Formal models of oscillation in rhythm, melody and harmony

Formalne modele oscylacji w rytmie, melodii i harmonii

Dafydd Gibbon

Universität Bielefeld, Germany
gibbon@uni-bielefeld.de

ABSTRACT

The aims of this contribution are to present (a) an intuitive explicandum of rhythm as iterated alternation, going beyond measurements of regularity; (b) a novel formal explication of rhythm in terms of Jassem's rhythm model seen from a computational point of view as a *Rhythm Oscillator Model*, (c) a computational model for two-tone Niger-Congo languages as a *Tone Oscillator Model* and (d) a pointer to the similarities of structure between these two and to a formal syllable model, suggesting new avenues of research and application.

STRESZCZENIE

Celem niniejszej pracy jest przedstawienie (a) intuicyjnej eksplikacji rytmu jako powtarzających się zmian, wychodząc poza pomiar regularności; (b) nowego formalnego wyjaśnienia rytmu w kategoriach modelu rytmu Jassema, postrzeganego z obliczeniowego punktu widzenia jako *Oscylacyjny Model Rytmu*, (c) modelu obliczeniowego dla dwutonalnych języków nigero-kongijskich (*Oscylacyjny Model Tonu*), oraz (d) wskazanie podobieństw strukturalnych pomiędzy tymi dwoma modelami a formalnym modelem sylaby. Uogólnienie nieodłącznych własności obliczeniowych modelu rytmu Jassema na inne dziedziny sugeruje nowe sposoby wyjaśnienia rytmu na poziomie stopy i sylaby jako oscylacji polegającej na naprzemiennym następowaniu postaci i tła, a także wskazuje na nowe obszary badań i zastosowań.

1. Rhythm, melody and harmony

Rhythm is a property of word sequences in utterances, perceived as regularly alternating values of an observable parameter in time or, metaphorically, in space. Some rhythms are forms alone – the rhythm of waves breaking, of pendulums swinging, many musical genres. Speech rhythms also have communicative functions, indicating semantic, pragmatic and social cohesion in the flow of utterances, and they share these communicative functions with conversational gesture types such as rhythmical 'beats' performed with hands, eyebrows, head-nodding. Measurement of the forms of speech rhythms cannot tell more than half the story, particularly since cohesive rhythm functions can – as with other patterns in speech – create top-down expectations which may not have any clear physical correlate.

In the architecture of language, speech rhythms are post-lexical syntagmatic patterns (pre-lexical from a perception point of view), and the present contribution touches on several varieties of such patterns.

The starting point is the rhythmic component of Wiktor Jassem's model of intonation, and the following areas of post-lexical patterning are addressed: (i) *rhythm* (of feet, syllables and combinations of these); (ii) *melody*, represented by the phonetic realisation of lexical tone in Niger-Congo languages; (iii) *harmony*, relating to alternations in the consonantal and vocalic timbres which enable syllable-based rhythms.

First, an overview of rhythm from an intuitive point of view is given, followed by a brief critical discussion of 'rhythm metrics', after which a formal generalisation of Jassem's rhythm model, the *Rhythm Oscillator Model*, is developed, with the aim of explaining how the tripartite (*pre-peak*) *peak* (*post-peak*) patterns of rhythm events emerge. It is shown how this oscillator model applies not only to stress feet but also to syllables and to tone terrace sequences; finally, pointers to future research in broader phonetic and multimodal contexts are given.

2. Models of rhythm

2.1. An explicandum for rhythm models and theories

Rhythmic form is a repeated, temporally regular iteration of alternating (perhaps binary) values of an observable parameter, e.g. *strong-weak*, *light-dark*, *loud-soft*, *consonant-vowel*, *hand raised vs. hand lowered*. The parameter may be a *single* feature type or a *complex* combination of many, it may be *hierarchical* in structure, it may appear in any modality, it may be just a physical *pattern* or have communicative *functionality*: visual (patterns of fences, waves, gestures); auditory (rhythms of speech, music, clocks ticking, train wheels); tactile (dancing, patting, stroking), perhaps even olfactory and gustatory.

