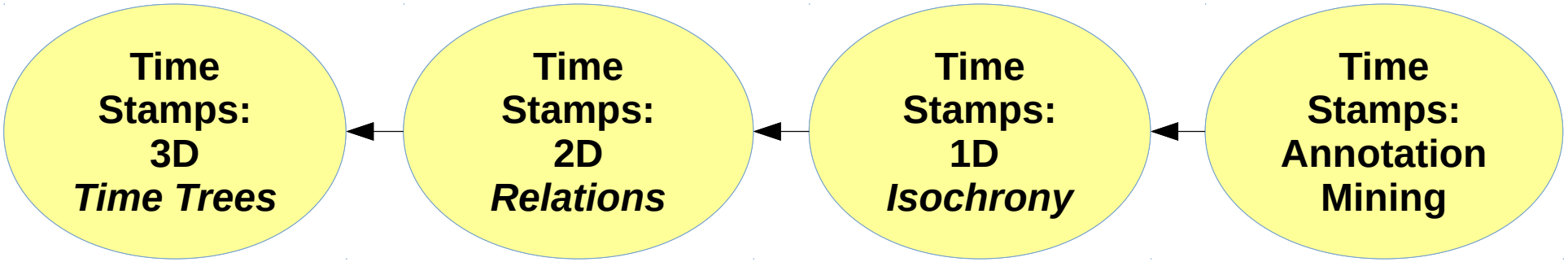
Three-dimensional because alternative trees are possible, depending on the algorithm settings:

- 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.



All these approaches

- are based on annotated time-stamps, not the signal
- model static durations
- use pairwise relations
- focus on *isochrony*, equal timing of durations

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

Phonetics, The Search for Rhythm II

Cycles and Frequencies: Oscillation Approaches

Rhythm Production as <u>Amplitude Modulation</u> (AM)
Rhythmic Amplitude Envelope + Carrier Frequency

There are many studies of oscillation in production

I will concentrate on ...

Rhythm Perception as <u>Amplitude Demodulation</u>
Rhythmic Amplitude Envelope – Carrier Frequency

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.

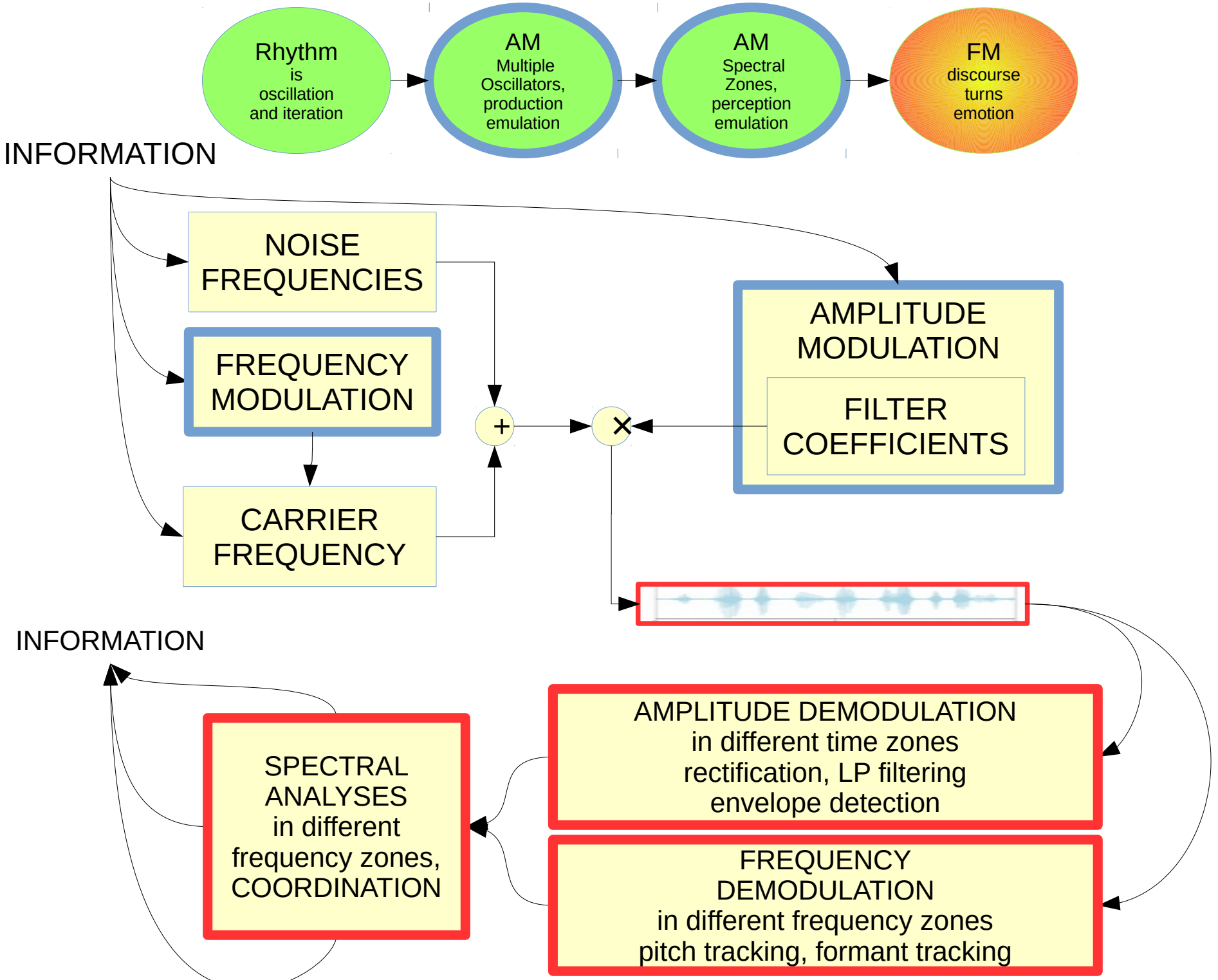
[8] Ojeda, Ariana, Ratree **Wayland**, and Andrew **Lotto**. “Speech rhythm classification using modulation spectra (EMS).” Poster presentation at the 3rd Annual Florida Psycholinguistics Meeting, 21.10.2017, U Florida. 2017.

☛ **[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]

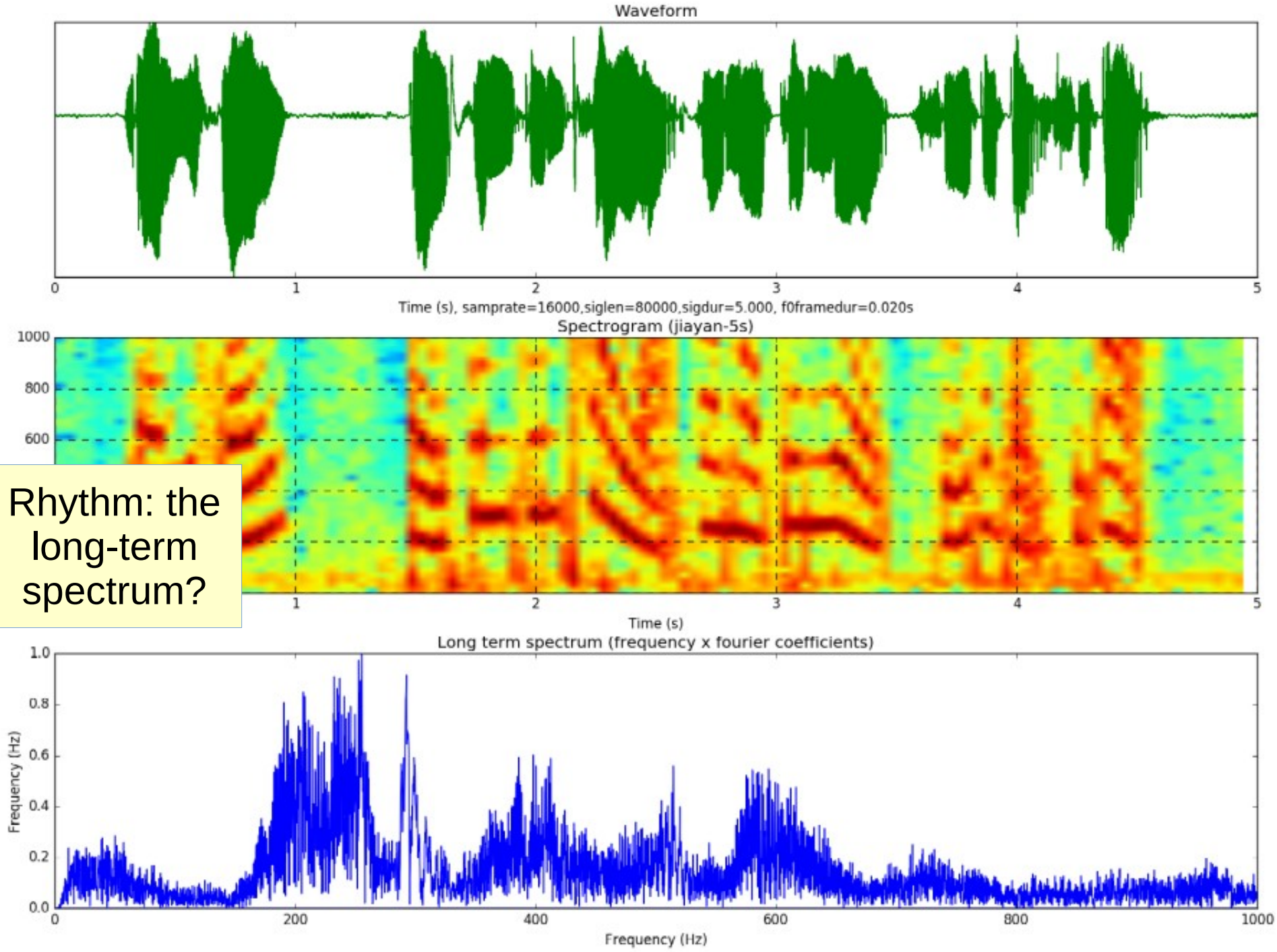
☛ **[10] Tilsen**, Samuel and Amalia **Arvaniti**. “Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages.” *The Journal of the Acoustical Society of America* 134, p. 628 .2013.

☛ **[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.

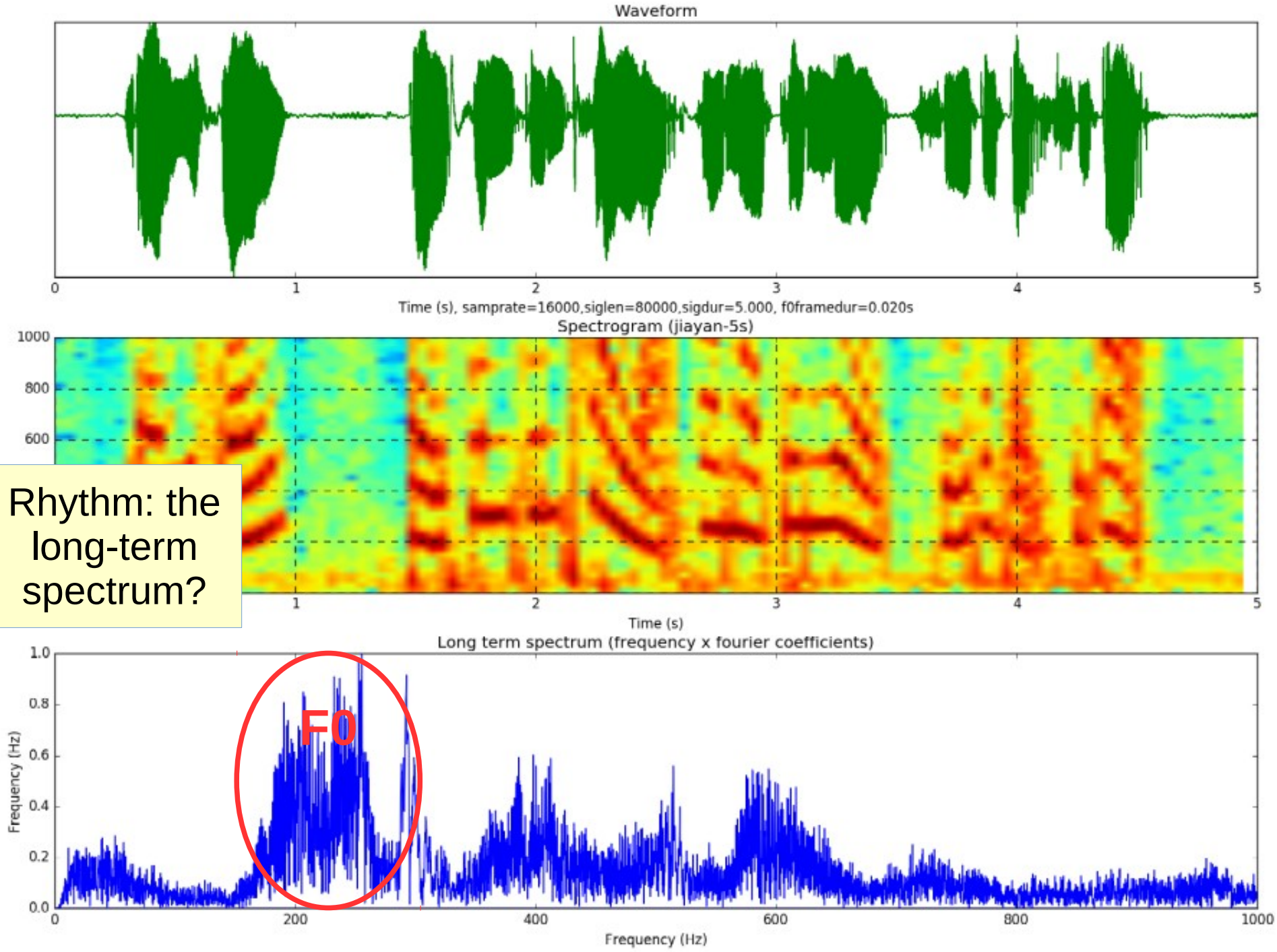

Rhythm
is
oscillation
and iteration
AM
Multiple
Oscillators,
production
emulation
AM
Spectral
Zones,
perception
emulation
FM
discourse
turns
emotion
INFORMATION
NOISE
FREQUENCIES
FREQUENCY
MODULATION
CARRIER
FREQUENCY
+
×
AMPLITUDE
MODULATION
FILTER
COEFFICIENTS
INFORMATION
SPECTRAL
ANALYSES
in different
frequency zones,
COORDINATION
AMPLITUDE DEMODULATION
in different time zones
rectification, LP filtering
envelope detection
FREQUENCY
DEMODULATION
in different frequency zones
pitch tracking, formant tracking

Modulation Regularities – Maybe the Long-Term Spectrum?



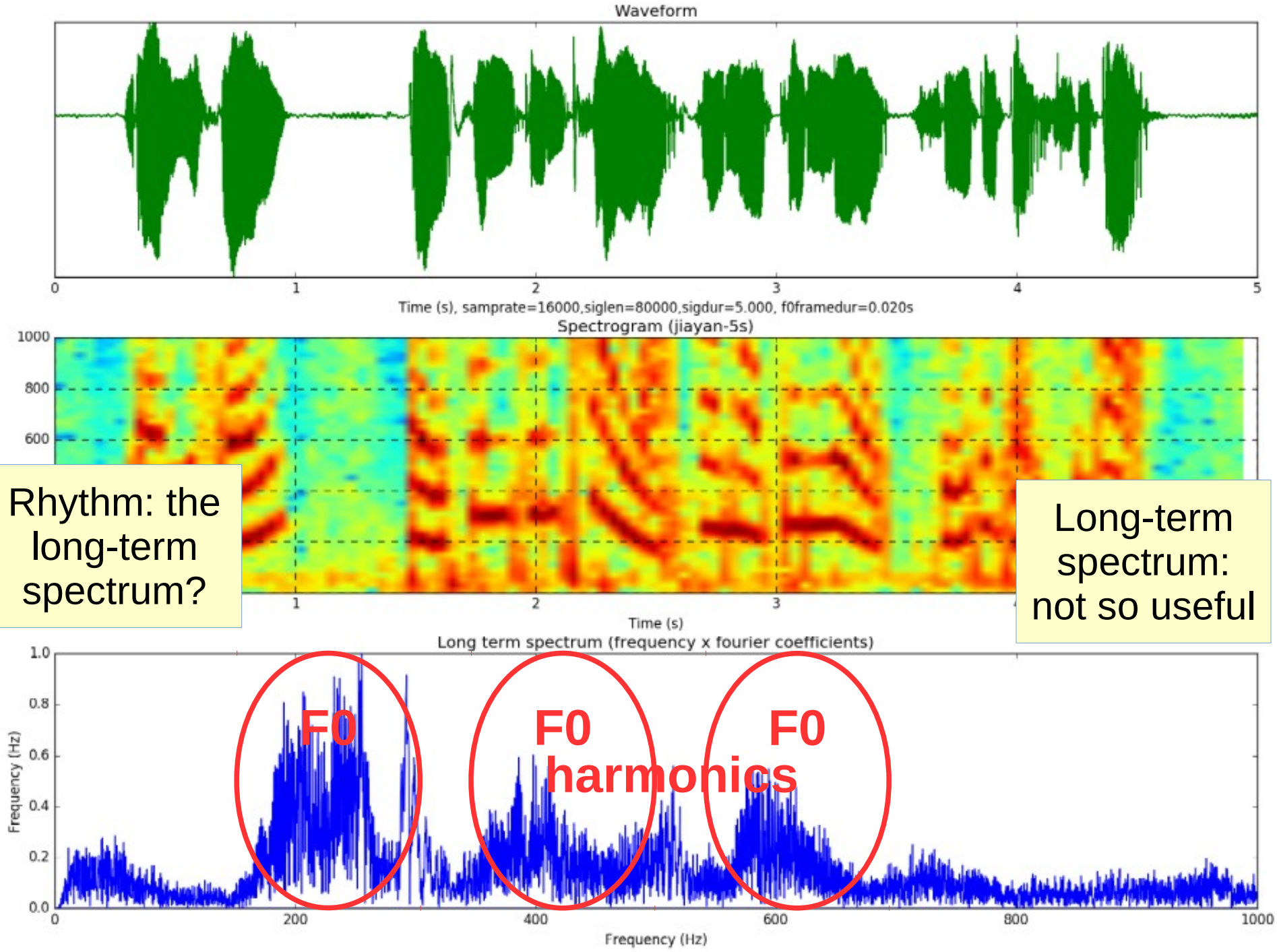
Rhythm: the long-term spectrum?

Modulation Regularities – Maybe the Long-Term Spectrum?



Rhythm: the long-term spectrum?

