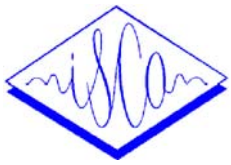
ISCA Archive
http://www.isca-speech.org/archive



International Symposium on Tonal Aspects of Languages: With Emphasis on Tone Languages
Beijing, China, March 28-31, 2004

Tone and timing: two problems and two methods for prosodic typology

Dafydd Gibbon

Universität Bielefeld, Germany
gibbon@spectrum.uni-bielefeld.de

Abstract

After an overview of some recent work on prosodic typology, both tonal and rhythmic, it is argued that both symbolic and numeric computational methods are required in order to advance the field by using large annotated resources and developing consistent, precise and evaluable models. Symbolic modelling is illustrated with reference to the lexical and grammatical tone systems of West African languages, and extensions to these are introduced in order to handle grammatical constraints on tone, affecting tonal templates and inflectional categories. A major motivating factor for the present discussion was tone generation for TTS system for Ibibio, a Lower Cross (Benue Congo) tone language spoken in South Eastern Nigeria. Numerical modelling is illustrated with reference to timing patterns, in particular rhythm, and a numerical method is introduced for inducing temporal hierarchies from annotated data, and for comparing these with syntactic hierarchies.

1. Overview

The present study[1] examines three main points:

1. Recent developments in prosodic typology in the areas of tone and rhythm.
2. The relevance of computational techiques for three main purposes:
   (a) precise operational modelling of prosodic systems,
   (b) empirical analysis of large annotated corpora in a variety of languages,
   (c) application of computational modelling to prosodic components of spoken language systems.
3. Illustration of these techniques in two different areas:
   (a) Tone modelling for the symbolic generation component of a TTS system for Ibibio, a West African tone language.
   (b) Induction of temporal structures from annotated speech data as a contribution to an empirical theory of rhythm.

Issues of tone typology are dealt with first, followed by a brief discussion of rhythm typology and the induction of temporal structures.

[1]This contribution has profi ted greatly from discussions with the late Eddy Gbery, and with Firmin Ahoua, Will Leben, Eno-Abasi Urua, Thorsten Trippel, Bruce Connell, Moses Ekpenyong, Ksenia Shalonova, Etienne Barnard and Gerald Gazdar. The work is partly funded by *Outside Echo Ltd.* in the LLSTI 'Local Language Speech Technology Initiative' consortium coordinated by Roger Tucker and Ksenia Shalonova; the Ibibio partners are Eno-Abasi Urua and Moses Ekpenyong.

2. Prosodic typology: tone

A comprehensive discussion of typology presupposes a comparison space determined by many typological parameters. This also applies to prosody, including so-called lexical prosody.

A number of typological prosodic distinctions have been mooted in the literature: tone languages vs. intonation languages, tone languages vs. accent languages, and within the tone languages Pike's well-known distinction between *contour tone language* and *register tone language*. Among the register tone languages there are the *terraced tone languages* and the *discrete level tone languages*. These distinctions are useful, but somewhat limited in scope:

1. It is gradually emerging that tone languages also have intonation, though differently constituted from the intonation systems of accent languages.
2. For this reason the term *accent language* will be considered here to be preferable to the term *intonation language*, using 'accent' to cover a variety of pitch accent and stress systems; issues of intonational and accentual typology have been well covered in (Hirst and Cristo, 1998).
3. The term 'tone language' is frequently limited to the area of lexical prosody, with the meaning of 'lexical tone language'; grammatical tone is, however, just as important.

In the literature on tone languages, most attention has been paid to phonological and phonetic dimensions of typology. Leben and Ahoua have recently published several fine-grained studies on the typology of Kwa languages (Niger-Congo, spread over the near-coastal areas of Ivory Coast, Ghana, Togo and Benin), in which they have discussed issues such as tone inventories, allotone patterning (downtrends and upsweep), but have also included the grammatical domains of tonal rules (Ahoua and Leben, 1997b), (Ahoua and Leben, 1997a). These studies show that the autonomy of phonology is illusory in regard to tonal systems, as it is for morphophonological contexts. But the combined phonological and grammatical parameter space is even more extensive than the division of labour between compositional grammatical domains and fixed lexical patterning suggests. The area of grammatical prosody is important for large classes of languages, of both the accentual and the tonal types. The term 'grammatical prosody' as used here includes *configurative* and *delimitative* functions, but includes *indexical emotional* functions. Specifically, it covers the following areas:

1. Negation scope: Sentence accent placement in English may delimit the scope of negation in English; by contrast, final lengthening of the phrasal constituent (without an explicit negation marker) may do this in Ega (Kwa, Ivory Coast).