Intuitively, a rhythmic form has several key properties, being (a) a *time series* of (b) *rhythm events*, with (c) each event containing (at least) a pair of *different observable values of a parameter* over (d) *intervals of time of relatively fixed perceived duration*. Finally, (e) *'it takes (at least) two to make a rhythm'*: one alternation of parameter values is not yet a rhythm. These features are visualised for a basic binary rhythm model in Figure 1; it will be shown, however, that not all speech rhythm is binary.

A model of rhythm needs to capture these rhythm event specifications. A basic formal model (as a kind of regular expression enhanced with linguistic notations) is shown in (1), with 'parameter' and 'duration' abbreviated as *PAR* and *DUR*).

(1) RHYTHM_SEQUENCE = <<*PAR*(*a*);*DUR*(*s*)>, <*PAR*(*b*);*DUR*(*w*)>>^{2}

The model defines sequences of at least two pairs of rhythm events, each rhythm event consisting of constituent observable simple event pairs <*PAR*(*a*);*DUR*(*s*)> and <*PAR*(*b*);*DUR*(*w*)>, with similar values for *a*, *b* and for *s*, w respectively in each iteration.

However, speech rhythm is much more complex than this intuition-based binary model permits, as plausible pronunciations of the following examples, with comments in the terminology of poetic metre, demonstrate:

(2) This | fat | bear | swam | fast | near | Jane's | boat. (Singlets, syllable-timed, *dum dum dum dum*.)

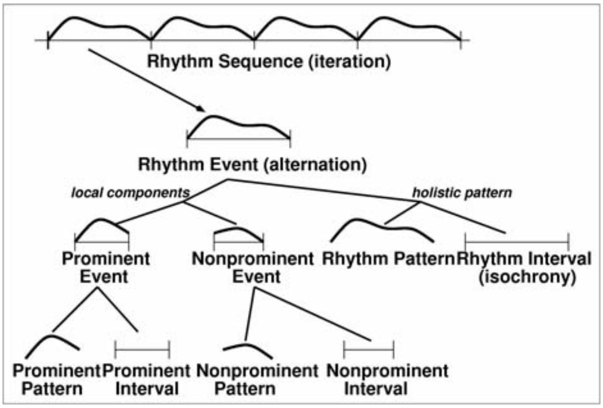

Figure 1: Visualisation of a binary rhythm model.

(3) And then | a car | arrived. (Iambs, *de-dum*.)
(4) This is | Johnny's | sofa. (Trochees, *dum-de*.)
(5) Jonathan | Appleby | carried it | awkwardly. (Dactyls, *dum-de-de*.)
(6) It's a shame that he fell | in the pond. (Anapaests, *de-de-dum*.)
(7) A lady | has found it | and Tony | has claimed it. (Amphibrachs, *de-dum-de*.)

The foot structures illustrated in examples (2) to (7), and more, may be freely mixed in spoken English. The upshot is: rhythms are neither more nor less binary in everyday speech than they are in poetry and music, a fact which has consequences for the structure of rhythm models.

2.2. A rhythm model typology

In the literature, e.g. [1, 2, 3], three main kinds of rhythm model are described: (1) the phonological type (not treated here), which is concerned with sequential and hierarchical rhythm relations between categories such as syllable, foot, and uses tree and grid patterns for this purpose; (2) the phonetic type, which is concerned with quantitative metrics for real-time temporal relations between speech events; (3) the oscillator type, which combines the phonological and phonetic types by defining temporal phonetic relations within a loop-shaped event pattern.

2.3. Quantitative phonetic models of rhythm

The quantitative 'rhythm metrics' of many studies over the past two decades have essentially just measured the variability of durations in utterances and lack the essential factor of iterated alternation ('*dum-de-dum-de-dum*' or '*de-dum-de-de-dum*' etc.) which characterises rhythm [1].

It is easy to show that variability metrics average globally across whole utterances and show a form of temporal 'smoothness': durations of different lengths can be randomly ordered, ordered longest-to-shortest or shortest-to-longest, and still have the same variability index (e.g. variance) as the durations in a genuinely rhythmical sequence. It is useful to know about the smoothness factor which may accompany rhythm, but it is not rhythm.

An apparent exception to the variability metric type is the *normalised Pairwise Variability Index* (*nPVI*), which determines the mean normalised duration difference between neighbours in an utterance, as shown in (8):[1]

$$(8)\quad nPVI = 100 \times \mathrm{MEAN}_{k=1;m\text{-}1}(\,|\,\mathrm{DIFF}(d_k, d_{k+1}) \,/\, \mathrm{MEAN}(d_k, d_{k+1})\,|\,)$$

where m is the number of rhythmic components (syllables, feet etc.) in sequence, k is the ordinal variable over the sequence.