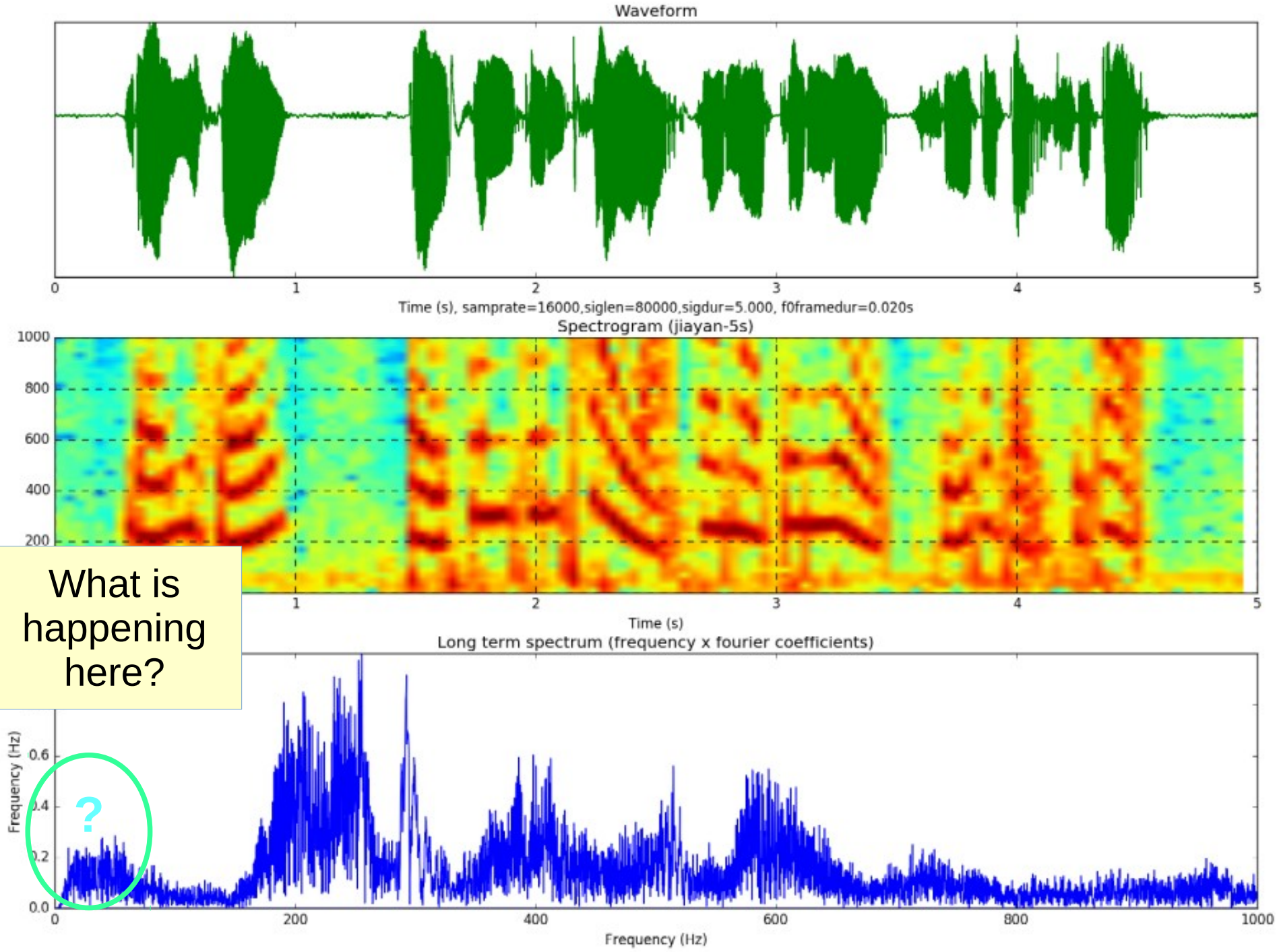
Modulation Regularities – Maybe the Long-Term Spectrum?



Rhythm: the long-term spectrum?

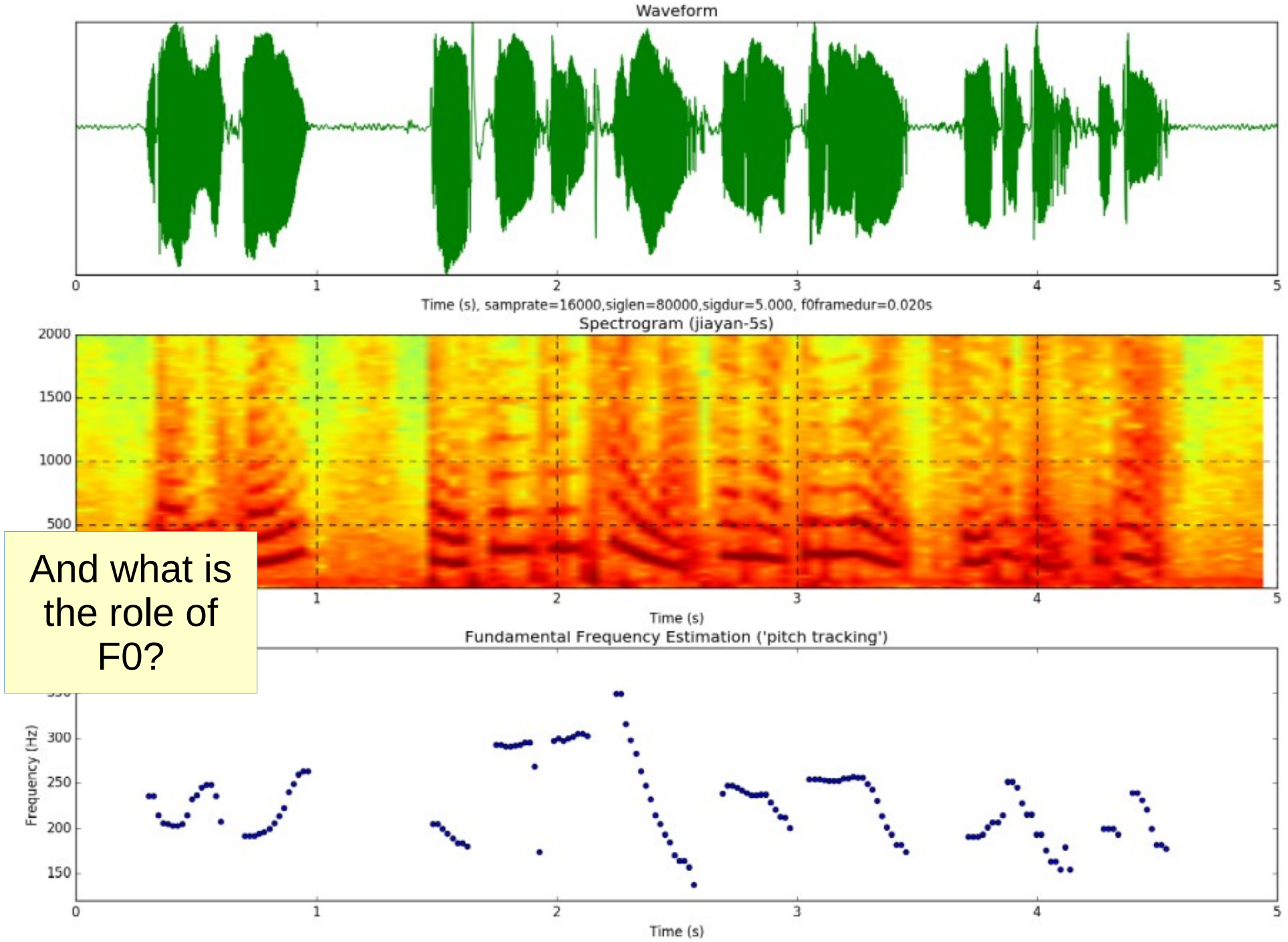
Long-term spectrum: not so useful

Modulation Regularities – Maybe the Long-Term Spectrum?



What is happening here?

Modulation Regularities – Maybe F0?



And what is the role of F0?

The Reality of Rhythm!

Amplitude Modulation, Demodulation and Envelope Spectrum

Amplitude Envelope Modulation

↓

Amplitude Envelope Demodulation

absolute value of Hilbert transform

(or rectification & peak-picking / LP filtering)

↓

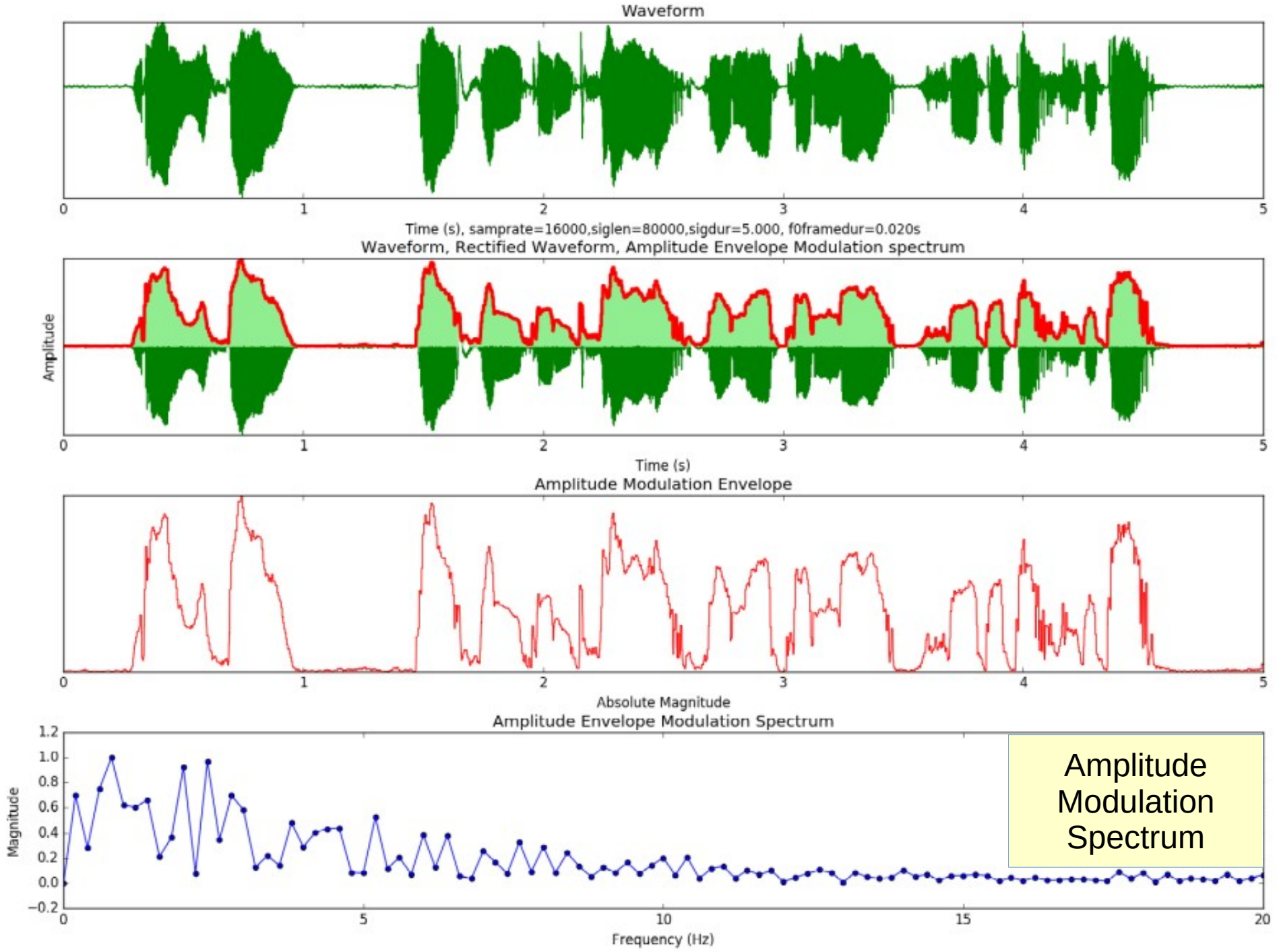
Spectral slice (FFT)

↓

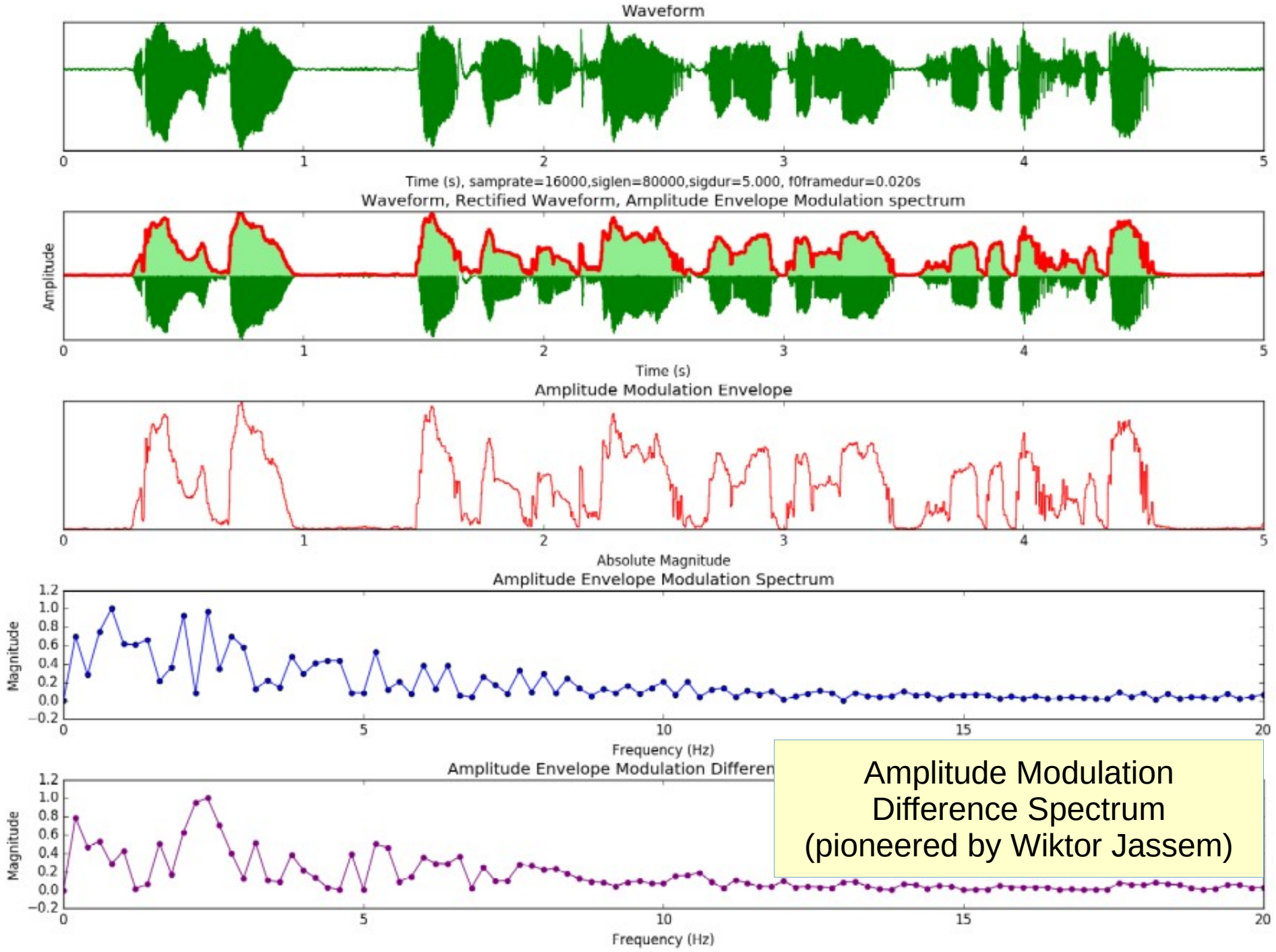
Spectral Zone Edge Detection

Amplitude Envelope Modulation Spectrum (AEMS, AMS, EMS) Frequency Zones

Amplitude Demodulation and the Amplitude Envelope Spectrum



AM Spectrum and the AM Difference Spectrum



Amplitude Modulation Difference Spectrum (pioneered by Wiktor Jassem)

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?

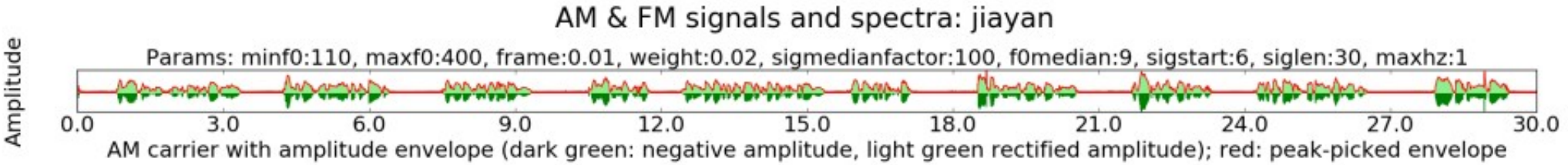
Preliminary answer:

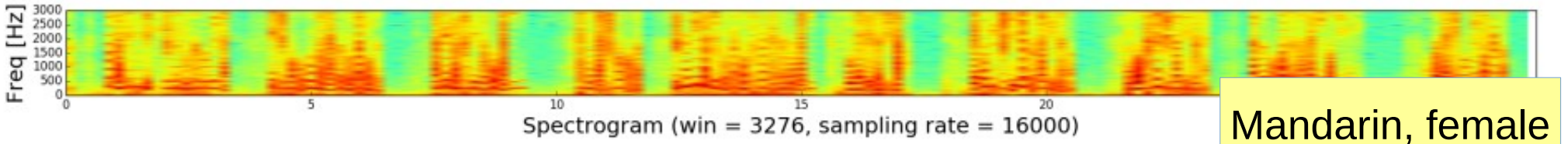
Yes, they do correlate to some extent, but not overwhelmingly strongly!

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

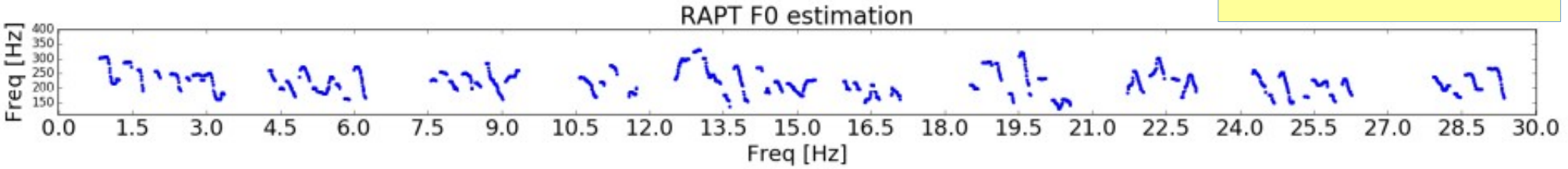
I will look at both AM and FM spectra.