2. Mood: The distinction of form between sentence types *declarative*, *interrogative*, *imperative*, etc. (as opposed to sentence functions or speech acts such as *statement*, *question*, *command*, etc., which relate in complex ways to distinctions of form). This category is discussed exhaustively in (Hirst and Cristo, 1998) and does not figure in the present study.
3. Inflection: Tonal inflection-related grammatical category realisation (e.g. of tense, aspect, modality, person, number) occurs in many Niger-Congo languages; the language focussed in the present study, Ibibio (Lower Cross, Nigeria), among other things, a tonal distinction between distal and proximal tenses.
4. Structure template: It is very common for nominal compound formation to be prosodically marked. Compound stress in English, as in *INcome tax*, is well-known. McCawley (McCawley, 1978) notes a Japanese dialect where nominal compounds have a fall in pitch before the second element. Ibibio has an obligatory H tone on the first vowel of the second element of a nominal compound: ènò 'gift' and àbàsì 'God' but ènò-ábàsì 'proper name'. Similar templates are found for grammatical constituents of noun phrases in African languages.

If these issues are considered, the parameter space for prosodic typology turns out to be quite rather high-dimensional:

3. Lexical defaults, grammatical overrides

Africa, and particularly the Niger-Congo language family in West, Central and South Africa, is well-known for its large number of tone languages, in which the main tones are terraced (see Section 4) register tones, but contour tones are also found.

Contour tones in these languages can, however, often be reduced historically or synchronically to sequences of register tones, as Ahoua and others have shown (Ahoua, 1996). A number of mechanisms contribute to the tonogenesis of contours, for example:

1. insertion of a grammatical tone;
2. deletion of a tone-bearing unit such as a vowel, reducing the syllable count but leaving the associated tone as a 'floating tone',
3. surviving floating tone displaces a neighbouring tone;
4. surviving floating tone forms a contour with a neighbouring tone.

Standard overviews are found in (Fromkin, 1978).

Without going into further details in the present context, it appears that the relation between lexical and other factors has, from a logical point of view, the character of a 'default and override' or 'if-then-else' system:

1. Lexical tones are available as defaults in case no overriding constraints are present (sometimes the case in citation forms or other isolating contexts),
2. In compositional contexts, three kinds of override may 'hide' lexical tone:
   (a) phonetic: terracing sandhi, which may lead to tone neutralisation with total downstep or other forms of tonal assimilation;
   (b) grammatical: word formation and nominal verbal phrasal tone templates, which mark syntagmatic constructions;
   (c) discoursal: focus, speech act and turn structure, emotion.

The following sections deal with some of the phonetic and grammatical constraints which operate on tone systems.

4. FST model for terraced tone

For the computational modelling of tone patterns it is necessary to define appropriate data structures and operational devices which can recognise and generate these structures. In metrical phonology, analyses of tone tended to use tree structures, right-branching or left-branching. However, no attempt was made to define algorithms for processing these tree structures (though sometimes informal rules were called algorithms).

But it is well-known in the theory of automata and formal grammars that

1. Right-branching parse trees can be generated by means of right-linear regular grammars.
2. Regular grammars generate regular languages (language in the sense of set of strings over an alphabet).
3. For any right-linear regular grammar there is a left-linear grammar which generates the same regular language.
4. Hence the choice of left or right branching trees is immaterial, though for a given regular language the one or the other might be simpler.
5. For any regular grammar there is a finite state automaton (FSA) which accepts the same regular language.

Based on these principles it was shown in (Gibbon, 1987) that the sequences of tone-allotone pairs which occur in terraced tone sequences are easily represented by a finite state (FS) device, specifically by a Finite State Transducer (FST), which in its basic form processes string pairs from two regular languages, and can be interpreted as

1. defining a *binary regular relation* between strings of regular languages;
2. translating one string into another: accepting first string of a pair as input and generating second as output;
3. reverse translating: accept second string of a pair as input and generating first as output;
4. accepting both strings simultaneously;
5. generating both strings simultaneously.

The standard reference for the linguistic use of FSTs is (Kaplan and Kay, 1994), where the relation between FSTs and standard phonological rule types is also shown; the advantage of using FSTs is that the overall system is clearly illustrated, and verification of the description by exhaustive generation is supported, while collections of isolated rules are not always perspicuous, and are certainly not evaluation-friendly.

