Formal models of oscillation in rhythm, melody and harmony

Dafydd Gibbon

Universität Bielefeld, Germany
Fakultät für Linguistik und Literaturwissenschaft
Postfach 100131, 33501 Bielefeld
gibbon@uni-bielefeld.de

ABSTRACT

The aims of the present 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 the Jassem Rhythm Model seen from a computational point of view as a Rhythm Oscillator Model, (c) a computational model for two-tone Niger-Congo languages (the Tone Oscillator Model) and (d) a pointer to the similarities of structure betwen these two and to a formal syllable model. The generalisation of inherent computationational properties of the Jassem Rhythm Model to other domains suggests a new way of explicating both foot and syllable level rhythm as an oscillation involving iterated alternation of figure and ground properties, and suggests new avenues of research and application.

1. Rhythm, melody and harmony

Rhythm will be taken here, intuitively, to be the regular temporal patterning of alternating observable events. In speech, rhythm may be occur with syllable sequences, foot sequences, or larger units, perhaps hierarchically. Melodies are organised rhythmically, as are the harmonies of speech sounds.

The aims of the present contribution are to present an intuitive explicandum of rhythm as alternation, going beyond measurements of regularity, a novel explication of rhythm in terms of the Jassem Rhythm Model seen from a computational point of view as a Rhythm Oscillator Model, a computational model for two-tone Niger-Congo languages (the

Tone Oscillator Model) and a pointer to the similarities of structure betwen these two and a formal syllable model. These cross-domain similarities suggest a common basis for foot and syllable level rhythm types, and open new avenues for research and application.

2. Models of rhythm

2.1. And explicandum for rhythm models and theories

Beyond the initial intuitive characterisation, rhythm is a temporally regular variation between alternating (e.g. strong-weak, light-dark, loud-soft) values of some parameter. The parameter may be a single parameter or a combination of many, and in any of the senses: visual (patterns of moiré silk, of waves); auditory (rhythms of speech and music, clocks ticking, train wheels); tactile (dancing, patting and stroking). As an explicandum for the modelling of rhythm, this informal extended definition is sufficient.

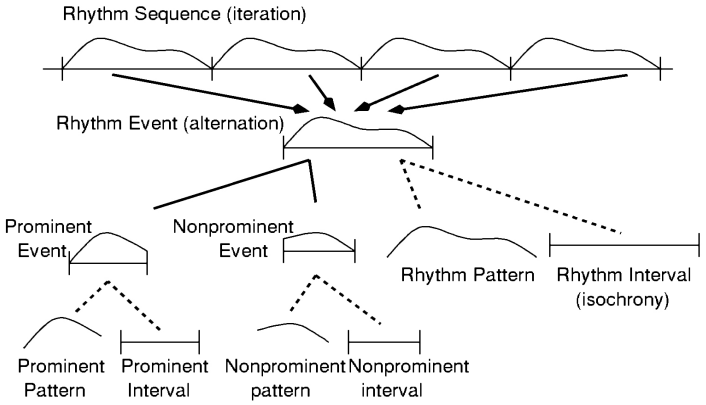

Figure 1: Visualisation of informal binary rhythm model.

The informal explicandum for rhythm immediately suggests (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 (c) *intervals of time* of *relatively fixed duration*, and (d) it crucially specifies that '*it takes two to make a rhythm*', at least: a single alternation of parameter values is not a rhythm. Example (1) summarises this set of conditions as a regular exrpression.

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

The regular expression describes a rhythm sequence as a sequence of at least two rhythm events, each consisting of constituent observable event pairs <PAR(*a*);DUR(s)> and <PAR(*a*);DUR(s)>, with similar values for *a*, *b*, *s, w*

respectively in each iteration. No position is taken here on whether a rhythm event is a physical event or a cognitively reconstructed epiphenomenal relation. This definition of rhythm event holds for binary rhythms only; definitions of ternary or other rhythms require specification of triples of constituent observable events, as well as pairs.

