Computational modelling of rhythm as alternation, iteration and hierarchy

Dafydd Gibbon

Universität Bielefeld

gibbon@spectrum.uni-bielefeld.de

ABSTRACT

Recently interest in rhythm has revived and a number of new empirical measures have been proposed, for example by Low and Grabe, and Ramus. Closer examination shows these models to be incomplete, however, with respect to both empirical and formal adequacy conditions. The present contribution addresses these issues in relation to hierarchical temporal structuring and proposes an integrated computational phonetic approach, introducing an empirical data mining heuristic for inducing rhythm timing trees from large quantities of time-annotated data. New algorithms are proposed for Timing Tree Induction (TTI), and for a Tree Similarity Index (TSI) to estimate the similarity between syntax parse trees and prosodic trees as predictors for the structure of the temporal TTI trees. A preliminary quantitative evaluation shows a preference for tail-heavy (iambic) tree branching in a read-aloud narrative. Applications in speech genre analysis and for duration modelling in speech synthesis are envisaged.

1 INTRODUCTION

A number of new quantitative models of rhythm timing have been advanced in recent years, and used to compare different languages [1, 2]. The present contribution aims to examine the adequacy of these models, on the basis of the results to propose and test a method for automatically inducing rhythm timing trees from annotated data, and to evaluate grammatical structures as predictors for these timing trees. The computational phonetic methodology is new, and consequently "clear case" data from a read-aloud narrative are used for this initial study.[1] Results for more spontaneous spoken genres will presumably be less clear-cut.

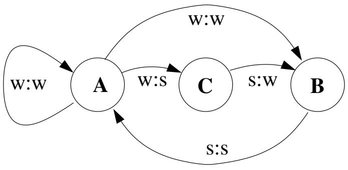

Figure 1: Basic rhythm filter FST (A, B are both initial and final nodes.)

2 RHYTHM TIMING MODELS

Current models of rhythm timing in speech are quite diverse but at the same time atomistic and selective, in that they focus on single parameters as different as global deviation of unit length, local unit length ratios, and consonant-vowel ratios [3, 1, 2], but do not cover non-binary rhythms or hierarchical factors in rhythm timing. More comprehensive approaches are emerging, however [4, 5, 6].

The dynamics of rhythm patterns include iteration (oscillation, from an operational perspective), and subordination to hierarchical timing patterns. The iterative component relates to Finite State (FS) phonology and prosody [7, 8, 4]. Both dimensions of hierarchical and iterative structure are found in Metrical Phonology [9, 10]. The hierarchical constraints are derived from grammatical structures, their core being the Nuclear Stress Rule (NSR). The earliest versions of the NSR were inner-bracket-removing operations on character strings containing well-formed nested bracketings. Liberman & Prince developed a tree-graph algorithm: a node-labelled binary tree has r on the root node, s on right daughter nodes, w on the others, and leaves are assigned abstract stress values equal to the depth of the first non-s node above the leaf. But the NSR is basically a recursive function:

```
nsr(t,n,m) { if (leaf? (t))
             then make-group(make-label(t,m))
             else concat(nsr(first(t),n+1,n+1),
                         nsr(rest(t),n+1,m))    }
```

Linear alternation constraints are expressed as

[1] Special thanks to Morten Hunke, Sara Johanning, Katrin Johannsen, Josef Raab, Alexandra Thies, Thorsten Trippel for syntax analyses, Ulrike Gut for the annotated data, Grazyna Demenko, Katarzyna Dziubalska-Kołaczyk, Ekaterina Iassinskaia, Ulrike Gut, Peter Ladkin, Zofia Malisz, and lecture audiences in Dublin, Bielefeld, Poznań and Tübingen, for much relevant discussion.

ISBN 1-876346-nn-n © 2003 UAB

histogram-like grids. Such constraints may be implemented Finite State Transducer (FST) filters, shown in the example in Figure 1, a filter with oscillatory loops which enforces the so-called Rhythm Rule (cf. *THIRteen MEN* but *thirTEEN*) and avoids metrical clashes. This FST applies to whatever phonostylistically determined "zoom" level of structural granularity is needed (syllable, foot, or larger unit) [11].

Roughly speaking, phonological models of rhythm have concentrated on the hierarchical component, while phonetic approaches have concentrated on the iterative component.