Application of the FST modelling approach to several West African tone languages showed that terracing patterns appear as *iterations*, *cycles*, *oscillations*, and that typological differences between the languages are clearly reflected in the structure of the FSTs used to describe them (cf. (Gibbon, 2003c) and Figure 1). The cases in Figure 1 have the following properties:

(a) Simplest tonal schema. In a basic two-level terraced system (possibly rare or non-existent in its purest form) the FST has a start state (at which initial H and L tone values are defined), and two states, one for H tone sequences

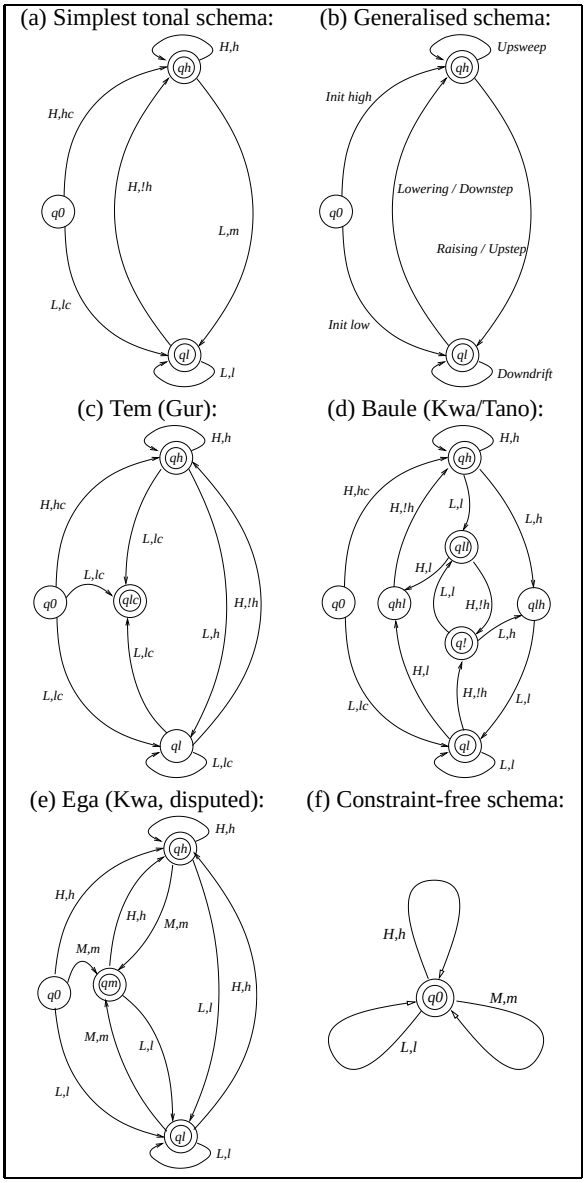


Figure 1: *Typologically distinct lexical tone automata.*

and one for L tone sequences. Between these states are transitions which are associated, language-specifically, with tonal assimilation relations, e.g. with total or partial automatic downstep (L followed by lowered H) or total or partial automatic upstep (H followed by raised L).

(b) Generalised schema. The two-level system can be generalised in terms of operations which have been introduced in the literature over the past 30 years. Additionally, a *terrace* can easily be defined as a complete cycle which includes both the H state and the L state (not necessarily in that order), and a *demiterrace* as a sequence of transitions leading to the same target state.

(c) In Tem (Gur, Togo) an additional state is introduced in order to handle the additional complexity of a final low tone which falls to a more or less constant level.

(d) Baule has a further complexity, in that the contexts for tonal assimilations are non-adjacent, necessitating longer paths between the main states, also with a non-deterministic element.

(e) Ega has a different complexity, namely a third tone, M. However, this complexity in fact appears on the basis of initial phonetic analyses to be associated with a lack of automatic downstep and therefore a lack of terracing. Consequently, this system can be reduced to

(f) a simple arbitrary iteration of all three tones. No doubt further investigation will show more complexity, but for present purposes this model will suffice to illustrate the situation with discrete-level (terracing free) tone languages, which typically have more than two tones.

The model generalises easily to other typologically related languages: case (b) in Figure 1 associates the linguistic terminology which is generally in use for the different tonal processes with transitions in the lexical tone automaton. It is also common to have special constraints associated with final states, either raising for final high tones or lowering for final low tones, as shown for Tem, a Gur language spoken in Togo, case (c) in Figure 1. This process is accommodated in the automaton simply by removing the final property from the low demiterrace state, and adding transitions to a separate final low state. Baule, a Kwa language spoken in Ivory Coast, has more complex contexts than most languages for tone sandhi; these can be formulated (non-deterministically) by adding additional oscillations; cf. case (d) in Figure 1.

5. FST enhancement: multi-tier constraints

The input and output of the terracing FST relate to grammatical and phonetic constraints respectively.

First, on the input side, constraints on grammatical tone are required. These will be discussed below.

Second, on the output side, phonetic constraints on the association of the tones with tone-bearing units are needed, including specification of segmental deletions with consequences for tone placement, syllable structures with depressor consonants, etc. These have been amply discussed in the literature.