2.2. A rhythm model typology

The speech domain to be discussed here is, in linguistic terms, that of 'post-lexical timing rules', in particular the phonetic realisation of syllable or word timing. The literature (cf. [1], [2], [3]) shows three main kinds of rhythm model, (1) the phonological type of rhythm model, which is concerned with sequential and hierarchical relations between categories such as the syllable, foot, rhythm unit etc. as components), (2) the phonetic type of rhythm model, which is concerned with quantitative real-time temporal relations between categories, and (3) the oscillator type of model, which combines the two by defining temporal relations within a loop-shaped categorial structure.

2.3. Quantitative phonetic models of rhythm

Various experimentally based quantitative models of rhythm have been proposed since the 1960s, but all except the oscillatory models have lacked the essential iterative alternating 'dum-de-dum-de-dum' feature which characterises rhythm (cf. [1]). Most such models reduce to a variance measure for syllable durations in the utterance. However, variance measures as such are not designed for ordered series but rather for unordered sets. Such global measures consequently simply show the degree of similarity in durations, saying nothing about the iterative rhythmical structure. Durations of different length could be randomly ordered, or ordered longest-to-shortest or shortest-to-longest and still have the same variance as a rhythmical sequence.

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

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

where m is the number of syllables in the utterance, and k is the ordinal variable over the set of syllables. The absolute differences between neighbouring durations are normalised by division with the mean of these durations, and the mean of all such normalised durations in an utterance is multiplied by 100. The *nPVI* index ranges asymptotically from 0 (for absolutely equal durations) towards 200 (for increasing variation).

1 The formula is written in a generalised form suitable for implementation (and for ease of writing); see [Gut, this volume] for the original formula.

Although relations between neighbours sound like a promising approach for capturing at least binary iteration, this hope is destroyed by taking the absolute value. If positive and negative differences are regarded as the same, then this is just another measure of evenness, since the same set of durations can be ordered randomly, or in increasing or decreasing order, and still yield the same *nPVI* as a genuinely iterative structure. The *nPVI* thus has the same properties as variance models and lacks a genuine iterating alternating structure which rhythm oscillation requires.

3. The Jassem Rhythm Oscillator Model

3.1. Jassem's Rhythm Model of English

Jassem illustrates his earliest theory of rhythm as follows [4]:

> in ðe sentns **ai 'hɜ:d ə moust pi'kjuljə 'saund** (wið hai pitʃ on **hə:d**) ðər ə θri: ʌnstrest siləblz a:ftə **hə:d** @nd jet ðat siləbl **hə:d** is noutisəbli lonŋgə ðeə ðən in ðə sentns : **ai 'hə:d im 'siŋ**, weə its ounli foloud bai wʌn ʌnstrest siləbl. ðe riðmikl dʒʌŋtʃə əkə:z a:ftə **hə:d** in ðə fo:mə ənd a:ftə **im** in ðə latə keis.[2]

The core structure is the *rhythmical unit* (in later work [5, 6, 7] renamed Narrow Rhythm Unit, RU), consisting of a stressed syllable and (optionally) a sequence of unstressed syllables terminated by a *rhythmical juncture* (RJ, rendered in the bold transcriptions by a space). 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 stress-only corresponds approximately initial (trochaic, dactylic etc.) foot: unstressed syllables after the juncture but before a stressed syllable do not belong to the rhythmical unit. These syllables are 'pronounced as short as possible' and named *anacrusis* (ANA) in later work (cf. also [3]). The traditional terms *ictus* and *remiss* will be used for the stressed syllable and the unstressed syllables of the NRU, respectively [8]. The traditional approach includes all unstressed syllables in the rhythm unit (RU) and ignores the anacrusis effect.

2 Original in IPA; orthographic rendering by D. Hirst: In the sentence /ai ˈhəːd ə moust piˈkjuːljə ˈsaund/ (with high pitch 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. The rhythmical junctures are (1) /aɪˈhəːd əmoustpiˈkjuːljə ˈsaund/, (2) /aiˈhəːdimˈsiŋ/.

3.2. The Jassem Rhythm Model as a Finite State Automaton

Using a traditional feature notation, example (3) uses a regular expression to show the structure of the Jassem Rhythm Model: an anacrusis (ANA) consists of zero or arbitrarily many (represented by an asterisk) unstressed syllables and is followed by a rhythmical unit (RU) consisting of an obligatory stressed syllable and zero or arbitrarily many unstressed syllables and the rhythmical juncture RJ:

(3) (BRU (ANA [+syll, -stress]*) (NRU [+syll, +stress] [+syll, -stress]*)) RJ

The conventional foot models can be formulated with a simpler regular expression, as in example (4).

