Dafydd Gibbon (Universität Bielefeld)*

Time Types and Time Trees: prosodic mining and alignment of temporally annotated data

1 Prosodic Annotation Mining: Objectives and Motivation

The objectives of the present contribution are theoretical, empirical and strategic:

- to outline a data–driven empirical method for exploiting the temporal properties of annotated speech data;
- to apply this method to the automatic construction of hierarchical models of speech timing (Time Trees) in the context of a Rhythm Periodicity Model (RPM) of rhythm;
- to demonstrate the potential of the methodology of computational phonetics as an empirical interface discipline between computational linguistics and phonetics by relating the Time Trees to grammatical hierarchies.

The contribution is structured as follows. The objectives and motivation for the prosodic data–mining of temporal annotations and the data–driven methodology are discussed in Section 1, followed by the introduction of Time Type Theory and relevant terminological clarification in Section 2. Representative data–driven Global and Local Linear Models are discussed in some detail in Section 3, and theory–driven hierarchical models are outlined in Section 4. In the following sections, a proof–of–concept description of the data–driven Time Tree Induction

* Thanks to Steven Bird, Doris Bleiching, Julie Carson-Berndsen, Nick Campbell, Fred Cummins, Kasia Dziubalska, Flaviane Romani Fernandes, Antonio Galves, Charlotte Galves, Esther Grabe, Ulrike Gut, Katja Jasinskaja, Klaus Kohler, Andras Kornai, Bob Ladd, Daniel Hirst, Peter Ladkin, Zofia Malisz, Francis Nolan, Franck Ramus, Alexandra Thies, Thorsten Trippel, Hans Tillmann, Ipke Wachsmuth, Petra Wagner, Rüdiger Weingarten and John Wells, to whom I am indebted for many discussions on related topics over a period of decades. I dedicate this study to my dear friend and esteemed colleague, Wiktor Jassem, for his pioneering contributions to the field for more than half a century.

(TTI) computational phonetic method for inducing tree–structures from temporally annotated data is given: in Section 5 the TTI approach is introduced, and the algorithm is presented; Section 5.3 describes a quantitative method for comparing prosodic time trees and grammatical phrase trees, and an appropriate Tree Similarity Index (TSI) alignment algorithm is given. Finally, the outlook for empirical contributions by computational phonetics such as the TTI–TSI induction and alignment method to the integration of phonetic and linguistic knowledge is considered.

1.1 Motivation

The reason for starting with the induction of temporal information is that temporal event sequence and overlap are the fundamental structuring principles of speech, define the domain into which phonetic interpretation maps syntagmatic phonological and other structures, and are the prerequisite for accurate time modelling in speech technology (particularly in speech synthesis, somewhat less in current methods of speech recognition).

Phonetic events, including prosodic events, are time functions; their phonological representations are prosodic, distinctive and conditioned features. The time functions are defined over temporal domains of different characteristic durations and are associated with different 'clock' frequencies in speech production and perception (cf. Tillmann and Mansell 1980). Levels in the discretely structured prosodic hierarchy (e.g. Selkirk 1984) from phones to discourse units can be phonetically interpreted in terms of such domains.

The phonetic time functions are transformations of the speech signal, and are simple (e.g. a pitch or vowel formant target) or complex in terms of sequential temporal trajectories (e.g. pitch contours, hierarchically larger units) or of overlap (e.g. co–functioning, partially simultaneous phonetic features). In phonology, complex trajectories are represented by the abstract concatenation operation, complex overlap is represented by feature bundles and autosegmental lattices. For foundational discussion of prosodic overlap issues such as temporal displacement, see Niebuhr and Kohler (2004) at the phonetic level, Clements and Ford (1979) at the phonological level, Bird and Klein (1990) for an explication of the issues in Event Phonology, and Carson-Berndsen (1998) for a computational phonetic approach, Time Map Theory (TMT), which interrelates the two levels and demonstrates a proof–of–concept application of TMT to speech recognition.

In the present study (Sections 5 to 5.3), words are chosen as the basic event type for temporal induction, mainly on the heuristic grounds that they constitute an ontologically basic linguistic rank and are a defining category for phonology and prosody, that they are small enough to provide enough data, that they are large and functionally clear enough to be immediately identified by labellers, and that they have fairly clear properties in both grammar and phonetics (and semantics, though this is not the concern of the present contribution). Words are thus well–suited to the present proof–of–concept study within a new compu-

tational phonetic methodological framework, in preparation for more extensive study of other, more fine–grained phonetic dimensions and feature spaces.

1.2 Data–Driven vs. Theory–Driven Modelling

A striking contrast between phonetic and phonological approaches to temporal modelling is that phonetic methods have generally been data–driven, and have resulted in linear models of timing in speech production, while phonological approaches are theory–driven, and have resulted in hierarchical models of timing. Temporal models for Text–To–Speech (TTS) synthesis have also often borrowed from phonology, are then theory–driven and, typically, hierarchical, though contemporary statistical unit–selection methods are clearly data–driven.

The present contribution tries to bridge this gap by refining phonetic data–driven methods to yield hierarchical rather than linear time models, in order to provide a *tertium comparationis* for both theory–driven and data–driven models. The kind of hierarchy addressed here is not the Classification and Regression Tree (CART) kind of classificatory or paradigmatic hierarchy which has often been used in time modelling. The present method is concerned with compositional, syntagmatic hierarchies over sequential and overlapping events.

The data–driven perspective is enhanced by the availability of large quantities of temporally annotated high quality speech data, both audio and visual, and by the relative ease of creating more of the same. Data of this quantity and quality puts empirical computational linguistic and phonetic methodologies potentially on a comparable footing with the natural and technological sciences. In fact, temporal annotation methodology, the basis of the broader 'Language and Speech Resources' paradigm,[1] originated in statistical methods in Speech Technology for Automatic Speech Recognition (ASR), later in unit-selection based TTS synthesis.

Resources are there to be used. The information in the annotations is *temporal*, in that event patterns are paired with time–stamped intervals, *categorial*, in that boundaries are discrete and have a symbolic rather than numeric basis, and *syntagmatic*, in that annotations define the criteria for the sequence and overlap based composition of complex events. This kind of information is optimal for the computational phonetic analysis of time structure induction and alignment which is reported in the present contribution.

[1] In the European context, most notably in the following consortia: *Speech Assessment Methodologies* (SAM), *Expert Advisory Groups for Endangered Language Systems* (EAGLES), *International Standards for Language Engineering* (ISLE), and the MATE (Multilevel Annotation Tools Engineering) project consortia from the late 1980s to the early 2000s, with publication outlets in the *Language Resources and Evaluation Conference* (LREC) series, and the new *Language Resources and Documentation* online journal.

2 Time, Time Type Theory and Event Alignment

Temporal structure is a traditional topic in philosophy as well as in linguistics and phonetics. Objective and subjective, rationalist and empiricist notions of the ontology of time are indirectly the concern of this contribution, in so far as different levels of temporal organisation are aligned with each other. In linguistics, most attention has been paid to time in semantics. Unsurprisingly, formal explications of time and event structure in terms of event logics and temporal calculi are valid in both semantic interpretation and phonetic interpretation, and the present contribution indeed uses concepts traditionally found in semantics.

2.1 Time in Phonetics

Studies of timing in phonetics, psychoacoustics, psycholinguistics and speech technology are too numerous to cite extensively here. They have covered many topics, of which the following are characteristic examples:

- correlates of rhythm patterns in terms of pitch, intensity and unit duration (syllable, consonant cluster, vowel, mora), addressing issues such as unit isochrony in which $duration(u_i) = duration(u_{i-1}) + N$, where N is an interval constant;
- partial alignment of parallel information streams in physical acoustic time;
- graph representations of word hypotheses in automatic speech recognition;
- psychological correlates of time in speech production and perception, including notions of subjective time;
- sequential and parallel ordering of units in time in prosodic phonologies;
- abstract notions of timeless structure, in which temporal sequentiality is represented by a general concatenation operation which is neutral with respect to temporal ordering in speech and handwriting processes, or to spatial ordering in printed and handwritten texts;
- iconic relations between temporal event ordering in speech and temporal event ordering in conventional semantic interpretation.

