Chapter 1

The linearity of prosody

Dafydd Gibbon
Bielefeld University

This contribution introduces the TRIM (Time Types, Rank-Interpretation, Modulation) framework for comparing phonologically based symbol-phonetic with signal-phonetic approaches to prosody analysis. The aim is to explain reasons for certain types of disagreement and to provide an integrative perspective on prosody with a methodological rather than a domain focus. One guiding thread in the argument is a real-time requirement for realistic models, and another is the pragmatist principle that 'the method defines the object'. As a complement to earlier criteria for explanatory, descriptive and observational adequacy, a criterion of operational adequacy is introduced, focussing on the linear, real-time capability of prosodic models. Specific directions in phonology and signal phonetics are examined in relation to these criteria, including three characteristic operational models: the IPO, Pierrehumbert and Fujisaki models. Finally, a proposal is made for closing a gap in time modelling, the analysis of physical speech rhythms, an issue which has received little attention in actual operational models. It is suggested that current large language models (LLMs), as 'finite machines with giant contexts', potentially fulfil the TRIM criteria and use them, for example for the analysis of prosodic meanings and realistic speech synthesis of discourse.

1 Linearity and time in speech modulations

The present metatheoretical contribution aims to bring together a number of theoretical and empirical directions in prosody research, with reference to both signal-phonetic approaches on the one hand, and phonology-based symbol-phonetic approaches on the other. The stance is critically integrative rather than confrontational with respect to the different approaches: the essence of scientific method is comparison. Several theoretical strands are brought together in the TRIM (Time Types, Rank-Interpretation, Modulation) framework: hierarchical Rank–Interpretation domain theory, the methodological theory of Time Types and the operational theory of Modulation, bringing together a range of topics from earlier work.

Each of these dimensions of the TRIM framework has a long history in phonetics and linguistics which can only be sketched in the present study. Earlier metatheoretical criteria such as explanatory, descriptive and observational adequacy (Chomsky 1964) are retained and extended with a criterion of operational adequacy (Figure 1), i.e. the specification of data structures, algorithms and their temporal and spatial complexity properties, including real-time behaviour and working memory size. The guiding thread through this discussion is provided by contrasting two concepts of linearity and introducing operationally adequate concepts of finite storage and real time.

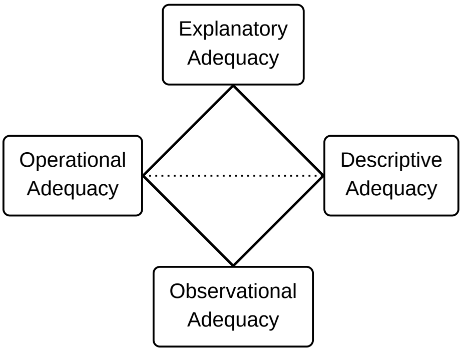

Figure 1: The diamond model of adequacy: explanatory, descriptive and observational adequacy criteria complemented by an operational adequacy criterion.

This chapter first establishes the theoretical foundations of prosodic linearity, time types and modulation (Sections 2-5), then examines traditional and derivational phonological approaches using TRIM criteria (Sections 6-7), analyzes three operational models that successfully bridge phonological and phonetic domains (Sections 8-9), addresses methodological gaps in physical rhythm modelling (Section 10), and presents conclusions (Section 11).

2 Operational adequacy

Speech production, transmission and perception take place in real time, an essential property of operational 'working models' and tools (Gibbon et al. 1997, Gibbon et al. 2000). Another essential requirement for operational adequacy is the modelling of speech as modulations of carrier signal baselines by information signals (Traunmüller 1994, Todd 1994, Cummins 1999, Barbosa 2002, Tilsen & Johnson 2008, Gibbon 2023), a bridge between the distinct methods of phonology and phonetics. The phonology-phonetics methodological distinction was innovated most clearly by the Prague School (Trubetzkoy 1939,[1] continued by Jakobson et al. 1952 and Chomsky & Halle 1968). The methodological difference from a contemporary perspective lies essentially between, on the one hand, the transcription oriented intuitive analytic methods of phonology and linguistic phonetics on the one hand, and, on the other, the abductive physical measurement methods of signal processing. Studies of language and speech are often ahistorical, rediscovering the wheel, so, where possible, a historical stance has been adopted, with a preference for the innovative older sources, rather than recent similar sources.

The field of phonology (and other areas of linguistics) have a perceptual basis in symbol phonetics and ostensibly has the same domain as SIGNAL PHONETICS. Phonology does notinclude the processing and real-time properties of speech: there is a 'time gap' in the models. There are exceptions in some areas of linguistics. For example, time is included in discourse analysis, as surveyed and extended in (Couper-Kuhlen & Selting 2018), and also in models for evaluating language proficiency in spontaneous and read speech, defined in terms of parameters such as tempo, length of 'run' (interpausal unit, IPU), pause distribution, local pitch accent shape, or global pitch slope over time (Hudson et al. 2005, Lin & Gibbon 2023, Wang & Gibbon 2024).

Phonological analyses model intuitively perceived categories, relations and structures of written and spoken language and interpret these models as metaphors for structures in language cognition. Temporal relations of sequence and overlap (Allen 1983), cf. Figure 2, are often not explicitly considered, but modelled in the visual domain with written transcriptions on paper or screen, and with no calibration in terms of real-time points and intervals, except in linguistic phonetic signal annotations. In generative and post-generative phonologies rooted in Chomsky & Halle 1968, temporal sequence is reflected in left-right character-string concatenation. Temporal overlap, in prosodic phonologies such as autosegmental and articulatory phonologies deriving from Leben 1973, Goldsmith 1976, Browman & Goldstein 1986, is modelled with lattices consisting of parallel concatenated strings (tiers), with temporal synchronisation reflected in links (association lines) between the parallel strings. In metrical phonologies based on Liberman & Prince 1977, perceived stress is modelled by column charts (grids), and string inclusion in the prosodic hierarchy is modelled by morphosyntactic tree graphs (Selkirk 1984).

