Corpus-based syntax-prosody tree matching

Dafydd Gibbon

Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld, Europe
gibbon@spectrum.uni-bielefeld.de

Abstract

Empirical study of the syntax-prosody relation is hampered by the fact that current prosodic models are essentially linear, while syntactic structure is hierarchical. The present contribution describes a syntax-prosody comparison heuristic based on two new algorithms: Time Tree Induction, TTI, for building a prosodic treebank from time-annotated speech data, and Tree Similarity Indexing, TSI) for comparing syntactic trees with the prosodic trees. Two parametrisations of the TTI algorithm, for different tree branching conditions, are applied to sentences taken from a read-aloud narrative, and compared with parses of the same sentences, using the TSI. In addition, null-hypotheses in the form of flat bracketing of the sentences are compared. A preference for iambic (heavy rightmost branch) grouping is found. The resulting quantitative evidence for syntax-prosody relations has applications in speech genre characterisation and in duration models for speech synthesis.

1. Hierarchical syntax, linear prosody?

The objective of this contribution is to provide a well-defined algorithmic approach to extracting complex prosodic information from speech corpora.[1]

Current empirical models of speech timing are based on a variety of algorithms, from single indices for timing patterns [1, 2] in psycholinguistics and phonetics to, in the speech synthesis domain, sum-of-products, CART and Bayesian classification approaches [3, 4], including models which use grammatical information. Campbell [5] has a model based on a strictly layered hierarchy, but in general duration models are linear, and hold for flat strings of words or syntactic categories. When syntagmatic grammatical information is used as a predictor for hierarchical structuring, in general the information used is also linear, based on strings of paradigmatic part-of-speech (POS) classes (grammatical categories) which provide weight factors for duration models. Wagner [6] uses a linear model for German speech synthesis based on five weighted POS sets. Grammatical categories imply at least local "hidden hierarchies", of course. Rarely, explicit hierarchical approaches have been used [7], and detailed approaches to the partially hierarchical description of timing are once more becoming available [8], [9, 10, 11];

But there is currently no technique available for data-driven investigation of more complex hierarchical duration models for syntagmatic prosodic relations, and the issue is not addressed in recent authoritative literature [12]. Classification methods need to include complex hierarchical timing information in addition to other to other phonetic and lexical properties of speech units. Further, it is a well-known phonostylistic effect that speech timing relations vary in highly complex ways depending on speech genre, including so-called *fast speech phenomena* [13]. Finally, other discourse factors such as focus and emotion are thought to affecting prosody, and thereby reducing the determining role of phrasal syntax, though these effects are currently not well understood (but see [14]).

[1]Thanks to Grazyna Demenko, Katarzyna Dziubalska-Kołaczyk, Ekaterina Iassinskaia, Peter Ladkin, Zofia Malisz and lecture audiences in Dublin, Bielefeld, Poznań and Tübingen for discussion and to Ulrike Gut, Katrin Johanning, Sara Johannsen, Josef Raab, Alexandra Thies, Thorsten Trippel for contributing data. The software developed for this work is in the public domain (GPL).

2. Linear timing measures

One set of approaches to investigating syntagmatic properties of timing is found in phonetic analyses of isochrony in syllable and foot timing.

In [15], tone unit duration is divided by the number of feet in the tone unit, yielding average or "ideal" isochronous foot duration, and normalised deviation from mean foot length is measured. 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. 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 isochrony it is (going beyond Roach's stated goals, of course) not a sufficient criterion for an adequate timing model.

Ramus, Nespor & Mehler [2] locate different languages in a timing space with the following parameters: $V\%$, percentage of vocalic intervals relative to overall utterance length; ΔC, variance of consonantal intervals; ΔV, variance of vocalic intervals. 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, though it is claimed to be a model of rhythm. Cummins has pointed out [9] that the model makes a statement about the evenness of the phonotactics of the language, rather than timing. The model possibly reflects necessary conditions on timing, 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.):

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

The model yields a minimal value of 0 (perfect isochrony), asymptotically approaching 200 for larger length differences. the variant used in [16] reverses the scale, and has a maximum of 100 for perfect isochrony.