Third, also on the output side, a numerical phonetic interpretation is necessary; in (Gibbon et al., 2003) the induction of numerical tone values is investigated using an exhaustive search model over a parameter space whose dimensionality is defined by the number of allotones of a given tone (i.e. the number of transitions in the FST on which the lexical input tone occurs) and the baseline and downtrend factors. This methodology is related to that of (Liberman and Pierrehumbert, 1984), except that an exhaustive search is used, without the downhill simplex search heuristic used in the latter study.

Focussing on the grammatical constraints, and continuing the FS approach, a first look is taken at simple, active, affirmative, declarative (SAAD, or kernel) sentences. For these sentences the following claims are made:

1. It should be evident that simple sentences do not require general context-free grammars, but form a finite and therefore trivially regular set which (given a finite vocabulary) can be generated by regular grammars and accepted by FSAs.
2. It is then straightforward to derive the claim that even discontinuous dependencies in simple sentences (e.g. agreement) can be captured with FS devices.
3. FS devices can cover iteration (recursion without centre embedding) and it should therefore also be clear

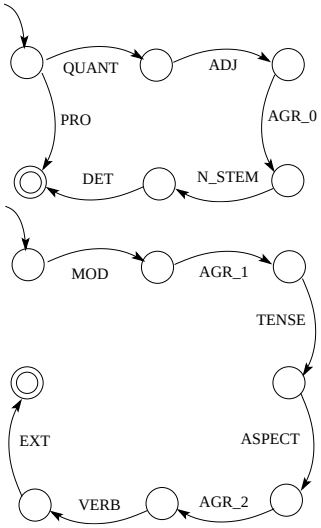


Figure 2: *Non-recursive Ibibio NP and VP FSTs.*

that conjunctions of simple sentences (or of their constituents) can be captured by FS devices.

Simple, nonrecursive FS networks describing Ibibio noun phrases and verbal expressions are shown in Figure 2. The networks characterise the two most interesting phrasal units of Ibibio and form the basis for their tonal interactions:

Noun phrases: The noun phrase may be pronominal (PRO), or consist of a sequence of optional quantifier (QUANT), obligatory person and number agreement prefix (AGR_0), optional adjective (ADJ), obligatory noun stem (N_STEM) and optional determiner or numeral (DET).

Verbal units: The verb has agglutinating prefixes, and consists of a sequence optional modality (MOD), agreement (AGR_1), tense (TENSE), aspect (ASPECT), a second agreement prefix (AGR_*2*) conditioned by (second) person and number.

Tonal properties and interactions: A number of elements are relevant for tone assignment.

The main tonal properties of grammatical constructions (Essien, 1990; **?**) are:

1. The structure of nouns: If the tone pattern of the second element of a compound noun contains a downstep (whether automatic or non-automatic), then the tone pattern is replaced by [H !H], otherwise by [H L].
2. The AGR features are relevant for tone assignment; as in English, AGR_*n* has the following internal structure:
$$\begin{bmatrix} \text{PERSON:} & \{\text{first, second, third}\} \\ \text{NUMBER:} & \{\text{singular, plural}\} \end{bmatrix}$$
3. TENSE has internal structure as follows, with the constraint that the SCALE feature only applies to past and future tenses:
$$\begin{bmatrix} \text{TEMP:} & \{\text{present, past, future}\} \\ \text{SCALE:} & \{\text{proximal, distal}\} \end{bmatrix}$$
where the future morpheme is ‘ya’, the past morpheme is ‘ma’, the distal morpheme is ‘ˇ’ (rising tone) and the proximal morpheme is ‘ˆ’ (falling tone).
4. The ARG_*2* position has a fixed H tone, regardless of the tone assigned to ARG_0 and ARG_1.
5. The other grammatical morphemes are associated with (more or less) regular tones, which are treated here as non-lexical.

6. Combining FSTs for different systems

The architecture of the generic model is shown in Figure 3 from a practical perspective as a possible design for a TTS system. This approach is being pursued in the Ibibio TTS project in the *Local Language Speech Technology Initiative* (LLSTI) consortium. The terracing FST has already been introduced and a specimen tape set of the grammar FST is illustrated in Table 1. The Unit Processor (whether diphone or corpus driven) is not discussed here; nor is the duration model, except to point out that the combinatoric complexity makes it unlikely that a sufficiently large corpus could be found for a data-driven unit selection approach to TTS, and that compositional duration and pitch models seem to be necessary for these language types.

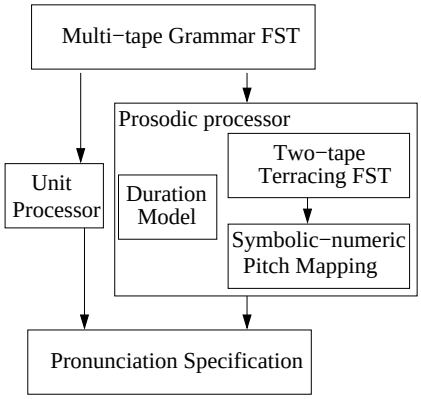


Figure 3: *Architecture of an FST based tone language processor.*

Regular languages are closed under concatenation, and in order to generate or accept sentences the NP and V FS devices can simply be joined by concatenation. If an object NP is required, or prepositional phrase, these can also be easily created in this fashion. Note that there is no recursion at this stage.