In phonetics, the issue of timing has long centred on the typological distinction of Pike (1945) between stress–timed and syllable–timed languages, on Abercrombie's notion of the foot in stress–timed languages (cf. Abercrombie 1967), and less frequently on more complex concepts of rhythm structure (cf. Jassem 1951, Jassem and Gibbon 1980, Jassem et al. 1984) which include foot–like structures, the pre–foot anacrusis constituent, and hierarchical structuring.

In phonology, structural and functional criteria are central, and the focus has been on abstract relational timing concepts such as contrastive and conditioned segment duration, complex temporal organisation in the prosodic hierarchy and in metrical grids, and on phonostylistic concepts such as tempo, pause and fast speech phenomena, which have also been taken up in many discourse analytic studies.

The old distinction between objective and subjective concepts of time is relevant here, as already noted. The analogue to the objective–subjective distinction in the study of speech is physical clock–measured time in data–driven phonetics, versus subjective perception–based time identification in interpretative phonetics, phonology and discourse analysis.

In the present contribution the topic therefore needs to be carefully delimited within the terms of reference of the linguist and phonetician, to concepts of time which are expressed or implied by models of language and speech. On the formal side, the linear, parallel and hierarchical structures of temporal domains are focussed. On the empirical side, the focus is on corpus phonetics, in particular on temporal annotations of the acoustic speech signal (whether manual, semi-automatic or automatic). The models discussed here are 'cognitively neutral': the data are speech productions, but the models represented by the algorithms are not claimed to be production models; this interpretation is left open. As inductive computational and data–driven models, the algorithms could conceivably be interpreted as emulations of temporal dimensions of speech perception, learning and information alignment, but further cognitive, linguistic phonetic modelling constraints are required in order for this to be plausible.

A systematic terminological clarification in terms of Time Type Theory and event–based patterning will be given below, as a basis for the computationally oriented further discussion, and as a contribution to the ontology of temporal concepts in phonetics.[2]

2.2 Time Type Theory

Three Time Types are needed as a basis for prosodic event alignment in the present analysis (cf. Gibbon 1992 and Carson-Berndsen 1998):

Absolute Time relates to signal–oriented phonetics, that is, to time points and intervals determined by calibrated physical measurement. For example, standard digital signal sampling techniques generate Absolute Time structures. In the Absolute Time domain, the quantitatively measured lengths of phones, syllables, etc., are important. Impressionistic phonetic judgments

[2] The term 'ontology' is used here as in Artificial Intelligence and Text Technology with respect to search strategy resources (cf. the 'semantic web'). It refers to the organisation of terms or concepts and their definitions in hierarchies and other complex structures, and is related to taxonomic lexical semantics and terminology theory.

on length and tempo, as practised in phonostylistics and discourse analysis, may be seen as coarse–grained and uncalibrated quantitative measures.

Relative Time relates to 'interpretative phonetics', phonology and prosody, and defines intervals and other relations between points in time with no explicit assignment to Absolute Time. Relative Time characterises the prosodic phonologies; the key relations are sequence, overlap and hierarchy, which are interpretable in terms of the Absolute Time domain.

Categorial Time relates to underlying lexical and grammatical levels, in particular to categories linked by algebraic operations such as concatenation. In the Categorial Time domain, there is only a notion of temporally uninterpreted structure; to include a notion of time, phonetic interpretations into the Relative Time and Absolute Time domains are required.

The three–level distinction between Time Types is supported by work in formal linguistic theory, in particular in Event Phonology (cf. Bird and Klein 1990; Kornai 1991), in Time Type Theory (cf. Gibbon 1992) and in the Time Map Phonology approach to alignment theory (cf. Carson-Berndsen 1998).

The key data–mining procedures in the exploitation of temporal annotations for prosodic induction and alignment can now be formulated in terms of the Time Types:

1. Analogue–digital transformation in the signal sampling process, between two subdomains of Absolute Time.
2. Annotation as mapping the quasi–continuous digital domain of speech signals into the discrete Absolute Time domain of annotation intervals.
3. Induction of temporal structures from the discrete Absolute Time subdomain of annotation intervals to linear and hierarchical Relative Time structures.
4. Mapping of Relative Time structures to Categorial Time grammatical and discourse patterns.

2.3 Event Alignment: Streams, Tracks, Tiers

Each of the three Time Types is associated with its own specific range of sequential and partially aligned parallel structures at different theoretical and heuristic levels of description. The relevant levels for the present study are distinguished as follows:

1. a set of parallel signal *streams* (time functions describing continuous or discrete sampled speech signals),

2. partially aligned with a set of parallel annotation *tracks* (time functions describing discrete, categorial sequences of events, as in a speech editor, for example, with sampled speech signal and parallel annotation tracks),

3. which are often derived from specific phonological *tiers* (linguistic constructs defining partially aligned trajectories through a feature space in Relative Time, as in autosegmental and other prosodic phonologies).

The 'stream–track–tier' terminology is intended to keep apart clean ontological levels which are often indiscriminately labelled with terms like 'tier', 'track', 'level', 'layer', 'stratum', 'stream'.

The following more detailed terminological overview is based largely on the related models of Event Phonology (cf. Bird and Klein 1990), Time–Map Phonology (cf. Carson-Berndsen 1998), and Annotation Graph theory (cf. Bird and Liberman 2001).

Event: A pair of an Absolute Time or Relative Time *interval* and a *pattern* of values in some phonetic dimension, parameter or feature. Examples:

- an interval of 120 ms and a phone segment, as a static or a dynamic time function in Absolute Time on an annotation track;
- an interval of 10 ms and a pitch value in an Absolute Time F_0 stream;
- an interval of 0.0208333 ms (corresponding to 48 kHz sampling rate) paired with an amplitude value in an Absolute Time signal stream;
- a pair of a phonological segment and its phonemic or feature–based properties in Relative Time.
- a signal annotation $<< x_{min}, x_{max} >, transcription >$, where x_{min} and x_{max} range over points, $transcription$ ranges over textual symbols, $< x_{min}, x_{min} >$ ranges over intervals ($(x_{max} - x_{min})$ ranges over durations); cf. the following (Brazilian Portuguese) syllable annotation:

```
xmin = 0.48473069812858305
xmax = 0.6301876830002222
text = "koN"
```

Transcription: The name of the pattern of an event.

Annotation: The name of an Absolute Time event, consisting of a set of pairs of transcriptions and either interval time–stamp pairs or point time–stamps.

Point: The undefined primitive for defining intervals as a pair of points (whether abstract points as in Relative Time, or clock time points as in Absolute Time), ignoring for present purposes the traditional discussion on whether points or intervals are primitives.

Time–stamp: The name of a point or a pair of points in Absolute Time, i.e. a calibrated quantitative designation of a relative to some pre–defined initial point (the term 'tick' is used in digital music and virtual machine technology). Examples:

- 24th December 1976
- Mon Mar 28 13:32:30 BST 2005
- 321.5 ms

Absolute interval: The difference between two time points. Examples:

- $0.145457 = 0.6301876830002222 - 0.48473069812858305$
- the time elapsed between two metronome beats.

Relative interval: A segment at an abstract phonological level, related to other intervals by relations of precedence and overlap. A relative interval has no absolute duration unless explicitly mapped into an absolute interval. Example:

- The epenthetic [t] in English [prints] "prince" arises when the end of the nasal event interval of [n] precedes the end of the occlusive event interval of [n].

Time–Map: A function within one Time Type or between Time Types, mapping one temporal representation into another. Example:

- speech signal digitisation (analogue signal sampling),
- annotation (aligns digital speech signal with eventlabel sequence),
- phonetic interpretation (mapping of lexico–syntactic representation of speech forms into a phonetic representation.