In psycholinguistics (Cutler 2012) and laboratory phonologies (Beckman & Kingston 1990, Keating 1990, Ohala 1990), and in the computational method of

[1] "Eine saubere Scheidung von Phonologie und Phonetik ist grundsätzlich notwendig und praktisch durchführbar." (p. 388), *A clear separation of phonology and phonetics is fundamentally necessary and feasible in practice.*

Pierrehumbert 1980 and Liberman & Pierrehumbert 1984, methods are provided which are aimed at bridging symbol-phonetic and signal-phonetic domains in terms of correlates between the domains and are compatible with the TRIM criteria. Bridging the different methods is also illustrated by speech signal annotation mining, with origins in speech technology, which pairs symbols with signal time-stamps (Gibbon & Fernandes 2005, Gibbon 2006, Yu et al. 2014).

Arguably, the neural networks of contemporary large language models, with their Gigatoken contexts and operational adequacy through auxiliary natural language processing (NLP) front-end models, are close to the goal of real-time validity, and certainly fulfil the traditional completeness criterion of modelling a language as *un système où tout se tient* ('a system in which everything is connected'), a requirement for any working model of language and speech.

3 Category linearity and system linearity

A second insight into speech production, transmission and perception within the TRIM framework is that real-time procedures require linear systems. The term 'linear' has somewhat different but related meanings in symbol phonetics and in formal and computational phonology on the one hand, and in signal phonetics and speech technology on the other.

In approaches related to symbol phonetics, 'linear' refers to grammars which recognise or analyse an input in an abstract time interval which is linearly proportionate to the length of the input, plus a grammar constant, and the term also refers to the class of languages (i.e. sets of strings) which such grammars define. In formal language theory, these grammars and languages are Type 3, the grammars which are equivalent to (and the languages which can be recognised by) finite state automata, FSAs, i.e. automata with finite working memory (Chomsky & Schützenberger 1963). FSAs are Markovian, with linear temporal properties: the future state of the system depends only on its present state, with limited working memory, not on its entire past history. There is evidence from anthropological and sociolinguistic studies that in unwritten languages and informal, unelicited speech these models are sufficient (Everett 2005, Gibbon & Griffiths 2017).

There are simpler grammars than Type 3, but there is a hierarchy of more complex grammars which have often been advocated for specific constructions in languages, including prosodic embedding, all of which are non-linear and not real-time capable in the general case. The more complex grammars are Type 2 (context-free) grammars which define strongly recursive centre-embedded syntagmatic tree structures, Type 1 (context-sensitive) for, *inter alia*, relations across trees such as cross-serial dependencies as in *Jill and Jen married Bill and Ben, respectively*, and Type 0 (unrestricted) for arbitrary relations between strings and substrings, used in early transformational grammar (Chomsky 1957). Types 2, 1 and 0 do not have linear real-time properties or finite memory, and are this operationally inadequate for the flat iterative (weakly recursive) property of speech, but are more adequate for written language, which has much looser time and space constraints which enable limited strong recursion.

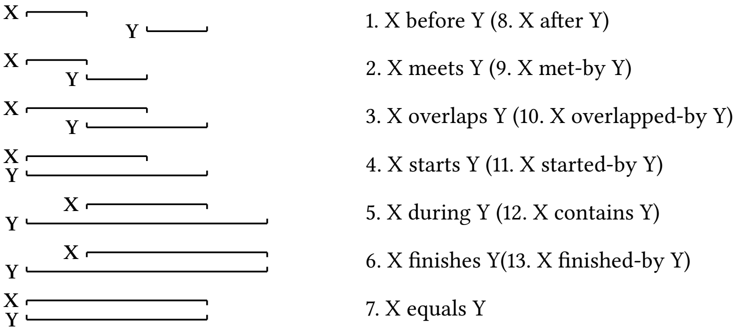
X
Y
1. X before Y (8. X after Y)
X
Y
2. X meets Y (9. X met-by Y)
X
Y
3. X overlaps Y (10. X overlapped-by Y)
X
Y
4. X starts Y (11. X started-by Y)
X
Y
5. X during Y (12. X contains Y)
X
Y
6. X finishes Y(13. X finished-by Y)
X
Y
7. X equals Y

Figure 2: The 13 temporal base relations of Allen Interval Algebra, demonstrating ways of explicitly formalising autosegmental tone-bearing, tone-spreading and other prosodic parallel patterns and signal-phonetic superposition models.

Models of simultaneity in phonologies such as autosegmental phonology (Goldsmith 1976) are also often referred to as 'non-linear', but are perhaps better referred to as 'parallel linear' or 'multilinear' (Gibbon & Griffiths 2017), as characterised by Allen interval algebra, Figure 2 and used in phonological (Bird & Klein 1990) and speech recognition (Carson-Berndsen 1998) contexts: the overlapping tiers are linear (Kay 1987). In a real-time phonetic model this would involve synchronisation and co-modulation of parallel streams. This point is important in view of the many parallel information signals which modulate speech: if each parallel channel is linear, and the system processor is genuinely parallel, then the overall system may be considered linear.

In signal phonetics, on the other hand, a linear system is a transfer function in which the output signal retains properties of the input signal: first, if two signals are mixed (added) in the input, both are preserved in the output; second, if the input is scaled (multiplied) by a certain amount, then the output is scaled by the same amount. Both of these conditions imply temporal linearity in a similar sense to the definition in computational phonology, but with explicit temporal detail. Both algebraic and numerical linearity definitions are necessarily combined in operationally adequate working models (Fujisaki & Nagashima 1969, Collier & ’t Hart 1971, Pierrehumbert 1980); cf. Sections 8 and 9.