A classic phonetic approach to rhythm timing is that of Roach [3]: tone unit duration is divided by the number of feet in the tone unit, yielding average or "ideal" foot duration approximating to isochrony, and the normalised deviation from mean foot length is measured. The idea, of course, is to measure *syllable isochrony*, rather than rhythm as such. Neither hierarchy nor linear alternation of timing units figure in the approach, which may be said to use a *Global Evenness* (GE) criterion as a measure of the isochrony property, rather than the alternation property. Any arbitrary re-sorting of the relevant segments in an utterance (random, shortest-to-longest, etc.) would yield the same index. Rhythm timing fulfils the GE criterion, in some sense, but it has other properties too, so while the GE criterion for timing is a necessary criterion for rhythm, it is not a sufficient one.

Ramus, Nespor & Mehler [2] locate different languages in a typologically distinctive timing space over the following parameters: $V\%$, percentage of V (vocalic intervals) relative to overall utterance length; ΔC, variance of consonantal intervals; ΔV, variance of vocalic intervals. The $V\%$ measure reflects preferences for certain phonotactic patterns (CV, CVC, vowel length) as corpus tokens rather than lexical types. The model also uses a variety of GE criterion: V stretches and C stretches would still yield the same results if randomly sorted (by length, longer consonant sequences first, etc.). Similar considerations apply to the ΔV measure, which reflects evenness of vowel sequence lengths, lower values tending to isochrony, and to the ΔC measure. The model does not have hierarchical and alternating timing components and is thus is incomplete as a model of rhythm timing. Perhaps a different measure, such as ΔCV, could be used to address the issues of hierarchical and iterative structuring. A perceptual control for rhythmicity is clearly needed. As Cummins has pointed out [5], the measure makes a statement about the evenness of the phonotactics of the language, rather than rhythm, rather like Roach's model; it reflects a possibly necessary condition on rhythm, but falls short of providing a sufficient condition.

Low, Grabe & Nolan [1] addressed the GE issue and developed the Pairwise Variability Index (PVI) in order to take iterative alternation into account. The PVI measures normalised differences between the durations of adjacent units (vowels, syllables, etc.):

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

Wetzel's [12] comment that the PVI factors out final lengthening is mistaken: the counter does not stop short of the final item — a sequence of length m just has $m-1$ differences between neighbours. The model yields a minimal value of 0 (perfect isochrony), asymptotically approaching 200 for larger length differences; the variant used in [13] reverses the scale, and has a maximum of 100 for perfect isochrony. The model has an empirical problem: it assumes *strictly binary rhythm.* Hence, alternations as in "*Little John met Robin Hood and so the merrie men were born.*" are adequately modelled, but not the unary rhythm (syllable timing) of "*This one big fat bear swam fast near Jane's boat.*" or ternary dactylic and anapaestic rhythms (or those with even higher cardinality) like "*Jonathan Appleby wandered around with a tune on his lips and saw Jennifer Middleton playing a xylophone down on the market-place.*" But the model unfortunately also has a formal problem: the PVI yields the same value for sets of alternating patterns and monotonic geometrical series, and for mixes of these ($n!$ patterns with identical PVI for series of a length n). It is easily verified that alternating sequences may receive the same PVI as exponentially increasing or decreasing series $PVI(2,4,2,4,2,4) = PVI(2,4,8,16,32,64)$. This is obviously not the desired result. Interesting though the resulting typological patterns are, it is not at all clear what they are patterns of. This model, too, is empirically and formally incomplete.

Cummins [5] discusses a number of additional factors involved in the production of rhythm in different styles, ranging from a paradigm of synchronous speaking designed to elicit maximally rhythmic utterances, to less constrained styles. He addresses both hierarchical and linear factors, and proposes a model for the more constrained styles with binary hierarchical structure, i.e. groupings of two-word feet, higher level groupings of two feet with four words, and so on. A new aspect of Cummins' experimental approach is the emphasis on the entrainment of different factors in the synchronous production of rhythm, particularly the interaction of discrete and gradient factors, with coupling between prosodic factors at foot level and a higher level.

Wagner [4] criticises the hierarchical NSR type approach of Metrical Phonology (without rejecting a grid filter component, however), and concentrates on the linear alternation criterion, using FSTs with local cycles to formalise metrical grid type linear filters. Wagner shows that better results for synthesis of German speech are given by a linear model based on five part-of-speech sets with different intrinsic weighted abstract stress values [4]: {Nouns, Numerals, Proper Names}, {Adverbs, Adjectives}, {Verbs, Demonstrative Pronouns, WH-Pronouns}, {Modal & Auxiliary Verbs, Affirmative & Negation Particles}, {Determiners, Conjunctions, Subjunctions, Prepositions}. Wagner reintroduces the idea that grammatical categories are predictors of rhythm timing. In fact, these also contain strong assumptions about syntax hierarchies. For example, in German, many "weaker" parts of speech alternate with stronger items on syntactic grounds alone, often preceding stronger items in a given construction, thus inducing shallow hierarchies and perhaps an iambic rhythm, and suggesting interactions between rhythm and grammar which are of interest for language history and typology.