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

In type (2) models, syllables are forced into the RU. While investigations of type (2) failed to find convincing evidence for the isochrony of the RU, later work [7] showed that if the anacrusis is not included, the NRU does tend to show a measure of isochrony, demonstrating the validity of the model.

3.3. Jassem's Rhythm Model as a Rhythm Oscillator

In the introductory section it was pointed out that regular iteration, i.e. oscillation, is an essential explicandum for rhythm. A rhythm model must handle this constraint explicitly. A dynamic rhythm model is introduced here as a Finite State Automaton (FSA). The expressions (3) and (4) can be stated formally as a Finite State Automaton (FSA) with three states, A, B, C. The FSA is shown in Figure 2 as a Finite State Network (FSN).

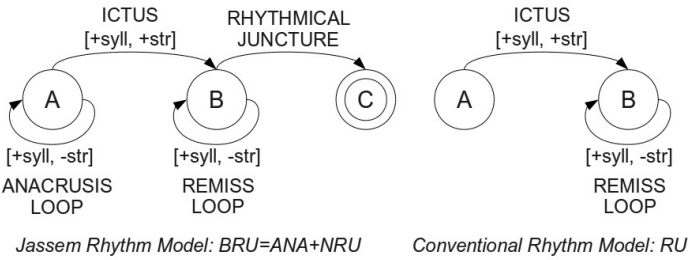

Figure 2: The Jassem rhythmical unit model (Foot model 1) and the traditional model (Foot model 2).

The Jassem Rhythm Model in Figure 2 is interpreted as a map defining a syllable pattern: starting at state A, the option is either to generate (or analyse) an anacrusis with a loop at state A permitting a choice between arbitrarily many unstressed anacrusis syllables, remaining at state A, or producing the ictus, a stressed syllable, and proceeding to state B. At state B, a optional remiss loop with an arbitrary number of unstressed syllables is permitted, remaining at state B, followed by a rhythmical juncture leading to

the final state C (marked by two concentric circles). The phonetically inadequate Foot Model 2 in Figure 2 is simpler, but interpreted similarly.

Another BRU can be started by iterating the entire rhythmical unit (Figure 5) with an RJ transition from B back to A. The advantage of this generalisation, the Jassem Rhythmical Oscillator Model, is that not only is the BRU unit well defined as ANA+NRU, but also a rhythmical juncture is licensed for in different contexts, either utterance internally or utterance finally, permitting different realisations (e.g. utterance final lengthening).

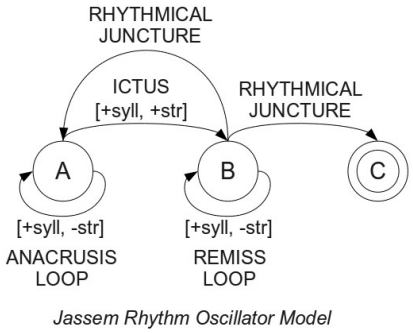

Figure 3: Jassem Rhythm Oscillator Model: Jassem Foot Model generalised to a foot sequence oscillator between two syllable oscillators.

The Jassem Rhythm Oscillator Model offers a more sophisticated foundation for quantitative and oscillatory models in phonetics and in speech technology (e.g. rhythm in corpus-based speech synthesis) and than the iterations of the traditional RU model (cf. also discussion by Hirst [3]).

3.4. Excursus on FSA syllable models

The FSA formalism is straightforward to deploy in a practical computational model, and, when enhanced by statistical information about probabilities of choice between transitions (i.e. as a Hidden Markov Model), can be used for both analysis and generation of rhythm. Further investigation reveals that an FSA oscillator model for the syllable shares a ternary formal structure with the Jassem Rhythm Oscillator Model, as example (5) shows.

(5) owe, coo, ski, scree, at, ask, asks, kick, skit, scrap, cant, pants

The corresponding English syllable patterns listed in (6) are a subset of the set of syllable patterns described by the regular expression C*VC*, with obligatory vowel nucleus and variable consonantal peripheries.