A central insight in the TRIM framework is that working memory size in speech production and perception is finite and small: fast real-time adequacy requires a small real-space working memory (Church 1980, Kornai 1985, 1987). In the symbol-based domains of linguistics and language philosophy this time constraint is excluded.

In the act of writing, too, this time constraint does not apply and the size of working memory is the size of the document or more. Transcriptions, for example, are visual writing-based models (Pike 1947) with spatial parallels to temporal structures. There is no principled limit to the time needed to re-analyse and finish a written sentence (note the page-long sentences in some *avant-garde* novels) or to include the centre-embedded subordinate clauses of the written registers of mathematics and practised public speaking, which are conspicuously absent in spontaneous, unpractised, rhetorically unpretentious speech.

The actual examples used in many phonological studies of prosody are often more related to transcriptions of writing-influenced formal reading styles than to informal speech styles and registers, showing the need to consider written and spoken ranks, styles and registers and not to restrict attention to ‘universal’ (but actually writing-based) varieties.

4 Time Types: syntagmatic hierarchies and time scales

The pragmatic principle, The method determines the object, provides a fundamental insight: different methods may start from similar premises, but they produce results which may or may not coincide, owing to the different methods used. A ‘starry starry night’ is not closely related to a radio map of the universe. A corollary of this principle is that there are also different models of time, depending on the methods used.

Table 1 shows the Time Types model (Gibbon 1992b, 2006), with less and less abstract concepts from top to bottom, illustrating differences which need to be considered when interpreting intuitively determined phonological constructs as measurable and analysable real-time signals. The most concrete Time Type is ‘cloud time’, the empirical usage basis for the higher level Time Types. Categorial time is a characteristic of most phonologies, rubber time of event phonologies (Bird & Klein 1990, Bird 1995), as well as articulatory phonology (Browman & Goldstein 1986) and other laboratory phonologies, which employ a combination

Table 1: Time Type Model of time abstractions in linguistics, phonetics and signal processing.

Time Type	**Domain**	**Structure and Role**
Categorial Time	Phonology	Categories, lattices of concatenated strings; spatial model
Rubber Time	Event Phonologies	Temporal relations between category intervals; logical model
Procedural Time	State Machines	Transition path times: linear, logarithmic, polynomial, non-polynomially hard
Clock Time	Measurements	Models of temporal points, intervals and signal magnitudes
Cloud Time	Realtime signals	Speech in the wild, analog systems

of rubber time and clock time.[2]

One characteristic of the Time Types which are more abstract than clock time is that they have no intrinsic relation to real time, except perhaps directionality: it does not matter, for a phonological analysis, whether a given sentence lasts three seconds or three hours (or more). The absence of a theory of Time Types in practically all phonological and phonetic studies has led to many misunderstandings and simplistic ideas about the relation between phonology and phonetics being a scale of detail. This may be true of phonology and of linguistic phonetics but there is a methodological gap: it is not true of the clock time of quantitative experimental phonetics and laboratory phonologies, which are based on very different empirical methods and thus yield different kinds of result.

Speech is characterised by empirically determined syntagmatic (i.e. compositional structure-defining) hierarchical time structures (Campbell 1992). There are three major functionally relevant hierarchical temporal domains for prosodic information and timing (Tillmann & Mansell 1980; cf. also Chao 1968, Gibbon 1992b), corresponding to the phonemic, word and supra-word ranks respectively (Table 1 and Table 2):

1. Micro-prosodic timing: <100 ms: subsyllabic chunks; F_0 perturbations, intrinsic pitch;

2. Prosodic timing: ca. 250-1500 ms: syllables, words, accent peaks, tones;

3. Macro-prosodic timing: >1500 ms, words in phrases, sentences, texts and discourses; pauses and intonation contours in larger contour hierarchies.

[2] ‘Rubber Time’ and ’Clock Time’ are due to Andras Kornai, ESSLLI summer school 1990.

Macroprosodic timing encompasses a scale with further divisions, with so-called 'paragraph intonation' or 'paratones', i.e. discourse intonation at higher ranks (Lehiste 1970, Gibbon & Richter 1984).

The hierarchical structure of speech is outlined in the semiotic Rank-Interpretation model of speech. The underlying model is a semiotic quintuple of semantic and pragmatic meanings, structure, space and time (Example 1.1):

$$<< Context, < Meaning, Structure, Modality >>, Time > \tag{1.1}$$

This semiotic structure applies to each rank of the entire ranked system of structures and their associated meanings, modalities and relevant contexts (Gibbon 1992a, Gibbon & Griffiths 2017). The Rank-Interpretation model is selectively summarised in Table 2.

Table 2: Semiotic Rank-Interpretation Model of Language and Speech illustrating the hierarchical and parallel association of prosodic structure, function and form.

Structural Ranks	Semantic and Pragmatic Meaning Interpretations	Metalocutionary Indexical and Iconic Prosodic Meaning	Auditory and Visual Modality Interpretations
Discourse	Denotation, reference	Metalocutionary framing of dialogues, calls, …	Long-term intonation and rhythm trends
Text	Speech act, argumentation, narrative	Metalocutionary rhetoric	Cohesive intonation and rhythm
Sentence	Proposition, information structure, focus	Metalocutionary cohesive contour marking	Tone group, pitch accent pattern
Word	Lexical meanings	Metalocutionary predication, pointing	Morphology and phonology

In Table 2 the formal metalocutionary semantic interpretation of prosody is shown separately from classical semantic and pragmatic categories: a prosodic sign is a METALOCUTION, in particular a metadeictic event which, in the sense of logical semantics, actually DENOTES locutions, i.e. words and phrases. In addition to the theoretical status of the Rank-Interpretation approach, for many practical purposes such as language teaching and speech engineering an overall picture of this kind is needed for operational models and their evaluation: *un système où tout se tient*, a coherent landscape which can be viewed from a general vantage point or close-up, as needed.