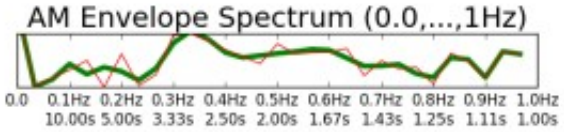
AM and FM Demodulation

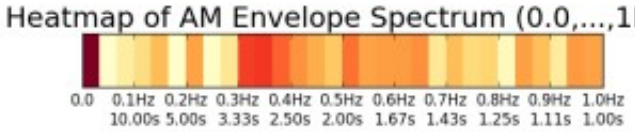


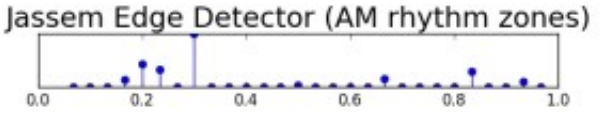


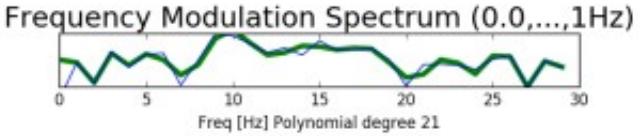
Mandarin, female
30 sec, < 1Hz





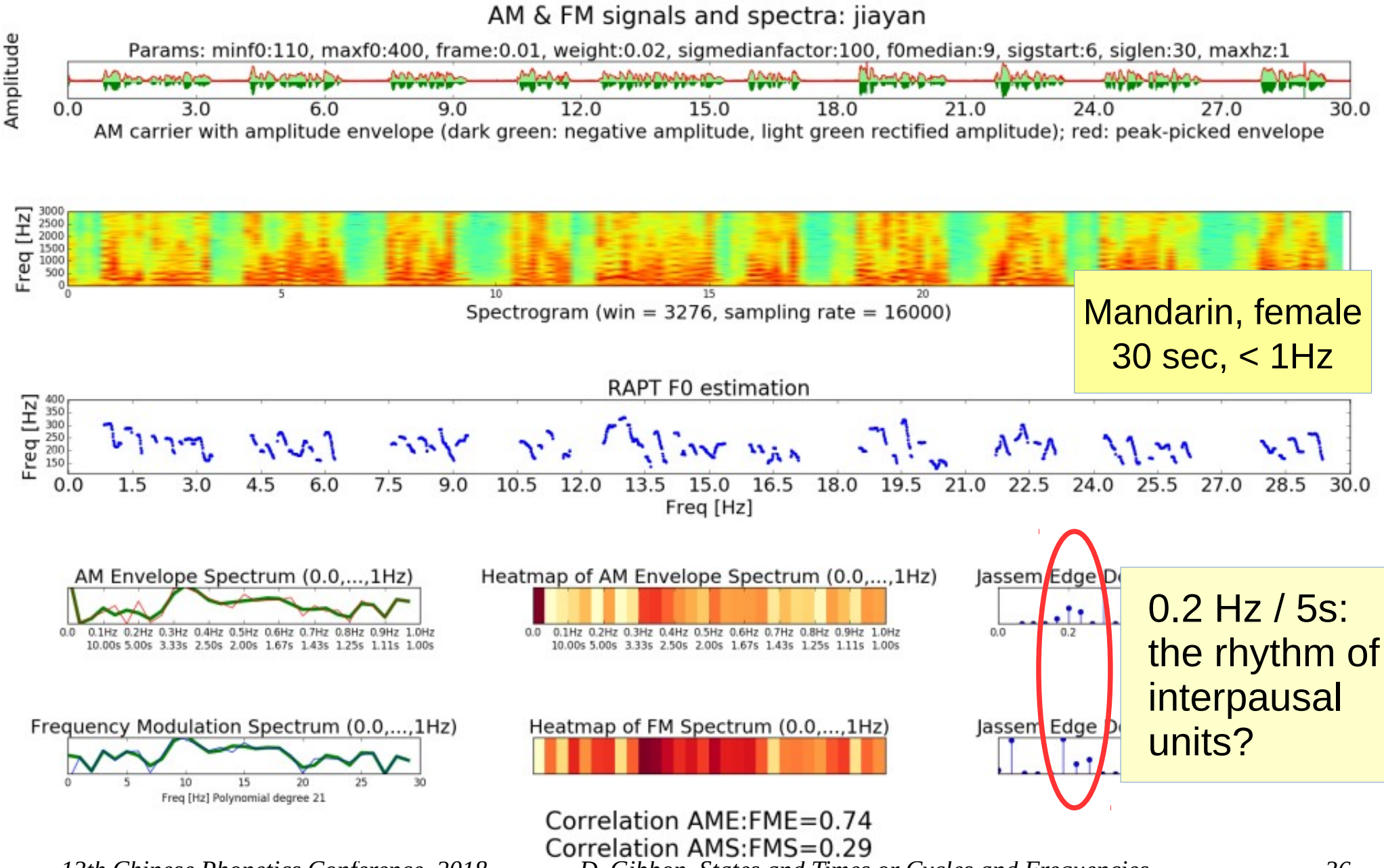






Correlation AME:FME=0.74
Correlation AMS:FMS=0.29

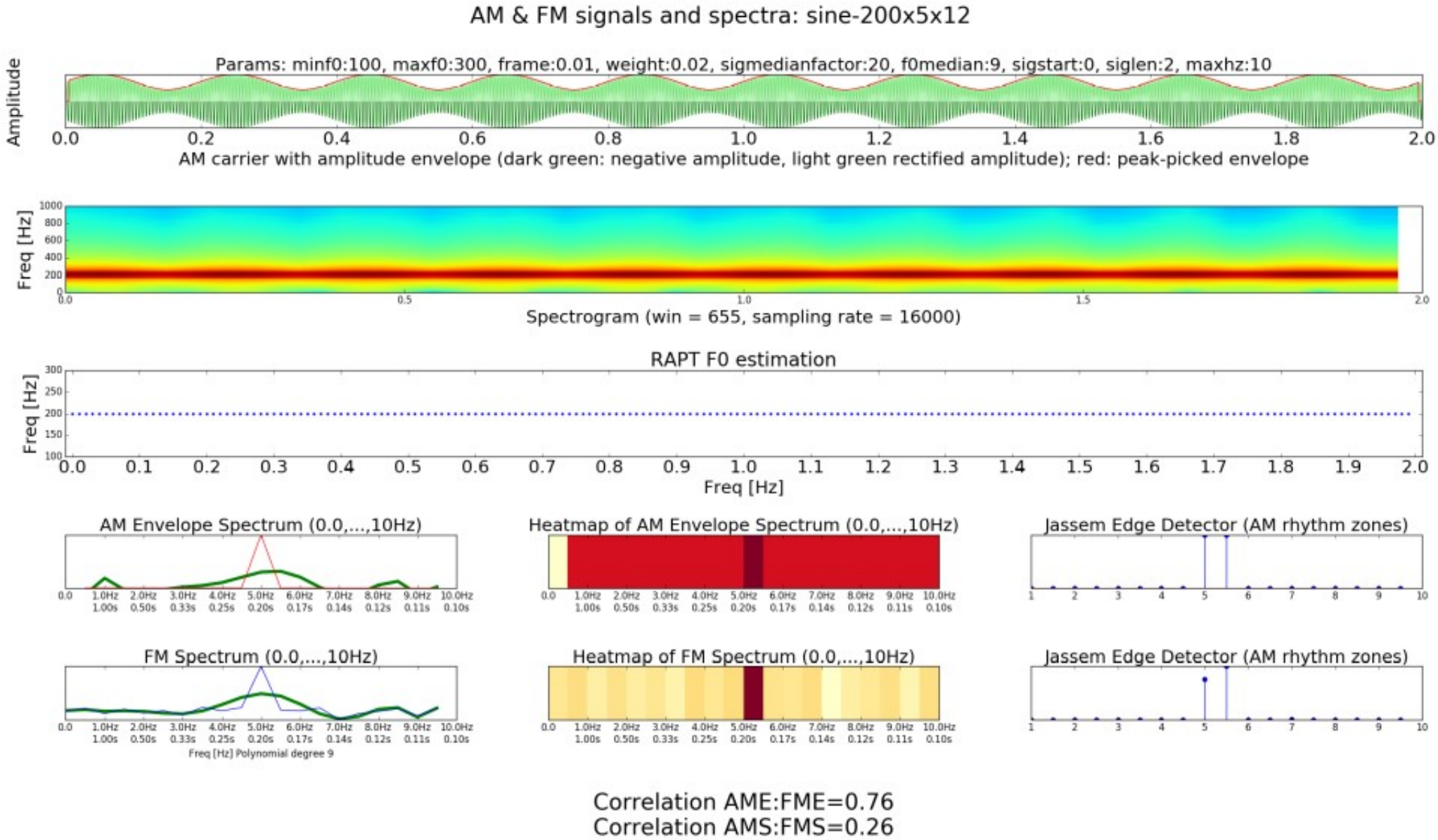
AM and FM Demodulation



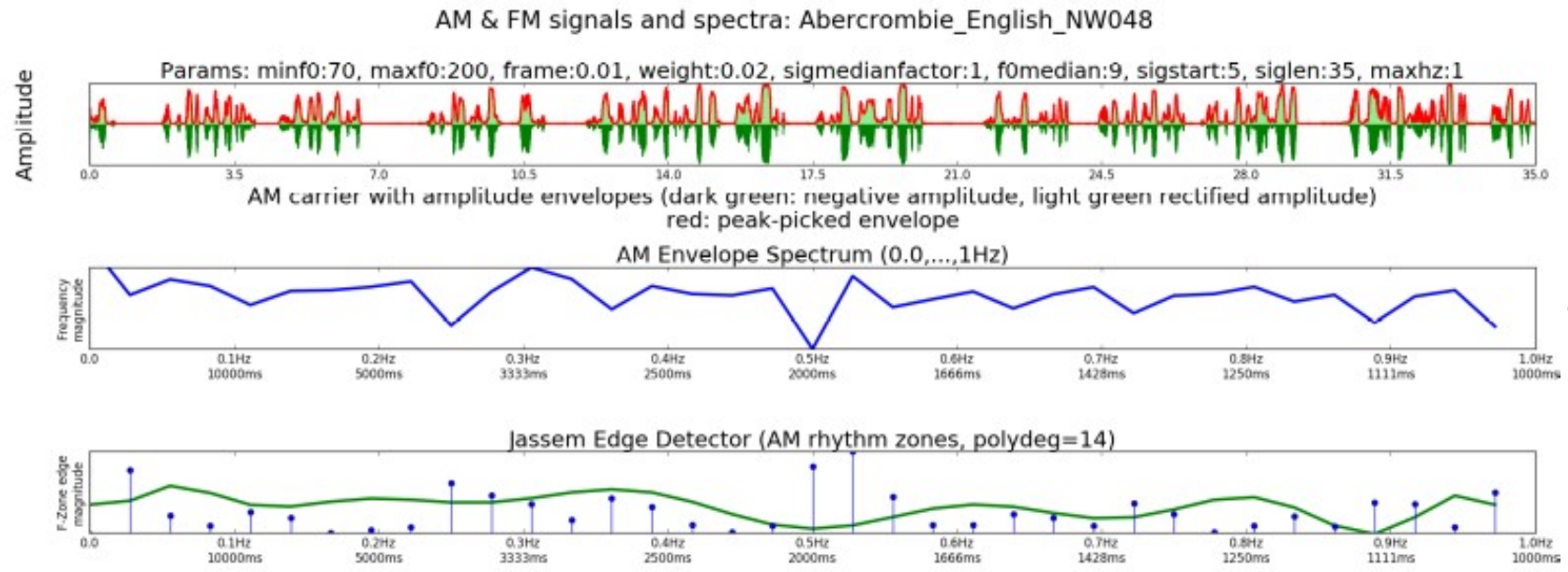
Mandarin, female
30 sec, < 1Hz

0.2 Hz / 5s:
the rhythm of
interpausal
units?

AM and FM Demodulation and Spectrum Analysis: Calibration

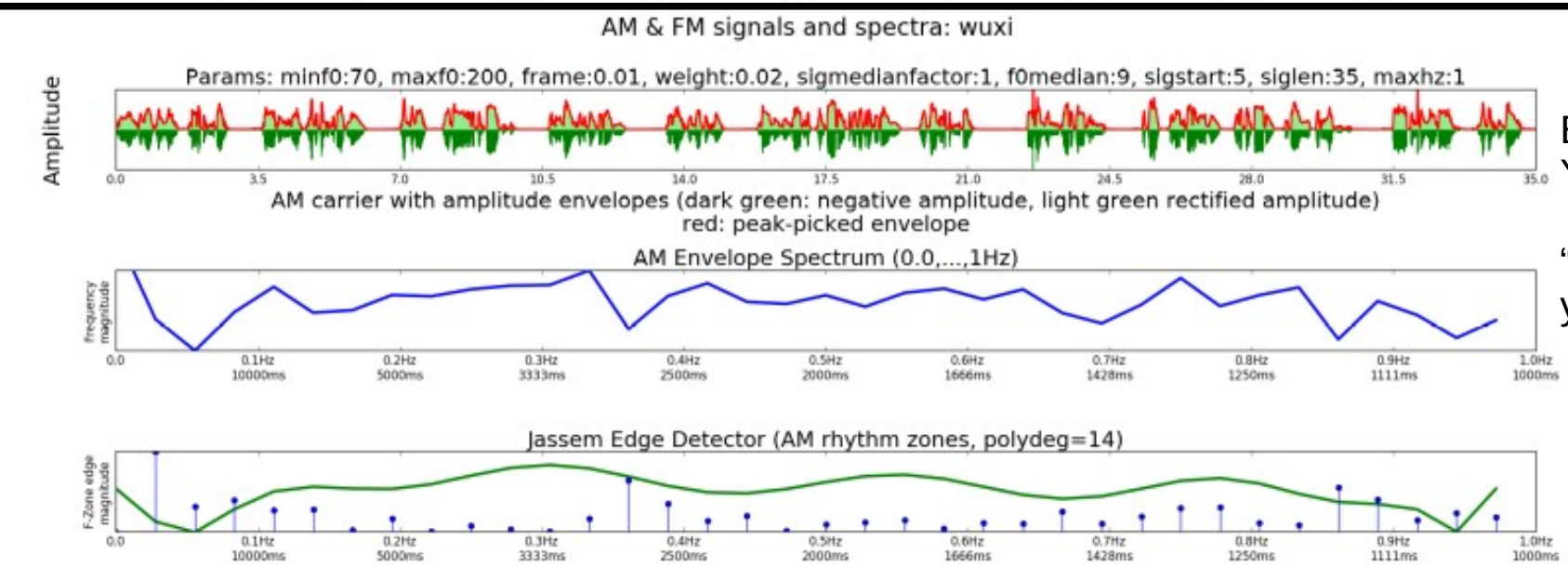


Envelope Demodulation: Extending to Discourse Spectra



English (RP)
Edinburgh corpus

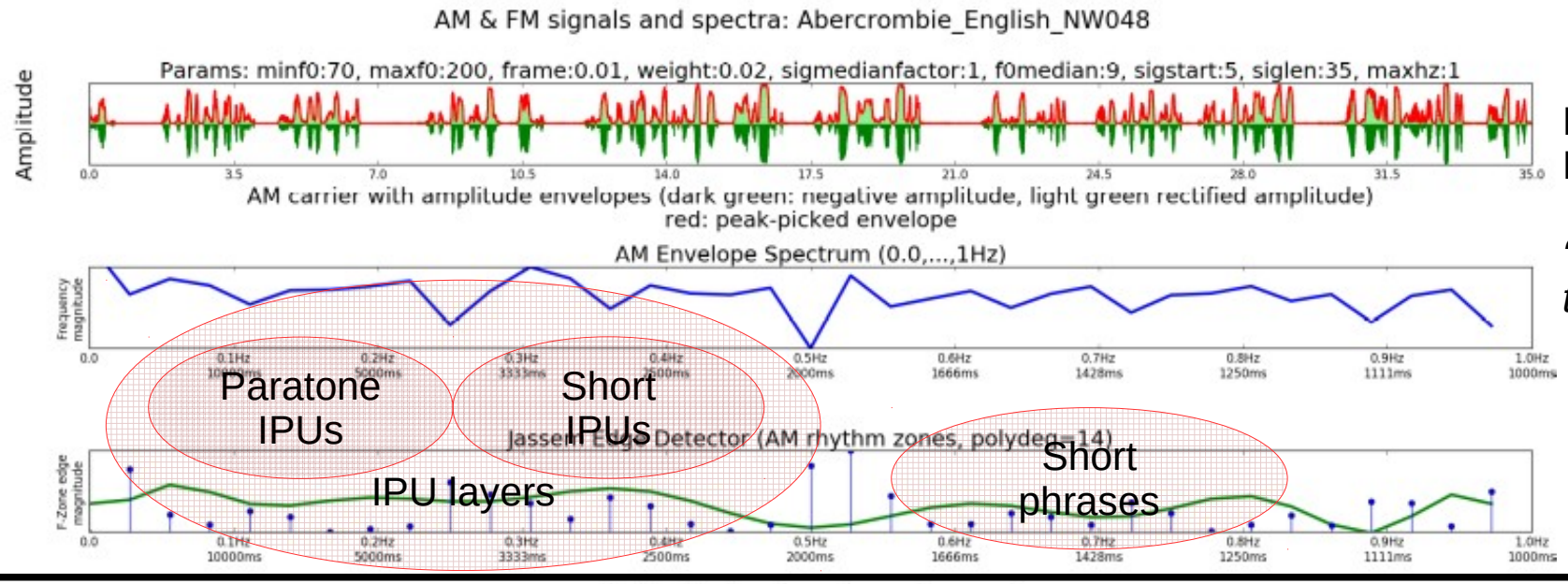
"The North Wind and the Sun"



Beijing Mandarin
Yu corpus

"bei3 feng1 gen1 tai4 yang2"

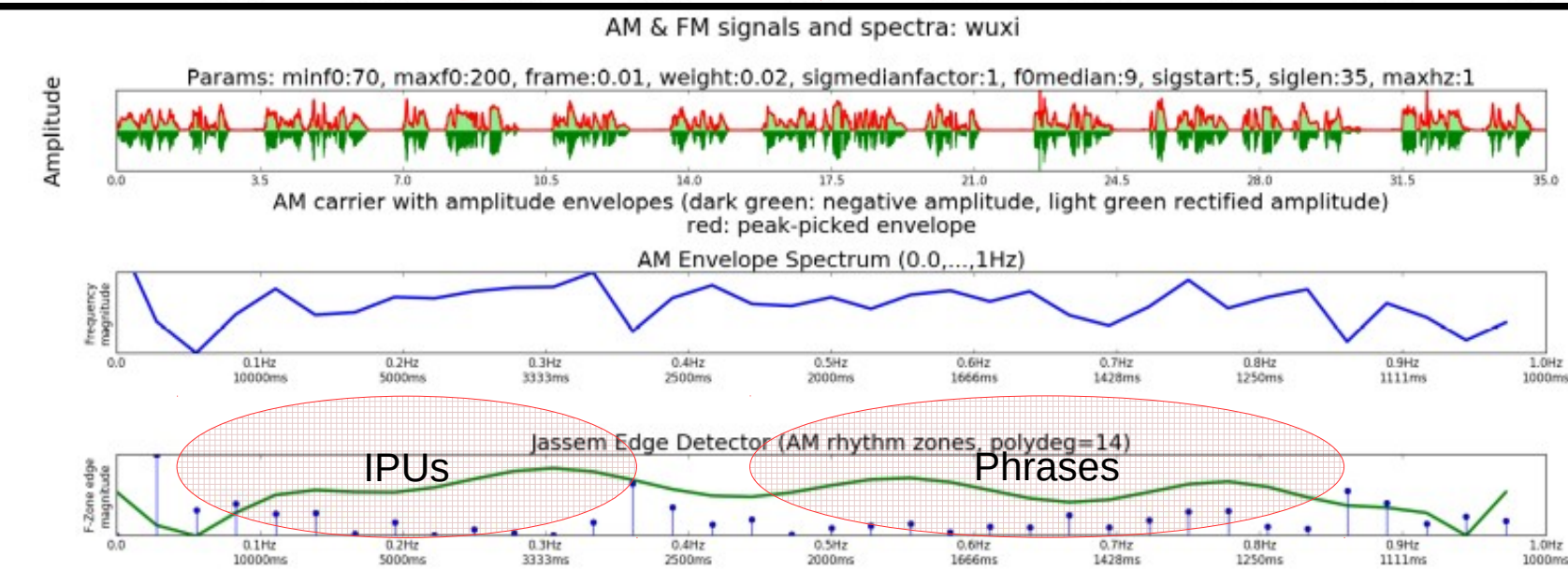
Envelope Demodulation: Extending to Discourse Spectra