The questions which arise here are, first, how to combine the FSTs for grammatical constructions with FSTs for lexical tone, and, second, how to apply the downtrend rules formalised in the tone sandhi FSTs which have already introduced: Tentative answers are:

1. The two levels of lexical and grammatical tone are combined in a multi-tape transducer. This transducer was implemented in an exhaustive generation mode, and yielded a total of 735 simple sentence templates (with single dummy lexical items). A characteristic sentence is given in Table 1, showing the five transducer tapes, representing five annotation tiers:
   (a) LEXTONE: Lexical tone, associated only with content items.
   (b) GRAMTONE:Grammatical tone, associated with grammatical affixes, partly conventionalised, partly autonomous.

Table 1: *Automatically generated underlying lexical and grammatical tone tiers for 'mmè àfjá édòng é yáà é bēd Ìmé'.*

LEXTONE:	[ ]	[L H]	[ ]	[ ]	[ ]	[ ]	[H]	[L H]
GRAMTONE:	[L L]	[ ]	[H HL]	[H]	[H L]	[H]	[ ]	[ ]
TONE:	[L L]	[L H]	[H HL]	[H]	[H L]	[HF]	[H]	[L H]
SEGMENT:	[mme]	[afja]	[edON]	[e-]	[yaa-]	[e-]	[bed]	[Ime]
GLOSS:	[3 plur]	[white]	[sheep]	[agr]	[fut prox]	[agr]	[await]	[Ime]

(c) TONE: Combined lexical and grammatical tone output.

(d) SEGMENT:Segmental items (underlying level, before phonological rules).

(e) GLOSS:Literal gloss in English.

2. The TONE and SEGMENT tiers form the input to the phonological association rules (mainly involving segment reduction and tone re-linking) and to the tone terracing FST (see Figure 3).

The next stage is clear, although it has not yet been implemented and therefore cannot be demonstrated. It is known that two FSTs operating in parallel on the same pairs of tapes can be *composed* into a single FST. Since the terracing FST and the grammar FST share the tone tiers, these two transducers can, if the present argument is correct, be composed into a single transducer which will produce the correct allotones for input to the numerical pitch mapping component. Intuitively the operation could alternatively be seen as an FST cascade, which is also known to be composable into a single FST:

1. Use the TONE output tape of the grammar FST as the input tape to the terracing FST.
2. Add the output tape of the terracing FST to the tape set of the grammar FST.
3. The tapes in the tape set of the grammar FST are synchronised, as are the terracing FST input and output, so the terracing FST output will also be synchronised with the grammar FST tapes.

The final operation to be performed, as noted above, is the mapping of the allotone sequence to a sequence of target pitches (or other appropriate pitch objects).

7. The typology of rhythm and timing

Timing, and specific aspects such as the typology of rhythm, presents a rather different set of problems. An integrated computational phonetic approach is proposed as a data mining heuristic for rhythm timing analysis using large quantities of annotated data, with the long-term aim of providing a quantitative empirical foundation for prosodic typology and its applications, for example in TTS systems (cf. also (Gibbon, 2003b), (Gibbon, 2003a)).

Current models of rhythm timing in speech are 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 (Roach, 1982; Low et al., 2000; Ramus et al., 1999),

A classic phonetic approach to rhythm timing is that of Roach (Roach, 1982): tone unit duration is divided by the number of feet in the tone unit, yielding average or "ideal" foot duration approximating to isochrony, and the normalised deviation from mean foot length is measured. The idea, of course, is to measure *syllable isochrony*, rather than rhythm as such. Neither hierarchy nor linear alternation of timing units figure in the approach, which may be said to use a *Global Evenness* (GE) criterion as a measure of the isochrony property, rather than the alternation or hierarchy properties which are also necessary in a rhythm model. Any arbitrary re-sorting of the relevant segments in an utterance (random, shortest-to-longest, etc.) would yield the same index. Rhythm timing fulfils the GE criterion, in some sense, but it has other properties too, so while the GE criterion for timing is a necessary criterion for rhythm, it is not a sufficient one.

