The Music of Speech

Rhythm

Dafydd Gibbon

Mannheim Summer School, June – July 2019

The melody of rhythm

Dafydd Gibbon
Bielefeld University, Jinan University

Mannheim Summerschool, The Music of Speech, 2019

THE FOUNDATION: RHYTHM

- Rhythm is a central topic in many disciplines
  - most obviously in
    - spoken language
    - music: 3 / 4, 6 / 8, 4 / 4
    - dance: waltz, foxtrot, ...
  - generalised to ‘regularities in time’
    - ‘the rhythm of the tides’
    - ‘the rhythm of the seasons’
  - ordinary language metaphorical usage:
    - ‘out of rhythm’
    - ‘out of step’ (metonymy for uncoordinated action)

Summary: the argument

- We need a clear explicandum for rhythm:
  - not just a definition
  - a model
- We need to be clear about the relevant levels of analysis:
  - semantic
  - grammatical
  - phonological
  - phonetic
- We need to be clear about the relevant parameters:
  - interval duration
  - amplitude variation
  - frequency variation
- We need to be aware that rhythm is oscillation

Finding an *explicandum*

We all know what rhythm is ...

... or do we?

Let me just ask you a question:

Please define “rhythm”!

(I already gave you an ostensive definition.)

Preliminary definitions as *explicanda*

“An ordered recurrent alternation of strong and weak elements in the flow of sound and silence in speech.” (Webster web version)

“Rhythm is the directional periodic iteration of a possibly hierarchical temporal pattern with constant duration and alternating strongly marked (focal, foreground) and weakly marked (non-focal, background) values of some observable parameter.” (Gibbon & Gut 2001)

“Rhythm is viewed here as the hierarchical organisation of temporally coordinated prosodic units ... certain salient events (beats) are constrained to occur at particular phases of an established period” (Cummins & Port 1998)

Systematising the *explicandum*

At least our *explicandum* should be ostensively clear:

boom-di-boom-boom (Cummins)

I got rhythm ...

A first systematisation – three conditions:

structured events (as rhythm units)

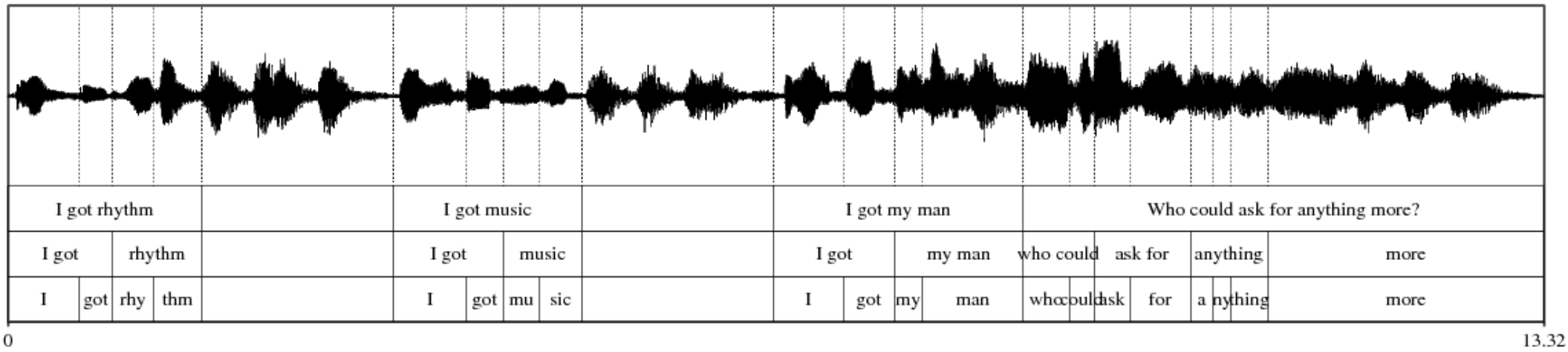
alternation within events

ordering of events as iteration (within rhythm unit sequences)

And, for rhythm units:

Two's company, three's a rhythm ☺

Clear cases: speech in song



Similarities:

- Freedom of the singer with final lengthening:
  - ***man***
  - ***more*** (whole bar)
- Focus accent:
  - ***my***

Differences:

- highly isochronous
- full synchronisation with accompaniment
- poetic features:
  - Jakobsonian ‘coupling’:
    - alliteration on ***m***
    - parallel syntax

Clear cases: recitation

Roland der Riese am Rathaus zu Bremen
Steht er als Standbild, tapfer und treu.

Our intuitions are clear in cases like this:

‘I have rhythm’

Jakobsonian coupling:

r

t

Contrast this with everyday speech, in which – as a rule – we do not have the immediate intuition of rhythm or Jakobsonian coupling.

I will return to this case later.

Finding an *explicatum*

There are many perspectives on rhythm

Music: *the beat*
Poetry: *the metre*
Writing: *reconstruction + re-production*
Speech: *hmm...??*

And we deal with it in ...

Phonology? *- Grids?*
Phonetics? *- Isochrony?*
Psycholinguistics? *- Percept or cognitive construct?*
Poetics? *- Metre (poem) vs. Rhythm (performance)?*
Musicology? *- Beat, phrasing, accentuation?*

There are many perspectives on rhythm

FORM:

- data: Phonetic? Duration? Pitch? Intensity? Production? Perception? Physical?
- structure: Pattern? Alternation? Hierarchy? Syllable, foot, phrase domain?
- timing: Isochrony? Periodicity? Oscillation?
- construct: Phonology? Prosody? Grammar? Text? Emergent cognitive construct? Neural clock? Multilevel entrainment?

FUNCTION:

- syntactic/semantic/pragmatic: Cohesion? Coherence? Configuration? Eurhythmy? Style? Coordination? Interaction? Alignment?

There are many percpectives on rhythm

THEORY:

ontology: Universal? Language specific?

epistemology: Innate? Maturational? Learned?

METHOD:

empirical-experimental-observational?

intuitive-analytic-structural?

holistic-interpretative-hermeneutic?

A first approximation:

rhythm as an emergent sign

Rhythm as an emergent sign

Rhythm is a sign

An emergent function of meaning, structure and realisation:

- Rhythm as a *sign* (Couper-Kuhlen)
- If so, we need to think about the *meaning*, etc., of rhythm
  - functionality of rhythm in discourse – coherence:
  - structure of rhythm:
    - alignment/association of rhythm – cohesion:
      - sentence structure
      - word structure
      - foot/syllable structure
    - rhythm as an autostructural pattern – synchronisation
  - realisation of rhythm
    - cognitive constraints on rhythm – emergent construct
    - phonetic correlates of rhythm - product/percept

Rhythm as an emergent sign

The meaning of rhythm:

- see Couper-Kuhlen & Auer (1999):
  - critique of detemporalisation of language
  - rhythm and coherence: turn-taking, interlocutor synchrony