3 Linear Timing Models

Linear models of timing generally relate to some characterisation of 'rhythm', as a sequencing of tendentially equi–temporal (isochronous) units of rhythmic temporal organisation, such as mora (a sub–syllabic timing unit), syllable, or foot (stress group). The present section addresses approaches of this type, and compares them with a design for a Rhythm Periodicity Model, RPM, defined as follows (modified from Gibbon and Gut 2001):

> *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.*

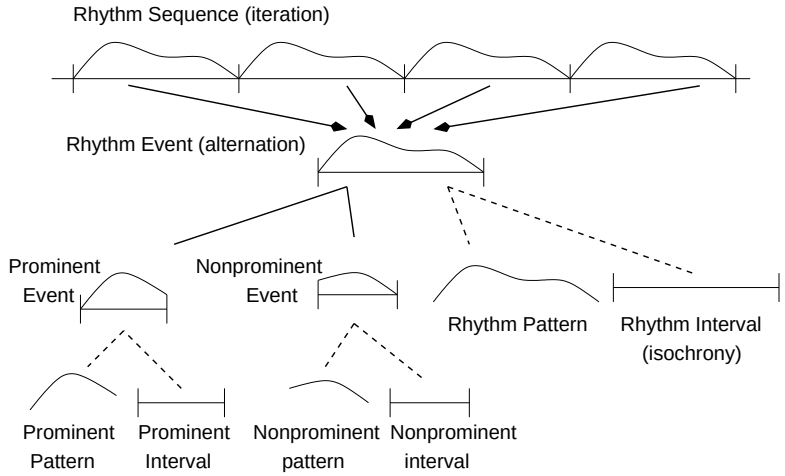

Figure 1: Decomposition of rhythmic temporal structure (arrows: generalisation of rhythm event from rhythm sequence; bold lines: sequence decomposition; dotted lines: overlap decomposition).

The condition 'constant duration' does not refer simply to Absolute Time constants: there are many factors which enter into judgments of constant duration at different levels of perception (for example, pattern similarities). Generic modelling conventions for rhythm structure based on this definition and the ontological clarification given previously are illustrated in Figure 1, which shows two levels of rhythm organisation: the focal and nonfocal structural components (traditionally: 'ictus' and 'remiss') each have internal alternating focal/non–focal structure. The structure shown in Figure 1 is syntagmatically decomposed along the two temporal dimensions of sequence and overlap.

Nothing is said in this definition about the epistemological status of rhythm as a complex emergent property of cognitive construction on the side of the listener and timing principles of speech production on the side of the speaker, or as a purely bottom–up physical pattern; cf. Gibbon and Fernandes (2005). It is unlikely to be just the latter. A strictly positivistic characterisation of rhythm in physical terms, often sought after in phonetic studies, is likely to fail: top–down factors, including grammatical and discourse patterns and cognitive expectations, play a significant role.

In the Rhythm Periodicity Model definition, three structural factors in the temporal organisation of rhythm are identified, and will be used as criteria for the adequacy of other rhythm models in subsequent discussion:

Pattern alternation: The internal focal–nonfocal rhythmic temporal pattern:

1. the time structure of the rhythm pattern can be a binary alternation,

or a more complex hierarchical rhythm, as found in metrical poetry, music, and formal speech, giving the impression of ternary or more complex rhythms;

2. non–semantic, structural terms 'focal' and 'nonfocal' applied to rhythm constituents are phonetically interpreted as sequential alternations of the following kinds (auditorily interpreted as prominence patterns):
   - Pitch: $[pitchpeak]^\frown[pitchtrough]$,
   - Syllable: $[long]^\frown[short]$
   - Segment: $[vowel]^\frown[consonant]$

   A focal component can occur anywhere within the rhythm pattern, not just initially or finally, but has to be consistently positioned in a given sequence.

Compositionality: The external rhythmic environment:

Compositional iteration: the regular directional periodic recurrence of the focal–nonfocal pattern within the rhythmic sequence domain as an iterated and alternating sequence (iteration is often represented as pure left or right recursion),

Compositional hierarchy: the recursive grouping of sequences of focal–nonfocal patterns (represented as mixed left, right and centre recursion).

Isochrony: The tendentially isochronous rhythmic domain, e.g. the syllable, the foot or some other unit (sometimes called 'rhythm unit', 'rhythm group').

The pattern alternation, compositional iteration and isochrony conditions together constitute the basic criterion of *periodicity* which characterises the Rhythm Periodicity Model.

Current Linear Models of rhythm timing in speech are quite diverse but at the same time tend to be atomistic and selective in that they focus on parameters as different as global deviation of unit length, local unit length ratios, and consonant-vowel ratios (cf. Roach 1982; Low et al. 2000; Ramus et al. 1999). These Linear Models are dicussed in more detail below. It will be demonstrated that while they address the isochrony condition, none covers all the necessary formal and empirical properties of rhythm, in particular in respect of the description of the pattern alternation and compositional iteration required for a Rhythm Periodicity Model, or of other aspects such as non-binary rhythms and rhythm hierarchies.

In the following subsections, formal and empirical aspects of linear time models are discussed. First, a formal characterisation of alternation and iteration in linear rhythm models is given in terms of Finite State Transducers (FSTs). Second, representative global and local linear modelling approaches are examined in terms of their formal and empirical modelling properties.

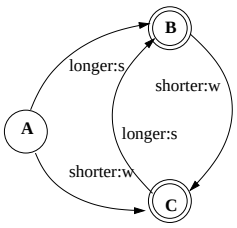

Figure 2: Strict binary rhythm acceptor.

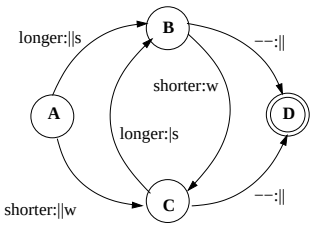

Figure 3: Rhythm acceptor with rhythm group boundaries.

3.1 Finite State Models of Linear Rhythm Organisation

The appropriate formal model for the alignment of linear structures involving alternation and iteration is the Finite State Transducer (FST), a translation device based on the Finite State Automaton (FSA); both are standard generic processing constructs in computational linguistics.[3] Finite State Transducers are used in the present study to define Time Map alignments between different levels of temporal organisation

Figure 2 shows a basic binary rhythm acceptor model; ‘s’ (strong) stands for a focal or prominent component of an alternation, ‘w’ (weak) for a nonfocal or nonprominent component. Rhythm group or foot boundaries are pre–defined according to the typological properties of the language as iambic or trochaic, and can be inserted automatically if required: cf. the pipe ‘|’ characters in Figure 3 for the trochaic case. The devices translate long–short sequences of units (syllables, feet, etc.) into string sets of the following types (the general case, the trochaic case, and the iambic case):

[3] The FSA represents the simplest form of grammar which can license infinite sets of finite strings over a finite vocabulary. FSAs permit iteration (equivalent to left or right recursion) but no centre–embedding, unlike phrase–structure grammars. Therefore they cannot model arbitrary tree structures (unless these are of finite depth, as in many prosodic hierarchy models). At most FSAs model flat structures (sometimes described with exclusively left or right branching trees) which are known to be sufficient for modelling phonological and morphological forms, and probably also prosodic forms.

An FSA network is a rooted directed cyclic or acyclic graph composed of nodes linked with transition arcs. The transition arcs are labelled with the elements of the vocabulary of which the strings to be accepted consist. The network topology defines the licensed set of strings; loops permit strings of arbitrary length and thus infinite sets of strings. The FST differs from the FSA in that transitions are labelled with symbol pairs, one symbol from an ‘input vocabulary’, one from an ‘output vocabulary’; the FST thereby not only accepts strings of symbols from the input vocabulary but also translates them into strings of symbols from the output vocabulary, an operation often used to model mapping operations expressed by phonological and tonological rules. FSTs are reversible; an FST can thus be used not only as a ‘parser’ but also as a ‘generator’.

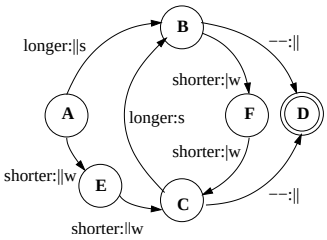

Figure 4: Strict ternary rhythm acceptor.

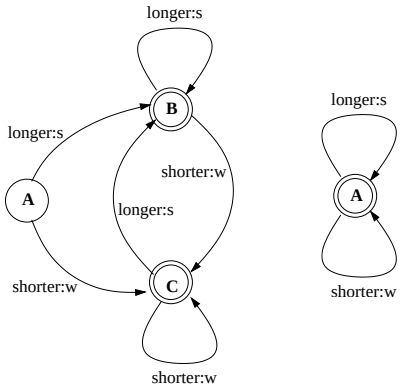

Figure 5: Naive and minimised general rhythm acceptors.

