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
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)

Finite depth linear 2-cycle iterative layered structure
(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
Three ‘hierarchies’, not one, with a ternary semiotic basis for signs at all ranks.

**Semantic Interpretation**
- Focus
- Contrast
- Emphasis

**Multimodal Interpretation**
- Hierarchical patterning
- Linear patterns

1. Denotation, Reference
   - semiotic relation
2. Symbols, Icons, Indices in Space-Time
3. Categorial simple and structured forms

**The Multilinear Rank Interpretation Model**
Discourse: Monologue, Dialogue

Utterance: turn, IPU, ...

Sentence, clause, phrase

Word: simple, inflected, compound, derived

The Multilinear Rank Interpretation Model
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

<table>
<thead>
<tr>
<th>Conversation Analysis</th>
<th>Phonetic Annotation Analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linguistic Analysis</td>
<td>Phonetic Signal Analysis</td>
</tr>
</tbody>
</table>
### 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

### Linguistic Analysis

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

### Phonetic Annotation Analysis

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

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

- Focus on isochrony

### 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

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

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

But linguistically the most interesting.
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, Average Magnitude Difference Function), not between mean and data, thus factoring out tempo variations to some extent.

<table>
<thead>
<tr>
<th>Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\text{Variance}(x_{1...n}) = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$</td>
<td>(or Standard Deviation)</td>
</tr>
<tr>
<td>$\text{PIM}(x_{1...n}) = \sum_{i \neq j} \left</td>
<td>\log \frac{I_i}{I_j} \right</td>
</tr>
<tr>
<td>$\text{PFD}(d_{1...n}) = \frac{\sum_{i=1}^{n}</td>
<td>\bar{d} - d_i</td>
</tr>
<tr>
<td>$\text{nPVI}(d_{1...n}) = \frac{\sum_{k=1}^{k-1}</td>
<td>d_k - d_{k+1}</td>
</tr>
</tbody>
</table>
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!

Three-dimensional because alternative trees are possible, depending on the algorithm settings:
- binary/nonbinary, lower/higher percolated
- related to phrasal and discourse patterns

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 Amplitude Modulation (AM)
Rhythmic Amplitude Envelope + Carrier Frequency

There are many studies of oscillation in production

I will concentrate on ...

Rhythm Perception as Amplitude Demodulation
Rhythmic Amplitude Envelope – Carrier Frequency
Selected Work on Amplitude Envelope Demodulation Spectra


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

FILTER COEFFICIENTS

AMPLITUDE MODULATION

SPECTRAL ZONES, perception emulation

Rhythm is oscillation and iteration

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

F0 harmonics
Modulation Regularities – Maybe the Long-Term Spectrum?

What is happening here?
And what is the role of F0?
The Reality of Rhythm!
Amplitude Envelope Modulation
\[ \downarrow \]
Amplitude Envelope Demodulation
absolute value of Hilbert transform
(or rectification & peak-picking / LP filtering)
\[ \downarrow \]
Spectral slice (FFT)
\[ \downarrow \]
Spectral Zone Edge Detection
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

AM & FM signals and spectra: jiayan

Params: minf0:110, maxf0:400, frame:0.01, weight:0.02, sigmedianfactor:100, f0median:9, sigstart:6, siglen:30, maxhz:1

AM carrier with amplitude envelope (dark green: negative amplitude, light green rectified amplitude); red: peak-picked envelope

Mandarin, female
30 sec, < 1Hz

Correlation AME:FME=0.74
Correlation AMS:FMS=0.29
AM and FM Demodulation

AM & FM signals and spectra: jiayan

Params: minf0=110, maxf0=400, frame:0.01, weight:0.02, sigmedianfactor:100, f0median:9, sigstart:6, siglen:30, maxhz:1

AM carrier with amplitude envelope (dark green: negative amplitude, light green rectified amplitude); red: peak-picked envelope

Spectrogram (win = 3276, sampling rate = 16000)

RAPT F0 estimation

Mandarin, female 30 sec, < 1Hz

0.2 Hz / 5s: the rhythm of interpausal units?
AM and FM Demodulation and Spectrum Analysis: Calibration

AM & FM signals and spectra: sine-200x5x12

Params: min0:100, max0:300, frame:0.01, weight:0.02, sigmedianfactor:20, f0median:9, sigstart:0, siglen:2, maxhz:10

AM carrier with amplitude envelope (dark green: negative amplitude, light green rectified amplitude); red: peak-picked envelope

Spectrogram (win = 655, sampling rate = 16000)

RAPT F0 estimation

AM Envelope Spectrum (0.0..., 10Hz)

FM Spectrum (0.0..., 10Hz)

Heatmap of AM Envelope Spectrum (0.0..., 10Hz)

Heatmap of FM Spectrum (0.0..., 10Hz)

Jassem Edge Detector (AM rhythm zones)

Jassem Edge Detector (AM rhythm zones)

Correlation AME:FME=0.76
Correlation AMS:FMS=0.26
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”

Beijing Mandarin
Yu corpus
“bei3 feng1 gen1 tai4 yang2”

1 Hz
Extending to Discourse Spectra: English Genres

A. Waveform Abercrombie English NW048-aems-5-35-1 (35s)

B. Amplitude Envelope Modulation Spectrum (Hz)

AEMS heatmap (log scale, darker: higher magnitude)

A. Waveform A0101B-aems-5-35-1 (35s)

Spectral Frequency Zone Boundaries

AEMS heatmap (log scale, darker: higher magnitude)
Extending to discourse spectra: English-Mandarin