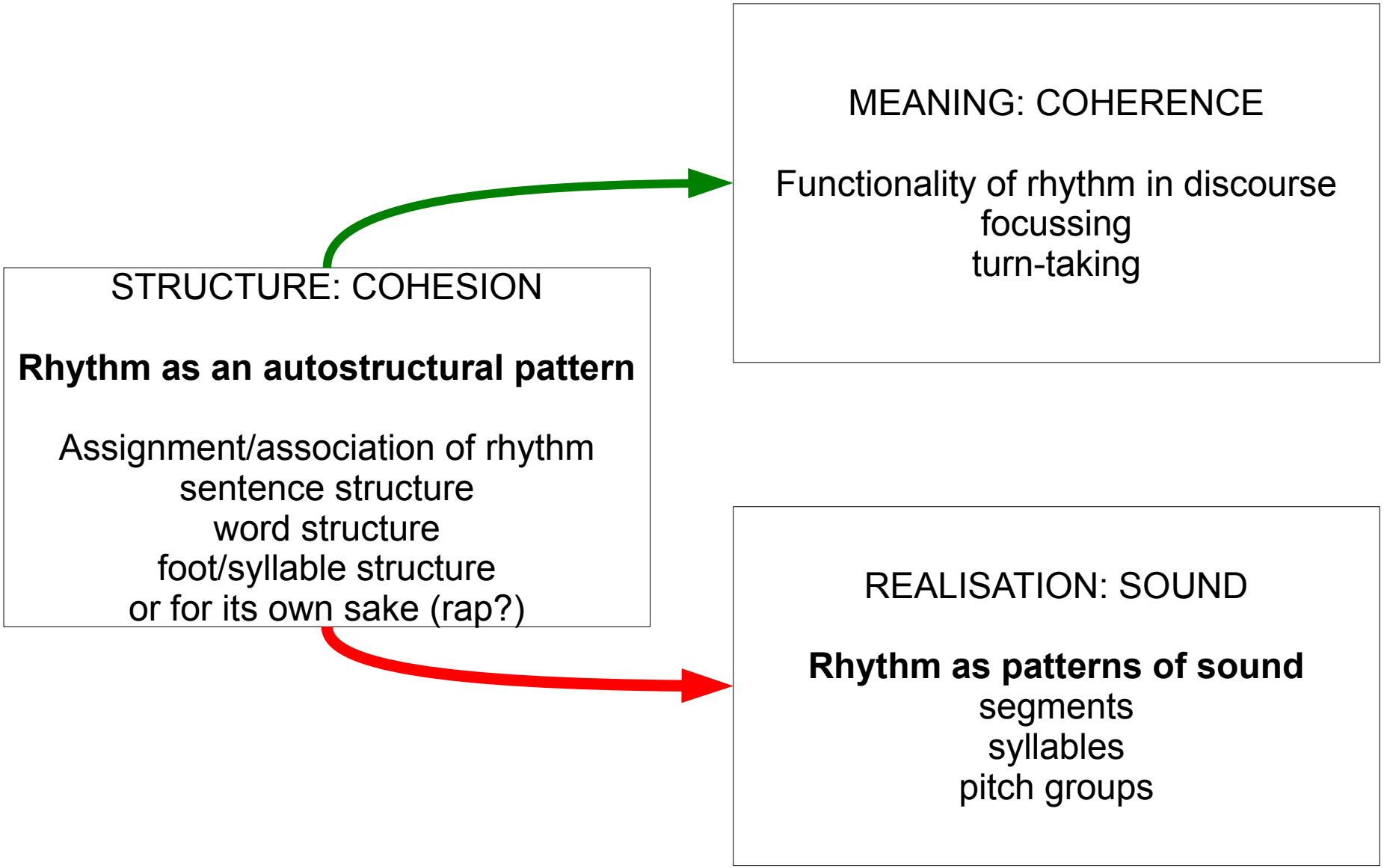
The structure of rhythm:

- categorial rhythm:
  - alignment/association algorithms: Generative Phonology
- relational rhythm:
  - tree & grid patterning algorithms: Metrical Phonology

The realisation of rhythm:

- absolute rhythm:
  - phonetic models

Rhythm as an emergent sign



Remember:

rhythm is temporal

‘Detemporalisation of language’

Couper-Kuhlen & Auer (1999):

- rightly criticise the ‘detemporalisation of language’ in structuralist and generative approaches
- claim to re-introduce time

But they leave a gap, in that they

- focus on the functionality of time patterns & rhythm
- but do not actually have a linguistic theory of time

This gap needs to be filled –

- note the terms previously used:
  - categorial
  - relative
  - absolute

Time Types

Categorial ('abstract') time:

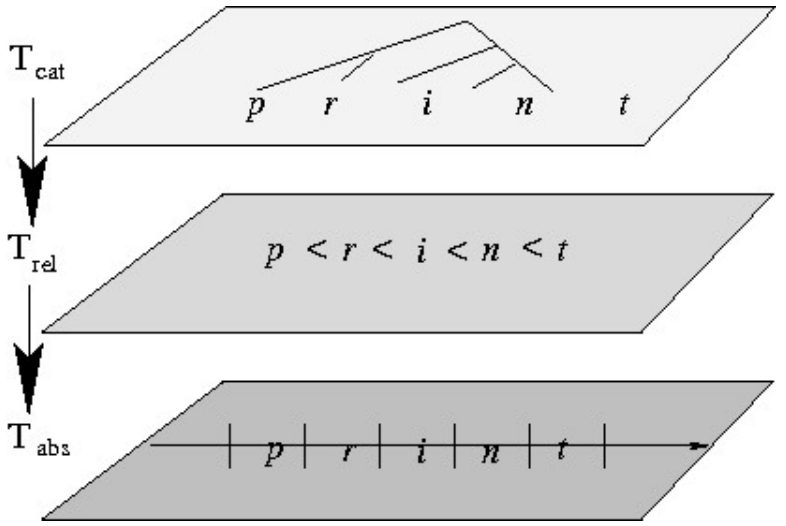
- category sequence as concatenation
- duration as property (e.g. [+/- long])

Relational ('rubber') time:

- point (or interval) events
- temporal precedence: $a <_t b$
- temporal overlap: $a \circ_t b$

Absolute ('clock') time:

- point (or interval) events
- time-stamps (time-stamp pairs)



Gibbon, Dafydd (1992). Prosody, time types and linguistic design factors in spoken language system architectures. In: G. Görz, ed., *KONVENS '92*. Berlin, Springer, S. 90-99

Figure due to Berndsen (1998).

Time Types

<u>Categorial ('abstract') time</u>:

- category sequence as concatenation
- category duration as property (e.g. [+/- long])

<u>Relational ('rubber') time</u>:

- point (or interval) events
- temporal precedence: $a <_t b$
- temporal overlap: $a \circ_t b$

<u>Absolute ('clock') time</u>:

- point (or interval) events
- time-stamps (time-stamp pairs)

grammatical approaches

phonological approaches

phonetic approaches

Gibbon, Dafydd (1992). Prosody, time types and linguistic design factors in spoken language system architectures. In: G. Görz, ed., *KONVENS '92*. Berlin, Springer, S. 90-99

PHONOLOGICAL APPROACHES

Generative + Metrical Phonologies

Generative phonology:

- stress patterns: encoding of tree structures as numbers
  - nuclear stress, compound stress
- 2 well-known algorithms:
  - Chomsky & Halle, Liberman
- inverse algorithm
  - Gibbon

Metrical phonology:

- prosodic hierarchy & alignment with segmental hierarchy
- addition of grid as filter
  - cf. Culicover & Rochemont's 'readjustment rules'
- interpretation (DG):
  - finite state filter over trees
  - declarative visualisation of oscillator output

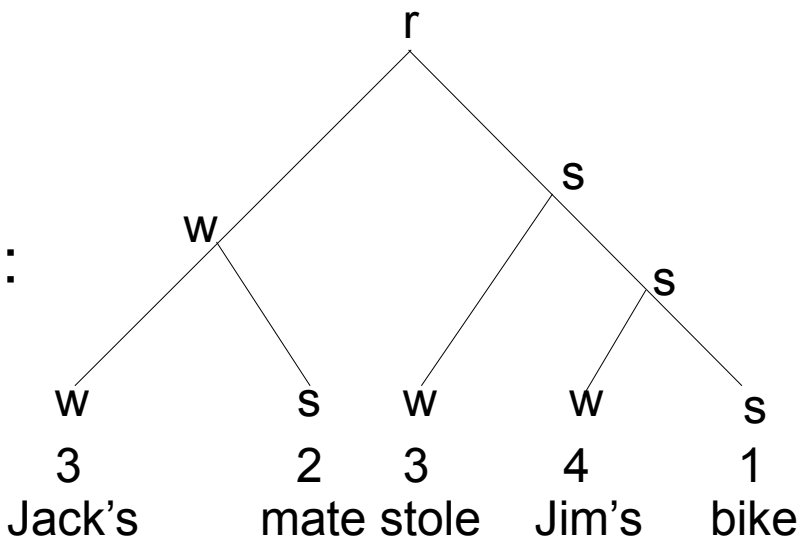
Liberman's algorithm (Metrical Phonology)

For each terminal node:

from (and including) the first 'w' node:

count the depth from the root

(whereby the root counts as 'w')



The inverse algorithm will be deployed later in a different context.