English (RP)
Edinburgh corpus

"The North Wind and the Sun"

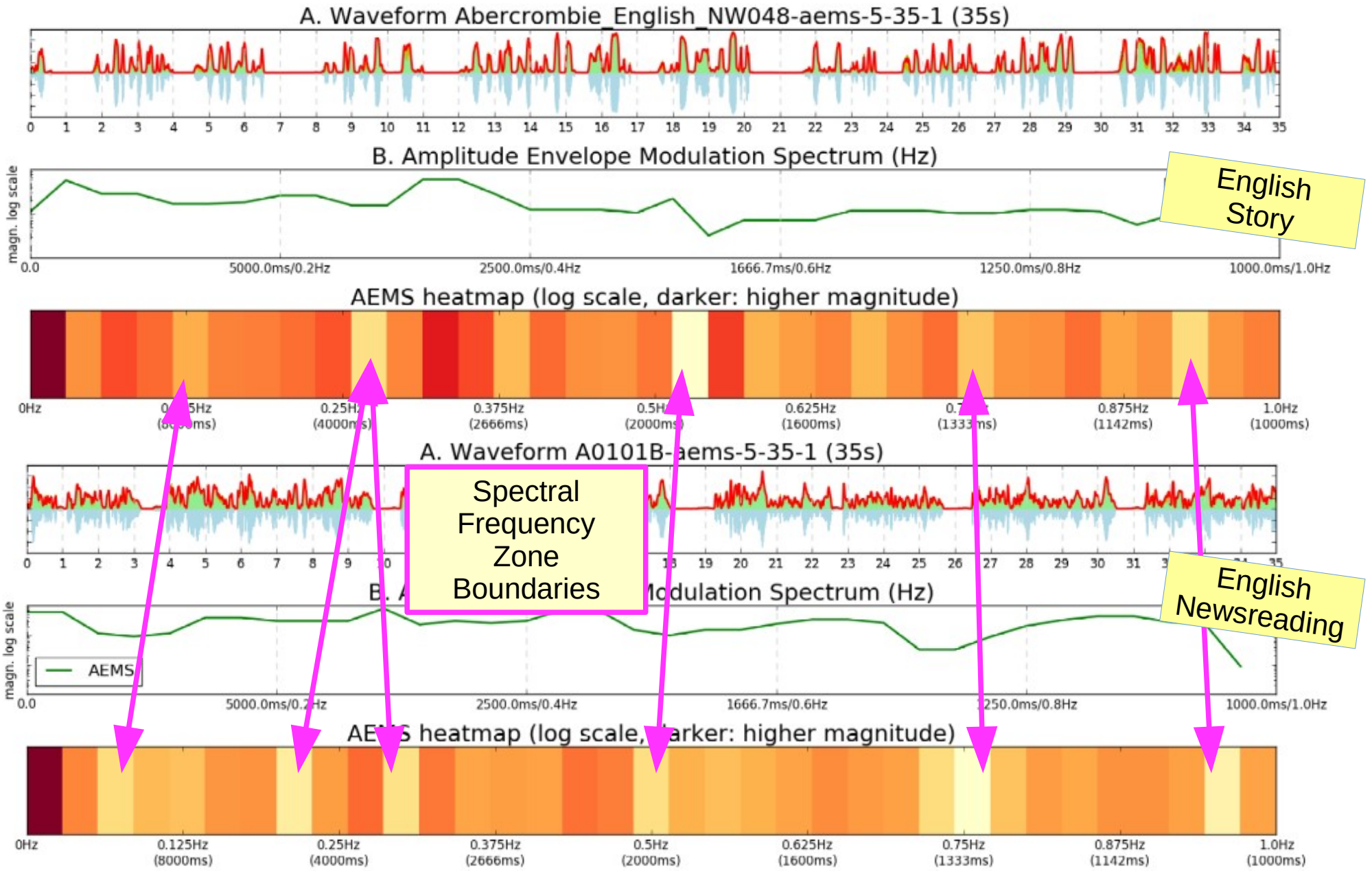
1 Hz



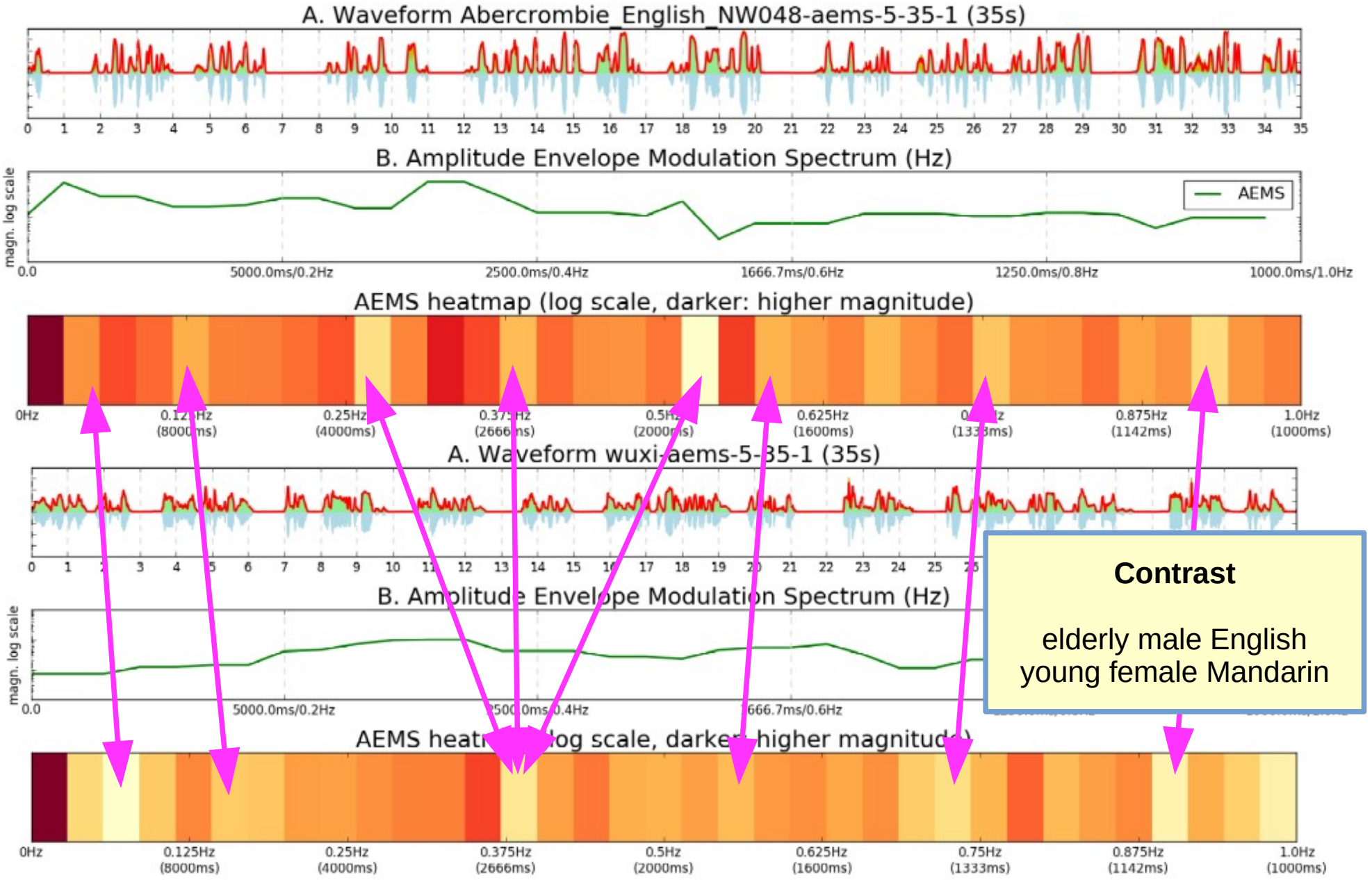
Beijing Mandarin
Yu corpus

"bei3 feng1 gen1 tai4 yang2"

Extending to Discourse Spectra: English Genres



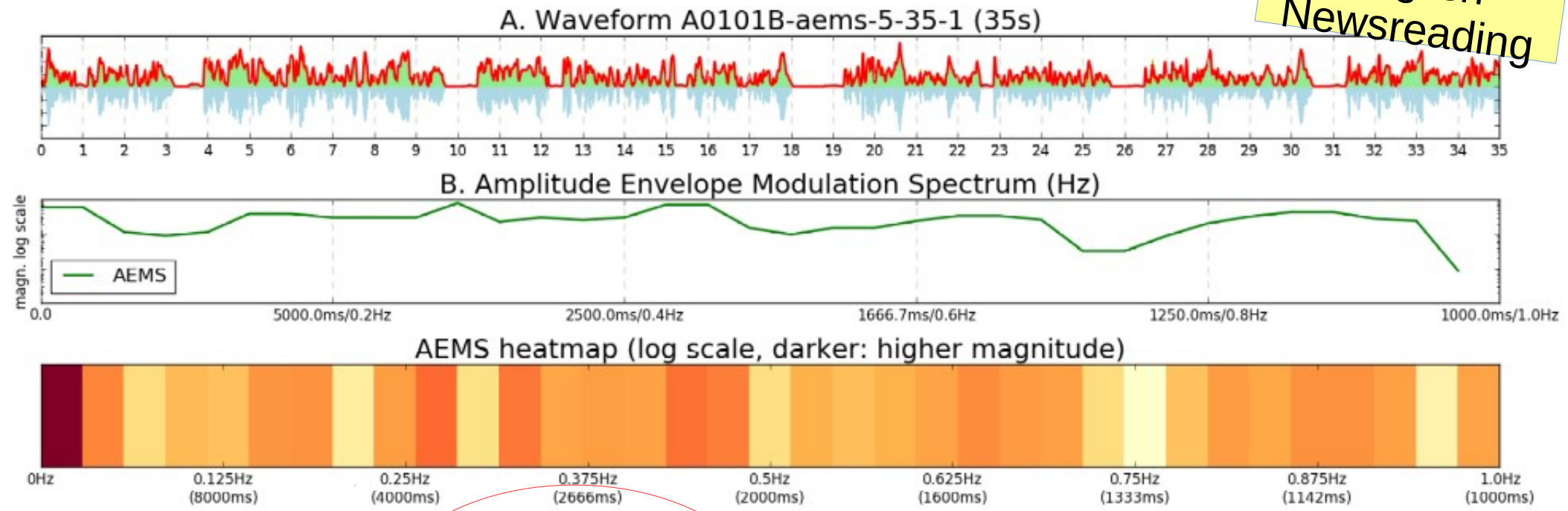
Extending to discourse spectra: English-Mandarin



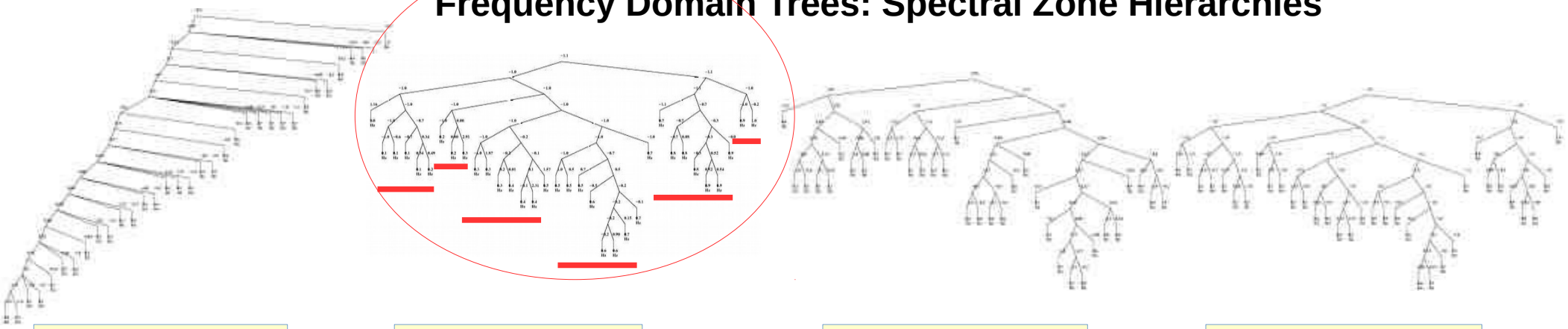
Rhythms in Syntagmatic, Compositional Spectral Trees

AM and FM Demodulation and Spectral Tree Induction

English Newsreading



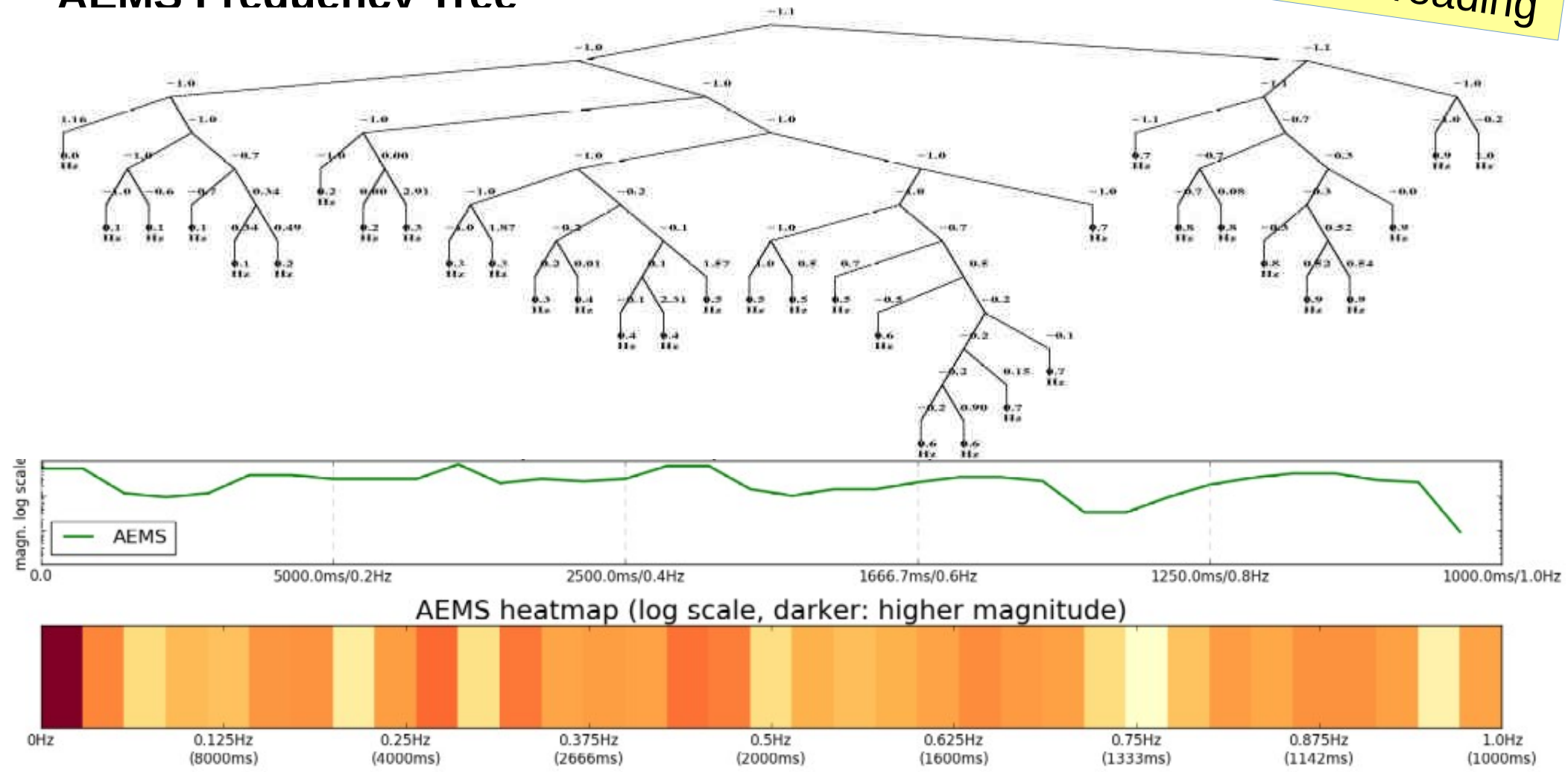
Frequency Domain Trees: Spectral Zone Hierarchies