The model has an empirical problem: PVI assumes *strictly binary alternation*. Hence, alternations as in "*Little John met Robin Hood and so the merrie men were born.*" are adequately modelled, but not unary rhythm (syllable timing) as in "*This one big fat bear swam fast near Jane's boat.*" or ternary dactylic and anapaestic rhythms (or those with even higher cardinality) as in "*Jonathan Appleby wandered around with a tune on his lips and saw Jenni fer Middleton playing a xylophone down on the market-place.*"

The model has worse a formal problem: the PVI is ambiguous and yields the same value for sets of alternating patterns, for monotonic geometrical series, and for mixtures of these, as shown by the following alternating and exponential series: $PVI(2,4,2,4,2,4) = PVI(2,4,8,16,32,64)$. A series of length n yields $n!$ patterns with identical PVI, obviously not the required result. So the model presupposes alternating input, and since this will not generally be the case it is not at all clear what the PVI is actually an index of.

3. Procedure

The empirical approaches examined in Section 2 are linear: the timing relations defined in the formulae are either *global*, and hold for arbitrary linear re-orderings, or *local*, and do not take global structures into account. One way of taking both types of property into account is to assume that timing is hierarchically structured, and to induce syntagmatic tree structures over the time-annotated sequence.

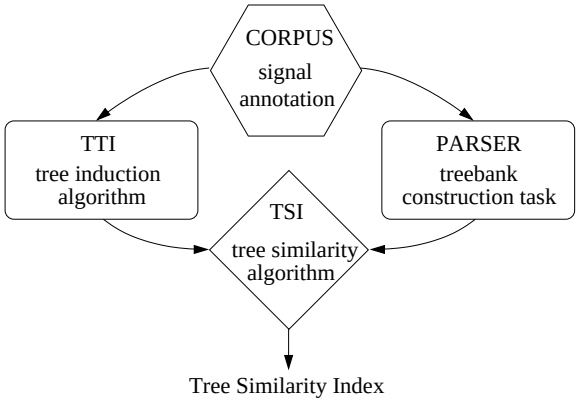

Figure 1: Corpus-based tree induction and comparison architecture.)

Hierarchical timing patterns would already be a useful source of information, but the timing trees still need to be related to other levels of description, in particular to hierarchical grammatical structure, in order to provide both useful and theoretically significant information about language processing. The kind of grammatical information required is of two kinds: first, purely structural, i.e. the hierarchy proper, represented by an unlabelled tree graph or bracketing; second, categorial, i.e. labels on the nodes or edges of the tree graph or in the bracketing, or an attribute-value structure. At the present stage, only the first goal, induction of unlabelled tree graphs, is pursued.

Timing, including the rhythmic factor, is a complex function of hierarchical and linear structuring, as already noted, and is combined here with local alternation criteria and with grammatical predictors for timing trees:

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

4. Parsing

The evaluation strategy for determining the predictive value of grammatical information is purely hierarchical, and does not use named categories, unlike Wagner [6]. In order to avoid the twin traps of theoretical and personal prejudice in automatic parsing, the unlabelled 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 (*subjective parsing*). A typical parse result is the following:

```
( ( there is
    ( nothing I
      ( can do ) ) )
  ( ( said
      ( the frog ) )
    and hopped away ) )
```

Deliberately, no attempt was made to ensure uniformity or theoretical consistency 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.

In large scale application, the unlabelled bracketed are taken from automatically constructed treebanks; however, cross-checking with the subjective parsing method seems desirable in order to have at least some operational criterion for naturalness during the development stage.

5. Time tree induction (TTI)

Timing trees, also as unlabelled bracketings, were extracted from readings of these sentences by a different subject, and hand-annotated at word level. The handmade annotations have tabular structure (in this case in esps/waves+ format):

```
42.799104  123
42.896017  123 there
42.977461  123 is
43.170525  123 nothing
43.336955  123 I
43.506263  123 can
43.730879  123 do
43.950073  123
44.116510  123 said
44.187593  123 the
44.534352  123 frog
44.976206  123
45.051352  123 and
45.286240  123 hopped
45.549465  123 away
46.708926  123
```

The TTI algorithm compares the durations of neighbouring items (in the present case, words), and groups sequences of monotonically increasing (or, in another variant, decreasing) durations together into a (not necessarily binary) local tree (represented by a branching node in a tree graph), and the first