In this model, the duration differences between neighbouring intervals are normalised by division with the mean of these durations, and the mean of the absolute values of the normalised durations is multiplied by 100.[2]

Although relations between neighbours sound like a promising approach for capturing at least binary patterns, this hope is dashed not only by the existence of non-binary speech rhythms but more fundamentally by taking the absolute value, which destroys any alternating property: if positive and negative differences are regarded as the same, then this is just another measure of smoothness, since the same set of durations can be ordered in increasing or decreasing order, or in any combination, and still yield the same nPVI as a genuinely iterative structure: *nPVI*(<1, 2, 1, 2, 1, 2>)=66.6' (alternation, 'real' rhythm), *nPVI*(<1, 2, 4, 8, 16>)=66.6' (geometrical series, not rhythm) and the same applies to any arbitrary mix: nPVI(<1, 2, 1, 2, 4, 8, 4, 2, 1, 2, 1>)=66.6'. This is easily verified in a simple spreadsheet implementation of the *nPVI*. The *nPVI* thus thus measures 'smoothness' like other variance models and lacks the iterating alternation pattern which rhythm oscillations require. Strictly speaking, therefore, none of the variability models are rhythm models, though they capture a useful overall 'smoothness' feature as a possible accompanying feature of rhythm.

3. A generalisation of Jassem's rhythm model

3.1. Jassem's rhythm model for English

The rhythm model proposed by Jassem in 1949 goes beyond intuitive binary models [3]:

> in ðe sentns **ai 'həːd ə moust pi'kjuːljə 'saund** (wið hai pitʃ on **həːd**) ðər ə θriː ʌnstrest siləblz aːftə **həːd** @nd jet ðat siləbl **həːd** is noutisəbli loŋɡə ðeə ðən in ðə sentns : **ai 'həːd im 'siŋ**, weə its ounli foloud bai wʌn ʌnstrest siləbl. ðe riðmikl dʒʌŋtʃə əkəːz aːftə **həːd** in ðə foːmə ənd aːftə **im** in ðə latə keis.[2]

The core structure is the *rhythmical unit*, later re-named *Narrow Rhythm Unit*, *NRU* [5, 6, 7], consisting of a *stressed syllable* and (optionally) a sequence of unstressed

[1] The formula is written in a compact equivalent style; see [2] for the equivalent original formula and critique of the *nPVI* measure.

[2] Values of *nPVI* range asymptotically from 0 (for absolutely equal durations) towards 200 (for random duration variation, 'noise'). If normalisation were by the sum, not the mean, the asymptote would be 100, like a percentage.

[3] "In the sentence /ai ˈhəːd ə moust pi'kju:ljə ˈsaund/ (with high F0 on /ˈhəːd/) there are three unstressed syllables after /ˈhəːd/ and yet that syllable /ˈhəːd/ is noticeably longer there than in the sentence /ai ˈhəːd im ˈsiŋ/, where it's only followed by one unstressed syllable. The rhythmical juncture occurs after /ˈhəːd/ in the former and after /im/ in the latter case." [Original in IPA]

syllables, terminated by a *rhythmical juncture*. The criteria for *RJ* placement are partly grammatical, as seen in (9) and (10).

(9) ai'həːd əmoustpi'kjuːljə 'saund
(10) ai'həːdim 'siŋ

NRU instances in a sequence tend towards approximately equal length (*NRU* isochrony): the more unstressed syllables in the *NRU*, the shorter the syllables. The *NRU* corresponds approximately to the traditional stress-initial (*trochaic*, *dactylic* etc.) foot.

However, *iambic*, *anapaestic*, etc., patterns with unstressed syllables after an *RJ* and before a stressed syllable do not belong to the *NRU* which begins with this stressed syllable; this type of sequence is "pronounced as short as possible" [4], and is termed in later work the *anacrusis* (*ANA*). The combination *ANA* + *NRU* is the *Total Rhythm Unit* (*TRU*) [5–7, 3].[4]

The conventional terms *ictus* (stressed syllable) and *remiss* (unstressed syllable sequence) [8] are not used by Jassem, but are convenient for referring to the constituents of the *NRU*, and for expressing the generalisation that for foot stress-accent languages the tripartite (*anacrusis*) *ictus* (*remiss*) sequence constitutes the (*pre-peak*) *peak* (*post-peak*) figure-ground pattern which underlies rhythmic alternation.