Ramus, Nespor & Mehler (Ramus et al., 1999) locate different languages in a typologically distinctive timing space over the following parameters: $V\%$, percentage of V (vocalic intervals) relative to overall utterance length; ΔC, variance of consonantal intervals; ΔV, variance of vocalic intervals. The $V\%$ measure reflects preferences for certain phonotactic patterns (CV, CVC, vowel length) as corpus tokens rather than lexical types. The model also uses a variety of GE criterion: V stretches and C stretches would still yield the same results if randomly sorted (by length, longer consonant sequences first, etc.). Similar considerations apply to the ΔV measure, which reflects evenness of vowel sequence lengths, lower values tending to isochrony, and to the ΔC measure. The model does not have hierarchical and alternating timing components and is thus also incomplete as a model of rhythm timing. Perhaps a different measure, such as ΔCV, could be used to address the issues of hierarchical and iterative structuring. A perceptual control for rhythmicity is clearly needed. As Cummins has pointed out (Cummins, 2002), the measure makes a statement about the evenness of the phonotactics of the language, rather than rhythm, rather like Roach's model; it reflects a necessary condition on rhythm models, but falls short of providing a sufficient condition.

Low, Grabe & Nolan (Low et al., 2000) 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.):[2]

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

The model yields a minimal value of 0 (perfect isochrony), asymptotically approaching 200 for larger length differences; the variant used in (Gut et al., 2001) reverses the scale, and has a maximum of 100 for perfect isochrony.

But the model has an empirical problem: it assumes *strictly binary rhythm*. Hence, alternations as in "*Little John met Robin Hood and so the merrie men were born.*" are adequately modelled, but not the unary rhythm (syllable timing) of "*This one big fat bear swam fast near Jane's boat.*" or ternary dactylic and anapaestic rhythms (or those with even higher cardinality)

[2] Wetzel's (Wetzels, 2002) comment that the PVI factors out final lengthening is mistaken: the counter does not stop short of the final item — a sequence of length m simply has $m-1$ differences between neighbours.

like "*Jonathan Appleby wandered around with a tune on his lips and saw Jennifer Middleton playing a xylophone down on the market-place.*"

And the model unfortunately also has a formal problem: the PVI yields the same value for sets of alternating patterns and monotonic geometrical series, and for mixes of these ($n!$ patterns with identical PVI for series of a length n). It is easily verified that alternating sequences may receive the same PVI as exponentially increasing or decreasing series $PVI(2,4,2,4,2,4) = PVI(2,4,8,16,32,64)$. This is obviously not the desired result. Interesting though the resulting typological patterns are, it is not clear what they are patterns of.

So this model, too, is empirically and formally incomplete. More comprehensive approaches to timing and rhythm modelling are emerging, however (Wagner, 2001; Cummins, 2002; Wachsmuth, 2002).

Cummins (Cummins, 2002), for example, discusses a number of additional factors involved in the production of rhythm in different styles, ranging from a paradigm of synchronous speaking designed to elicit maximally rhythmic utterances, to less constrained styles. He addresses both hierarchical and linear factors, and proposes a model for the more constrained styles with binary hierarchical structure, i.e. groupings of two-word feet, higher level groupings of two feet with four words, and so on. A new aspect of Cummins' experimental approach is the emphasis on the entrainment of different factors in the synchronous production of rhythm, particularly the interaction of discrete and gradient factors, with coupling between prosodic factors at foot level and a higher level.

Wagner (Wagner, 2001) criticises the hierarchical NSR type approach of Metrical Phonology (without rejecting a grid filter component, however), and concentrates on the linear alternation criterion, using FSTs with local cycles to formalise metrical grid type linear filters. Wagner shows that better results for synthesis of German speech are given by a linear model based on five part-of-speech sets with different intrinsic weighted abstract stress values (Wagner, 2001): {Nouns, Numerals, Proper Names}, {Adverbs, Adjectives}, {Verbs, Demonstrative Pronouns, WH-Pronouns}, {Modal & Auxiliary Verbs, Affirmative & Negation Particles}, {Determiners, Conjunctions, Subjunctions, Prepositions}. Wagner re-introduces the idea that grammatical categories are predictors of rhythm timing. In fact, these also contain strong assumptions about syntax hierarchies. For example, in German, many "weaker" parts of speech alternate with stronger items on syntactic grounds alone, often preceding stronger items in a given construction, thus inducing shallow hierarchies and perhaps an iambic rhythm, and suggesting interactions between rhythm and grammar which are of interest for language history and typology.