5 Speech modulation channels

The time, frequency, linearity and hierarchy properties of the parallel channels of spoken language combine within the modulation-theoretic model. The units of speech communication are very slow, <15 Hz, with syllable rates around 5 Hz, word rates around 1.5 Hz, and slower rates at sentence and discourse ranks. These frequencies are those of rhythmic beats, as with African telephonic drums, and as such are far below the range of hearing. Modulation transforms the structurally coded very low-frequency locutionary and prosodic information into amplitude modulation (AM) and frequency modulation (FM) of higher-frequency carrier signals, thus making the lower frequency information signals audible.

The pulse modulation of voice on/off switching is one type of AM, and the AM differences between the phone types obstruent, glide, liquid and vowel are another. The higher-frequency formants F_1, F_2 and F_3 are themselves amplitude modulations of the harmonics of F_0, and are in turn frequency modulated and thereby convey syllable patterns: phone formant AM is created by vocal tract configurations, and phoneformant FM by configurational changes of the vocal tract. An F_0 carrier signal between 70 Hz and 600 Hz (FM for tones, pitch accents and intonation), and F_1 FM around 500 Hz, F_2 FM around 1500 Hz, and F_3 FM around 2500 Hz. The low frequency FM and AM variations of the higher-frequency carrier signals enable audible transmission of the lower-frequency information signals.

The basic modulation model is shown in equation 1.2 and visualised in Figure 3 for linear signals, in which a linear timeline t is shared by F_0, F_1, F_2 and F_3 carrier signals which are modulated by AM and FM information signals.

$$s(t) = (1 + k_{\mathrm{AM}} \cdot a_{\mathrm{AM}}(t)) \cdot A_c \cdot \cos\left(2\pi\left[f_c + \Delta f \cdot a_{\mathrm{FM}}(t)\right] \cdot t\right) \tag{1.2}$$

where $a_{\mathrm{AM}}(t)$ and $a_{\mathrm{FM}}(t)$ are the amplitude and frequency modulating signals, A_c is the carrier amplitude, k_{AM} is the AM modulation index, and Δf is the FM deviation scale.

The AM information is typically constituted by sequences of vowels and consonants, syllables and phrases with regularly varying high and low magnitudes (corresponding to the informal concept of sonority in phonology) which create speech rhythms. The FM information is typically the melodic modulation of F_0 in

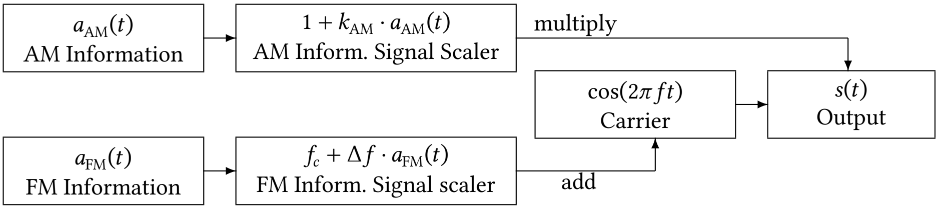

Figure 3: Schematic of AM and FM modulation of the speech signal by two information signals, showing the formal basis for synchronising prosodic and locutionary parallel and simultaneous streams.

tones, pitch accents and intonation, and of the formants F_1, F_2 and F_3 in vowels and consonants. The equation and the figure are formulated with a bias towards generation of modulation, but apply also in reverse to demodulation both in perception and in speech analysis. Prosody demodulation, for example, means extraction of (1) the AM envelope of rhythms as oscillation of consonant and vowel amplitude, and (2) the FM trajectories of rhythmically organised tones, pitch accents and intonations, for example as shown in Figure 5, in both cases as functions of time.

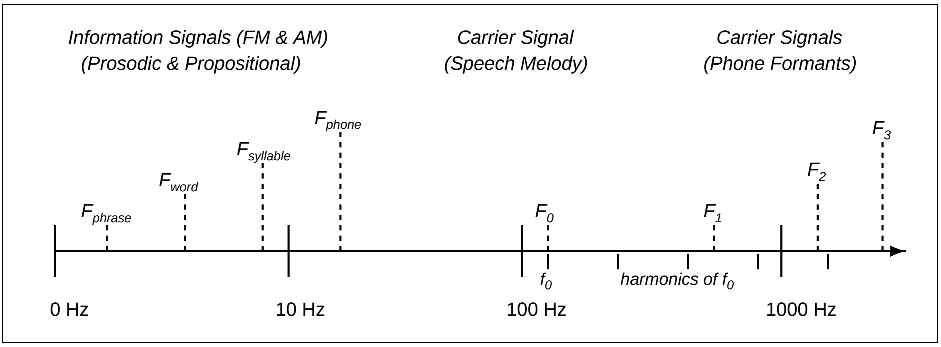

Figure 4: The Speech Frequency Scale of physical prosodic modulation hierarchies. F_0 denotes the fundamental frequency used functionally, f_0 denotes the physical fundamental frequency as the first harmonic of a series.

The scales of a temporal hierarchy such as those indicated in Table 2 can be translated ($f = 1/T$) into the Frequency Scale model shown in Figure 4, illustrating frequency relations between information signals and carrier signals.

After a first glance at this characterisation of speech modulation it might appear that written, spoken and gestural language varieties are too complex to be linear, in the sense of real-time capable, a caveat which seems to be confirmed by the existence of disfluencies, but is evidently empirically wrong. Even the apparent non-linearities which occur when there are time and memory conflicts due to interference from lexical access, pathology, or the communication scenario (Levelt 1983) are handled in real time.