Jassem's model thus includes all the foot structures shown in examples (2) to (7), and more, while the traditional rhythm model only recognises the *ictus – remiss* foot pattern, and ignores *anacrusis* and *RJ* as independent categories.

The same tripartite (*pre-peak*) *peak* (*post-peak*) pattern applies to many other kinds of speech rhythm, whether *accent-based* (for *foot-timing*) or *sonority-based* (for *syllable-timing*), or as hybrid combinations of accent and sonority based rhythm. Figure 2 visualises the tripartite pattern applied to a syllable and a *TRU* (syllable timing is not claimed for this illustration): the (*onset*) *nucleus* (*coda*) sequence, like the (*anacrusis*) *ictus* (*remiss*) pattern are both instantiations of the generic (*pre-peak*) *peak* (*post-peak*) template.

With the generic model of the (*pre-peak*) *peak* (*post-peak*) template in implemented in terms of accent or sonority or both, an explanatory framework emerges for integrating foot-timed and syllable-timed languages into a common framework, in a 'timing space' structured on the one hand by accent assignment conventions and on the other by phonotactic sonority patterns and post-lexical fortition-lenition conventions. The model also shows, incidentally, how a single syllable can work as an entire foot in 'syllable-timed' sequences in English with neither anacrusis nor remiss.

3.2. The Jassem's rhythm model as a Finite State Automaton

A structure which defines the formal properties of Jassem's rhythm model is shown in (11) as a computable regular expression, enhanced with linguistic notational conventions.

(11) (*TRU* (*ANA* [+syll, -stress]*) (*NRU* [+syll, +stress] [+syll, -stress]*)) *RJ*

[4] Hirst [4] integrates a Jassem type rhythm model into an overall intonation model (but with no *TRU* and with *ANA* as single syllables, not a sequence).

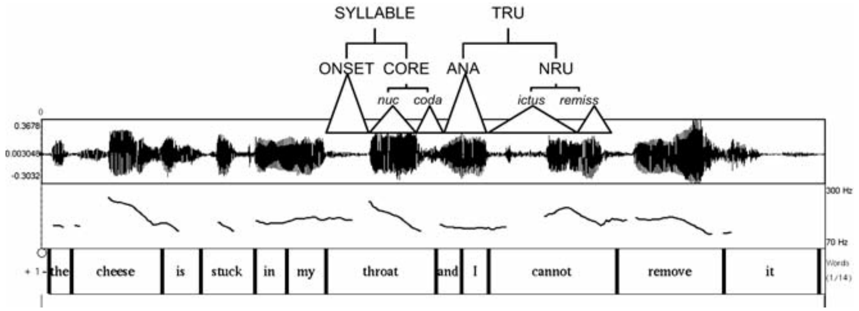

Figure 2: The generic (*pre-peak*) *peak* (*post-peak*) template applied to a syllabic sonority pattern (*thr* + *oa* + *t*) and an accent pattern (*and I* + *can* + *not*).

The bracketed sequence labelled *TRU* is followed by an *RJ*. The optional *ANA* is followed by the *NRU*, with obligatory *ictus* and optional *remiss*. The asterisk means 'none or arbitrarily many', and the exact specification of 'arbitrarily many' is narrowly constrained by language specific syllable and foot patterns. Conventional foot models can be formulated with a simpler regular expression (12).

(12) (*RU* [+syll, +stress] [+syll, -stress]*)

Studies of *RU* models have failed to find evidence for isochrony; however, for the *NRU* a tendency towards isochrony has been demonstrated [7].

3.3. Jassem's rhythm model as a rhythm oscillator

In order to capture the tripartite (*pre-peak*) *peak* (*post-peak*) template, a dynamic rhythm model is proposed here as an *oscillator* system, formalised as a *Finite State Automaton* (*FSA*), a formalism which has the advantage of being both simple to understand and easily computable (see [9] for an application to English syllable structure). Figure 3 shows an *FSA* as a transition diagramme for Jassem's *TRU* model (left, three states: *A*, *B*, *C*) and the conventional *RU* model (right, two states *A*, *B*).