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

AM and FM Demodulation and Spectral Tree Induction

English Newsreading

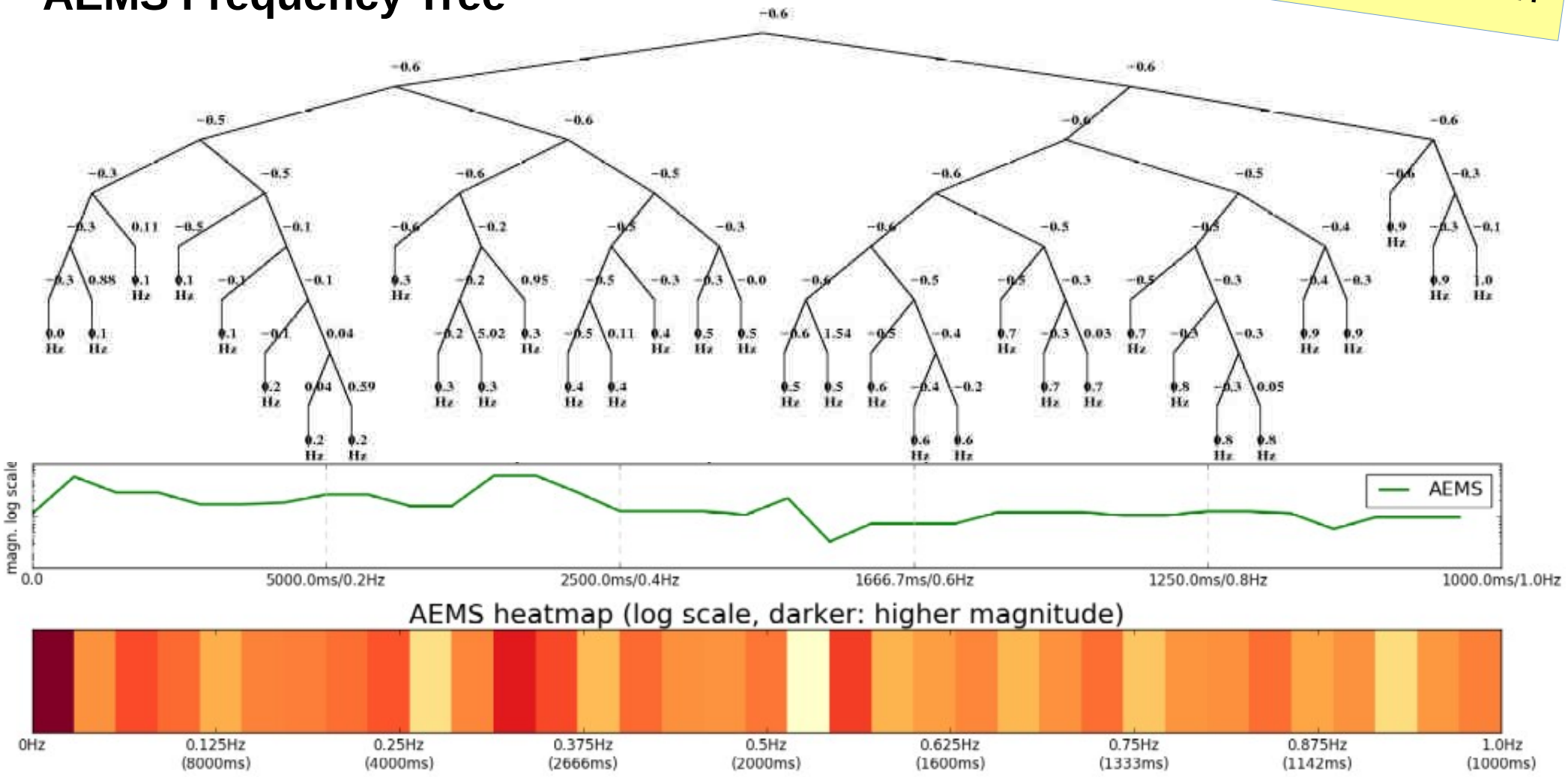


L-strong, <

AM and FM Demodulation and Spectral Tree Induction

English
North Wind & Sun

AEMS Frequency Tree

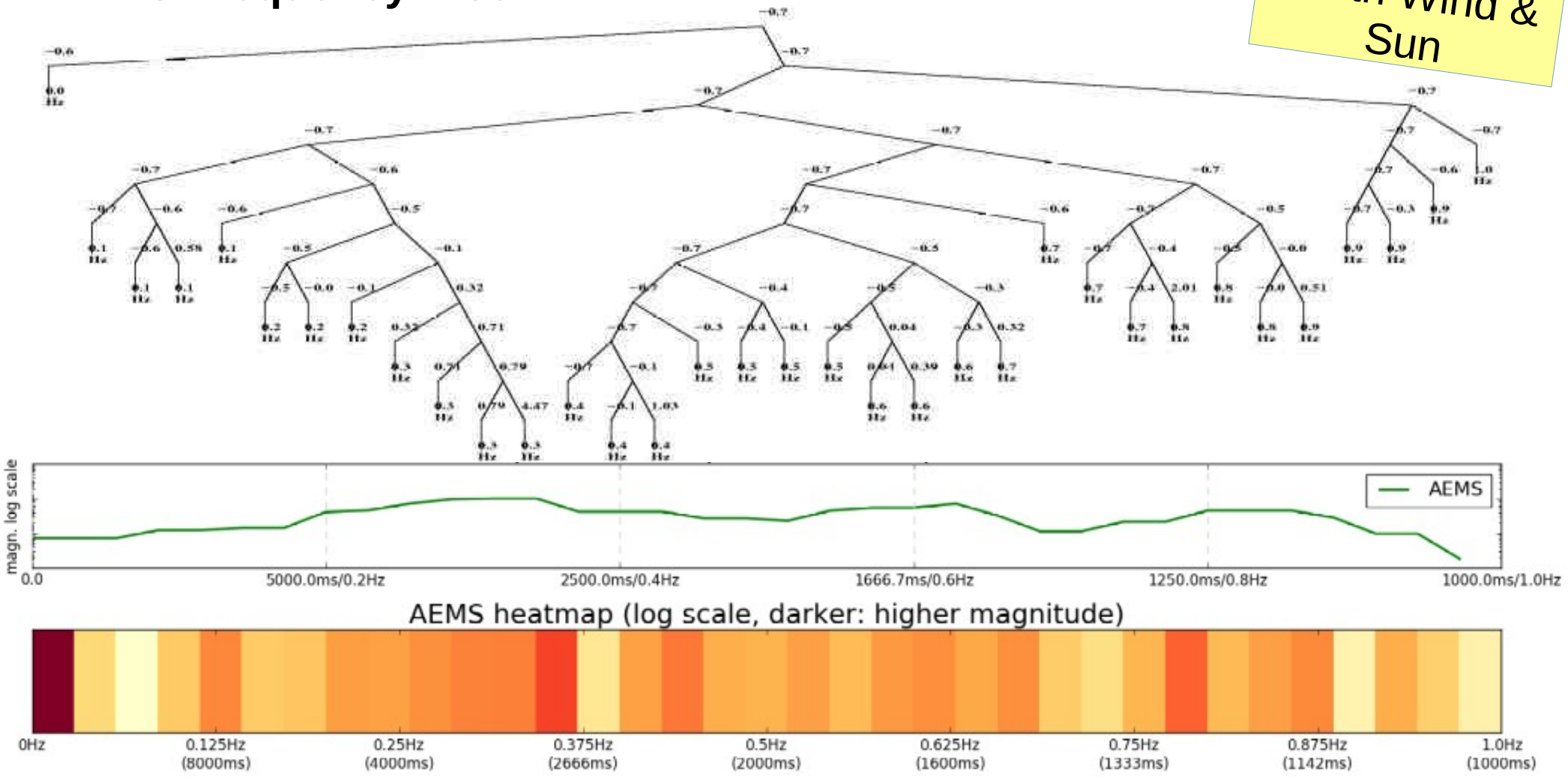


L-strong, <

AM and FM Demodulation and Spectral Tree Induction

Mandarin North Wind & Sun

AEMS Frequency Tree



L-

strong,

<

Rhythms in Paradigmatic, Classificatory Spectral Trees

Consistency of AM Spectra: Rhythm Classification

Data:

Chunks from “The North Wind and the Sun”

Male, English: 40s

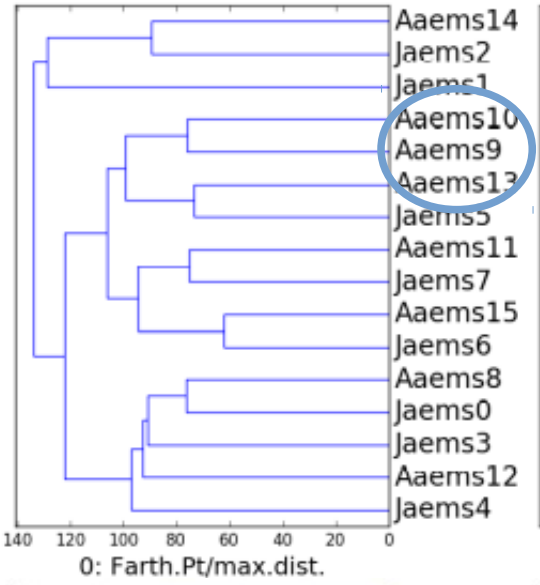
Female, Mandarin: 40s

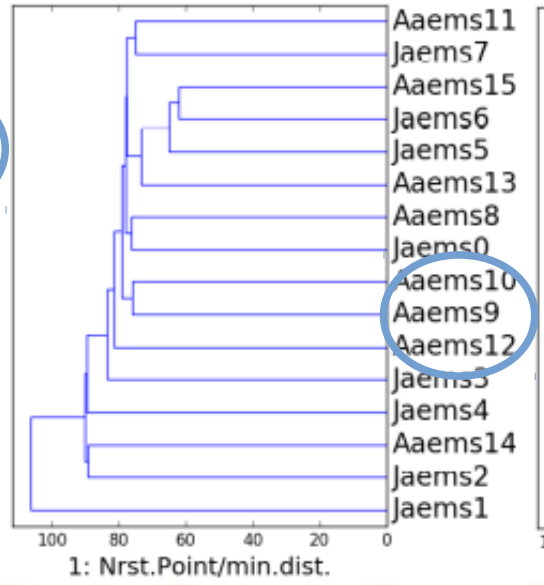
Method:

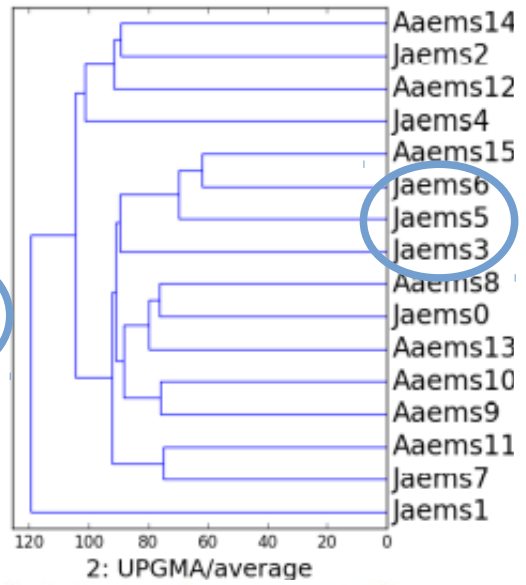
Comparison of non-overlapping adjacent 5s audio chunks

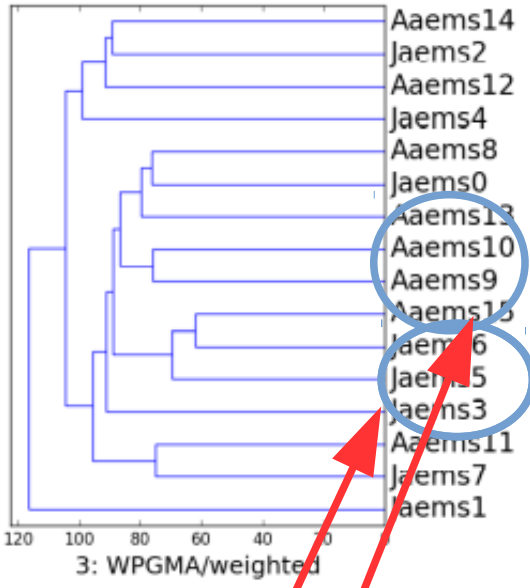
- offsets into recording: 0, 5, 10, 15, 20, 25, 30, 35
- AEMS for each chunk
- Inter-speaker comparison (AEMS pointwise means, r=0.82)
- Classification by hierarchical similarity / distance:
  - Pearson Distance: 1 – r , range 0 … 2
  - comparison of 7 classifiers
  - similar classification methods used in dialectometry, stylometry, language typology

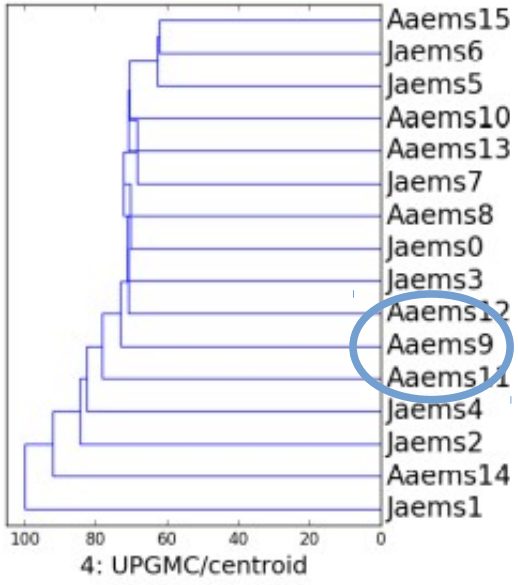
Consistency of Spectra: Rhythm Classification

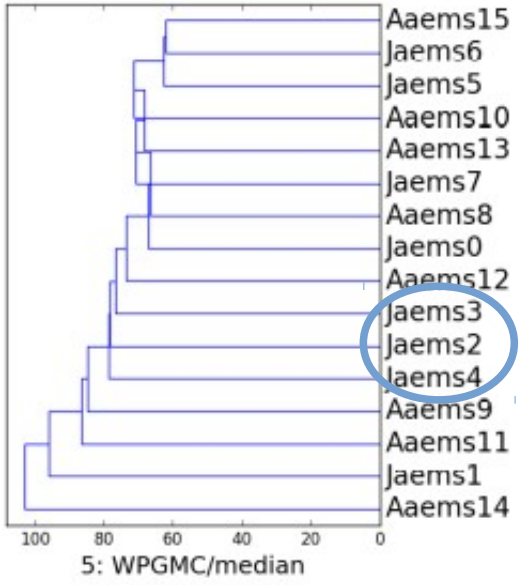


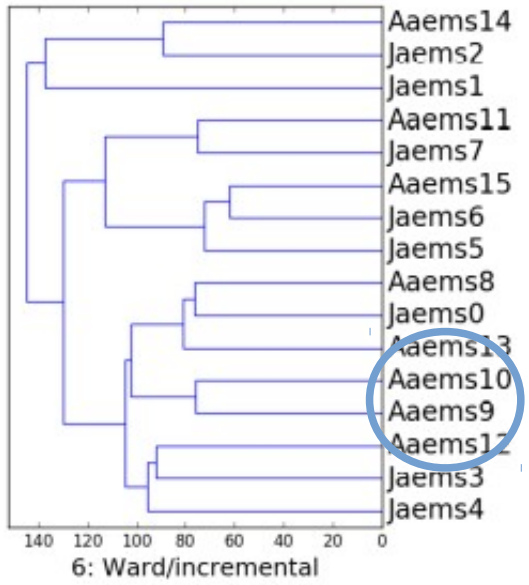










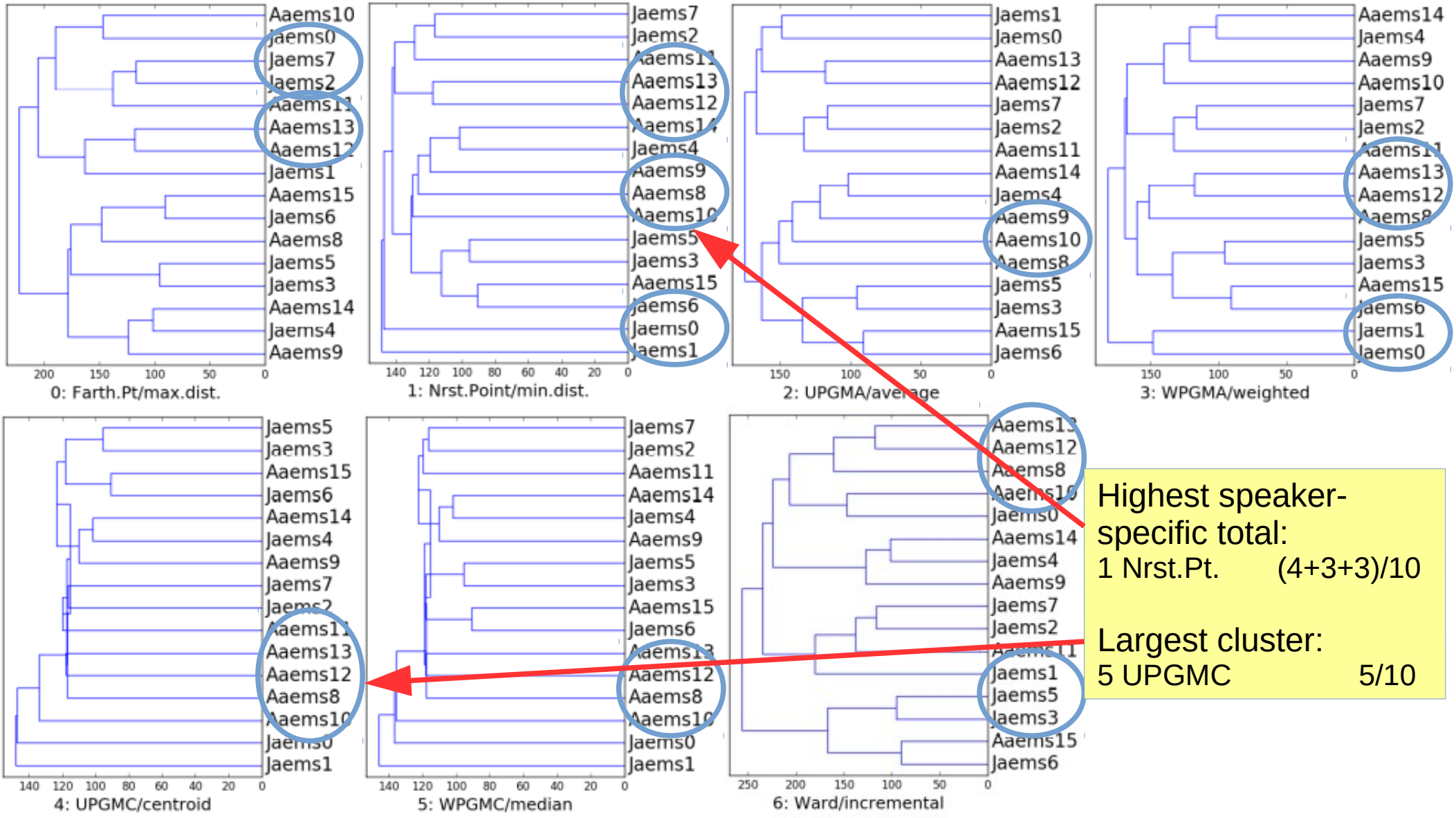


Similarity criterion:
>3 adjacent same-speaker settings

Largest per speaker score: (4+3)/16

Largest cluster: 4/16

Consistency of Spectra: Rhythm Classification



Highest speaker-specific total:
1 Nrst.Pt. (4+3+3)/10

Largest cluster:
5 UPGMC 5/10

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,

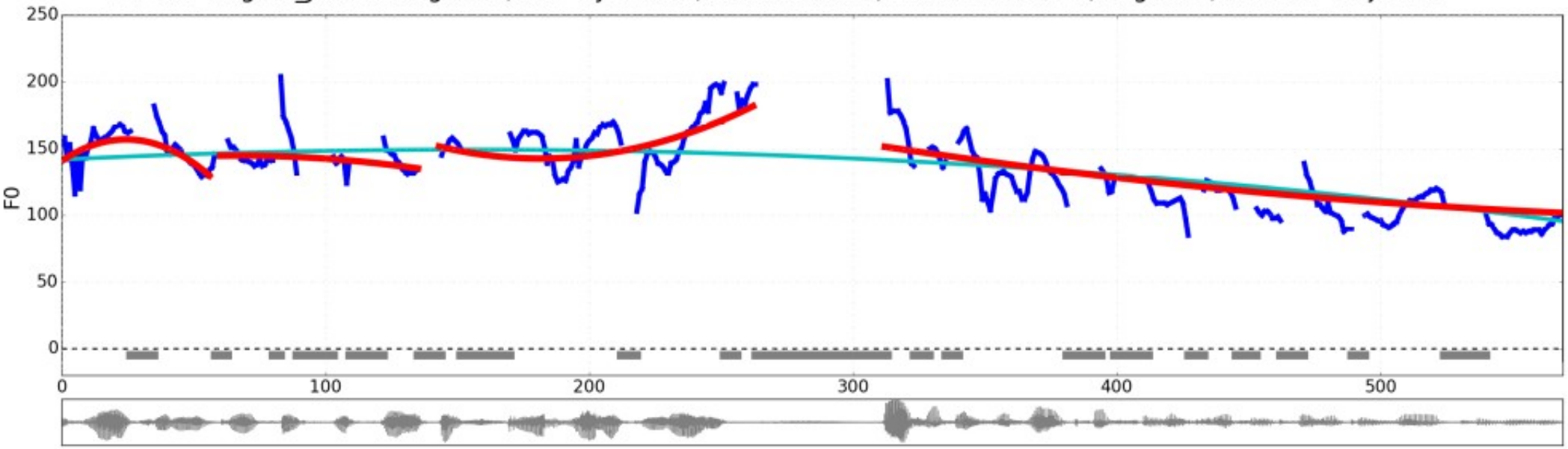
- **frequency modulation and rhythm also come first**
  - **emotional cries**
    Wermke, Sebastian-Galles
  - **turn-taking**
    cf. the ‘bootstrapping’ literature