(1) $\{s, sw, sws, swsw, ...w, ws, wsw, wsws, ...\}$

(2) $\{||s||, ||sw||, ||sw|s||, ||sw|sw||, ...||w||, ||w|s||, ||w|sw||, ||w|sw|s||, ...\}$

(3) $\{||s||, ||s|w||, ||s|ws||, ||s|ws|w||, ...||w||, ||ws||, ||ws|w||, ||ws|ws||, ...\}$

The binary model is too strict: durations in real speech do not follow a simple long–short–long–short pattern; nor are they strictly linearly organised. A ternary model is shown in Figure 4, here the trochaic (actually dactylic) case. Both binary and ternary cases (at least) must be catered for; this is handled by the union operation over FSTs, which combines the ternary model with the binary model. Effectively, this just means adding the transitions $< B, w, C >$ and $< C, |s, B >$ to the ternary model. But the real–speech situation requires a more general automaton, which could be reduced to an equivalent single–state device, as in Figure 5. Clearly this is not the end of the story: further constraints on timing and structure are needed. The unreduced models provide contexts for these: binary, ternary structures and their time behaviour need to be modelled; so–called 'stress clash' contexts need to be captured by mapping to abstract underlying lexico–syntactic stress patterns, and (a more complex issue) hierarchical constraints for more complex models (cf. Campbell 1992; Cummins 2002; Wagner 2001). This is still an open issue, but the present discussion will serve as basic background for an evaluation of existing Linear Models of rhythmic timing.

3.2 Global and Local Linear Models of Rhythm

The Linear Models of rhythmic temporal structure can be classified as follows:

1. Global Linear Models (GLM), i.e. variance–based models of 'regularity' and 'irregularity' in sequences of unit durations, focussing on isochrony,
2. Local Linear Models (LLM):
   a) Oscillation and Entrainment Models: Cummins (2002), O'Dell and Nieminen (1999), Barbosa (2002), Wachsmuth (2002),
   b) Pairwise variability models: Low et al. (2000), Gibbon and Gut (2001).

The reasons for the classification as 'global' and 'local' models will appear in the discussion.

Formal analysis of recent approaches in the Oscillation and Entrainment paradigm of rhythm modelling (cf. Barbosa 2002; Cummins 2002; Wachsmuth 2002) requires modification of the basic FST mechanism. Work is in progress on this; the approaches will not be considered in detail here, except for a brief account of structural aspects of the Cummins model in Section 4.

3.2.1 Global Linear Models of Rhythm

One class of linear rhythm models is global, in that the local alternation property of rhythm is ignored, and general 'evenness' properties of events involving a particular parameter are characterised by means of variance measures.

The simple and precise method introduced by Roach (1982) is a Global Linear Model: the sum of deviations from average foot duration is divided by total utterance length, yielding normalised foot duration deviation. A similar measure would be Standard Deviation of foot duration. The PFD is a measure of evenness of duration, i.e. isochrony:

(4) $\textit{Mean Foot Length (MFL)} = \frac{\Sigma_{i=1}^{n} |foot_i|}{n}$

$\textit{Percentage Foot Deviation (PFD)} = 100 \times \frac{\Sigma_{i=1}^{n} |\textit{MFL} - len(foot_i)|}{n \times \textit{MFL}}$

Roach interpreted his results as showing that PFD is actually higher in English than in French, Telugu and Yoruba, i.e. in languages classified as syllable-timed by Abercrombie (1967), which is contradictory to expectations (assuming that the category 'foot' is somehow applicable to these languages). While the global evenness or isochrony criterion for rhythm is modelled in Roach's approach, pattern alternation, and compositional iteration and hierarchy are not. Further, any arbitrary re-sorting of the units (random, shortest-to-longest, etc.) would yield the same global index. The model is thus too unconstrained, in that it only defines one necessary condition for a Rhythm Periodicity Model, isochrony, and not the other necessary conditions.

A quantitative measure for 'rhythmic irregularity' is discussed by Scott et al. (1986). This is an open-ended, normalised measure which is calculated pairwise for all intervals I in a sequence, i.e. globally over the whole sequence:

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

The absolute value of the logarithm ensures that the correct ratio is found, independently of the order of division. The more similar the durations of units are, the closer the value approaches 0. Like Roach (1982), Scott et al. (1986) deal with the isochrony condition, but do not take pattern alternation or compositional iteration into account.

Studies by Ramus et al. (1999), Ramus (2002) use subsyllabic variables to locate different languages in a typological timing space: $V\%$, percentage of V (vocalic intervals) relative to overall utterance length, ΔC, variance of consonantal intervals, and ΔV. Like the single variable *PFD* and RIM measures, these two–variable measures reflect preferences for certain phonotactic patterns (CV, CVC, vowel length), though as corpus tokens rather than lexical types. The model uses a form of global evenness or isochrony criterion, and also ignores pattern alternation and compositional iteration: V stretches and C stretches would still yield the same results if randomly sorted. The ΔV measure reflects evenness of vowel sequence lengths, lower values tending to isochrony; similarly the ΔC measure for consonants.

To summarise: the approaches of Roach (1982), Scott et al. (1986), Ramus et al. (1999) and Ramus (2002) are 'complete' in that they capture evenness of temporal structures, but are 'unsound' in that they only refer to isochrony and capture many unwanted structures because of the neglect of focal/non–focal alternation and the directional iteration of alternations.

3.2.2 Local Linear Models of Rhythm

The global evenness issue was addressed by Low et al. (2000), who developed a Pairwise Variability Index (PVI) to take account of pattern alternation and compositional iteration in rhythmic temporal structure. The PVI is an averaged distance measure for adjacent units (vowels, syllables, etc.):

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

Differences between consecutive pairs of durations are normalised by the average duration of the pair, and absolute values of the normalised differences are averaged and multiplied by 100. Normalisation is intended to handle speech rate variation in the utterance and between utterances.[4] Like the RIM the PVI is lower when the durations of vowels in adjacent syllables are similar. The RIM is open ended as irregularity increases, but the PVI ranges between the limits of 0

[4] Note that the comment by Wetzels (2002) that the $(m-1)$ component factors out final lengthening is mistaken: a sequence of length m simply has $m-1$ differences between neighbours.

and 200.[5] For example, in the artificial 'ideal' case of perfect isochrony the PVI for the duration sequence <100,100,100,100> is 0; in the artificial and impossible case of total non-isochrony, the PVI for <100,0,100,0> is 200. In a realistic case, using Brazilian Portuguese syllables (not vowel lengths),[6] a time–stamp series of length 8 <0.084, 0.143, 0.357, 0.426, 0.588, 0.727, 0.889, 1.267> yields a duration (neighbouring time–stamp difference) series of length 7 <0.06, 0.21, 0.07, 0.161, 0.139, 0.162, 0.378>, which in turn yields a (rounded) PVI of 67 (at the syllable level), corresponding to an average neighbouring syllable duration ratio of 2:1 (i.e. moderately non–isochronous).

A variant used in Gut et al. (2001), the Rhythm Ratio (RR), reverses the scale, uses division rather than subtraction, and has a maximum of 100 for perfect isochrony, but otherwise has the same fundamental properties as the PVI:

The PVI and RR models have three serious inherent empirical problems:

1. Contrastive vowel length is ignored: *pretty Sally tickled Tim*, with short stressed vowels, may well behave very differently from *tiny Davey danced with Joan*, with long stressed vowels.

2. The pattern measured is vowel length differences, but Dauer (1983), Ramus et al. (1999) stress the role played by variation in consonantal duration; when Ramus (2002) compared his variance measure and the PVI for both C and V, he discovered that plotting $PVI(C) \times PVI(V)$ and $\Delta C \times \Delta V$ yields similar results. Clearly, other units need to be considered.

3. The PVI and RR models assume strictly binary rhythm, as in "*Little John met Robin Hood and so the merrie men were born.*". But while the focal—nonfocal property of rhythm is always binary, the nonfocal section may be internally more complex, and the PVI and RR do not handle this. Cf.

    a) unary rhythms (effectively: stress–based syllable timing with CV alternation), as in "*This one big fat bear swam fast near Jane's boat.*";

    b) ternary dactylic and anapaestic rhythms (or rhythms with even higher cardinality), as in "*Jonathan Appleby wandered around with a tune on his lips and saw Jennifer Middleton playing a xylophone down on the market-place.*"

[5] Case 1: The units in each pair have equal length. In this case, the difference between adjacent units is 0, the normalised difference is 0, the average multiplied by 100 is 0. This is the lower limit. Case 2: The units in each pair have very different length, with the length of one approaching zero to all intents and purposes, and the other being much longer. Then the difference will be approximately the same as the duration of the longer unit and the average duration will be approximately half this, so the normalised difference will be approximately 2 and the average multiplied by 100 asymptotically approaches 200.