In the *TRU* model, the *anacrusis* option (absent in the *RU* model) is expressed as a loop, and is either ignored or traversed arbitrarily many times. Then comes the obligatory

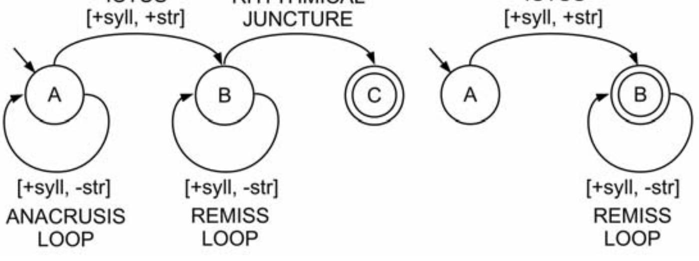

Figure 3: The Jassem TRU model and the conventional RU model.

ictus, proceeding to state *B*, where a *remiss* loop is optionally traversed, followed (in the *TRU* model) by the obligatory *RJ*, ending at state *C*, the final state (marked with concentric circles).

Sequences of *TRU*s are modelled by iterating the entire *TRU* (Figure 6) by means of a *RJ* transition from *B* back to *A*, generalising the *TRU* model as a *Rhythm Oscillator Model*. The generalisation also provides a free ride for defining different *RJ* contexts, permitting context-specific realisations such as utterance final lengthening on the B-C transition.

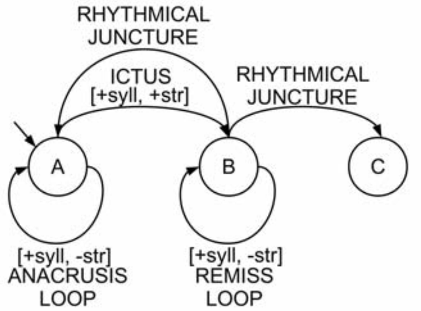

Figure 4: Jassem Rhythm Oscillator Model as two local oscillators linked in a general oscillator.

4. The tone sequence oscillator model

It turns out that post-lexical phonetic reflexes of lexical tone follow a similar oscillator template to post-lexical rhythm organisation. Tchagbale's Tem data [10] (Niger-Congo, Gur; ISO 639-3: *kdh* Togo), show two tones and language-specific assimilatory tonal sandhi rules which lead to stepwise falling 'terraces' of high-low tone sequences: (a) a low tone and high lexical tones are realised low and high, respectively, by default, except that (b) a low after a high is raised to the level of the high, (c) a low tone triggers lowering ('automatic downstep') of the following high in relation to the previous high.

In the lexical representation in (13) the syllables /lɪ/, /ka/ and /ɲɟɔ/ have lexical high tone; /bɛ/ and /jɪ/ have lexical low tone (acute accent before vowel indicates lexical high, no accent means lexical low).

(13) /bɛlˈɪ jɪkˈaɲɟˈɔ/ 'cut near the horn'

In contrast, the post-lexical, phonetic tone representation of (13) has a slightly different pattern (14).

(14) [bɛl!ˈɪ jˈɪk!ˈaɲɟˈɔ]

Figure 5 shows spectrum and F0 trace for an utterance of (14) with the 'terraced' pattern predicted by the tone rules. The low on /bɛ/ remains low, the following lexical high on /lɪ/ is partially assimilated (rises slowly) after the lexical low. After the lexical

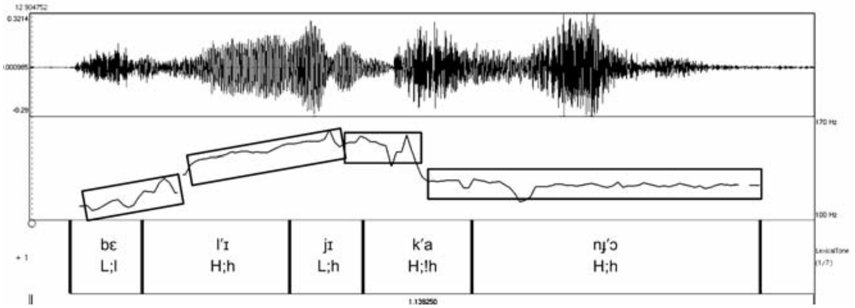

Figure 5: Tem /bɛl'ɪ jɪk'anɟ'ɔ/ (Z. Tchagbale, 2012-04-22). Rectangles delineate phonetic tone realisations; syllable labels show lexical-postlexical tone pairs.

high on /lɪ/ the low on /jɪ/ is fully assimilated and surfaces as phonetic high. Lexical low on /jɪ/ triggers downstepped high on /ka/, shown by [!']. High on /ɲɟɔ/ retains the F0 of the preceding high.