    the infant ‘twin-talk’ videos on YouTube ☺

Discourse Rhythms: Long FM contours

Question: rising utterance contour

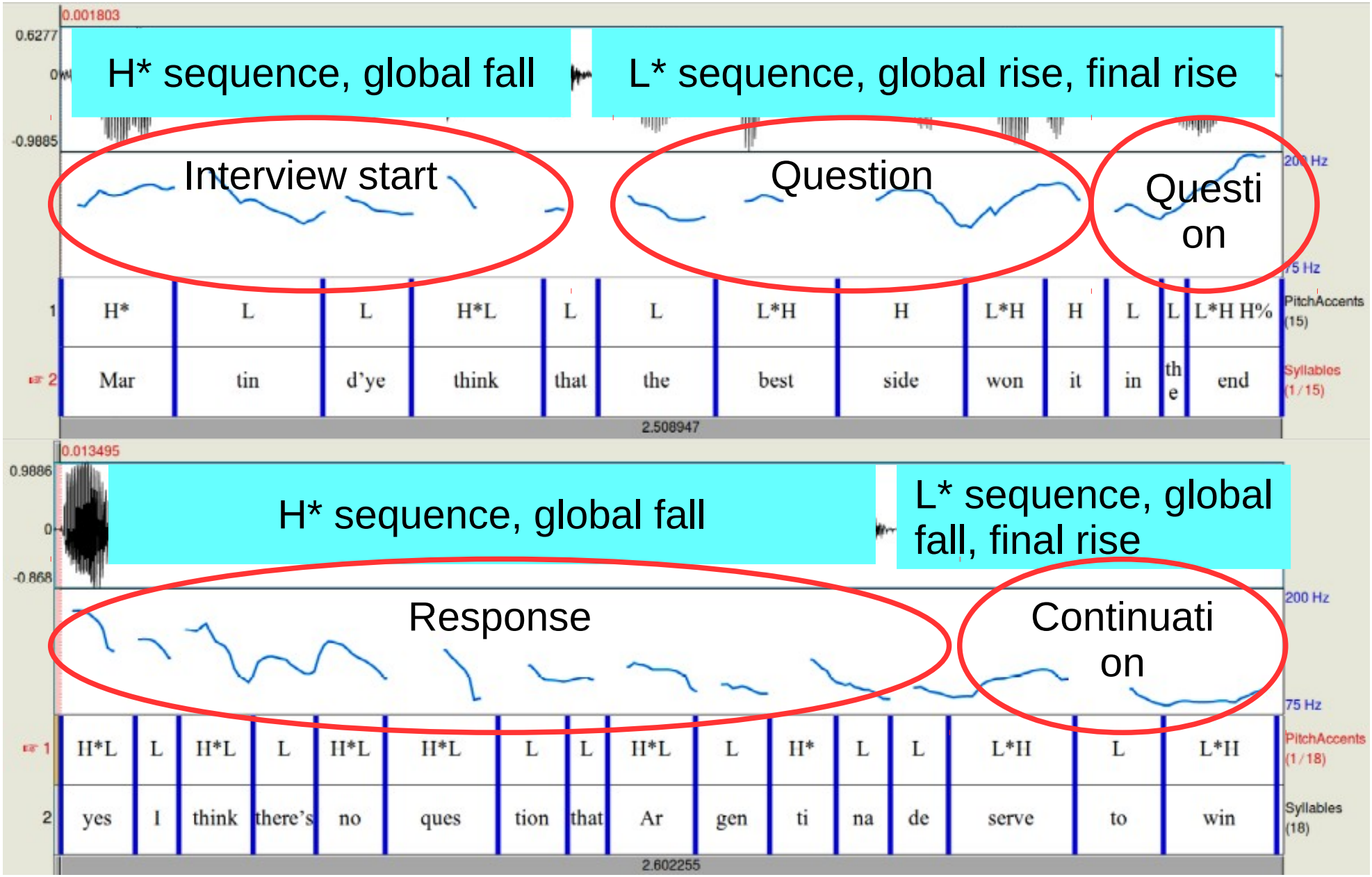
Answer: falling utterance contour



Question+Answer: rising-falling adjacency pair contour

syntagmatic entrainment

Discourse Rhythms: Long FM contours



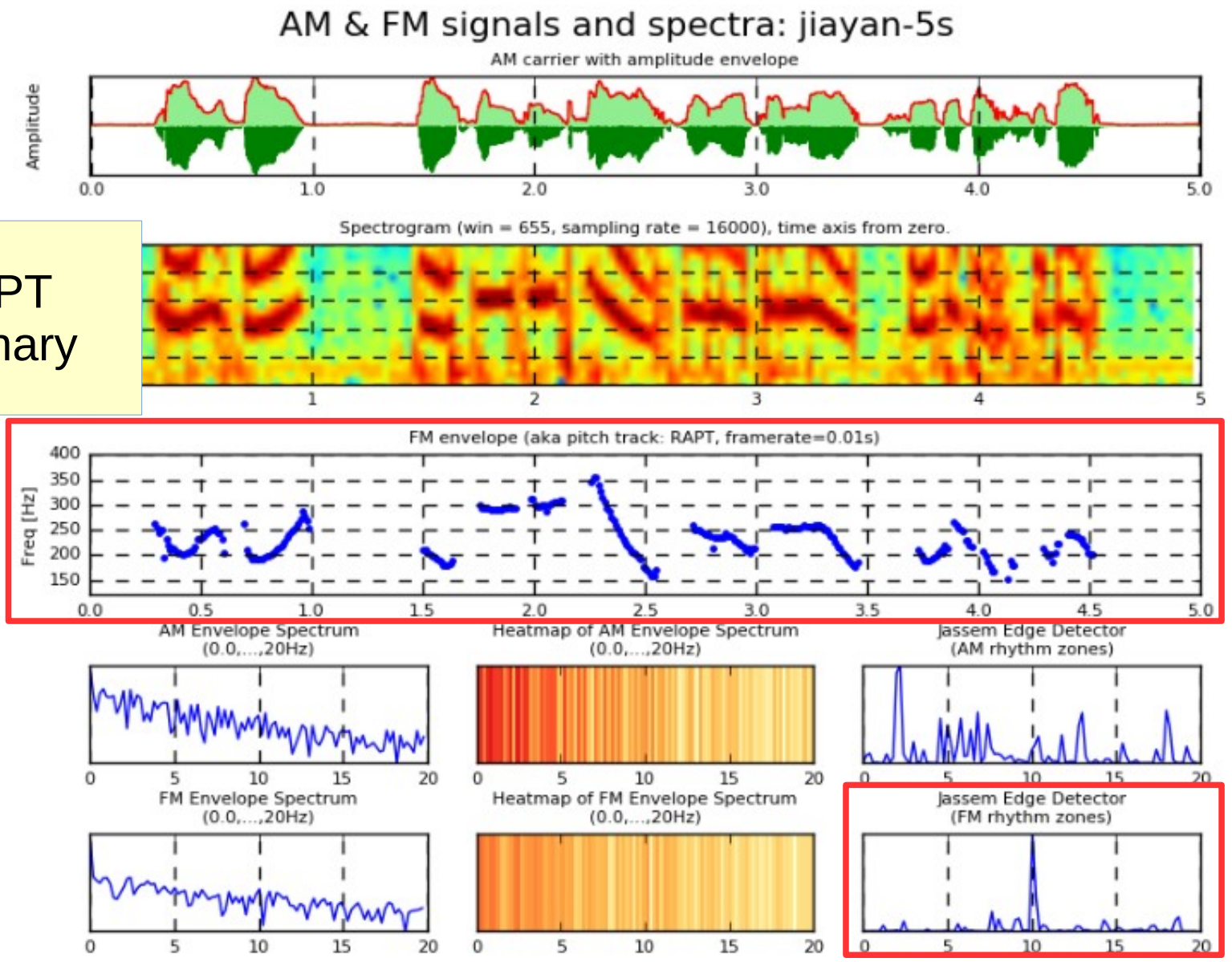
But there are Methodological Problems in F0 / Pitch estimation

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

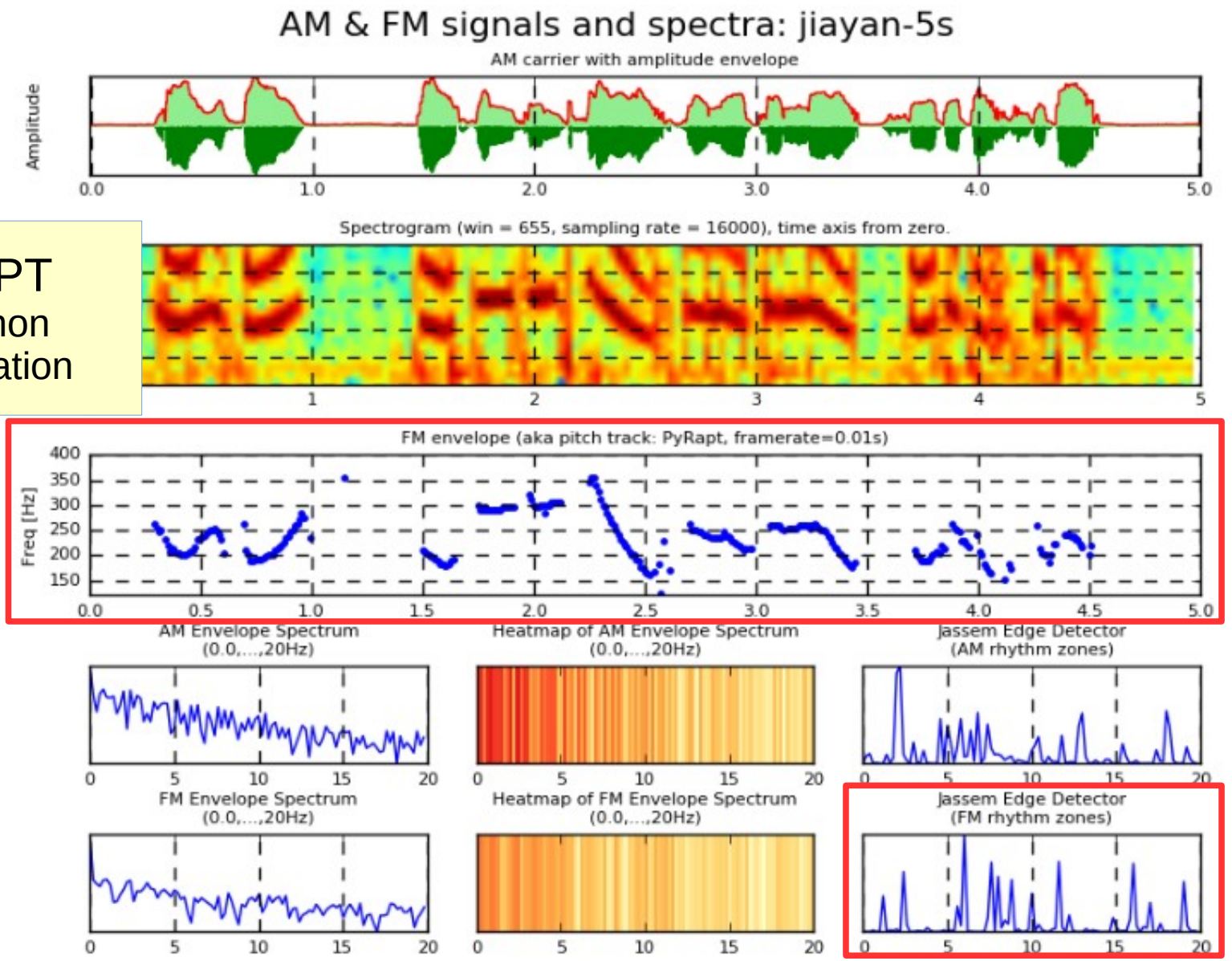
RAPT
C binary



RAPT (Python emulation)

Daniel Gaspari

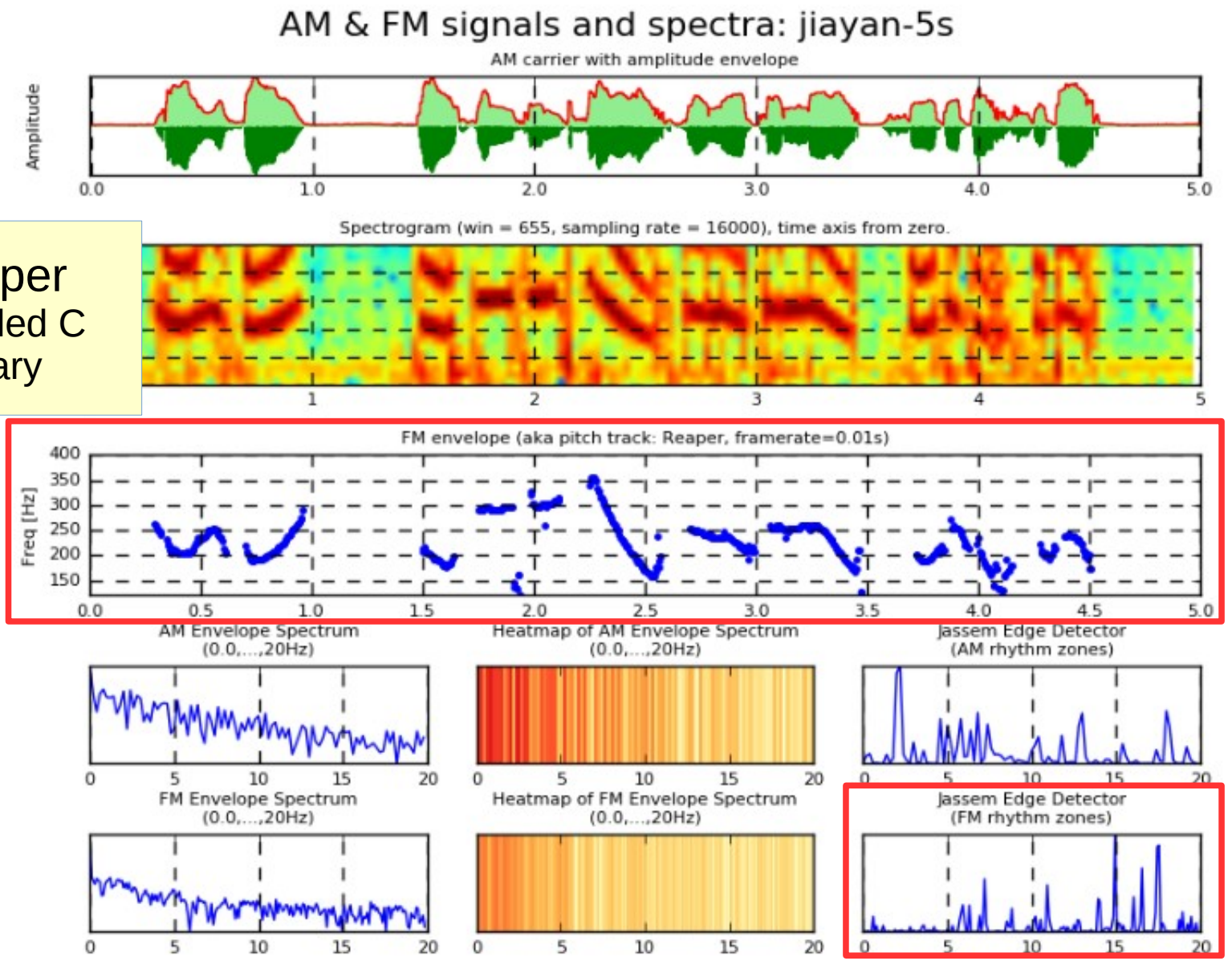
RAPT
Python
emulation



Reaper (Robust Epoch And Pitch EstimatoR)