3 INDUCTION OF TIMING TREES

Rhythm timing as a complex function of hierarchical and linear structuring (cf. also Campbell's timing model [14]) is combined here with local alternation criteria and with grammatical predictors for timing trees. The approach is operationalised in two stages (both algorithms were first prototyped in the LispMe Scheme dialect, then ported to MIT-Scheme for large datasets):

1. automatic Timing Tree Induction (TTI) from long-short local duration differences in annotated speech signal data,
2. automatic calculation of a Tree Similarity Index (TSI) between the resulting timing trees and grammatical trees.

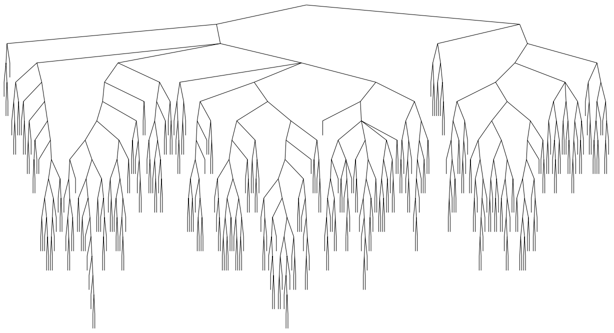

Figure 2: TTI tree induced over a narrative.

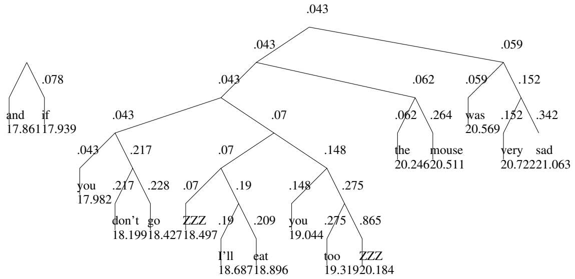

Figure 3: Zoom into the narrative timing tree.

In abstract terms, the TTI algorithm is the inverse of the NSR function, not used for abstract stress values but modified to handle value differences between real data values as weighting operations. The weighted values percolate upwards, adjoining larger and larger units into a (not necessarily binary) timing tree. Four variants of the algorithm exist, and two were used in this study: TTI-A, grouping short-long, left-hand (short) value percolates up, TTI-B, grouping long-short, right-hand (long) value percolates up. Figure 2 shows a tree induced from the narrative. The smallest units are words whose durations are projected into a tree spanning the entire narrative, reflecting interesting divisions of the text which cannot be dealt with here. Figure 3 zooms into the tree, showing a syntax-timing correspondence, and bottom-up percolation, e.g. of the value .043 (zzz denotes a pause).

4 QUANTITATIVE EVALUATION

The evaluation strategy for determining the predictive value of grammatical information is purely structural, and does not use named categories, unlike Wagner's approach. In order to avoid the twin traps of theoretical and personal prejudice in automatic parsing, the syntax trees were obtained by dividing a narrative into a set of 20 consecutive sentences, and requesting six linguistically literate subjects to group expressions in the sentences by bracketing them. No attempt was made to ensure uniformity of bracketing. Some formally improper bracketings resulted, which were normalised by adding additional brackets left or right of the entire bracketed sentence. A total of 120 bracketings were elicited. Timing trees, also as unlabelled bracketings, were extracted from readings of these sentences by a different subject, and hand-annotated at word level. The timing and syntactic trees were then compared automatically, yielding the TSI. The decision on which kind of tree similarity measure to use is not trivial. A suitable measure is the number of subtrees spanning the same leaf sequence in each tree (in the present case, words), divided by the mean of the total numbers of nodes in the trees being compared. Summarising:

1. Compare tree pairs with identical leaf sequence spans; uniquely rename leaves.
2. Count subtrees with identical leaf sequence spans.
3. Calculate TSI as the number of matches divided by the average number of nodes in the trees; calculate mean TSI over all subjects and sentences.