This post-lexical tone patterning (tone sandhi, tone terracing) has been captured previously with *Finite State Transducer*, *FST* models [11, 12]. A *FST* is a *FSA* with single symbols on transitions replaced by pairs of symbols (here: lexical/post-lexical tone pairs). An *FST* based application for computing F0 in Ibibio (Niger-Congo, Delta Cross; ISO 639-3 *ibb*) is shown in [13]. Figure 6 shows the Tem *FST* as an oscillator with two local loops for high and low default sequences, linked by a global loop which handles tone assimilations, thus matching Tchagbale's data (and the F0 trace of Figure 5).

It is striking that the *Tone Oscillator Model* in Figure 5 has essentially the same tripartite oscillator structure as the *Rhythm Oscillator Model* in Figure 4, Interpretation as a (*pre-peak*) *peak* (*post-peak*) sequence is initially not so plausible in the case of tones, unless high tones are defined as 'marked', in which case the downstepping low-

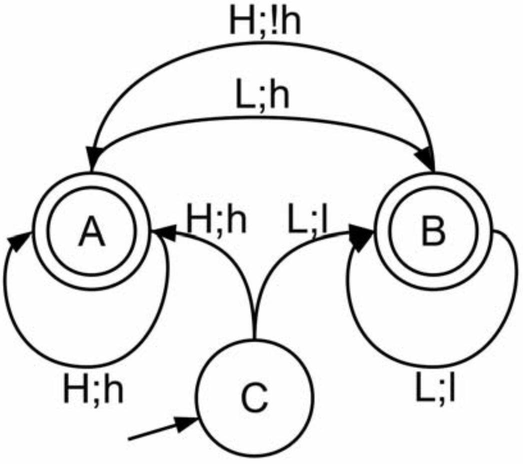

Figure 6: Tone Oscillator Model for Tem, based on data in [10].

high transition has *peak* status. States *A* and *B* in the *Tone Oscillator Model* follow the same pattern as States *A* and *B* in the *Rhythm Oscillator Model*. Formal differences (transition labels aside) are the labelling of syllables with tones, not stresses, the marking of both *A* and *B* as final states, and in replacing the terminal state *C* with an initial state *C* which feeds *A* and *B*. Functional differences between the rhythm and tone models are (a) the tone model needs no strict constraints on sequence length, (b) durations of the tone 'terraces' will not necessarily tend towards isochrony. This is a research avenue which has not previously been described.

5. Conclusion and outlook

A generic oscillator model based on a tripartite (*pre-peak*) *peak* (*post-peak*) template was developed in order to provide a common formal framework for different post-lexical domains: accent based rhythm ('foot timing'), sonority based rhythm ('syllable timing') as well as terraced tone sandhi. The starting point is Jassem's rhythm model, generalised as the *Rhythm Oscillator Model*. The oscillator model is formalised as a *Finite State Automaton* and shown to capture post-lexical phonetic implementations of sequences in each of the post-lexical domains, with both domain and language specific differences in the details of the respective models. An explanation for these similarities may lie in the temporal figure-ground *gestalt* (cf. [15]) represented by the (*pre-peak*) *peak* (*post-peak*) template.

Application of the generic model to multimodal domains, for example to iterative 'beat' gestures in conversation [14], is one obvious way forward. Hyman [16] has already suggested close relationships between accent and tone systems at the lexical level, and the generic oscillator model suggests a way of expressing these relationships post-lexically. Another future line of research is harnessing the generic tripartite *FSA* oscillator model to quantitative oscillator models of rhythm [17].

Whatever the explanation for these similarities may turn out to be: Jassem's rhythm model has become a catalyst for a much wider range of post-lexical oscillatory patterns than could have been anticipated when Jassem developed its foundations in the mid 20th century.