(6) V, CV, CCV, CCCV, VC, VCC, VCCC, CVC, CCVC, CCCVC, CVCC, CVCCC

The C*VC* model is a strong constraint on possible combinations of consonants and vowels, and is still *complete* (in the technical sense) in covering all English syllable structures. However, it permits arbitrarily long strings of consonants and is therefore not *sound*, since these strings are strictly limited in number, a disclaimer which also applies to the Foot Models. (For a more detailed Finite State Automaton for both English C*V*C patterns and also the required further constraints, see [9].)

4. The tone sequence oscillator model

The speech domain to be discussed in this section concerns the post-lexical phonetic reflexes of lexical tone. Tchagbale's analysis of Tem (Niger-Congo, Gur ISO 639-3: *kdh*, spoken mainly in Togo but also in Ghana and Benin) detailed in [10], shows that Tem has two tones and a well-known kind of assimilatory tonal sandhi in which a low tone (a) is realised low by default, (b) after a high tone is raised to the level of that tone, (c) triggers slight lowering of the following high tone.

The sentence in example (7), meaning "cut near the horn", is shown with an acute accent before the vowel for lexical high tone, no accent indicating lexical low tone): the syllables /lɪ/, /kan/ and /ɟɔ/ have lexical high tone, /bɛ/ and /jɪ/ have lexical low tone.

(7) /bɛlˈɪ jɪkˈanɟˈɔ/

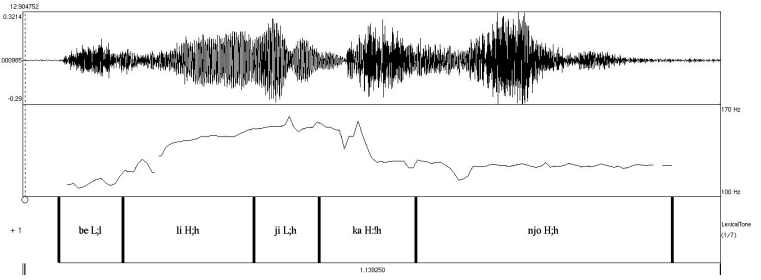

Figure 4: Praat pitch trace of Tem /bɛlˈɪ jɪkˈanɟˈɔ/, with tone domains shown by rectangles, and labels with syllables and lexical-postlexical tone pairs. (Recording: Z. Tchagbale 2012)

In example (8) and in Figure 3, a Praat pitch trace, the post-lexical tone representation is shown.

(8) [bɛlˈɪ jˈɪk!ˈanɟˈɔ]

The first low tone on lexical /bɛ/ is rendered low, but low tone on /jɪ/ is realised post-lexically with a high tone, after the lexical high tone on /lɪ/. After the lexical low tone (note: not the post-lexical high) /jɪ/ the lexical high tone on /kan/ is realised post-lexically with a downstepped high tone,

symbolised by [!′], slightly lower than the previous high. The pitch trace shows a ratio of 142:124 Hz, i.e. 1:0.873, just over two semitones, close to a supermajor second in musical terms, i.e. 8:7 or 1:0.875).

The pitch trace shows additional phonetic features which are not relevant to the present discussion (rise on initial low; rise on first high; level pitch on second high and the sequence beginning with downstepped high; delay of tones relative to associated syllables).

In [11] and [12] formal models of Niger-Congo tone sandhi were formulated, capturing the post-lexical tonal sandhi as transitions in a Finite State Transducer (FST). An FST is an FSA, but with pairs of symbols on the transitions, in this case pairs of lexical and post-lexical tones. Sequences of identical high or low tones are captured with a loop on the high tone or low tone state of the FST, respectively. To capture sequences of tones, and thus enable the formulation of the tone assimilation constraints, the low and high tone states are linked with transitions on which the tone assimilations take place (Error: Reference source not found). The phonetic interpretabililty of the model has been demonstrated in a number of places, for instance in Gibbon et. al. 2001].

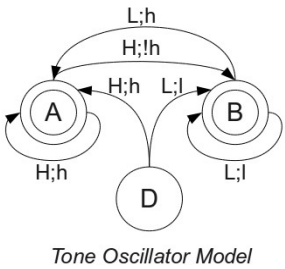

Figure 5: Tone oscillator [11], [12], based on data in [10].