[6] Data and annotations (speaker MC) due to Flaviane Fernandes, U Campinas, Brazil.

Worse, the PVI and RR models have two serious inherent formal problems. The distance measure is *undirected*, like the global measures, and does not distinguish between the 'greater than' and 'less than' orderings needed by the pattern alternation condition: PFD uses absolute differences, RIM uses the absolute log value of ratios, PVI uses the absolute value of differences, and the RR uses an explicit condition to merge the two relations. The formal consequence is that for an utterance containing n units, the PVI function generates by 2^{n-1} combinations of increasing and decreasing duration relations. PVI is therefore massively ambiguous. For example, for utterances of length 3 (i.e. 2 durations), the PVI function generates the same index for $2^2 = 4$ patterns:

(7) $pvi(2,4,2) = pvi(4,2,4) = pvi(2,4,8) = pvi(8,4,2) = 66.6667$

The Local Linear models are thus effectively just as unconstrained as the Global Linear Models. Though formally 'complete' in that they handle all the required patterns, the Linear Models discussed here are 'unsound' in that they produce spurious identical values for exponentially many alternating and non–alternating patterns. As with the Global Linear Models, the 'isochronous' end of the scale is meaningful, and the further away the indices get from isochrony, the less is known about what 'non–isochrony' means. The use of an undirected distance measure means, despite the original intention, that isochrony is addressed, but not pattern alternation.

When Ramus (2002) compared plots of $\Delta C \times \Delta V$ $PVI(C) \times PVI(V)$, very similar results for a range of very different languages appeared. This apparently provides mutual confirmation for consistency in both approaches. With hindsight the similarity is not too surprising: Ramus data are empirically controlled for speech rate, whereas the PVI model normalises locally for speech rate. The models are also conceptually similar: the variance model calculates (by definition) sums of squared duration differences from global duration means, while the PVI model calculates sums of local duration differences divided (normalised) by the local duration mean.

Summarising: Both Global Linear Models and Local Linear Models fail as models of rhythm, though they are valid as models of isochrony. The Local Linear Models additionally use the iteration property of rhythm, though they fail formally on the alternation property and basically reflect phonotactic structures; cf. also Cummins (2002).

Nevertheless, all of these results show that there are finely tunable computational measures which can be used to provide partial models of linear rhythmic temporal organisation in the Absolute Time domain. The Local Linear Models add a distance measure which normalises utterance–internal speech rate variation; on the other hand, speech rate variation is in itself an interesting factor in temporal analysis, in that it underlies fast speech phonostylistic phenomena which affect rhythm, as well as individual differences.

4 Higher Order Temporal Patterning

Non–phonetic approaches to rhythmic temporal organisation tend to be hierarchical, with the hierarchies derived directly or indirectly from grammatical structures. A large number of approaches to prosody in TTS synthesis also use explicit or implicit hierarchical models (cf. for example Campbell 1992; Wagner 2001). In Cummins (2002) a number of additional factors involved in the production of rhythm in different styles are discussed, 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.

The hierarchical approaches derive ultimately from the recursive approach to abstract stress hierarchy definition with the Nuclear Stress Rule (NSR) in Generative Phonology (cf. Chomsky and Halle 1968). From the start (cf. Bierwisch 1966) the problem of arbitrary recursive depth was addressed, and 'flattening' measures were introduced in the form of readjustment rules (cf. Culicover and Rochemont 1983). Metrical Phonology (cf. Liberman and Prince 1977; Selkirk 1984), one of the offspring of Generative Phonology, introduced a flattening component in the form of an iterative, alternating linear filter, the 'metrical grid', to readjust non–rhythmical 'stress clashes' in some contexts (cf. *thirTEEN* but *THIRteen MEN*). The metrical grid is closely related to the FST models introduced earlier, and explicates Relative Time pattern alternation and compositional iteration rhythm constraints, but not the isochrony constraint, which would require a Time Map into Absolute Time.

The NSR recursively defines a hierarchical iambic temporal structure; a similar rule for English nominal compounds defines a trochaic structure. The idea underlying NSR recursion is that there is an alternation or relative prominence relation between focal and non–focal constituents of a domain (the NSR says nothing about empirical correlates such as duration or pitch). The relative prominence relation holds recursively and compositionally at all levels of a hierarchy and is used to interpret syntactic structures and assign 'stress' values to words. The NSR will be discussed in some detail (despite its controversial status) because in Section 5 a generalised inverse NSR function will be used as the core of the induction algorithm.

It is the general recursive rather than simply iterative (either right or left recursive) compositionality property of the NSR function which distinguishes it from Local Linear Models: "the form of a complex expression is determined by a fixed set of processes that take account of the form of its parts" (Chomsky and Halle, 1968, 20). The same principle underlies compositional semantic interpretation, in which 'the meaning of the whole is a function of the meanings of the parts', (noted explicitly in Footnote 7 on the same page). Similarly, then, the pronunciation of the whole is a function of the pronunciations of the parts'.

Several equivalent formulations of the NSR have been proposed in the literature. A number are outlined here in order to make the point that 'there is more than one way to do it', the important issue being the underlying formal property of general recursion; the detail of the algorithm is purpose–dependent.

The NSR as introduced by Chomsky and Halle (1968) is usually described in the following operational terms:

> Generate a sentence as a string of words with a structural description in the form of a (labelled) bracketing, and assign the lexical stress value 1 to each word.
> Iterate, for all sequences with no intervening brackets, until all brackets have been removed: Add 1 to all except final existing stress values in innermost bracket domains, and remove innermost brackets.

A tree–based formulation was given by Liberman and Prince (1977), briefly:

> Generate a sentence as a string of words with a structural description in the form of a (labelled) bracketing. Label the root with 'r' and assign 's' to each rightmost branch in the tree corresponding to the bracketing, and 'w' to all the others.
> Then, for each path in the tree from leaf to root:
>
> > Beginning with the first non-'s', count the number of nodes from this node to the root (including this node) and assign these numbers to the leaves as their stress values.

From a formal point of view it is simpler to define the NSR as a recursive function which maps tree–structures and initial numerical values into a sequence of pairs of 'stress' integers and leaves of the tree:[7]

(8) NSR: $structure, nonfinalvalue, finalvalue \rightarrow stresspattern$
Definition: If $structure$ is a leaf/atom, then pair $structure$ with $finalvalue$. If $structure$ is a sub–tree, re–adjoin the left branch of $structure$ after recursively applying NSR with higher values to it with the right branch of $structure$ after recursively applying NSR with higher $nonfinalvalue$ and the same $finalvalue$ to it.

An incremental algorithm for strings representing well–formed bracketings (the original NSR definition was formulated in similar terms) is also possible, but something of a curiosity:[8]

---

[7] For computational linguists, a proof–of–concept translation into Scheme:

```
(define (nsr t n m)
        (if (symbol? t)
                (cons t m)
                (list (nsr (car t) (+ n 1) (+ n 1)) (nsr (cdr t) (+ n 1) m))))
```

e.g. *(nsr '((caring . kate) . (phoned . (darling . dave))) 1 1)* yields the hierarchical 'stress pattern' *(((caring . 3)(kate . 2))((phoned . 3)((darling . 4)(dave . 1))))*, i.e. ^{3}*Caring* ^{2}*Kate* ^{3}*phoned* ^{4}*darling* ^{1}*Dave*.

[8] This sketch can be implemented directly in Perl, awk, etc., and the curious reader is invited to do so as an exercise. The input "((big john) (saw ((very little) (mary smith))))" generates "3big 2john 3saw 5very 4little 5mary 1smith".

Initialise depth counter to 1, word store to empty, last constituent flag to ”off”, iterate through the symbols in the string left–to–right:

if current symbol is a left bracket, the last constituent flag is set, and a word is stored, then print the depth counter and the word.
if current symbol is a left bracket, then increment the depth counter and turn off the last constituent flag.
if the current symbol is a right bracket, then decrement the depth counter and turn on the last constituent flag.
if current symbol is a word, and a word is stored, then print the depth counter and the word, reset the word store.
if the current symbol is a word, then store it.