References

[1] Gibbon, D. 2006. Time Types and Time Trees: Prosodic Mining and Alignment of Temporally Annotated Data. In: S. Sudhoff, D. Lenertová, R. Meyer, S. Pappert, P. Augurzky, I. Mleinek, N. Richter, J. Schließer, eds. *Methods in Empirical Prosody Research*. Berlin: Walter de Gruyter. 281–209.

[2] Gut, U. (this volume). Rhythm in L2 speech. In: In D. Gibbon, D. Hirst and N. Campbell, eds, *Rhythm, Melody and Harmony in Speech. Studies in Honour of Wiktor Jassem*. Poznań: Polskie towarzystwo Fonetychne/Polish Phonetics Association.

[3] Jassem, W. 1949. indikeiʃn əv spiːtʃ riðm in ðə traːnskripʃn əv edjukeitid sʌðən inɡliʃ (Indication of speech rhythm in the transcription of educated Southern English). *Le Maître Phonétique*, III (92), 22–24. [Republished in *Journal of the International Phonetic Association* with the original IPA version and a new orthographic version by D. Hirst].

[4] Hirst, D. (this volume). Empirical models of tone, rhythm and intonation for the analysis of speech prosody. D. Gibbon, ed. *Rhythm, Melody and Harmony. Festschrift for Wiktor Jassem on the occasion of his 90th birthday. Speech and Language Technology 14*.

[5] Jassem, W. 1952. *Intonation of Conversational English (Educated Southern British). Prace Wrocławskiego Towarzystwa Naukowego (Travaux de la Société des Sciences et des Lettres de Wrocław)*. Seria A. Nr. 45. Wrocław: Nakładem Wrocławskiego Towarzystwa Naukowego.

[6] Jassem, W. 1983. *The Phonology of Modern English*. Warsaw: Państwowe Wydawnictwo Naukowe.

[7] Jassem, W., D. R. Hill and I. H. Witten. 1984. Isochrony in English Speech: its Statistical Validity and Linguistic Relevance. In D. Gibbon and H. Richter, eds. *Intonation, Accent and Rhythm: Studies in Discourse Phonology*. Berlin: Mouton de Gruyter. 203–225.

[8] Abercrombie, D. 1967. *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

[9] Gibbon, D. 2001a. Preferences as defaults in computational phonology. In K. Dziubalska-Kołaczyk, ed., *Constraints and Preferences*. Trends in Linguistics, Studies and Monographs 134. Berlin: Mouton de Gruyter. 143–199.

[10] Tchagbale, Z. 1984. Tonologie: 16 – TEM (Gur, Togo). In Z. Tchagbale, ed. T.D. *de Linguistique: exercises et corrigés*. Abidjan: Institut de Linguistique Appliquée. Université Nationale de Cote-d'Ivoire, No. 103.

[11] Gibbon, D. 1987. Finite state processing of tone languages. In: *Proceedings of European ACL*, Copenhagen.

[12] Gibbon, D. 2001b. Finite state prosodic analysis of African corpus resources, *Proceedings of Eurospeech 2001, Aalborg, Denmark*, I: 83–86.

[13] Gibbon, Dafydd, E.-A. Urua, U. Gut 2003. A computational model of low tones in Ibibio. In: *Proceedings of the International Congress of Phonetic Sciences, Barcelona*, 2003, I, 623–626.

[14] Gibbon, D. 2011. Modelling gesture as speech: A linguistic approach. In *Poznań Studies in Contemporary Linguistics (PSiCL)* 47, 470f.

[15] Dziubalska-Kołaczyk, K. 2002. *Beats-and-Binding Phonology*. Frankfurt: Peter Lang.

[16] Hyman, L. 2009. How (Not) to Do Phonological Typology: The Case of Pitch-Accent. *Language Sciences* 31 (2–3), 213–238.

[17] Barbosa, P. and W. da Silva. 2012. A New Methodology for Comparing Speech Rhythm Structure between Utterances: Beyond Typological Approaches . In H. Caseli, A. Villavicencio, A. Teixeira, F. Perdigao, eds. *Proceedings of Computational Processing of the Portuguese Language: 10th International Conference, PROPOR 2012*, Coimbra, Portugal, April 17–20, 2012. Berlin: Springer.