Gibbon, Dafydd (2003). Computational modelling of rhythm as alternation, iteration and hierarchy. Proceedings of the International Congress of Phonetic Sciences, Barcelona, August 2003, III: 2489-2492.

PHONETIC APPROACHES:

the interval duration method

Interval duration based approaches

Focus on regularity/irregularity/isochrony of intervals:

Static approaches:

- Top-down phonological structure-oriented approaches:
  - hierarchical (e.g. Metrical Phonology: metrical trees)
  - linear (e.g. Metrical Phonology: metrical grid)
- Data-driven phonetic isochrony-oriented approaches
  - global: (e.g. Roach; Scott & al., Ramus)
  - local: (e.g. Grabe & al.; Gibbon & Gut)

Dynamic process-oriented approaches:

- Finite machines (e.g. Wagner; Wachsmuth)
- Oscillators: (e.g. Cummins, Barbosa)
- Entrainment: (e.g. Cummins, Barbosa)

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 (?)

Strange: if all *percentage deviations* happen to be the same, whether large or small, the *variance* will be 0 ☺

This is a global measure:

ignores alternation and iteration criteria

Isochrony as ratio: Scott et al.

$$Rhythmic\ Irregularity\ Measure\ (RIM) = \Sigma_{i \neq j} \left| \log \frac{I_i}{I_j} \right|$$

The Rhythmic Irregularity Measure (RIM) for individual utterances calculates the sum of the ratios of each 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.

Isochrony as local distance: Grabe & al.

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

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

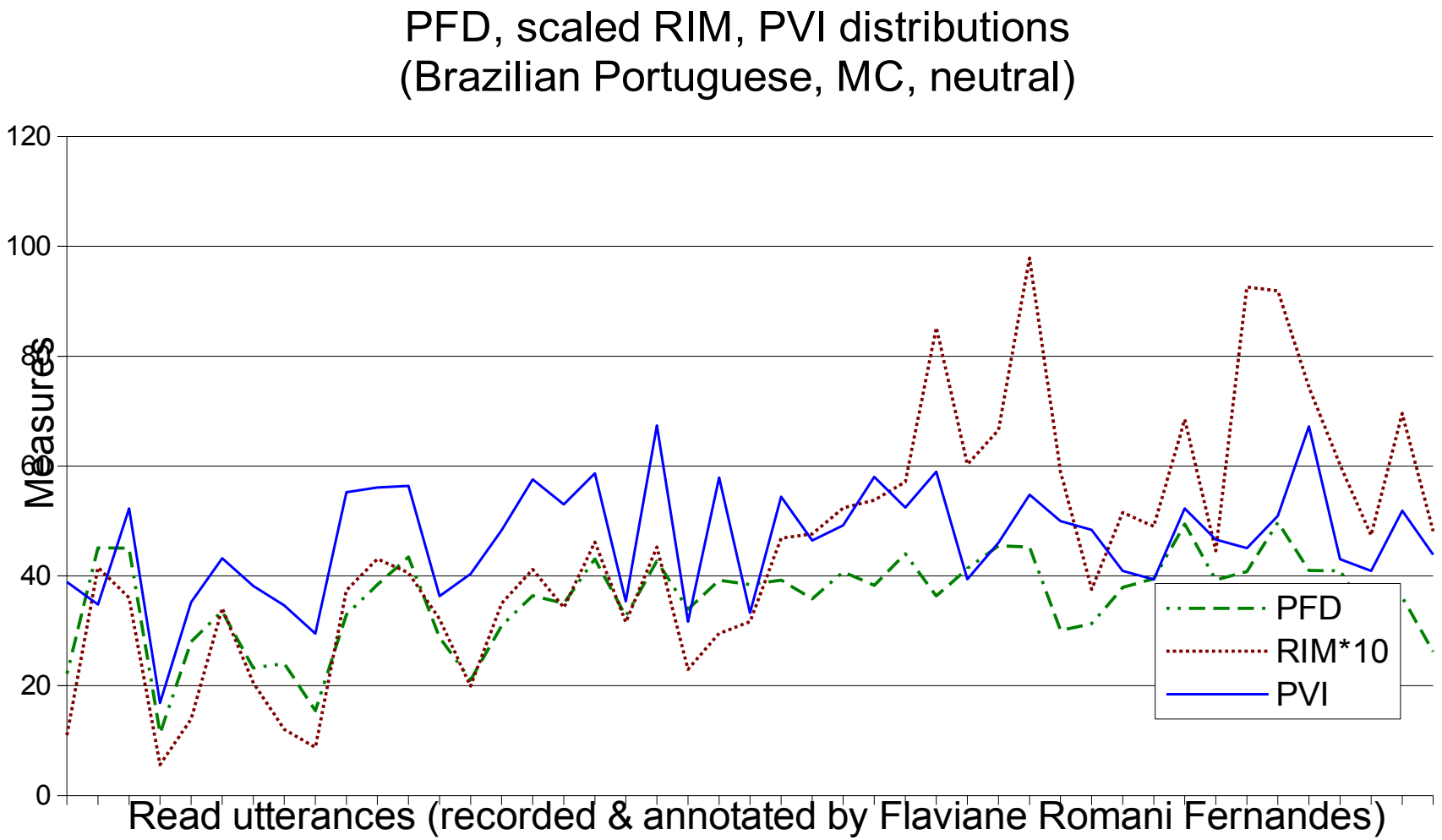
DISTANCE$_i$ = | INT$_i$ – INT$_i$+1 | / AVG(INT$_i$, INT$_i$)

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:

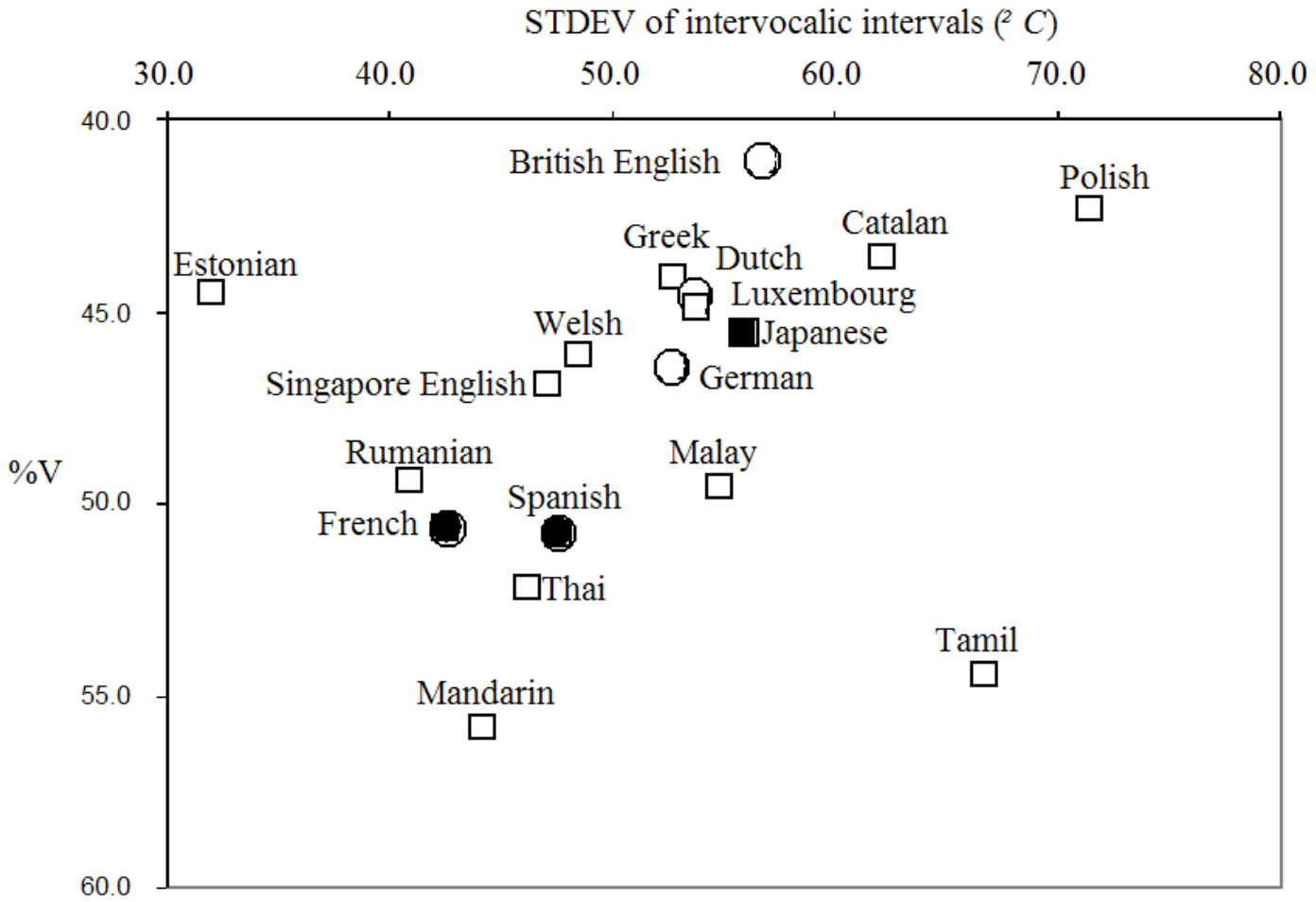
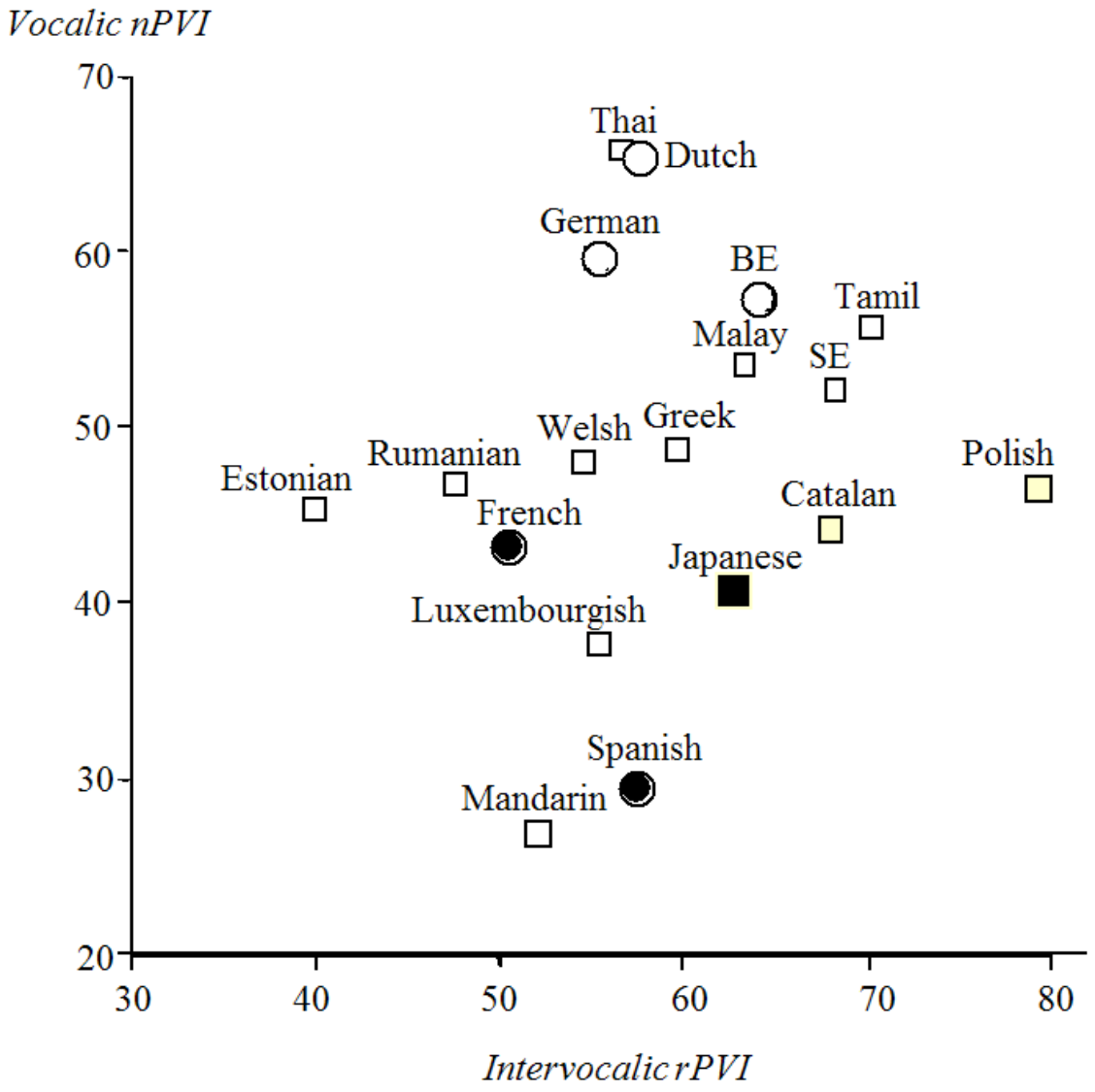


Figure 3. The measure *%V* is plotted on the y-axis, in reverse order. The standard deviation of intervocalic intervals ΔC , is given on the x-axis.

Interval duration approaches and typology

Grabe & al.:

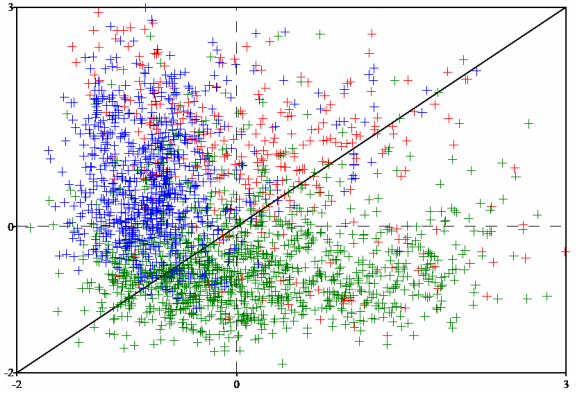


Interval duration approaches and typology

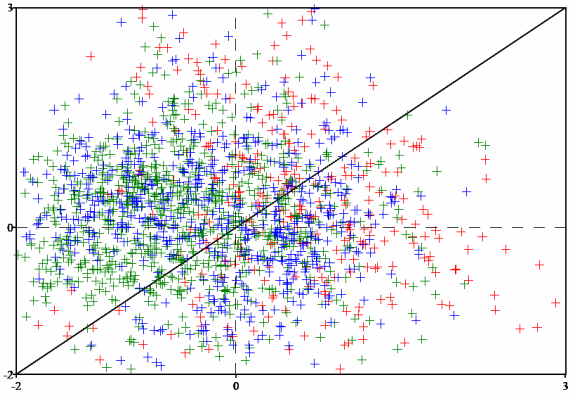
Wagner:

- recognises the normalisation to magnitude (absolute value)
- plots $duration_i$ x $duration_{i+1}$
- creates typologically interpretable clusters
  - also for other variety types?