In this way, post-lexical local phonetic structure of tonal assimilation is explicitly generalised to cover not only local automatic downstep but also the overall terraced downtrend which occurs in very many two-tone Niger-Congo languages.

It is very striking that the technique used for the Tone Oscillator in Figure 5 is the same as that independently developed for the Jassem Rhythm Oscillator in Figure 3. The structure involving states A and B in the Tone Oscillator is topologically the same as the structure involving states A and B in the Jassem Rhythm Oscillator. The formal differences (apart from the transition labels) are in the labelling of the syllables with tones rather than stresses, in the marking of both B and C as final states, and in the introduction of a start state D which links immediately to either state A or

state B. Functional differences are that the tone model needs no strict constraints on the length of sequences with the same tone, unlike the corresponding sequences in the rhythm and syllable models, and that the durations of the oscillations will not necessarily be regular. This is a research issue which has not yet been addressed in the literature.

The post-lexical pattern similarities for rhythm and tone sandhi, and their similarity with syllable structure, constitute a rather novel and surprising result in languages like English and post-lexical tone sandhi in languages like Tem, and between these and syllable structure in general.

5. Conclusion and outlook

In the preceding sections, the organisation of post-lexical prosody was discussed in terms of rhythm (the sequencing of accented syllables) and tone sandhi (the sequencing of syllables with contrastive tone). It was discovered that the basic computable formal models for these two cases are isomorphic, i.e. the same, barring detailed interpretation of the labels of the graphs concerned and details of constraints on sequence length and type.

This is not the first time that a close relationship between the organisation of tonal and accentual systems has been postulated: Hyman [13], in a conspectus of tone systems, pitch accent systems and other types of accent system, has suggested close relationships between the systems at the lexical level. The present contribution suggests close relationships at the post-lexical level (i.e. phonetic) level, too, and also a relationship with syllable organisation.

It is perhaps not too daring to suggest that a ternary oscillatory structure, with two very local oscillators and one main oscillator encompassing both of the local oscillators, is a characteristic and possibly (in a weak sense of the term) universal processing pattern for this level of speech organisation. The FSA oscillator models can conceivably be harnessed to quantitative oscillator models of rhythm such as that of Barbosa [15]. An explanation for these similarities across different domains is not obvious, but in general terms the models provide computable accounts of the temporal figure-ground *gestalt* relations [16] which underlie the iterative alternations of foot and syllable rhythm patterns in English and related languages, as well as of post-lexical terracing in the Tem type of Niger Congo language.

Whatever the explanation may turn out to be: the Jassem Rhythm Oscillator Model has become a catalyst for a 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., pp. 281-209.

[2] Gut, U. (in press). Rhythm in L2 speech. In: D. Gibbon, ed. *Rhythm, Melody and Harmony. Festschrift for Wiktor Jassem on the occasion of his 90th birthday. Speech and Language Technology 14.* [This volume.]

[3] Hirst, D. (in press). 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.* [This volume.]

[4] Jassem, W. 1950. indikeiʃn əv spi:tʃ riðm in ðə tra:nskripʃn əv edɜukeitid sʌðən ingliʃ (Indication of speech rhythm in the transcription f educated Southern English). *Le Maître Phonetique*, 22-24.

[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 Dafydd Gibbon and Helmut Richter, eds. *Intonation, Accent and Rhythm: Studies in Discourse Phonology*. Berlin: Mouton de Gruyter, pp. 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, pp. 143-199. Berlin: Mouton de Gruyter.

[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: pp. 83-86.

[13] Hyman, L. 2009. How (Not) to Do Phonological Typology: The Case of Pitch-Accent. *Language Sciences*, v31 n2-3 p213-238 Mar-May 2009.

[14] 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.
[15] 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. *Procedings of Computational Processing of the Portuguese Language: 10th International Conference, PROPOR 2012*, Coimbra, Portugal, April 17-20, 2012. Berlin: Springer.
[16] Dziubalska-Kołaczyk, K. 2002. *Beats-and-Binding Phonology*. Frankfurt: Peter Lang.