6 Precursors of prosodic phonologies

The first comprehensive descriptive models of intonation are centuries old but their principles are still used in applied linguistics: the icon-based 'tonetic' family of intonation models (Gibbon 1976, Hirst 2024).

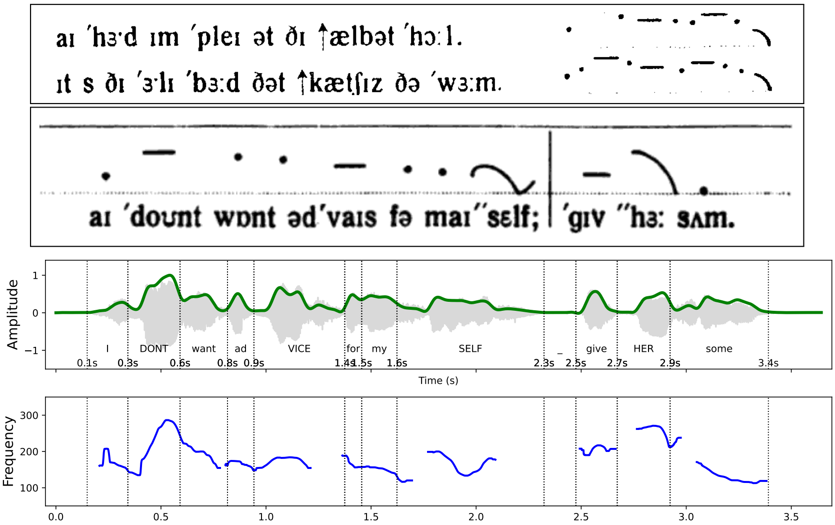

Figure 5: Examples of the Armstrong and Ward tonetic model: row 1, examples of tonal reset; row 2: example of tone group structures; row 3: annotated reading aloud with AM demodulation (row 3); row 4: FM demodulation.

The tonetic models were developed primarily for intonation teaching on a solid empirical basis and have explicit concepts of pitch accents (tones), hierarchical structure and pitch reset (Armstrong & Ward 1926, p. 19, cf. Figure 5 row 1; later Gibbon 1981 and others). Their practical purpose in applied linguistics is teaching, motivated by the need for students to obtain operationally adequate and verifiable results. The main representatives are Armstrong & Ward 1926, Kingdon 1958, O'Connor & Arnold 1961, with linguistic extensions in Halliday

1967 and phonetic extensions in Crystal 1969); these approaches provide a wealth of prosodic information.

The representation shown in Figure 5 row 2, from Armstrong & Ward 1926 p.75, example 1, shows one of the early examples of tonetic transcription of a read sentence. The icon-based transcription distinguishes between accented and unaccented syllables and has pause boundary marking, a minimal hierarchy. The transcription was taken as a prompt for reading aloud: the AM and FM tracks of a reading are shown in Figure 5 rows 3 and 4, respectively. The reason why a century-old textbook example is chosen is the innovative character for its time, with an explicit, detailed, consistent and comprehensive model driven by operational teaching goals.

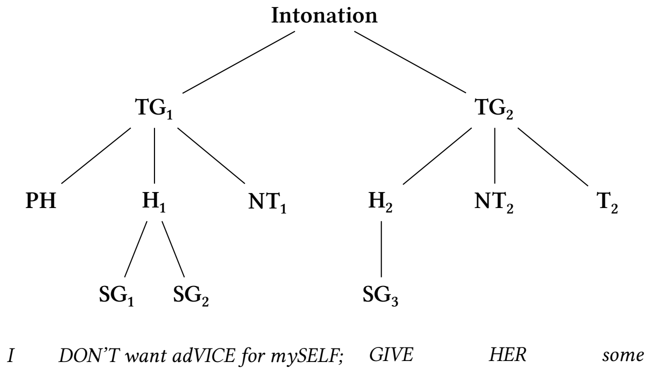

Figure 6: Tonetic intonation structure of the Armstrong and Ward example with approximate text alignment (upper case marks pitch accents), showing the explicit analysis of a prosodic hierarchy in an applied phonetic paradigm.

One way of showing the structure of the tonetic models, based on later developments, is provided in Figure 6 as a tree graph representing a conflation of several later tonetic models, and following terminology used in later applied phonetics textbooks (Kingdon 1958, O'Connor & Arnold 1961). The overall intonation pattern is explicitly hierarchical, with the Tone Group constituents Pre-Head, Head (or Body), Nucleus with Nuclear Tone (or Tonic Syllable) and Tail, with the Head defined as a sequence of similar Stress Groups, also with a Major Tone Group and Minor Tone Group hierarchy (Trim 1959), precursors of the intermediate phrase and intonation phrase of later models. The structure of the hierarchical tonetic models foreshadows the prosodic hierarchy, formally introduced by Selkirk 1984. Each of the four Head categories (O'Connor & Arnold 1961 — High, Low, Rising, Falling) contains a sequence of accents of the same shape, depending on the Head type, predating the accent sequence uniformity constraint which is often discussed in the context of declination in later studies (Dilley 2005).

In the TRIM framework, the hierarchical structure represents sequential and parallel organisation in both categorial time and rubber time (Figure 2 and Table 1). The hierarchies have finite depth and are internally iterative, not strongly recursive, fulfil the Type 3 linearity condition, and are real-time capable.