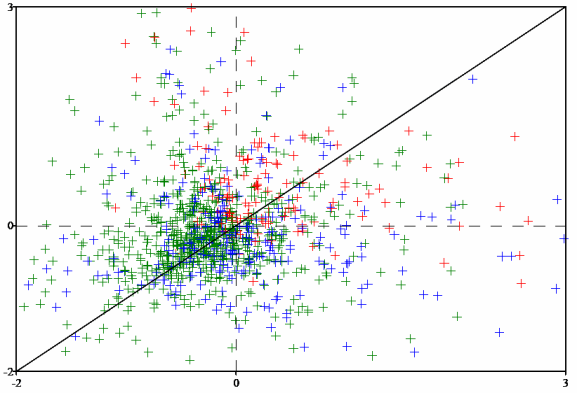
English



French



Polish



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

Critique of the interval duration method

Summary: 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

Which intervals, which durations?*

Roland der Riese am Rathaus zu Bremen
Steht er als Standbild, tapfer und treu.

Definitely not spontaneous speech – but that is the point...

- But definitely rhythmical
- Jakobsonian coupling: ***r, t***
- Clear syntactically determined proclitic anacrusis:
  - Roland | *der* Riese | *am* Rathaus | *zu* Bremen ||
  - Steht er | *als* Standbild | tapfer | *und* treu ||

So what are the results of duration analysis?

* *Special thanks to Anna Kutscher BA, Bielefeld, for example + analysis!*

Which intervals, which durations?

Duration measurements (pauses underlined):

foot lengths are relatively similar:

718 778 945 705 300 790 1047 295 665 1031

syllable lengths are relatively dissimilar

336 382 227 238 313 193 394 358 133 295 277 300 455 335 206 411 430 295 176 489 447 584

pauses are roughly of syllable length

Which intervals, which durations?

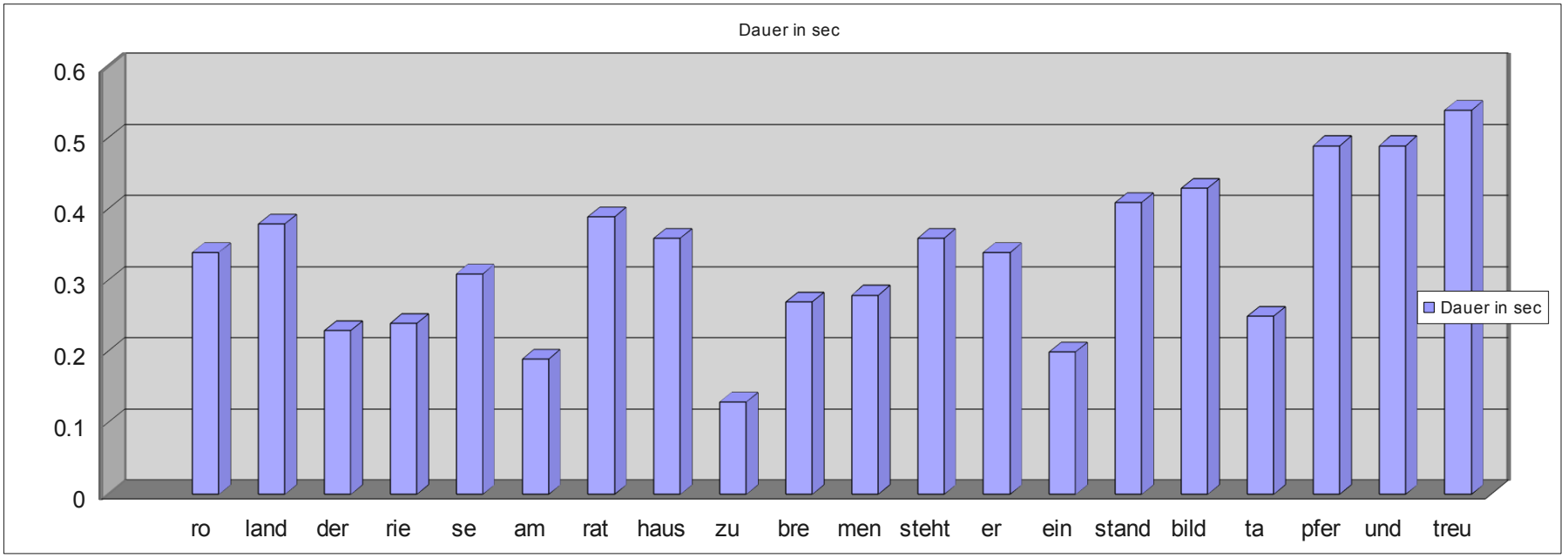
Foot properties:

- Regularity of foot lengths with pauses:
  - mean = 681, sd = 279, nPVI = 60
- Regularity of foot lengths without pauses:
  - mean = 835, sd = 142, nPVI = 14
  - close to isochronous ☺ (as predicted for clear case)

Syllable properties:

- Regularity of syllable lengths with pauses:
  - mean= 331, sd = 112, nPVI = 38
- Regularity of syllable lengths without pauses:
  - mean = 334, sd = 118, nPVI = 42
  - not close to isochronous ☺ (strong-weak structure of foot)

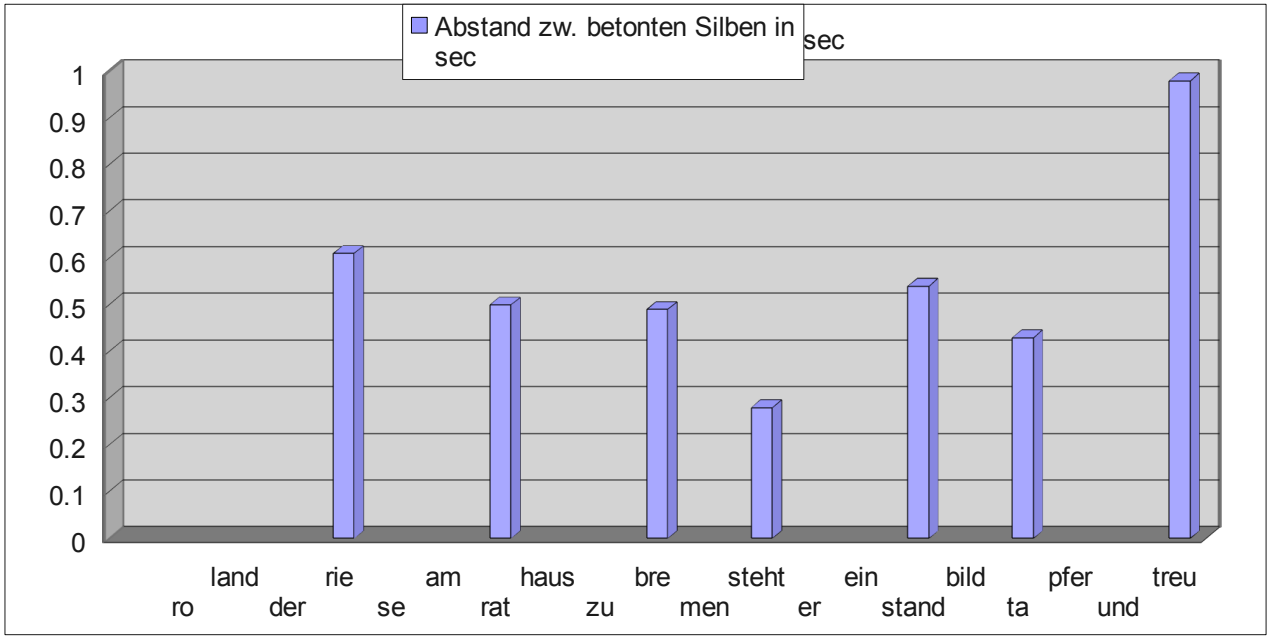
Which intervals, which durations?



Note:

- check the stressed and unstressed syllables
- a fundamental hypothesis of previous phonetic methods does not apply: *strong≠long* and *weak≠short*

Which intervals, which durations?