All these NSR algorithms are equivalent; it does not really matter which one is used to map tree–structures to numerical values. Since a mapping of this kind is often seen as one of the tasks to be solved in TTS synthesis, it is not surprising that the basic NSR idea (in some iambic or trochaic parametrisation) has frequently been used in the TTS application domain.

There is an empirical problem with the NSR, however. As noted above, the naive assumption of direct interpretation of grammatical structures does not work: grammatical hierarchies are not only different from but also deeper and more complex than temporal organisation hierarchies, so there can be no simple alignment procedure for deeply embedded phrase structures. Nevertheless, this does not detract from the validity of the basic insight of compositional phonetic interpretation (often referred to by the term ‘cyclic’), which remains attractive, particularly for the flatter structures of informal speech.

The hierarchical NSR type approach in Metrical Phonology is criticised in the context of TTS synthesis by Wagner (2001). Wagner does not reject the linear alternation and iteration criteria provided by the metrical grid filter component, however, and uses FSTs with local cycles to formalise linear metrical grid filters (cf. Section 3). She retains the idea that grammatical parts–of–speech are useful predictors for alignments with rhythmic timing and demonstrates that better results for synthesis of German speech are given by a linear model based on five part–of–speech (POS) sets with different intrinsic stress weights:

1. Nouns, Numerals, Proper Names;
2. Adverbs, Adjectives;
3. Verbs, Demonstrative Pronouns, WH-Pronouns;
4. Modal & Auxiliary Verbs, Affirmative & Negation Particles;
5. Determiners, Conjunctions, Subjunctions, Prepositions.

It must be pointed out, however, that the POS weights contain strong assumptions about syntax hierarchies. For example, in German (and in English), many ‘weaker’ parts of speech alternate with stronger items on syntactic grounds alone, often preceding stronger items in a given construction. This induces shallow hierarchies and provides likely candidates for deriving iambic rhythms. In

the study of tree induction which follows in Section 5, a similar relation between function words and non–function words appears in the induced hierarchy, confirming that this kind of relation yields iambic timing structures.

The recursive compositionality property of the NSR function invites the question whether there might be an inverse function, NSR^{-1}, which would permit the (re–)construction of a tree from a sequence of numerical values. In fact it is possible to prove formally that there is indeed an inverse function; the proof will be presented elsewhere. An algorithm based on a generalisation of NSR^{-1} is used in the induction procedure presented in Section 5.

5 Tree Induction, Alignment and Comparison

The following sections are concerned with a data–driven approach to the modelling of pattern alternation in hierarchical timing patterns. The approach does not yet constitute a complete Rhythm Periodicity Model, because compositional hierarchy is modelled at the expense of compositional iteration, and isochrony is not yet incorporated. The term 'induction' is used here in the sense in which it is used in Machine Learning (ML; cf. recent linguistic applications by Sporleder 2004) and in data–mining. For linguistic purposes it is useful to make a distinction between paradigmatic and syntagmatic induction (though both are based on generalisation and data compression techniques). Paradigmatic induction, as in decision tree induction, or classfication and regression tree (CART) construction, generalises over sets of units which have shared properties. Syntagmatic induction is the compositional induction of part–whole generalisations over sequences of units, as in automaton induction or grammar learning. In the present computational phonetic context of the TTI–TSI induction and alignment methodology, 'induction' refers specifically to the syntagmatic hierarchical grouping of units according to empirical dependency and constituency criteria, and 'alignment' refers to the identification of phonetic events in speech signal recordings by assigning time–stamps to units (phones, syllables, words etc.) in transcriptions; 'comparison' refers to a distance measure for trees with identical leaf sequences.

The tree–building procedure in the present study can also be interpreted as a kind of parsing procedure, in which the numerical *greater* than or *less than* relations define 'categories', and the trochaic and iambic orderings define 'rules' in a 'grammar'. Parse trees are constructed with this grammar using top–down or bottom–up parsing schedules: the procedure parses the annotation sequence into Time Trees which are locally normalised for speech rate (as in the Local Linear Models), and generates alternating and hierarchical timing structures.

The induction procedure is demonstrated here using word–level annotations; it can be also used with temporally annotated data from any other time domain and other types of numerically labelled data, for instance to investigate interfaces between prosody and lexical, grammatical and discourse patterns.

5.1 The Induction, Parsing and Comparison Procedures

The Local Linear Models examined in Section 3 represent progress beyond the Global Linear Models in that they incorporate local speech rate normalisation and (at least in intention) an alternation component. They fail on the alternation property of rhythmic temporal organisation, however, because they use a non–directional distance measure which ignores the difference between *greater than* and *less than* numerical relations. In the present approach, based on the Rhythm Periodicity Model, the directionality of *greater than* and *less than* is utilised in the tree–building procedure in order to distinguish between hierarchical levels in building time trees, i.e. syntagmatic tree structures over time-annotated sequences with pattern alternation; compositional iteration and isochrony criteria are not yet incorporated.

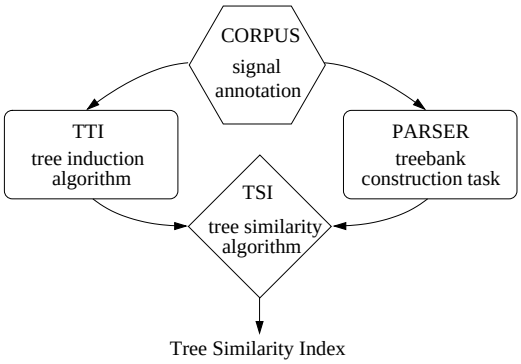

Figure 6: TTI–TSI tree induction, alignment and comparison architecture.

The Time Trees are compared with tree structures over the same strings from other domains, in the present case hierarchical grammatical structure. For this purpose, only unlabelled tree graphs are used; future developments will need to take the categories of the nodes into account. The procedure, with two tree analysis components and a comparison component, is as follows (cf. Figure 6; see also Gibbon 2003a,b):

1. Time Tree Induction (TTI) from long-short local duration differences in annotated speech signal data,
2. Parsing of the annotated transcriptions into syntax trees,
3. Calculation of a distance/proximity measure yielding a Tree Similarity Index (TSI) between the Time Trees and grammatical trees.

5.2 Time Tree Induction (TTI)

The data used in this procedure are from a narrative read by an English native speaker and hand–annotated at word level. The annotation relations have the following structure (in esps/waves+ format: time–stamps in the left–hand column refer to ends of words; the centre number is irrelevant for present purposes):

(9)
```
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
```

The tree–building algorithm[9] used here is essentially a deterministic bottom-up (shift–reduce) parser with a shift vs. reduce criterion derived from comparing neighbouring duration pairs $< \Delta t_i, \Delta t_{i+1} >$, like the PVI and RR algorithms but without the absolute values of duration differences, and also comparing values assigned to sub–trees. Various parametrisations of the algorithm are possible: *greater/equal* vs. *less*, *greater* vs. *less/equal*, or their association with conditions may be reversed, or an *equality* condition may be introduced, to cover non-binary structures. Formally, the TTI algorithm is a generalisation of NSR^{-1}, the inverse of the NSR function, where differences are not restricted to integer increments; the generalisation enables the algorithm to handle arbitrary numerical duration values. Given a *greater/equal* vs. *less* parametrisation, either of two

[9] For computational linguists, a proof–of–concept Scheme implementation:

```
(define (induce t stack)
        (cond
                ((null? t) (reduce stack))
                ((null? (cdr t)) (reduce (cons (car t) stack)))
                ((<= (caar t) (caadr t))
                        (reduce (induce (cdr t) (cons (car t) stack))))
                ((if (not (null?stack)) (¡= (caar t) (caar stack)))
                        (induce (cons (car stack) t) (cdr stack)))
                ((> (caar t) (caadr t))
                        (induce (cons (list (caadr t) (car t) (cadr t)) (cddr t)) stack))))
    (define(reduce stack)
        (cond
                ((null? stack) stack)
                ((and (not (null? stack)) (null? (cdr stack))) stack)
                (#t (cond
                        ((> (caar stack) (caadr stack))
                        (reduce (cons (list (caadr stack) (cadr stack) (car stack)) (cddr stack))))
                        ((<= (caar stack) (caadr stack))
                        (reduce (cons (list (caar stack) (cadr stack) (car stack)) (cddr stack))))))))
```

conditions may arise:

1. A *greater/equal* relation is found: a subtree is created with these two durations, and the duration of the longest unit (or shortest, depending on the parametrisation) percolates up to the root of the new subtree as a basis for recursive hierarchy construction by comparison with other neighbours.