8. Time tree induction

We now retreat from over-ambitious claims about rhythm to timing. Rhythm timing as a complex function of hierarchical and linear structuring (cf. also Campbell's timing model (Campbell, 1992)) combines with local alternation criteria and with grammatical predictors for timing trees.[3] The present approach exclusively addresses the problem of identifying prosodic timing hierarchies and their relation to syntactic hierarchies. The approach is operationalised in two stages:

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

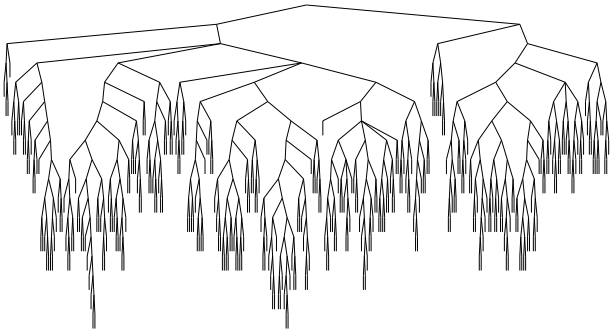


Figure 4: *TTI tree induced over a narrative.*

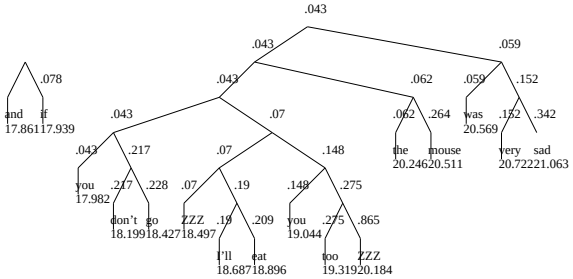


Figure 5: *Zoom into the narrative timing tree.*

In abstract terms, the TTI algorithm resembles the inverse of the NSR function of classical Generative Phonology, though it is not used for abstract stress values but to handle value differences between real data values as weighting operations. The weighted values percolate upwards, adjoining larger and larger units into a (not necessarily binary) timing tree. Four variants of the algorithm exist, and two were used in this study: TTI-A forms short-long groups, and the left-hand (short) value percolates up; TTI-B forms long-short groups, and the right-hand (long) value percolates up. Figure 4 shows a tree induced from an entire narrative. The smallest units in this example are words (any size unit can be investigated with this method, of course) whose durations are projected into a tree spanning the narrative, reflecting interesting divisions of the text which cannot be dealt with here. Figure 5 zooms into a tree which was induced for an entire narrative, showing the bottom-up percolation of values, e.g. of the value .043, and intuitively showing a syntax-timing correspondence (ZZZ denotes a pause).

9. Comparison with grammatical trees

The evaluation strategy for determining the predictive value of grammatical information is purely structural, and does not use named categories, unlike Wagner's approach. In order to avoid the twin traps of theoretical and personal prejudice in automatic parsing, the syntax trees were obtained by dividing a narrative into a set of 20 consecutive sentences, and requesting six linguistically literate subjects to group expressions in the sentences by bracketing them. No attempt was made to ensure uniformity

[3] Material in the following sections was published in (Gibbon, 2003b; **?**).

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. The following example shows a prettyprint of a subjective parse:

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

Timing trees, also as unlabelled bracketings, were extracted from readings of these sentences by a different subject, and hand-annotated at word level. A typical annotation file has the following (simplified) structure:

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

The following example shows the output of one parametrisation of the tree induction algorithm for the annotation file:

```
(.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 (Tree Similarity Index) algorithm:

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

The timing and syntactic trees were then compared automatically, yielding the TSI. 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.

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 the TSI the number of shared nodes spanning the same substring is simply divided by the mean node count of the two trees:

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

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. In brief, the measure is the number of subtrees spanning the same leaf sequence in each tree (in the present case, words), divided by the mean of the total numbers of nodes in the trees being compared. Summarising:

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

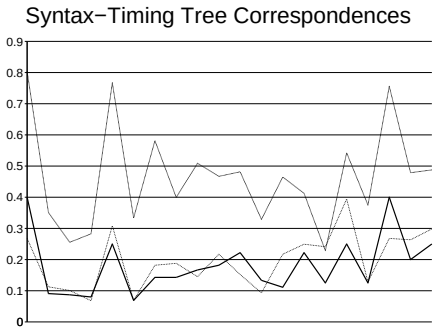


Figure 6: *Syntax-prosody correspondences in read narrative (X: syntax/TTI tree pairs, Y: TSI).*

The results of the study are visualised in Figure 6. The thick solid line shows correspondence between timing trees and unparsed (UP) ssentences, the higher thin line shows mean TSI for TTI-A short-long (iambic) grouped trees, the lower thin line shows mean TSI for TTI-B long-short (trochaic) grouped trees. Both TTI-A (0.85) and TTI-B (0.89) TSI sequences correlate highly with the UP sequence, probably due to shallow bracketing, but the TSI levels differ considerably. Averaged over all subjects and sentences: TTI-A mean TSI = 0.47; TTI-B mean TSI = 0.2; UP condition: mean TSI = 0.19. Clearly, mean TSI for A (short-long) is much higher than for B (long-short) or UP, which are indistinguishable. Syntax and TTI trees are thus more similar under TTI-A than under TTI-B. The methodological orientation of the study and the number of subjects do not currently justify further statistical evaluation.

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

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

10. Conclusion and outlook

Two very different areas of prosody were considered from the point of view of future developments in modelling prosodic typology: tone language prosody (including lexical and grammatical tone) and rhythm (as part of a general timing model). Of course this kind of enterprise courts the danger of trying to be two papers in one, and this danger has not been avoided here.

But the real point is a methodological one: the contribution advocates the importation of computational linguistics and AI methodology into phonetics of prosody. In agreement with Hirst [this conference] the present approach works on the thesis that complexity of the problems which confront anyone who attempts to model prosody – whether at discourse, sentence, phrasal, or word level, and whether symbolically or numerically – dooms any attempt to failure which does not use these computational tools to help develop and evaluate the models.

Many computational models of prosody are of course available, particularly for intonation. Not so many empirically valid models of tone or of timing are available, however. In the cases of tone and timing discussed here it was argued, following this premise, that computational methods are required in order to produce quantitatively and qualitatively adequate models. The two areas were subjected to complementary computational modelling strategies in order to illustrate the point:

First, symbolic modelling with complex systems of Finite State Transducers was proposed as a method for handling the complex interactions of lexical and grammatical tone in West African tone languges.

Second, numerical modelling of trees induced from annotated corpus data was proposed as a method for handing the kind of fuzzy data involved in the timing of speech utterances.

In both cases new and previously used techniques are combined in such a way that prosodic differences and similarities between languages may be better understood, and if possible used in applications of information and communication technologies to improve the situation of local languages around the world by taking their typological specificities into account.

11. References

Firmin Ahoua and William Leben. 1997a. Comparative Phonology of Kwa Languages in Côte d'Ivoire. *Papers of the Linguistic Society of Ghana*.

Firmin Ahoua and William Leben. 1997b. Prosodic domains in Baule. *Phonology*, 14.

Firmin Ahoua. 1996. *Prosodic Aspects of Baule*. Rüdiger Köppe Verlag, Köln.

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

Fred Cummins. 2002. Speech rhythm and rhythmic taxonomy. In *Proceedings of Speech Prosody 2002*, 121–126, Aix-en-Provence.

Okon E. Essien, editor. 1990. *A Grammar of the Ibibio Language*. University Press Limited, Ibadan.

Victoria A. Fromkin, editor. 1978. *Tone: A Linguistic Survey*. Academic Press, New York, San Francisco, London.

Dafydd Gibbon, Eno-Abasi Urua, and Ulrike Gut. 2003. A computational model of low tones in ibibio. In *Proceedings of the International Congress of Phonetic Sciences*, 623–626.

Dafydd Gibbon. 1987. Finite state processing of tone languages. In *Proceedings of EACL 3*, 291–297, Copenhagen.

Dafydd Gibbon. 2003a. Computational modelling of rhythm as alternation, iteration and hierarchy. In *Proceedings of the International Congress of Phonetic Sciences, Barcelona, August 2003*, 2489–2492.

Dafydd Gibbon. 2003b. Corpus-based syntax-prosody tree matching. In *Proceedings of Eurospeech 2003*.

Dafydd Gibbon. 2003c. Finite state prosodic analysis of African corpus resources. In *Proceedings of Eurospeech 2003*.

Ulrike Gut, Sandrine Adouakou, Eno-Abasi Urua, and Dafydd 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.

Daniel Hirst and Albert Di Cristo, editors. 1998. *Intonation Systems: A Survey of Twenty Languages*. Cambridge University Press, Cambridge.

Ron Kaplan and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378.

Mark Y. Liberman and Janet B. Pierrehumbert. 1984. Intonational invariance under changes in pitch range and length. In M. Aronoff and R. Oehrle, editors, *Language and Sound Structure*, 157–233. MIT Press, Cambridge MA.

Ee Ling Low, Esther Grabe, and Francis Nolan. 2000. Quantitative characterisations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech*, 43(4):377–401.

James McCawley. 1978. What is a tone language? In Victoria Fromkin, editor, *Tone: a Linguistic Survey*, 113–131. Academic Press, New York, San Francisco, London.

Franck Ramus, Marina Nespor, and Jacques Mehler. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3):265–292.

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

Ipke Wachsmuth. 2002. Communicative rhythm in gesture and speech. In Paul McKevitt, Conn Mulvihill, and Sean O'Nuallain, editors, *Language, Vision and Music*, 117–132. John Benjamin, Amsterdam.

Petra Wagner. 2001. Rhythmic alternations in German read speech. In *Proceedings of Prosody 2000*, 237–245, Poznan.

Leo Wetzels. 2002. Comments on Low and Grabe. In Carlos Gussenhoven and Natasha Warner, editors, *Laboratory Phonology*. Mouton de Gruyter, Berlin.