Distances between stressed syllables are close to isochronous:

When pause lengths are included, the distances increase:

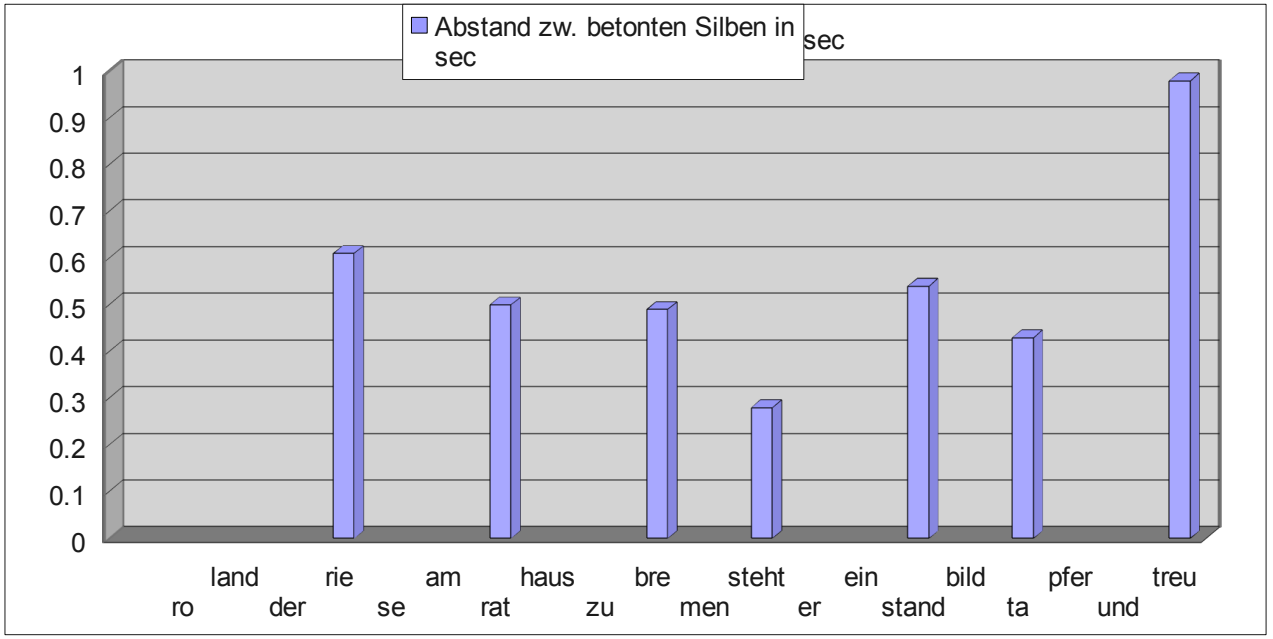
bre ... steht: 600ms

stand ... ta: 725ms

mean = 635, sd = 159, nPVI = 16

So we can say more than simply feet are isochronous

Which intervals, which durations?



close copy re-synthesis

equal syllable length re-synthesis

Distances between stressed syllables are close to isochronous:

When pause lengths are included, the distances increase:

bre ... steht: 600ms

stand ... ta: 725ms

mean = 635, sd = 159, nPVI = 16

So we can say more than simply feet are isochronous

So what if there is no rhythm

in other speech styles?

Time-trees: data-driven method

General strategy:

- take the local 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

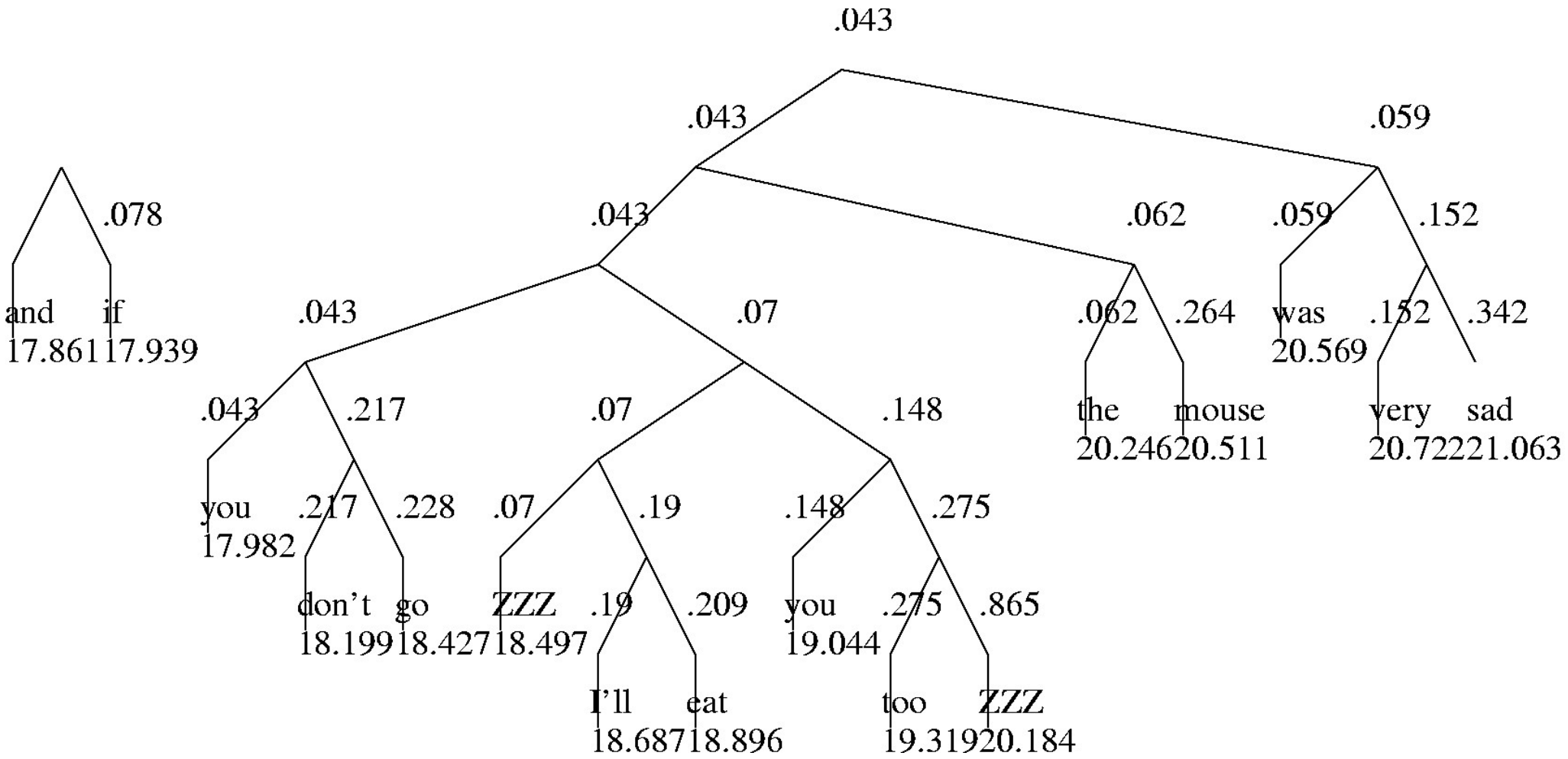
Data – reading style presumed optimal

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



Grammar: “Subjective Parsing”

Six linguistically trained subjects were asked to

bracket separate sentences (tree-equivalent notation)

without category labels

to show grammatical grouping

ill-formed bracketings completed at beginning or end

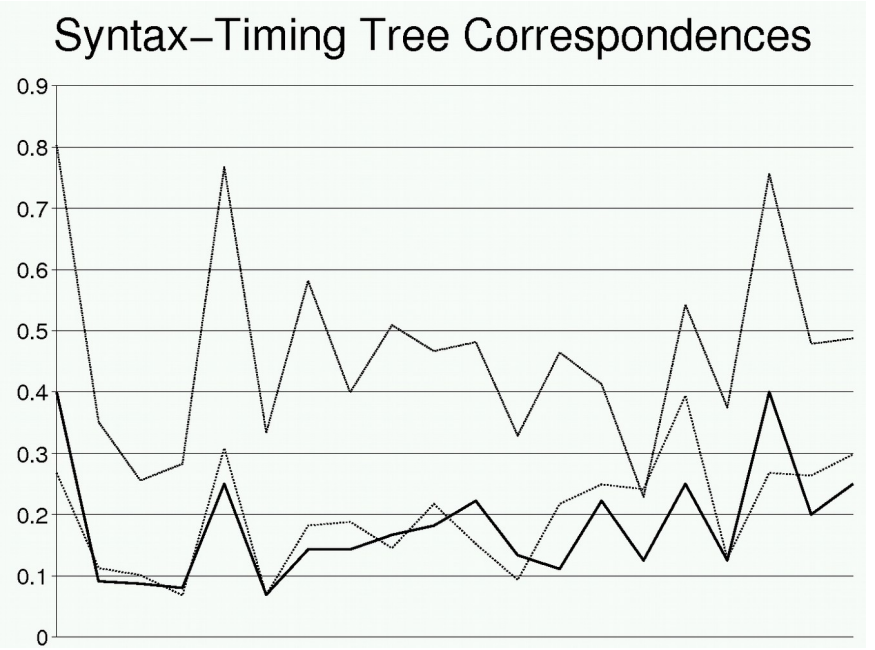
Example:

English:

(((a tiger) and (a mouse)) ((were walking) (in (a field))))

Result of 3 comparison conditions

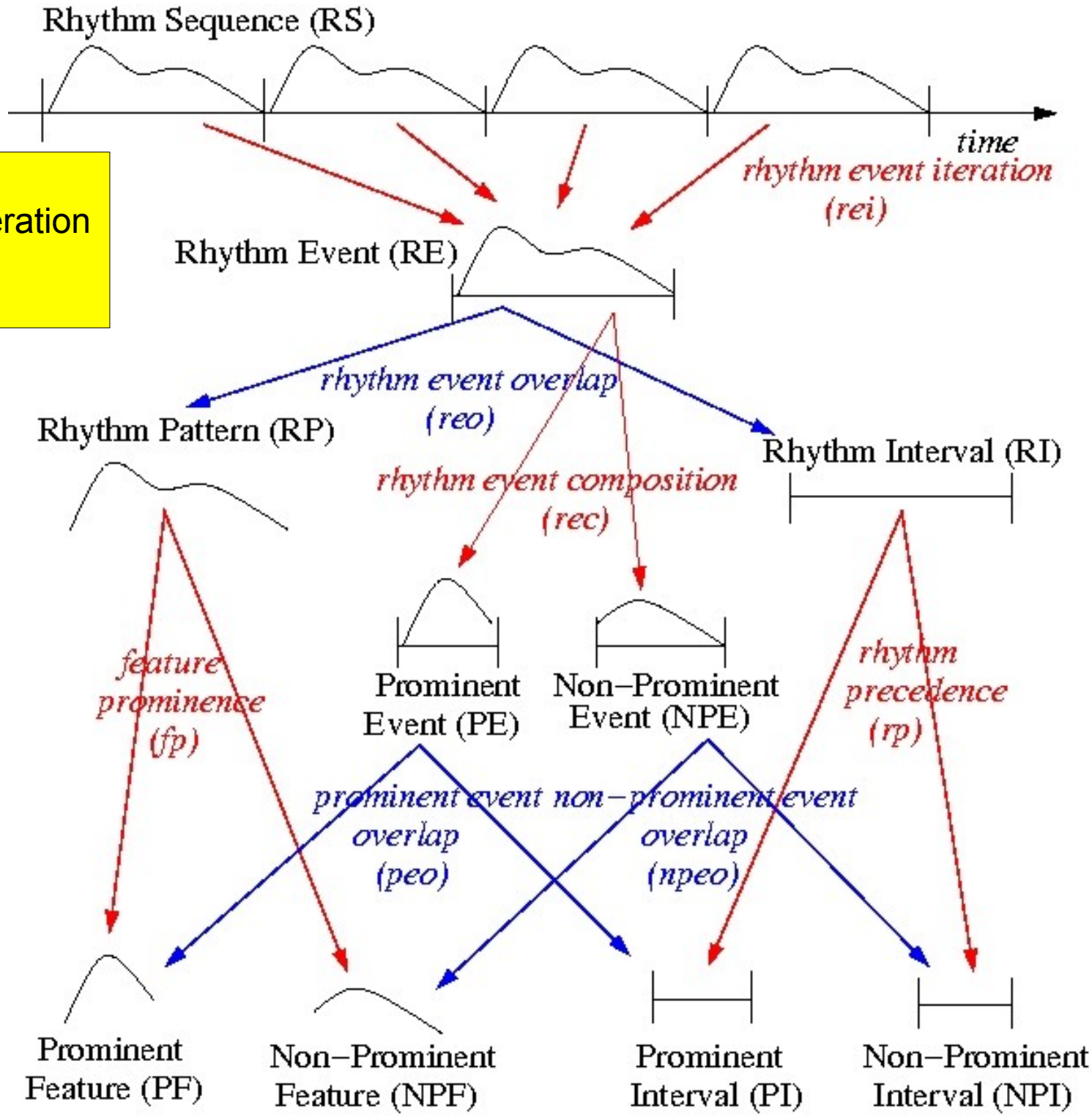
- Correspondence timing trees &
  - unparsed sequences (thick),
  - parsed sequences
    - iambic grouping (upper thin)
    - trochaic grouping (lower thin)
- Structural correlation in all cases
  - shallow bracketing?
  - short sentences?
- Absolute indices differ:
  - iambic: higher index
  - trochaic: lower index
  - due to right-headedness? NSR? Little to do with rhythm!



Back to the boom-di-boom-boom

Two core properties:
structured event + iteration
= alternation
interval recurrence

A basic model for rhythm

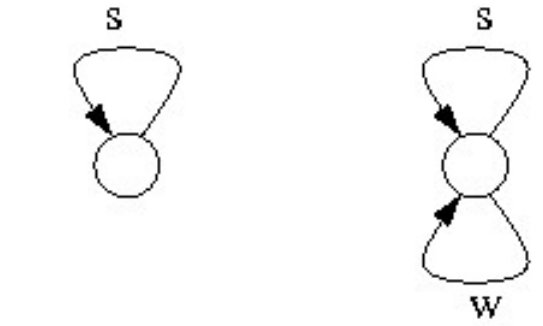


From model to procedure

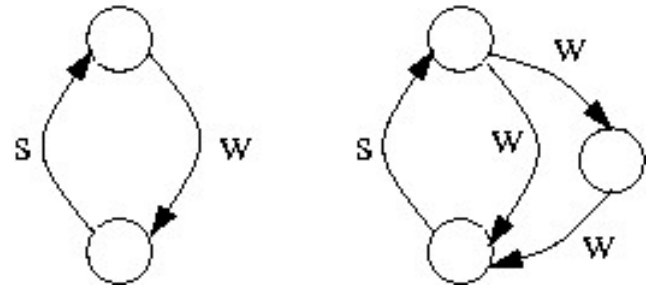
The obvious formalisation of iteration is a finite machine (finite state automaton)

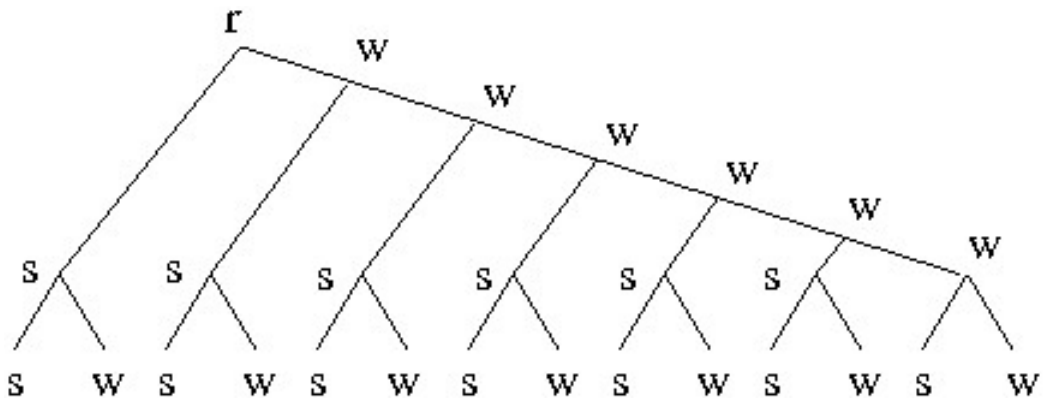
Note that finite state machines are sufficient for generating right-branching trees