2. A *less* relation is found: subtrees are constructed from following duration pairs in the sequence, and adjoined later.

Each time a new subtree is created, the duration label of the longest daughter percolates up to the top of this subtree, so 'neighbour' may refer either to a neighbouring leaf, or to a neighbouring subtree which has already been parsed. Duration percolation is used recursively to build larger subtrees until the entire sequence has been mapped into a tree; the longest duration in the sequence will label the root of the tree.[10]

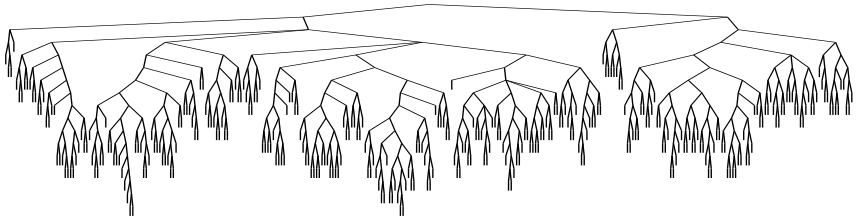

Figure 7: TTI tree over word durations in a complete narrative, illustrating divisions into constituents (e.g. the small leftmost main branch represents the title, others represent other episodes in the narrative). The largest constituents are determined mainly by pause durations and final lengthening.

Figure 7 shows a tree induced from the whole narrative. The durations of the smallest units (words) are projected into a tree spanning the entire narrative. Text structure is heterogeneous; the tree is correspondingly 'noisy'. But visual inspection also shows a number of interesting relations with discourse structure: for instance, the small leftmost subtree corresponds to the title of the narrative; other larger subtrees correspond to episodes in the story, and are largely determined by the length of pauses between these episodes. Figure 8 zooms into the tree, showing a syntax-timing correspondence (ZZZ denotes a pause) and bottom-up duration percolation (cf. the value .043) in an iambic parametrisation. An example of the output format of the algorithm is:

[10] John Wells has conjectured that the overall duration of the pair might be used instead. This is worth investigating, but the duration of a deep subtree may then not be directly comparable with the durations of shallower neighbouring subtrees or leaves.

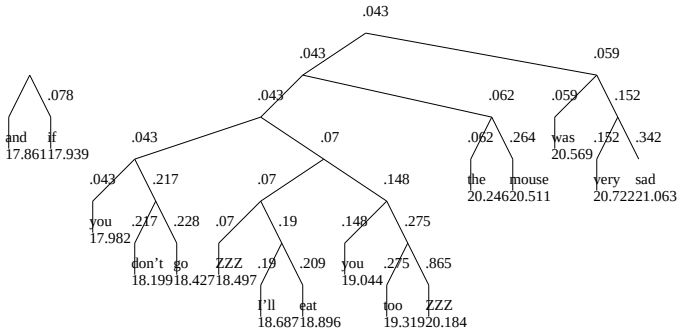

Figure 8: Zoom into a tree labelled with leaf strings and duration values, generated by an iambic parametrisation of the TTI algorithm (longer items right, shorter values promoted).

(10)
```
(.081
   (.081 (.081 "is:42.977") (.193 "nothing:43.171"))
   (.166
      (.166 "I:43.337")
      (.169
         (.169 (.169 "can:43.506") (.225 "do:43.731"))
         (.219 "ZZZ:43.950"))) ... )
```

The numerical labels following the left parentheses show durations; those following the colons are annotation time-stamps. The bracketing illustrates numerical value percolation from leaf to root, in this case with promotion of the shortest duration: the root value 0.081 is promoted from the value of the leaf 'is', here yielding an iambic pattern.

5.3 Parsing and Grammar–Prosody Correlation

The information required for determining the predictive value of grammatical information for timing, and vice versa, is purely structural, with no grammatical tags. In the long term, a treebank creation procedure is required for this purpose, but as an interim measure, and also in order to avoid falling into the trap of fashionable theoretical prejudice in automatic parsing, a 'subjective parsing' procedure was developed: unlabelled syntax trees were obtained by dividing a narrative into consecutive sentences, and requesting non-linguistics graduates to group expressions by bracketing them. A typical subjective parse result (not too different from conventional parsing wisdom) is the following, with top–down NSR 'stress pattern' generated by the bracketed string algorithm:

(11) ((there is (nothing I (can do))((said (the frog)) and hopped away)))
3there 3is 4nothing 4I 5can 3do 5said 6the 4frog 4and 4hopped 1away

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 bracketed sentence. A total of 120 subjective parses were elicited.

The induced time–trees and the syntax trees were compared automatically[11] with a distance/proximity measure, yielding a Tree Similarity Index (TSI):

(12) $TSI = \frac{|SUBSTR(T_1) \cap SUBSTR(T_2)|}{(|SUBSTR(T_1)| + |SUBSTR(T_2)|)/2}$

Each leaf in each tree is uniquely labelled (skolemised) before the algorithm is applied, non-branching nodes are pruned, and for each tree T_i, the set of substrings spanned by nodes in the tree $SUBSTR(T_i)$ is collected.
The two substring sets are intersected and the cardinality of the intersection $|SUBSTR(T_1) \cap SUBSTR(T_2)|$ is normalised in relation to the total number of nodes, in this case by calculating the mean of the node counts of the two trees:

The results of the study are visualised in Figure 9. The thick solid line shows correspondence between timing trees and unparsed (UP) sentences, the higher thin line shows mean TSI for the iambic condition, the lower thin line shows mean TSI for the trochaic condition. Both the iambic (0.85) and the trochaic (0.89) results correlate well with the unparsed sequence, probably due to the shallow bracketing and dependence on constituent lengths, but the absolute index levels differ considerably. Averaged over all subjects and sentences, the

[11] For computational linguists, a proof–of–concept Scheme implementation:

```
(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))))
```

iambic condition yields a mean TSI of 0.47, the trochaic condition yields a mean TSI of 0.2, while the unparsed condition yields a mean TSI of 0.19. The trochaic and unparsed conditions are practically indistinguishable. Syntax and TTI trees are thus interpreted as more similar under the iambic condition than under the trochaic condition (the proof–of–concept orientation of the pilot study and the number of samples involved did not justify further statistical evaluation).

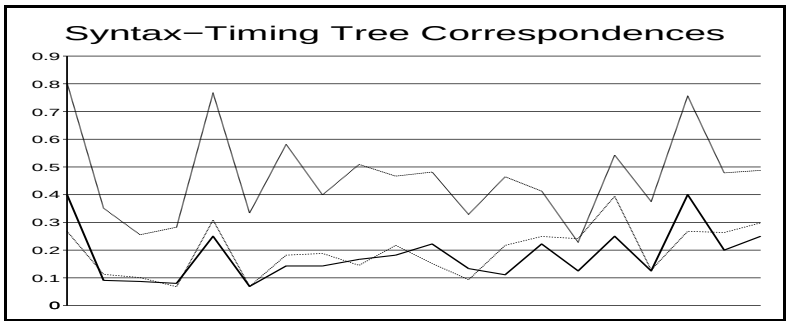

Figure 9: Syntax–timing tree correspondences in read-aloud narrative (X: syntax/ TTI tree pairs, Y: TSI).

Table 1: Summary of main results.

	mean UP-correlation	*mean* TSI
parsed + iambic:	0.85	0.47
parsed + trochaic:	0.89	0.2
unparsed + iambic:		0.19
unparsed + trochaic:		0.19

The results in Figure 9 and Table 1 show a preference for a match between grammatical structures and iambic groups, with short-long constituent pairs. Examination of the sentences indicates that the measure provides substantive and relevant information: the iambic pattern corresponds to the typical 'short–long' relation between function (closed class) words, which tend to be unstressed and short, and lexical or content (open class) words, which tend to be stressed and long. The potential which this comparison algorithm holds for the examination of discourse hierarchies (cf. Figure 7) has not yet been exploited.