7 Derivational prosodic phonologies

The earliest formally explicit derivational phonology (apart from the formal historical phonologies developed in the 19th century) is SPE (Chomsky & Halle 1968), *The Sound Pattern of English*, the common ancestor of a clan of well-known and widely used MIT phonologies: autosegmental, metrical, and optimality theoretic. From the complexity point of view, the core of the SPE model has been demonstrated to have linear phonotactics and phonetic interpretation, and can thus be modelled with FSAs (Johnson 1972, Kornai 1985). Operational finite state implementations of the SPE rule system were developed (Kaplan & Kay 1994) and have been used to model the morphophonology of many languages (Beesley & Karttunen 2003).

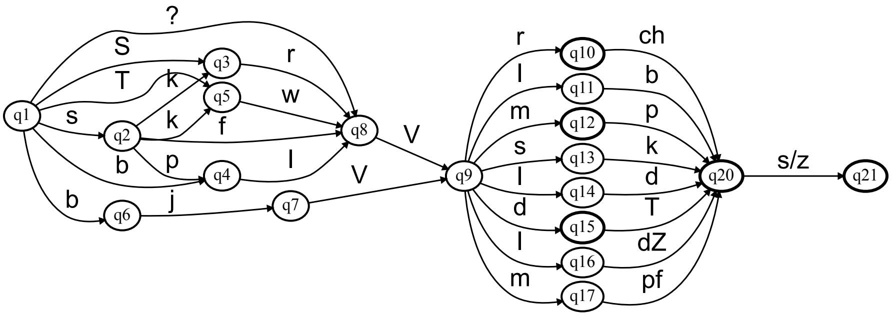

Figure 7: Rendering of Whorf's regular-expression-like linear grammar for English strong syllable phonotactics (Whorf 1940) as a nondeterministic finite-state automaton (to save space, each edge label stands for a natural class of phonemes).

In fact, the regular-expression-like linear phonotactic grammar of Whorf 1940, 16 years before regular expressions and FSAs were invented, straightforwardly converts to a nondeterministic finite state transition diagram, shown in Figure 7.[3]

SPE has nothing to say about intonation, but represents stress patterns as numerical encodings of lexical and grammatical tree graphs, which led to the development of a theory of ‘linguistic rhythm’ (Liberman & Prince 1977) and to metrical phonologies. In metrical phonologies, the numerical encodings are visualised as icon-like column charts (grids), and provide a visualisation of the numerical encodings and their context-determined modifications.

Autosegmental phonology essentially took up the traditional Firthian concept of a prosody as a sound feature trajectory in parallel with words, syllables and phonemes (Firth 1948), also taking tones in Niger-Congo languages as the starting point, and later applying the principle to intonations. The parallel streams and inter-stream associations of autosegmental morphophonology have been modelled formally as regular relations and implemented as finite state transducers (FSTs) by Kay 1987, fulfilling the linearity condition.

Perhaps the currently most widely represented and productive phonology paradigm is the family of optimality theoretic (OT) phonologies, a set of increasingly complex and diverse phonologies which share a component GEN which generates hypotheses about linguistic phonetic representations, a component CON with ordered violable constraints and a component EVAL which uses the constraints to filter out the most suitable surface candidate; the components function in a kind of generate-and-test strategy. The prosodic constraints described in OT studies have mainly dealt with marked and unmarked alignment of prosodic and sentence boundaries, with few exceptions such as tone-sequencing constraints (Cassimjee & Kisseberth 1998), reflecting tonal assimilation constraints related to terracing (cf. Gibbon 1987, Connell 2001) and OCP, the Obligatory Contour Principle (Leben 1973, Goldsmith 1976).

The optimality phonologies are often characterised as ‘declarative’ or ‘non-derivational’, meaning that an ordered rule system, as in SPE, is not used and constraints are applied simultaneously. However, this claim is an artefact of a high level of abstraction from precise detail and not relevant for empirical embodiments of the theory in processing, measurement and implementation in terms of operational adequacy. On closer inspection, the optimality phonologies turn out to be derivational. First, the constraints are ranked, and changing the order changes the output, a clear demonstration of a derivational procedure. Second, the GEN component applies rules to generate output candidates, another derivational property. Third, the application of SPE-like pattern-matching rules

[3] The FS diagram was automatically generated with single phonemes representing distributional classes, for reasons of space).

(Karttunen 1998) which parse and pair constraints and candidates, resulting in an acceptance or rejection score, also constitutes a derivational subsystem. Fourth, more recent versions of OT have a cycle of repeated generation and filtering, clearly derivational.

Sometimes critique of this kind is dismissed as ‘engineering’ in contrast to ‘science’, without consideration of the roots of engineering in formal theories or its roles in many disciplines for theory support and validation. Or such points are deprecated as ‘implementation details’. But, first, algorithmic operational adequacy is not yet implementation, and, second, implementation means empirical validation with hardware and software and actual speech: ‘if it ain’t tested, it don’t work’, as the saying goes. A system without operational adequacy is by definition untested except in terms of subjective intuitions in the domain of categorial and rubber time (cf. Table 1).

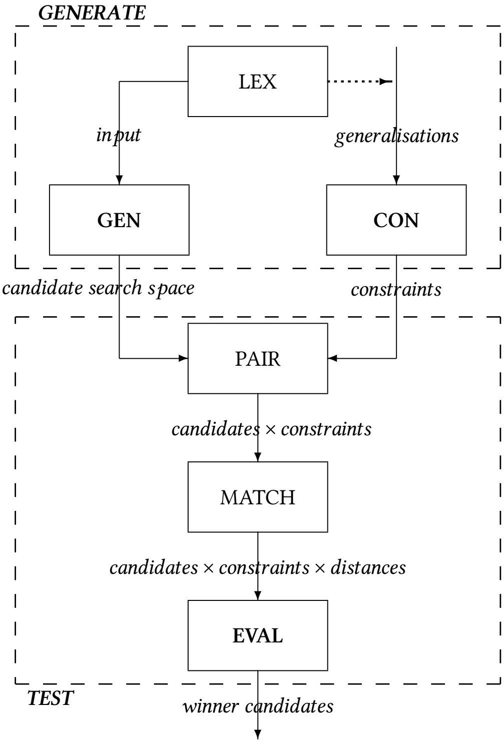

Figure 8: Operational outline of Optimality Phonology as Generate and Test Search.