A. Waveform Abercrombie English NW048-aems-5-35-1 (35s)

B. Amplitude Envelope Modulation Spectrum (Hz)

AEMS heatmap (log scale, darker: higher magnitude)

A. Waveform wuxi-aems-5-35-1 (35s)

B. Amplitude Envelope Modulation Spectrum (Hz)

AEMS heatmap (log scale, darker: higher magnitude)

Contrast
elderly male English young female Mandarin
Rhythms in Syntagmatic, Compositional Spectral Trees
AM and FM Demodulation and Spectral Tree Induction

A. Waveform A0101B-aems-5-35-1 (35s)

B. Amplitude Envelope Modulation Spectrum (Hz)

AEMS heatmap (log scale, darker: higher magnitude)

Frequency Domain Trees: Spectral Zone Hierarchies

L-strong, >  L-strong, <  R-strong, >  R-strong, <
AM and FM Demodulation and Spectral Tree Induction

AEMS Frequency Tree

AEMS heatmap (log scale, darker: higher magnitude)

L-strong, <
AM and FM Demodulation and Spectral Tree Induction

AEMS Frequency Tree

L-strong, <
AM and FM Demodulation and Spectral Tree Induction

AEMS Frequency Tree

Mandarin North Wind & Sun

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
Similarity criterion: >3 adjacent same-speaker settings
Largest per speaker score: (4+3)/16
Largest cluster: 4/16

Consistency of Spectra: Rhythm Classification
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

H* sequence, global fall

L* sequence, global rise, final rise

Interview start

Question

H* sequence, global fall

L* sequence, global fall, final rise

Response

Continuation
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

---

**AM & FM signals and spectra: jiyan-5s**

![Spectrogram](image)

Spectrogram (win = 655, sampling rate = 16000), time axis from zero.

---

**FM envelope (aka pitch track: Reaper, framerate=0.01s)**

![FM envelope](image)

---

Reaper compiled C binary
**Praat**

Paul Boersma

- AM & FM signals and spectra: jiyan-5s
  - AM carrier with amplitude envelope
  - Spectrogram (win = 655, sampling rate = 16000), time axis from zero.
  - FM envelope (aka pitch track: PraatDefault, framerate=0.01s)
  - AM Envelope Spectrum (0.0...,20Hz)
  - Heatmap of AM Envelope Spectrum (0.0...,20Hz)
  - Jassem Edge Detector (AM rhythm zones)
  - FM Envelope Spectrum (0.0...,20Hz)
  - Heatmap of FM Envelope Spectrum (0.0...,20Hz)
  - Jassem Edge Detector (FM rhythm zones)
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: Methodological Problems

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

So let’s do something 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
FFTpeak (Simple F0 Tracker, Python)
Dafydd Gibbon*

* inspired by a snippet from ‘Jonathan’, gist.github.com/endolith/255291
Comparing F0 estimators with RAPT as ‘gold standard’

<table>
<thead>
<tr>
<th>F0 estimator pair</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>RAPT:PyRAPT</td>
<td>0.8902</td>
</tr>
<tr>
<td>RAPT:Praat</td>
<td>0.8657</td>
</tr>
<tr>
<td>RAPT:FFTpeak</td>
<td>0.9096</td>
</tr>
<tr>
<td>RAPT:f0AMDF</td>
<td>0.8605</td>
</tr>
<tr>
<td>FFTpeak:RAPT</td>
<td>0.9096</td>
</tr>
<tr>
<td>FFTpeak:Praat</td>
<td>0.8409</td>
</tr>
<tr>
<td>FFTpeak:PyRAPT</td>
<td>0.8352</td>
</tr>
<tr>
<td>FFTpeak:f0AMDF</td>
<td>0.8016</td>
</tr>
</tbody>
</table>

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

AM & FM signals and spectra: jiayan-5s

Spectrogram (win = 655, sampling rate = 16000), time axis from zero.

FM envelope (aka pitch track: RAPT, framerate=0.01s)

AM Envelope Spectrum (0.0,...20Hz)
Heatmap of AM Envelope Spectrum (0.0,...20Hz)
Jassem Edge Detector (AM rhythm zones)

FM Envelope Spectrum (0.0,...20Hz)
Heatmap of FM Envelope Spectrum (0.0,...20Hz)
Jassem Edge Detector (FM rhythm zones)
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

Cantonese region (Guangzhou)  Wu region (Shanghai)

Tone 4  ‘Tone 6’ 😊
Emotive Exclamations

AM & FM signals and spectra: yilia-aaah-oooh

Params: minf0:120, maxf0:300, frame:0.01, weight:0.02, sigmedianfactor:20, f0median:13, sigstart:0, siglen:6, maxhz:20

Cantonese region (Shenzhen)
FM Rhythms in Teleglossia

Params: minf0:70, maxf0:200, frame:0.01, weight:0.02, sigmedianfactor:20, f0median:21, sigstart:0, siglen:8, maxhz:100

0.0  0.8  1.6  2.4  3.2  4.0  4.8  5.6  6.4  7.2  8.0

Frequency [Hz]

0  1  2  3  4  5  6  7  8

Time [sec]

Params: minf0:70, maxf0:200, frame:0.01, weight:0.02, sigmedianfactor:20, f0median:21, sigstart:0, siglen:8, maxhz:100

0.0  0.8  1.6  2.4  3.2  4.0  4.8  5.6  6.4  7.2  8.0

AM carrier
Thank you!
谢谢！