Thus: a formal explication of the Generative Phonology 'readjustment rules' and the Metrical Phonology 'linear grid filter'



FSMs with temporal constraints on transitions



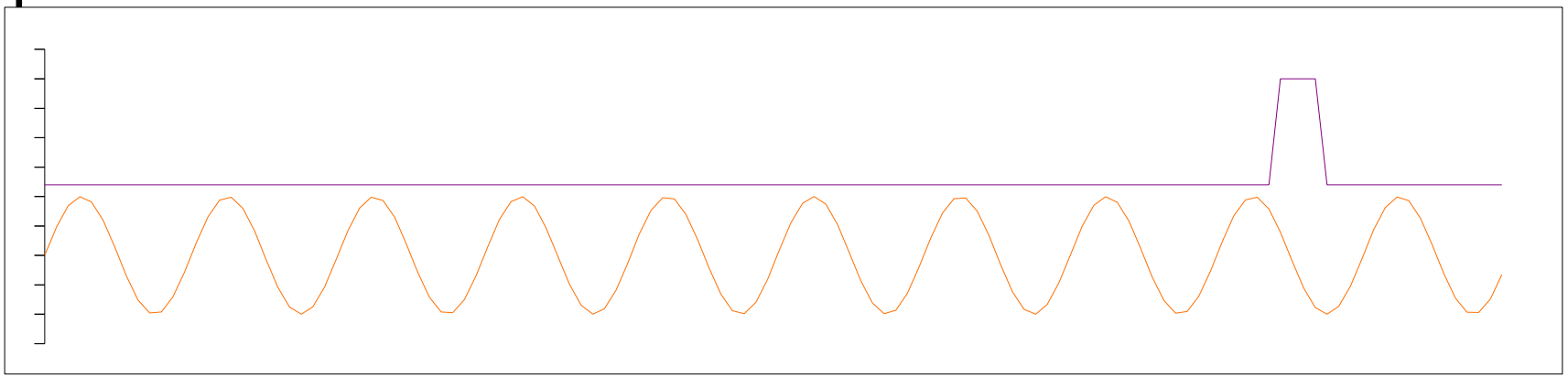


From procedure to process

Barbosa's two level model: lexical stress + 2 oscillators:

- phrase oscillator: pulses
- syllable oscillator: sine

Entrainment of syllable oscillator through attraction by phrase oscillator



Query *en passant*:

- How well does the Barbosa model relate formally to the Fujisaki model of intonation (note: syllables~accents)?
- Could the Barbosa timing model provide a timing dimension for the Fujisaki model?

The Rhythm Comb Model

An iterative low frequency ‘spectral comb’ filters input:

cf. Tillmann’ ‘prosodies’, three clocks:

A-prosody:

phrasal - intonation, pause structure (controlled)

several seconds

B-prosody:

words, syllables: rhythmic structure (given by structure of a language)

approximately ‘heartbeat’ length ≈ 1 Hz ♥ ☺

C-prosody:

segments - CV sequencing, transitions, allophones (maybe universal)

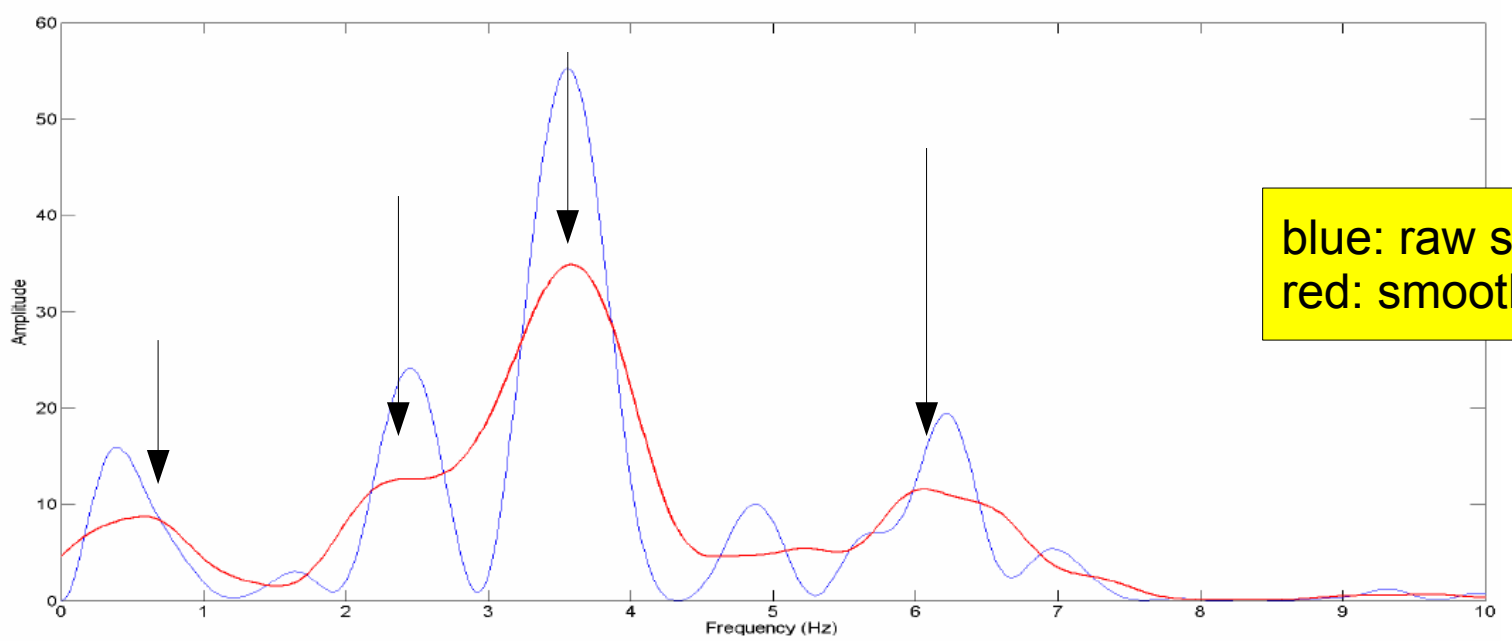
approximately 10-15 per second

Each level yields different correlations:

specifically: B-prosody yields a ‘Rhythm Comb’ (cf. Barbosa)

analogy: Fourier analysis of speech signal spectra

The Rhythm Spectrum



blue: raw spectrum
red: smoothed spectrum

Fig. 4. Raw (blue) and smoothed (red) spectrum. L = 31 points. N = 2048.

Tilsen & Johnson (2007):

- Low frequency Fourier analysis of speech rhythm
  - Here, peaks in smoothed spectrum at:
    - 0.5 Hz (2.0 s) - phrases?
    - 2.2 Hz (0.45 s) - feet?
    - 3.7 Hz (0.27s) - syllables?
    - 6 Hz (0.2s) - segments?

Envoi

Summary: the argument

We need a clear explicandum for rhythm:
- not just a definition
- a model

We need to be clear about the relevant level of analysis:
- grammatical
- phonological
- phonetic

We need to be clear about the relevant parameters:
- interval duration
- amplitude variation
- frequency variation

We need to be aware that rhythm is oscillation
- variants of the Rhythm Comb Model
  - including low frequency Fourier analysis

Conclusion

- Rhythm is an emergent product/percept
  - a function of
    - many regularities in language
    - and in the production / perception of speech
- Sometimes rhythm – like pitch patterning – may be
  - stylised, as in
    - song
    - recitation
  - focussed rhetorically, as in
    - public speaking
    - emphatic speech
- But usually the factors are so complex that rhythm as an objective measure does not emerge
- So let us look for TIMING PATTERNS of many kinds ...

And maybe one day we can make sense of the temporal structure of whole texts ...