To illustrate this point, the infographic in Figure 8 was designed, partly based on published OT diagrams, partly as an interpretation based on formal computational criteria, resulting in a schematic which is more detailed than the sketches in the literature. The figure makes explicit several components: an underlying lexicon, as the source of both input for candidate generation by ordered rules in GEN and of generalisations for the CON constraint set. The most abstract component is EVAL, for which candidates have to be paired, parsed and matched, and assigned a cost score determining acceptance or rejection in a procedure akin to the use of language models in speech recognition.

The overall structure of the optimality phonologies resembles the classic search strategy 'generate and test', though the details obscure this characteristic: solutions (candidates) are proposed simultaneously or sequentially, and rejected until an acceptable solution is found. Constraint resolution is termed 'parallel' or 'simultaneous', but from a search-theoretic perspective this means just memory-intensive and breadth-first as opposed to time-intensive and depth-first.

It is hard to decide if this complex system has linear properties. If, in principle, the cascade of rules and constraints consists of finite state automata, then it can be collapsed into a single deterministic automaton (Karttunen 1998). But there are studies of the finite state properties of optimality theories (Hao 2019, 2024), which indicate that some optimality phonologies do not satisfy linearity conditions.

8 IPO and Pierrehumbert: sequential working models

There have been numerous acoustic models of different aspects of prosody, from intonation through pitch accents and tones to prominence parameters. Three stand out as being interesting from symbol-phonetic, signal-phonetic and acoustic modelling perspectives, and for meeting engineering criteria of clear objectives, clarity of specification, realistic models, reproducibility and validation. These approaches are discussed briefly in relation to the TRIM criteria of linearity and real-time properties in a working model. The IPO model and the Pierrehumbert model are discussed first, then, in a separate section, the Fujisaki model.

The IPO (Instituut voor Perceptie Onderzoek) model (Collier & 't Hart 1971, 't Hart et al. 1990) is based on units of speech perception and represents sequences of intonation contours phonologically as pitch shape categories, modelled as straight lines on a perceptually motivated log frequency scale, and expresses 'bottom-up' empirical generalisations about basic units, whose beginnings and ends may overlap in signal-phonetic interpretation. The model is formalised as a finite state automaton ('t Hart et al. 1990), cf. Figure 9 (top). In earlier work (Collier & 't Hart 1971) the system was already formalised in FSA-equivalent rules.

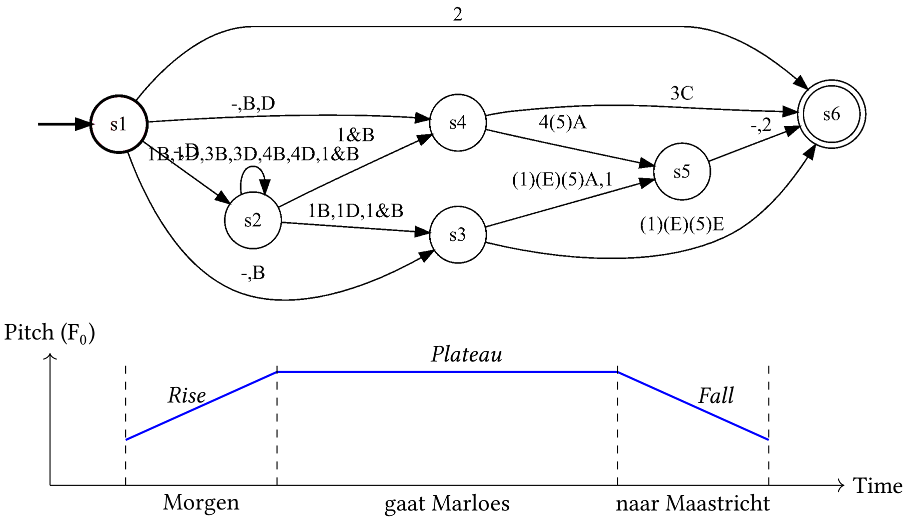

Figure 9: Top: IPO FSA for Dutch intonation (reformatted from the original non-standard format (two edge labels overlap), showing the linear characteristic of the model. Bottom: The IPO model of the Dutch 'hat' pitch pattern, consisting of a rise and a fall component connected by a level plateau, an instance of a linear intonation (specifically: intonotactic) pattern.

The main accent sequences are shown as iteration on state s2. The FSA is nondeterministic but can be straightforwardly compiled into a deterministic FSA, thus fulfilling the real-time, linearity and working model criteria for operational adequacy. The FSA can be generalised by formulating further constraints and also treating the contours as lexical items. A simple example of one pattern, the fez-like 'hat pattern', is shown in Figure 9 (bottom).

The Pierrehumbert model (Pierrehumbert 1980) is also formulated as a nondeterministic FSA, but as a 'top-down' phonological model, with abstract phonological pitch accents or 'tones' which are interpreted phonetically as high or low F_0 targets, with single or paired high or low components. Each pitch accent is marked with '*', denoting the component associated with a lexically or phrasally stressed syllable in the parallel locution. In the case of tone pairs (contour accents) either the first or the second member may receive the asterisk, e.g. $\mathrm{L}^{-}H^{*}$, $\mathrm{H}^{*}\mathrm{L}^{-}$. The model has been extended into the signal-phonetic domain (Pierrehumbert 1980, Liberman & Pierrehumbert 1984) with consideration of the relation between pitch accents and a carrier-like declining baseline.