6 Conclusion: Toward an Integrated Timing Model

Concepts of time and temporal organisation in phonetics and neighbouring disciplines were examined from a computational phonetic perspective, with the aim

of developing distributional data–driven hierarchical prosodic analyses. Possible 'cognitively real' dimensions of these analyses in terms of speech production and perception were not investigated; at a later stage, the distributional and correlationist methodology could perhaps be supplied with a cognitively relevant interpretation. Consequently, the methodology has rather positivistic traits. A purely positivistic account of timing structures may or may not be possible; it seems unlikely, so the present distributional methodology will need to be enhanced with information from other sources in order to create a more differentiated picture of emergent timing patterns, promising a future Emergent Rhythm Theory; cf. Gibbon and Fernandes (2005). Such sources are complex grammatical structures, discourse patterns, timing and alignment in dialogue interaction, cognitive expectations and neurological timing mechanisms, with Time Trees for each of these data streams.

The computational phonetic TTI–TSI methodology seems to be a suitable starting point for such enterprises in integrating alternating, iterating and hierarchical timing patterns, however. First results appear plausible, for instance in identifying the iambic, NSR–type prosodic structures associated with a certain kind of right–heavy syntactic structure: automatic induction and alignment of Time Trees produces a result which harmonises with linguistic expectations. Much remains to be done to develop Emergent Rhythm Theory, of which the Rhythm Periodicity Model will be a part, including generalisation to other speech genres and languages, deeper bracketing, weighting of categories, normalisation for sentence length effects, interpretation of the tree structures in terms of 'eurhythmic' criteria, incorporating compositional iteration and isochrony, methodological improvements concerning the size of the subject set, the use of treebanks, statistical treatment with more data, and the use of different empirical paradigms such as the investigation of rhythmicity by means of perception experiments.

The computational phonetic TTI–TSI methodology is thus still in its infancy, and very much a basic research activity. Nevertheless, on the basis of the Rhythm Periodicity Model, computational phonetics holds promise for the deployment of prosodic data mining strategies which will help to exploit the enormous quantities of annotated speech resources which have been amassed in many language resource projects all over the world. Potential fields of application of this distributional analysis and alignment methodology in descriptive linguistics and phonetics include the investigation of prosodic patterns in existing temporal annotations of endangered and extinct languages (particularly in the latter case, in the permanent absence of native speakers). Potential fields of application in speech technology are many; perhaps the most obvious area of application is prosodic pattern learning for TTS synthesis.

References

Abercrombie, D. (1967): Elements of General Phonetics. Edinburgh: Edinburgh University Press.

Barbosa, P. A. (2002): Explaining Cross-Linguistic Rhythmic Variability via a Coupled Oscillator Model of Rhythm Production. In: Proceedings of Speech Prosody 2002, Aix-en-Provence, 163–166.

Bierwisch, M. (1966): Regeln für die Intonation deutscher Sätze. Studia Grammatica 7, 99–201.

Bird, S. and E. Klein (1990): Phonological events. Journal of Linguistics 26, 33–56.

Bird, S. and M. Liberman (2001): A formal framework for linguistic annotation. Speech Communication 33(1,2), 23–60.

Campbell, N. (1992): Multi-level timing in speech. Ph.D. thesis, University of Sussex.

Carson-Berndsen, J. (1998): Time Map Phonology: Finite State Models and Event Logics in Speech Recognition. Dordrecht: Kluwer Academic Publishers.

Chomsky, N. and M. Halle (1968): The Sound Pattern of English. New York etc: Harper and Row.

Clements, G. and K. Ford (1979): Kikuyu Tone Shift and its Synchronic Consequences. Linguistic Inquiry 10, 179–210.

Culicover, P. and M. Rochemont (1983): Stress and focus in English. Language 59, 123–165.

Cummins, F. (2002): Speech Rhythm and Rhythmic Taxonomy. In: Proceedings of Speech Prosody 2002, Aix-en-Provence, 121–126.

Dauer, R. M. (1983): Stress–timing and syllable–timing reanalysed. Journal of Phonetics 11, 51–62.

Gibbon, D. (1992): Prosody, Time Types, and Linguistic Design Factors. In: Proceedings of KONVENS 92, Nürnberg, 90–99.

Gibbon, D. (2003a): Corpus-based syntax-prosody tree matching. In: Proceedings of EUROSPEECH 2003, Geneva.

Gibbon, D. (2003b): Computational modelling of rhythm as alternation, iteration and hierarchy. In: Proceedings of ICPhS 2003, Barcelona.

Gibbon, D. and F. R. Fernandes (2005): Annotation–mining for rhythm model comparison in Brazilian Portuguese. In: Proceedings of EUROSPEECH 2005, Lisbon.

Gibbon, D. and U. Gut (2001): Measuring speech rhythm in varieties of English. In: Proceedings of EUROSPEECH 2001, Aalborg, 91–94.

Gut, U., S. Adouakou, E.-A. Urua, and D. Gibbon (2001): Rhythm in West African tone languages: a study of Ibibio, Anyi and Ega. In: Proceedings of "Typology of African Prosodic Systems 2001" (TAPS), 159–165.

Jassem, W. (1951): Intonation of Conversational English (Educated Southern British). Wrocław: Wrocławskie Towarzystwo Naukowe.

Jassem, W. and D. Gibbon (1980): Re-defining English stress. Journal of the International Phonetic Association 10, 2–16.

Jassem, W., D. R. Hill, and I. H. Witten (1984): Isochrony in English speech: its statistical validity and linguistic relevance. In: Gibbon, D. and H. Richter (eds.), Intonation, Accent and Rhythm: Studies in Discourse Phonology., Berlin: Walter de Gruyter, 203–205.

Kornai, A. (1991): Formal Phonology. Ph.D. thesis, Stanford University.

Liberman, M. and A. Prince (1977): On stress and linguistic rhythm. Linguistic Inquiry 8, 249–336.

Low, E. L., E. Grabe, and F. Nolan (2000): Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English. Language and Speech 43(4), 377–401.

Niebuhr, O. and K. Kohler (2004): Perception and Cognitive Processing of Tonal Alignment in German. In: International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages, Beijing.

O'Dell, M. L. and T. Nieminen (1999): Coupled oscillator model of speech rhythm. In: Proceedings of the International Congress of Phonetic Sciences, San Francisco 1999.

Pike, K. (1945): The Intonation of American English. Ann Arbor: University of Michigan Press.

Ramus, F. (2002): Acoustic correlates of linguistic rhythm: Perspectives. In: Proceedings of Speech Prosody 2002, Aix–en–Provence, 115–120.

Ramus, F., M. Nespor, and J. Mehler (1999): Correlates of linguistic rhythm in the speech signal. Cognition 73(3), 265–292.

Roach, P. (1982): On the distinction between ‘stress-timed’ and ‘syllable-timed’ languages. In: Crystal, D. (ed.), Linguistic Controversies: Essays in Linguistic Theory and Practice, London: Edward Arnold, 73–79.

Scott, D. R., S. D. Isard, and B. de Boysson-Bardies (1986): On the measurement of rhythmic irregularity: a reply to Benguerel. Journal of Phonetics 14, 327–330.

Selkirk, E. (1984): Phonology and Syntax. The Relation between Sound and Structure. Cambridge: Cambridge University Press.

Sporleder, C. (2004): Discovering Lexical Generalisations. A Supervised Machine Learning Approach to Inheritance Hierarchy Construction. Ph.D. thesis, School of Informatics, University of Edinburgh.

Tillmann, H. G. and P. Mansell (1980): Phonetik. Stuttgart: Klett-Cotta.

Wachsmuth, I. (2002): Communicative rhythm in gesture and speech. In: McKevitt, P., C. Mulvihill, and S. O’Nuallain (eds.), Language, Vision and Music, Amsterdam: John Benjamin, 117–132.

Wagner, P. (2001): Rhythmic alternations in German read speech. In: Proceedings of Prosody 2000, Poznan, 237–245.

Wetzels, L. (2002): Comments on Low and Grabe. In: Gussenhoven, C. and N. Warner (eds.), Laboratory Phonology, Berlin: Mouton de Gruyter.