(in another variant, the last) value percolates up to become the node's durational value. This value is used recursively to build larger trees until the entire sequence has been mapped into a tree. Formally, the TTI algorithm is a modification of the inverse of the Nuclear Stress Rule of Generative Phonology, though it handles real timing values, not abstract stress numbers. The algorithm will be described in detail elsewhere. Four variants of the algorithm exist, of which two are 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. In this study, A and B conditions were used; the implementation will be described elsewhere. The output of the TTI-A algorithm for the annotation file is:

```
(.071
 (.081
  (.097 "there:42.896")
  (.081 (.081 (.081 "is:42.977")
        (.193 "nothing:43.171"))
        (.166 (.166 "I:43.337")
              (.169 (.169 "can:43.506")
                    (.225 "do:43.731")))
        (.166 "said:44.117")))
 (.071
  (.071 (.071 "the:44.188")
        (.347 "frog:44.534"))
  (.075 (.075 "and:45.051")
        (.235 (.235 "hopped:45.286")
              (.263 "away:45.549")))))
```

The numerical labels following the left parentheses show durations; those following the colons are annotation timestamps. The bracketting illustrates numerical value percolation from the leaves to the root. The temporal labels output are filtered out of the tree before passing it to the TSI algorithm:

```
(
 (
  ( there )
  ( ( ( is ) ( nothing ))
    ( ( I ) ( ( can ) ( do )))
  ( said )))
 (
  ( ( the ) ( frog ))
  ( ( and )
    ( ( hopped ) ( away )))))
)
```

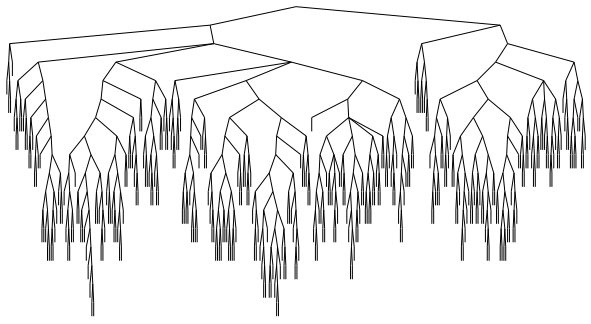

Figure 2: TTI tree over word durations in a narrative.

Figure 2 shows a tree induced from the whole narrative. The durations of the smallest units (words) are projected into a tree spanning the entire narrative. Figure 3 zooms into the tree, showing a syntax-timing correspondence (ZZZ denotes a pause) and bottom-up duration percolation (cf. the value .043).

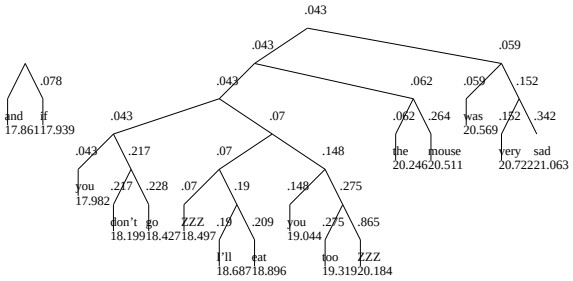

Figure 3: Zoom into the TTI tree.

6. Tree similarity indexing (TSI)

The task for the second stage of the procedure is to compare syntactically parsed trees with the duration trees and calculate an index of similarity. It is not immediately obvious how to do this, as trees have many properties which could be used as sources of criteria: number of nodes, number of edges, branching factor (binary or n-ary), branching tendency (right vs. left vs. centre branching), homomorphism or strict isomorphism.

Table 1: TSI algorithm as Scheme code.

```
(define (treecomp t1 t2 n)
 (if (pair? t1)
  (if (pair?  (car t1))
   (begin
    (treecomp-1 (leaves (car t1)) t2 (+ 1 n))
    (treecomp (car t1) t2 (+ 1 n))
    (treecomp (cdr t1) t2 n))
   (treecomp (cdr t1) t2 n))))

(define (treecomp-1 ll t2 n)
 (if (pair? t2)
  (if (pair? (car t2))
   (begin
    (if (equal? ll (leaves (car t2)))
     (set! *count-sim* (+ 1 *count-sim*)))
    (treecomp-1 ll (car t2) (+ 1 n))
    (treecomp-1 ll (cdr t2) n))
  (treecomp-1 ll (cdr t2) n))))

(define (leaves t)
 (if (pair? t)
  (append (leaves (car t)) (leaves (cdr t)))
  (if (null? t)
   t
   (list t))))
```

The basic requirement has already been defined, however: comparison in respect of the way trees are used to represent syntagmatic structuring (parsing) of sentences and prosodic series. Consequently, a new but conceptually simply similarity measure was defined, based on the number of nodes in each tree which span the same substring of the annotated and parsed sequence, i.e. the same leaf node sequence. Each leaf is uniquely labelled before the algorithm is applied, and non-branching nodes are pruned. To derive an index of similarity (TSI, Tree Similarity Index) the number of shared nodes spanning the same substring is simply divided by the mean node count of the two trees:

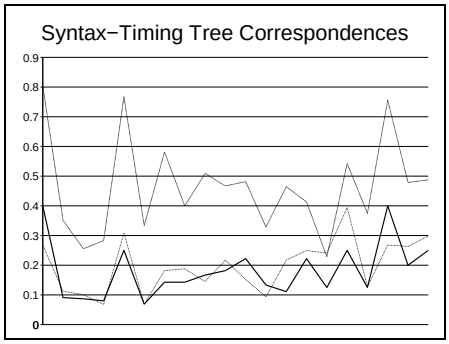

Figure 4: Syntax-prosody correspondences in read-aloud narrative (X: syntax/TTI t ree pairs, Y: TSI).

$$TSI = \frac{2 \times NC_{shared}}{NC_i + NC_j}$$

The recursive algorithm for calculating NC_{shared} climbs the trees, comparing pairwise the leaf sequences spanned by the nodes in each tree, and incrementing a counter if nodes share a leaf sequence. The algorithm is implemented in Scheme; the code itself is given in a portable style in Table 1. This implementation of the algorithm is overly greedy (nodes may be vacuously compared) but has the merit of simplicity.

7. Tentative results

The results of the study are visualised in Figure 4. The thick solid line shows correspondence between timing trees and unparsed (UP) sentences. For parsed (P) sentences, 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, maybe due to shallow bracketing. TSI levels differ considerably, however, as summarised in Table 2 (averaged over all subjects and sentences). The mean TSI for TTI-A trees is much higher than for TTI-B trees or UP strings which are indistinguishable. Syntax trees are thus more similar to iambic timing trees than to trochaic timing trees.

Table 2: Overview of main results.

	mean UP-correlation	*mean TSI*
P + TTI-A:	0.85	0.47
P + TTI-B:	0.89	0.2
UP + TTI-A:		0.19
UP + TTI-B:		0.19

8. Summary and prospects

The results show a preference for a *match between grammatical structures and iambic groups*, with short-long constituent pairs, indicating that the measure provides substantive and relevant information related to patterns (such as the iambic Nuclear Stress Rule) which figures in traditional descriptions of the intonation of West Germanic languages. Work in progress includes: generalisation to other speech genres and languages, deeper bracketing, weighting of categories, normalisation for sentence length effects, size of subject set, use of treebanks, full statistical treatment. The available software is suitable for application in larger scale applications, and these questions are currently being addressed in cooperation with specialists in a number of European and African languages.

9. References

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

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

[3] J. P. H. van Santen, "Assignment of segmental duration in text-to-speech synthesis," *Computer Speech and Language*, vol. 8, no. 3, pp. 95–128, 1994.

[4] O. Goubanova and P. Taylor, "Bayesian belief networks for model duration in text-to-speech systems," in *CD-ROM Proceedings of ICSLP2000, Beijing*, 2000.

[5] N. Campbell, "Multi-level timing in speech," Ph.D. dissertation, University of Sussex, 1992.

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

[7] K. Alter, J. Matiasek, and G. Niklfeld, "VIECTOS: The Vienna Concept-to-Speech System," in *Natural Language Processing and Speech Technology*, D. Gibbon, Ed. Berlin: Mouton de Gruyter, 1996, pp. 156–165.

[8] W. Jassem, D. R. Hill, and I. H. Witten, "Isochrony in English speech: its statistical validity and linguistic relevance." in *Intonation, Accent and Rhythm: Studies in Discourse Phonology.*, D. Gibbon and H. Richter, Eds. Berlin", year = 1984,: Walter de Gruyter, pp. 203–205.

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

[10] D. Gibbon, "Measuring speech rhythm in varieties of English," in *Proceedings of EUROSPEECH 2001*, 2001, pp. 91–94.

[11] I. Wachsmuth, "Communicative rhythm in gesture and speech," in *Language, Vision and Music*, P. McKevitt, C. Mulvihill, and S. O'Nuallain, Eds. Amsterdam: John Benjamin, 2002, pp. 117–132.

[12] R. I. Damper, Ed., *Data-driven Techniques in Speech Synthesis*. Boston: Kluwer Academic Publishers, 2001.

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

[14] S. Mozziconacci, *Speech Variability and Emotion: Production and Perception*. Eindhoven: Technische Universiteit Eindhoven, 1998.

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

[16] U. Gut, S. Adouakou, E.-A. Urua, and D. 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.