David Talkin

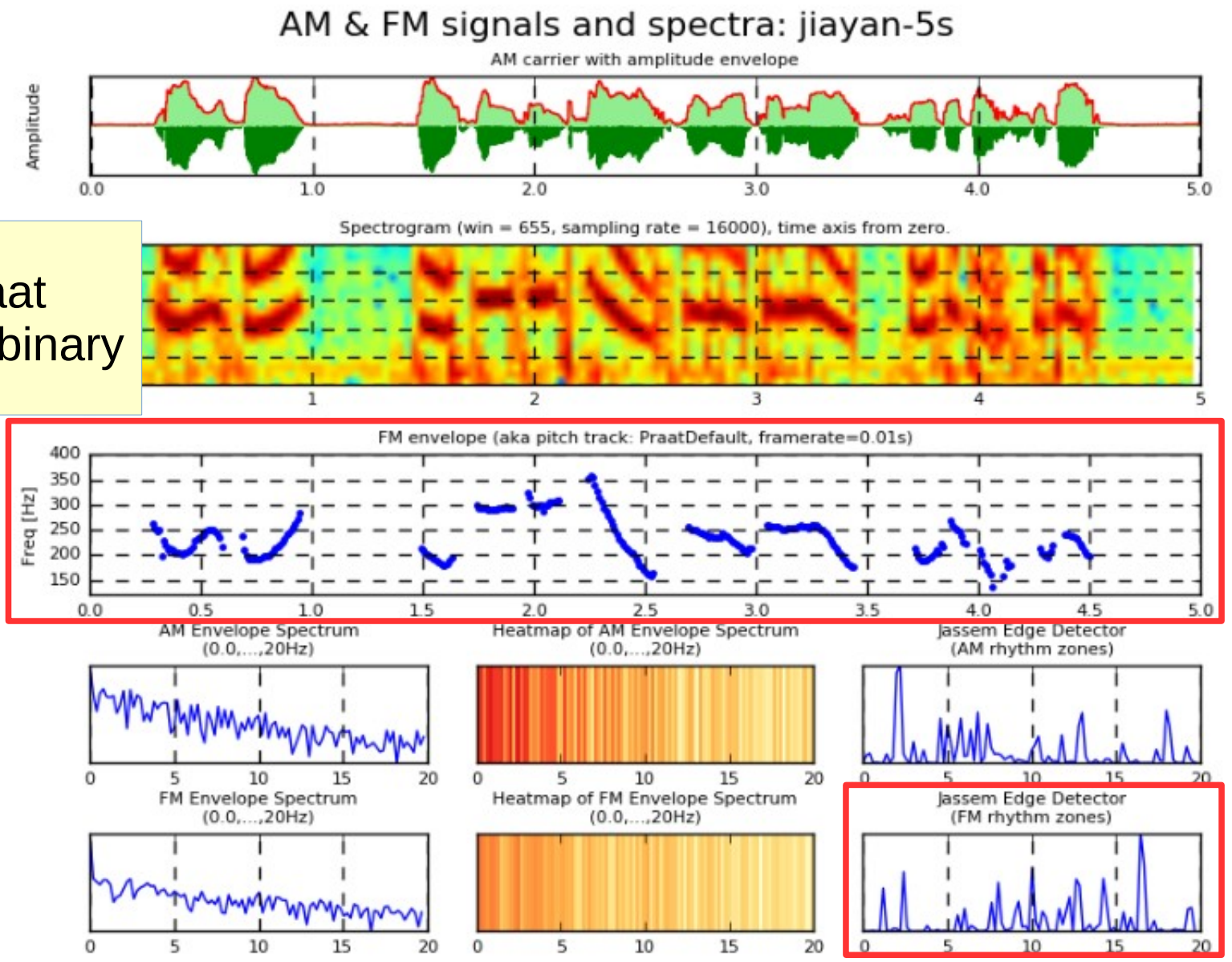
Reaper
compiled C
binary



Praat

Paul Boersma

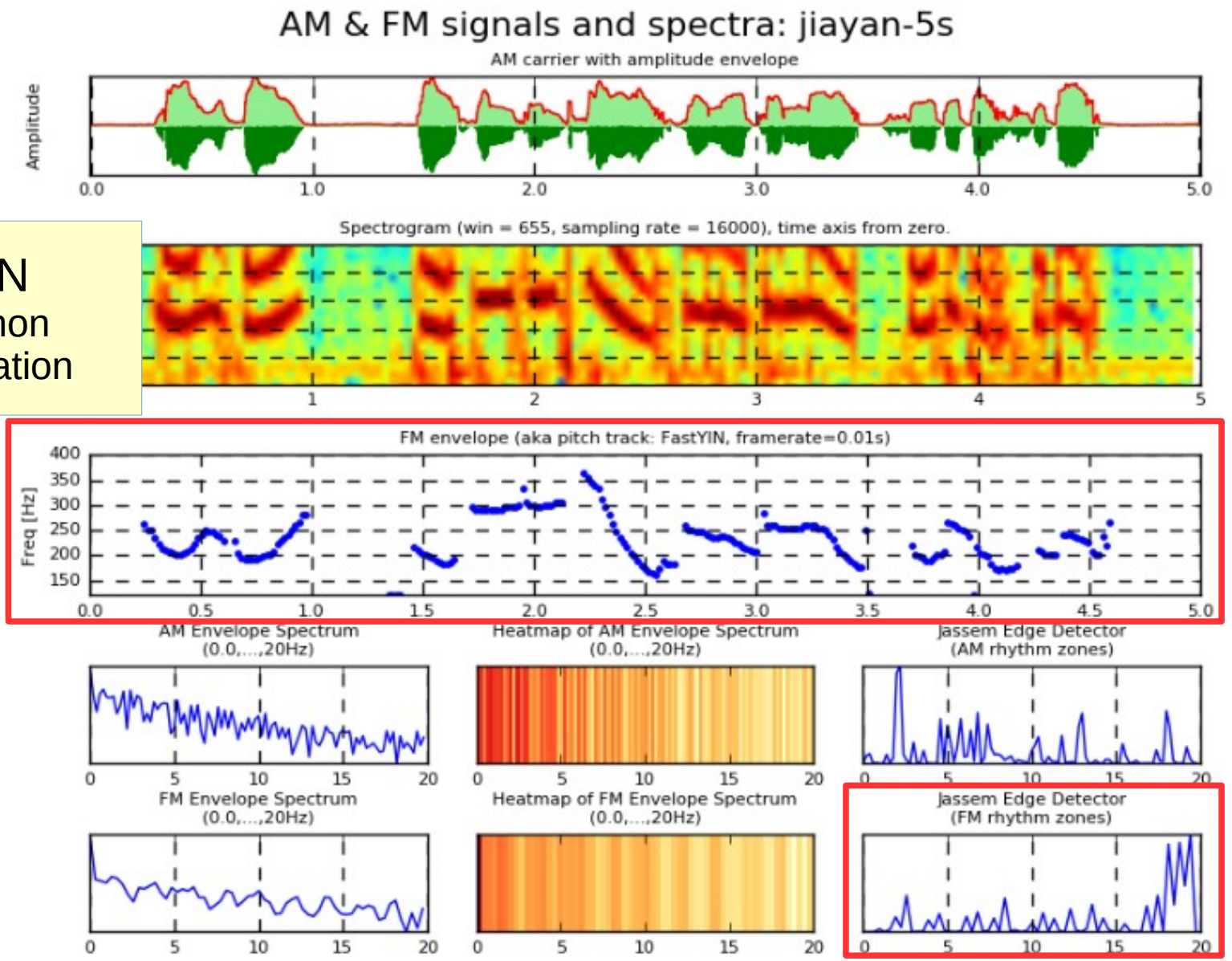
Praat script+binary



YIN (as opposed to YANG, Python emulation)

Patrice Guyot

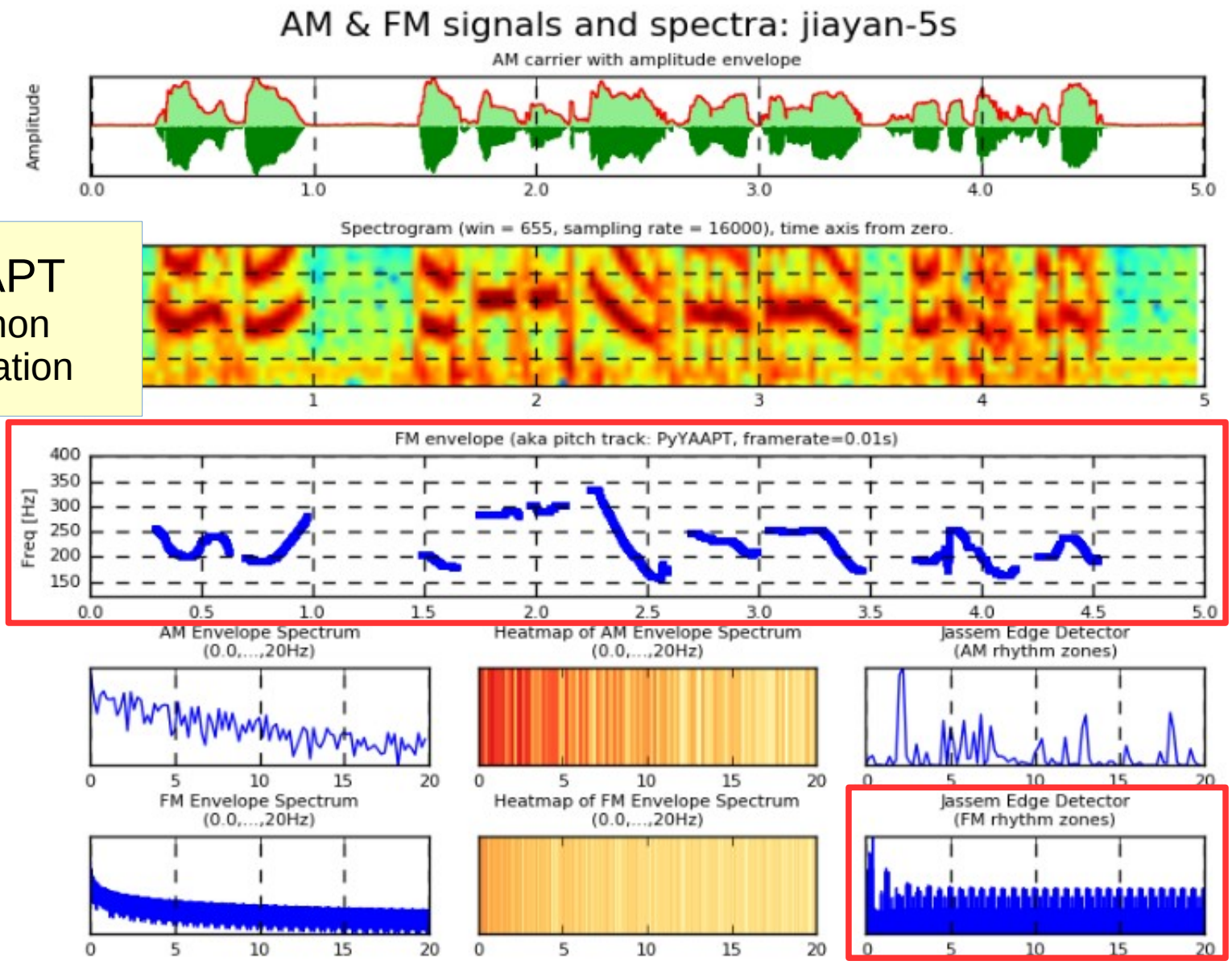
YIN
Python
emulation



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

Bernardo J. B. Schmitt

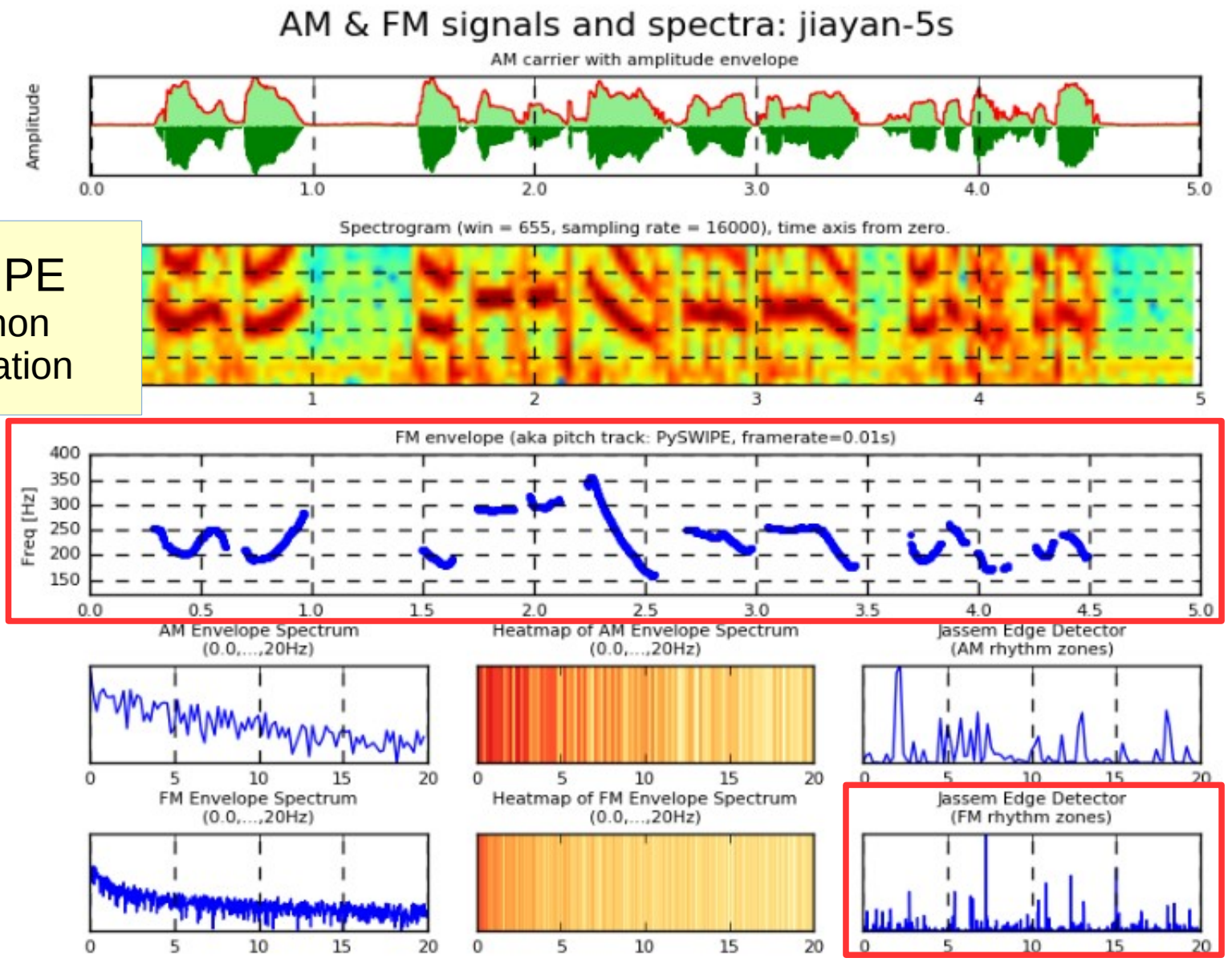
YAAPT
Python
emulation



SWIPE (Square Wave Inspired Pitch Estimator, Python emulation)

Disha Garg

SWIPE
Python
emulation



F0 – Pitch: Methodological Problems

1. Terminology:
   - *articulation rate* (production)
   - *F0* (acoustic transmission)
   - *pitch* (perception)
2. M
   - results
     - Average Magnitude Difference Function (AMDF)
     - FFT peak detection
     - Cepstrum
   - Environment differences
     - Preprocessing: low-pass filter; centre-clipping
     - Postprocessing: moving median

So let's do somthing about it

- do-it-yourself F0 estimation
- with full control over all parameters.

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

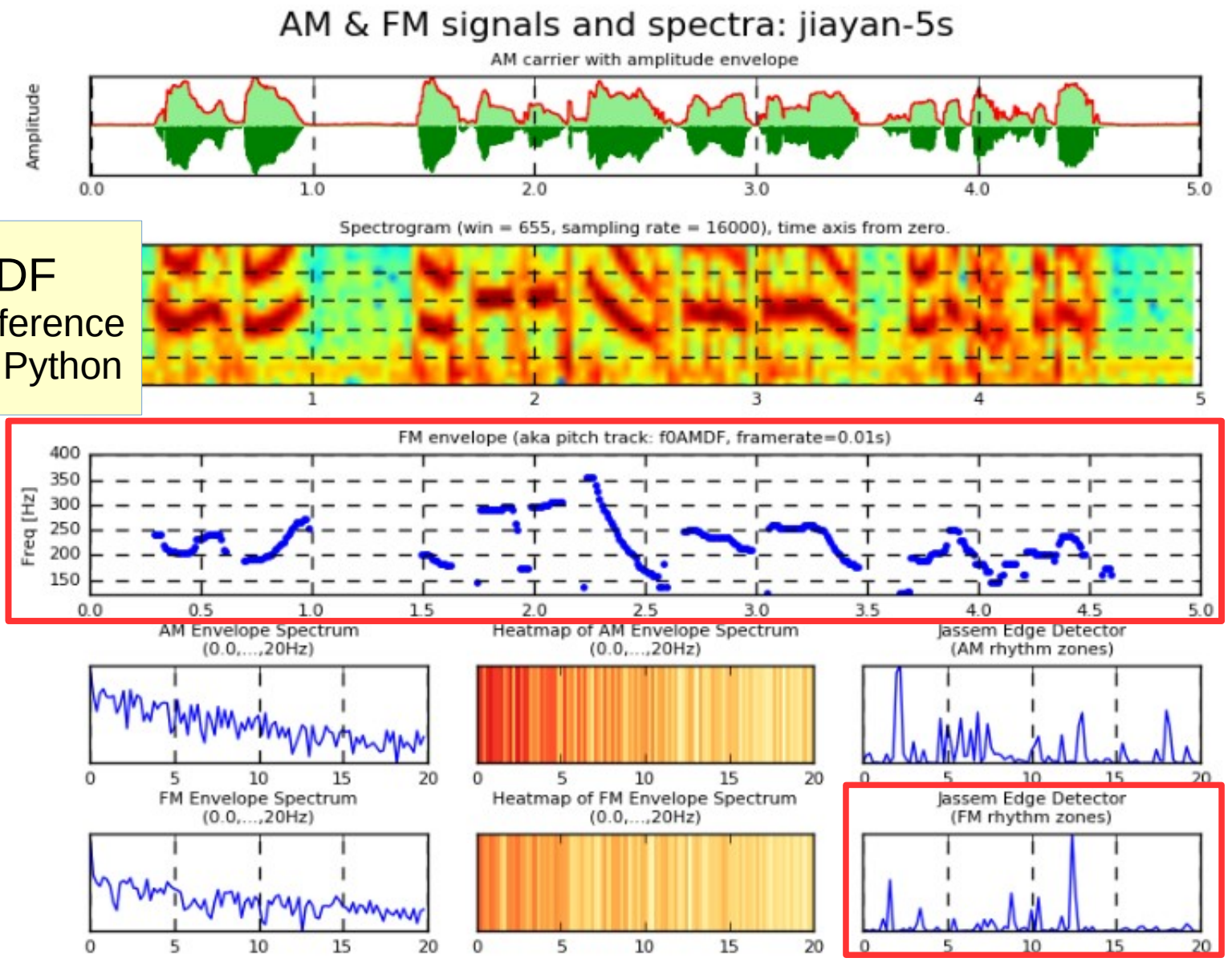
Simple - no
- voice detection
- candidate weighting

(code on GitHub)

How about AMDF (Average Magnitude Difference Function, Python)