The results of the study are visualised in Figure 4. The thick solid line shows correspondence between timing

ISBN 1-876346-nn-n © 2003 UAB

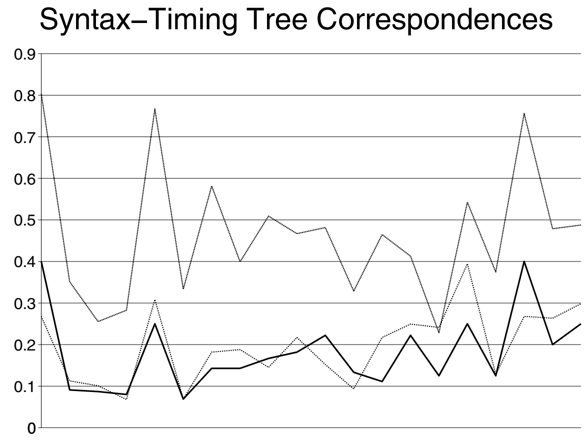

Figure 4: Syntax-prosody correspondences in read-aloud narrative (X: syntax/TTI tree pairs, Y: TSI).

trees and unparsed (UP) ssentences, the higher thin line shows mean TSI for TTI-A short-long (iambic) grouped trees, the lower thin line shows mean TSI for TTI-B long-short (trochaic) grouped trees. Both TTI-A (0.85) and TTI-B (0.89) TSI sequences correlate highly with the UP sequence, probably due to shallow bracketing, but the TSI levels differ considerably. Averaged over all subjects and sentences: TTI-A mean TSI = 0.47; TTI-B mean TSI = 0.2; UP condition: mean TSI = 0.19. Clearly, mean TSI for A (short-long) is much higher than for B (long-short) or UP, which are indistinguishable. Syntax and TTI trees are thus more similar under TTI-A than under TTI-B. The methodological orientation of the study and the number of subjects do not currently justify further statistical evaluation.

5 TOWARD AN INTEGRATED RHYTHM TIMING MODEL

The visualisation shows a preference for a *match between grammatical structures and iambic groups*, with short-long constituent pairs. An interesting result: the structure is like the end-weighted (iambic) Nuclear Stress Rule, not the trochaic structures often proposed for English rhythm. A number of points remain open: generalisation to other speech genres, deeper bracketing, normalisation for sentence length effects, use of a broader selection of subjects, statistical treatment. This research programme is facilitated by the non-language-specific TTI and TSI algorithms, and an implementation for arbitrary time-annoted data.

Nevertheless, the results are encouraging, and suggest that TTI and TSI could form the core of a prosodic data mining strategy for utilising the enormous quantities of annotated speech resources amassed in European and national projects, for instance in training hierarchical duration models for speech synthesis.

REFERENCES

[1] Ee Ling Low, Esther Grabe, and Francis Nolan, "Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English," *Language and Speech*, vol. 43, no. 4, pp. 377–401, 2000.

[2] Franck Ramus, Marina Nespor, and Jacques Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.

[3] Peter Roach, "On the distinction between 'stress-timed' and 'syllable-timed' languages," in *Linguistic Controversies: Essays in Linguistic Theory and Practice*, David Crystal, Ed., pp. 73–79. Edward Arnold, London, 1982.

[4] Petra Wagner, "Rhythmic alternations in German read speech," in *Proceedings of Prosody 2000*, Poznan, 2001, pp. 237–245.

[5] Fred Cummins, "Speech rhythm and rhythmic taxonomy," in *Proceedings of Speech Prosody 2002*, Aix-en-Provence, 2002, pp. 121–126.

[6] Ipke Wachsmuth, "Communicative rhythm in gesture and speech," in *Language, Vision and Music*, Paul McKevitt, Conn Mulvihill, and Sean O'Nuallain, Eds., pp. 117–132. John Benjamin, Amsterdam, 2002.

[7] Janet Pierrehumbert, *The Phonology and Phonetics of English Intonation*, Ph.D. thesis, M.I.T., 1988.

[8] Dafydd Gibbon, "Finite state processing of tone languages," in *Proceedings of EACL 3*, Copenhagen, 1987, pp. 291–297.

[9] Mark Liberman and Alan Prince, "On stress and linguistic rhythm," *Linguistic Inquiry*, vol. 8, pp. 249–336, 1977.

[10] Elizabeth Selkirk, *Phonology and Syntax. The Relation between Sound and Structure*, Cambridge University Press, Cambridge, 1984.

[11] Katarzyna Dziubalska-Kołaczyk, *Beats-and-Binding Phonology*, Peter Lang, Frankfurt, 2002.

[12] Leo Wetzels, "Comments on Low and Grabe," in *Laboratory Phonology*, Carlos Gussenhoven and Natasha Warner, Eds. Mouton de Gruyter, Berlin, 2002.

[13] Ulrike Gut, Sandrine Adouakou, Eno-Abasi Urua, and Dafydd Gibbon, "Rhythm in West African tone languages: a study of Ibibio, Anyi and Ega," in *Proceedings of "Typology of African Prosodic Systems 2001" (TAPS)*, 2001, pp. 159–165.

[14] Nick Campbell, *Multi-level timing in speech*, Ph.D. thesis, University of Sussex, 1992.