Dafydd Gibbon

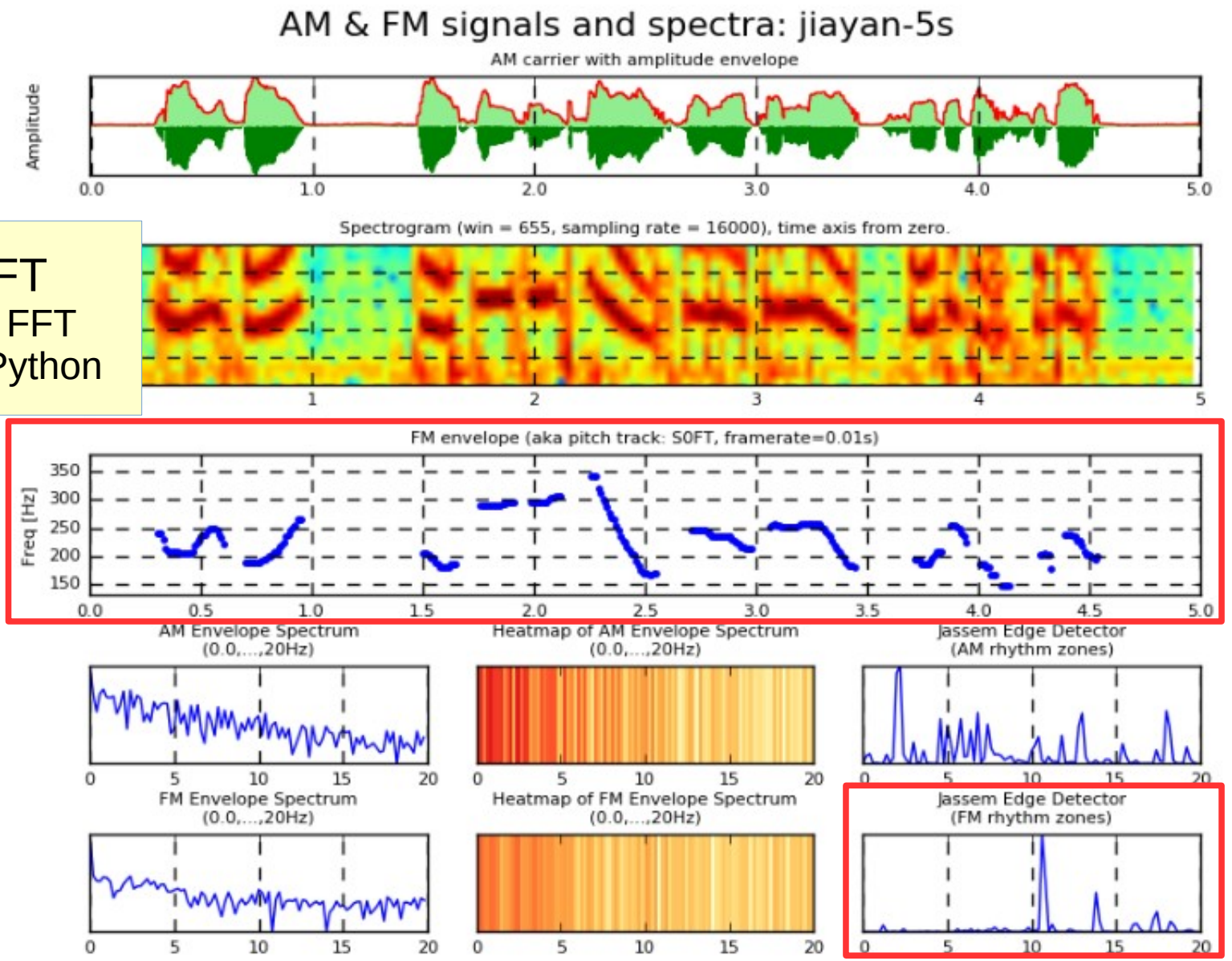
AMDF
naive difference minima, Python



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

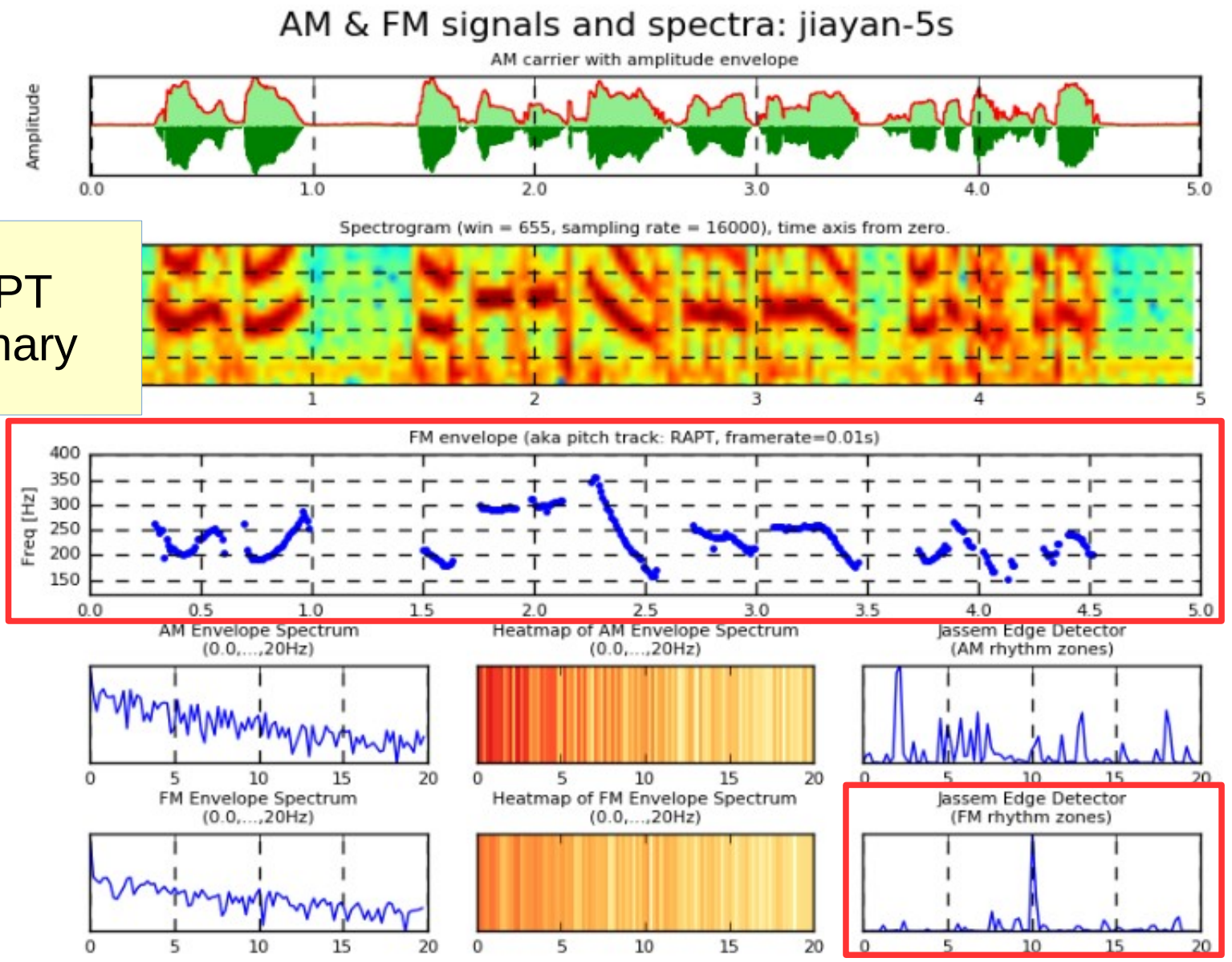
Encouraging for FFTpeak (problems remain, of course):

- 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

RAPT (Robust Algorithm for Pitch Tracking)

David Talkin

RAPT
C binary



Continuing with FM anyway: 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:

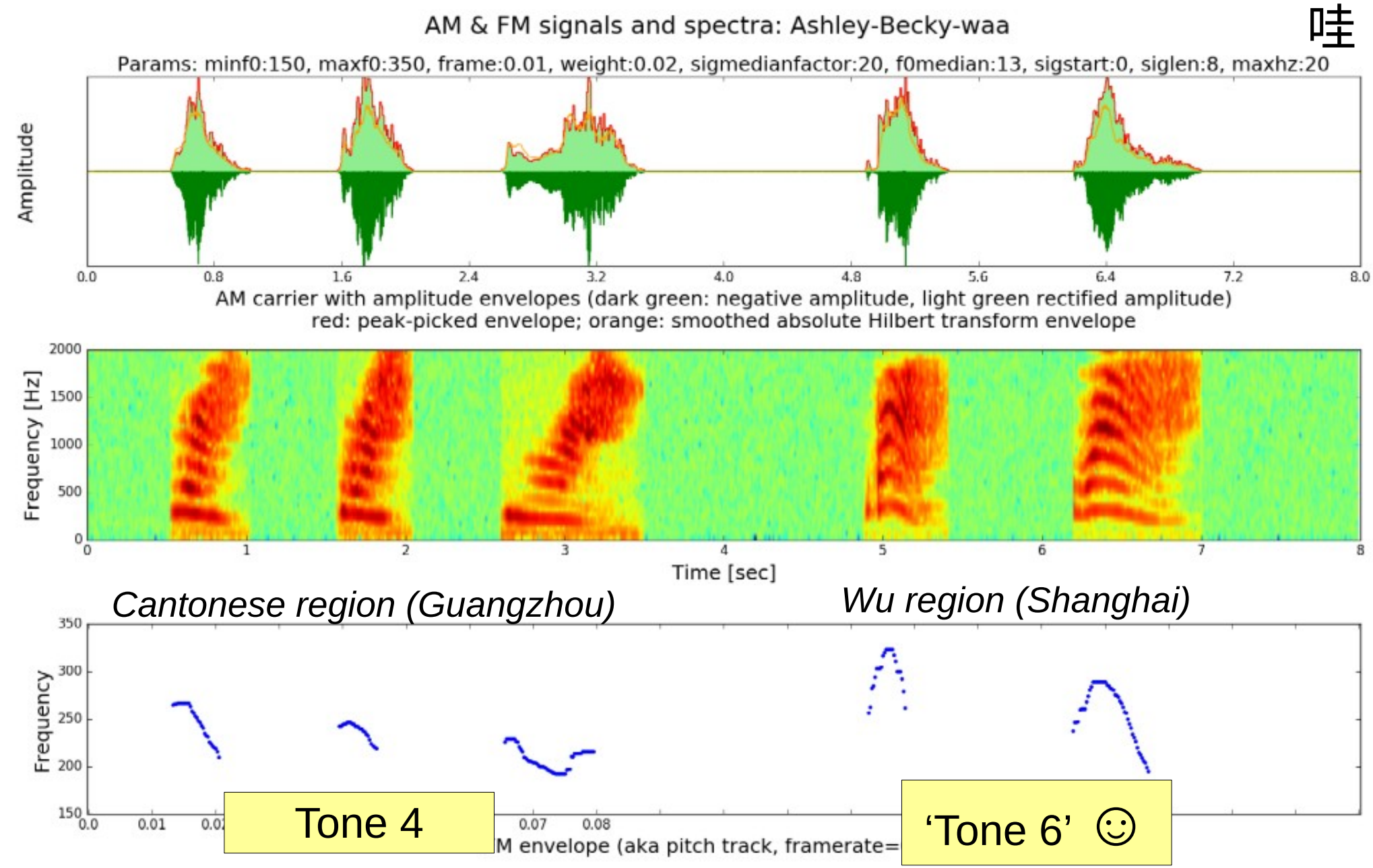
Other primates wowed, aahed and whistled first – we humans continued the custom

Is this why in some societies there is a taboo on whistling?

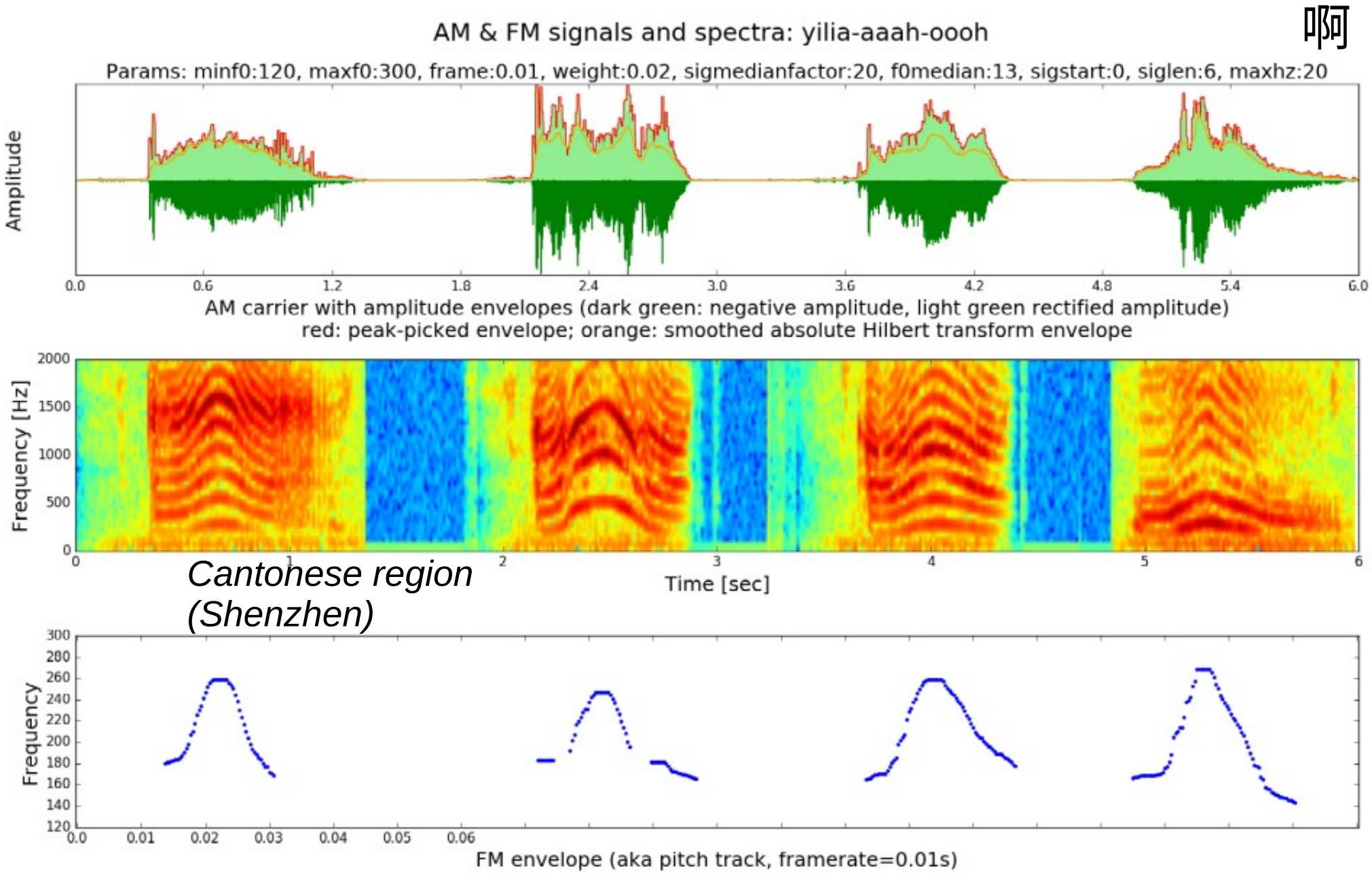
Many Uses of Frequency Modulation in Discourse Rhythms

Alibaba FM

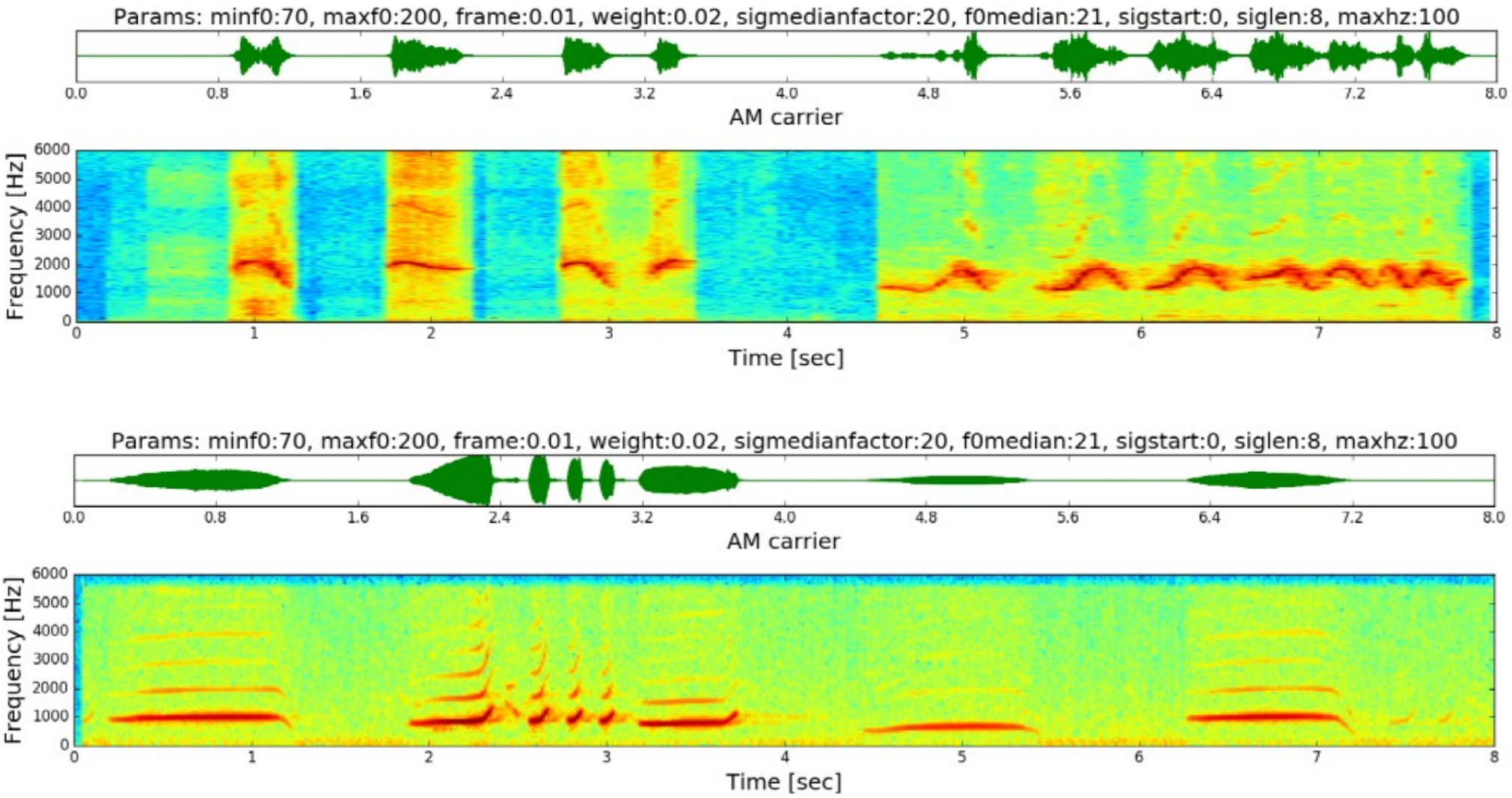
Emotive Exclamations



Emotive Exclamations



FM Rhythms in Teleglossia



Thank you!

谢谢！