The Pierrehumbert model was not the first finite state model of intonation, but the first phonological model, using traditional symbol-phonetic pitch-height target categories (Pike 1945, Trager & Smith 1951). Intonation was in fact discussed already by Chomsky 1965 (Chapter 1, p. 12), in connection with typical linear regular language features such as right-branching and left-branching structures. A similar but more detailed analysis is given by Reich 1969.

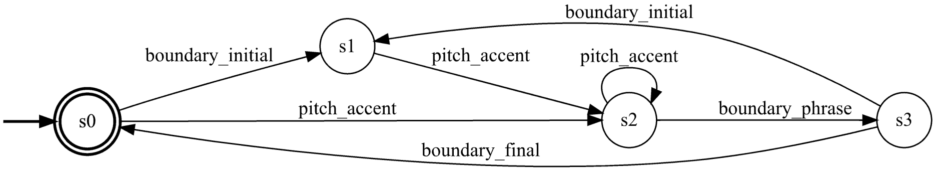

Figure 10: Determinised version of Pierrehumbert's FSA assuming a lexicon of tone categories, with tone/accent iteration and both intermediate and intonation phrase iteration, showing potential for further integration into both linguistic and modulation-theoretic contexts.

The Pierrehumbert FSA is represented in Figure 10 in a generalised and determinised format which fulfils the linearity and real-time conditions and assumes that the tonal pitch accents are collected in a lexicon and accessed on the FSA transitions. In an approximation to traditional terminology, the initial boundary tone would be referred to as a property of the Prehead, the iterated pitch accents as the Head, the intermediate phrase tone as a feature of the Nucleus and Tail of a minor tone group and the intonation group tone as a feature of the Nucleus and Tail of a major tone group (Trim 1959).

A prosodic FSA model for tone-terracing in 2-tone Niger-Congo languages, such as Baule and Tem, has been demonstrated (Gibbon 1987, 2001), in which the upper and lower sections of a tone terrace are represented as cycles around an H node and an L node respectively, with downstepped H on the transition from the L to the H node, and upstepped L on the transition from the H node to the L node.

9 Fujisaki: superpositional working model

In contrast to the strictly concatenative IPO and Pierrehumbert models, the Fujisaki model is superpositional, based on the physiology of speech production, and with modulation-theoretic properties: a carrier signal is defined as a baseline which is modulated with a superimposed phrase contour and a sequence of pitch accents (Hirose et al. 2016). Like the IPO model, the approach is 'bottom up' and data-driven, rather than phonological, though it is designed to be compatible with phonological constructs. The sequence of pitch accents and the final terminal depression has been formalised from the start in FSA-equivalent terms (Fujisaki & Nagashima 1969, Fujisaki et al. 1979, Fujisaki & Hirose 1984).

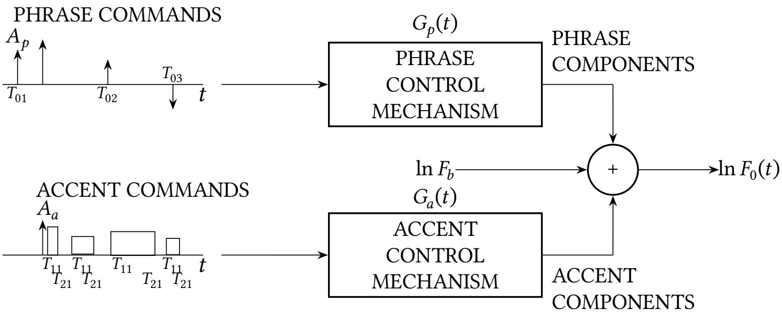

Figure 11: Schematic of the Fujisaki model superpositional architecture, showing the formal principle of parallel patterns in an explicit modulation-theoretic context.

The Fujisaki model is outlined in standard fashion in Figure 11. The Fujisaki model is very closely related to the architecture of Modulation Theory (Figure 3): a neutral, speaker specific F_0 base frequency $\ln F_b$, interpretable as a carrier signal, is simultaneously modulated by two synchronised information signals, $G_p(t)$, the phrase information signal, and $G_a(t)$, the accent information signal. The modulated output is $\ln F_0(t)$. The two synchronised information inputs to the phrase and accent modulators which produce the information signals (the phrase control mechanism and the accent control mechanism) consist of grammatically motivated and phonologically positioned phrase commands and accent commands with timing information (Möbius 1993).

Figure 11 reveals a coherent theoretical structure with explicit components. The information component of the Fujisaki model has an explicit time component and the information signals can be modelled as a regular set, with contours and accents in a regular relation, and thus linear.

The Fujisaki model dates back more than half a century. Since its inception it has been widely applied in signal phonetics and speech technology (Chien & Furui 2004, Fujisaki & Hirose 1984, Hirose et al. 2016, Mixdorff 2000, Möbius 1993), mainly in speech synthesis, also in applied phonetics, with various implementation techniques, for example Hidden Markov Models (HMMs). The claim can be made with reasonable justification that the Fujisaki model still sets a standard for the most empirically complete modulation-theoretic working model and for the most theoretically coherent currently available intonation theory.

10 Filling a gap: linguistic rhythm and real-time rhythm

Key candidates for empirical observation, description and explanation, and for operational modelling within the TRIM framework, are the physical rhythms of spoken language, which contrast starkly with the patterns observed in categorial models and those which occur in the output of operational systems. There is an interesting divergence between symbol-phonetically based phonological and signal-phonetic studies. As already noted, prosodic studies in phonology have a tendency to concentrate on "stress and linguistic rhythm" (Liberman & Prince 1977), and less on intonation (though tones are extensively dealt with). Indeed there is widespread scepticism among linguists whether rhythm is detectable at all in the signal, as opposed to rhythm as an emergent impression in perception.

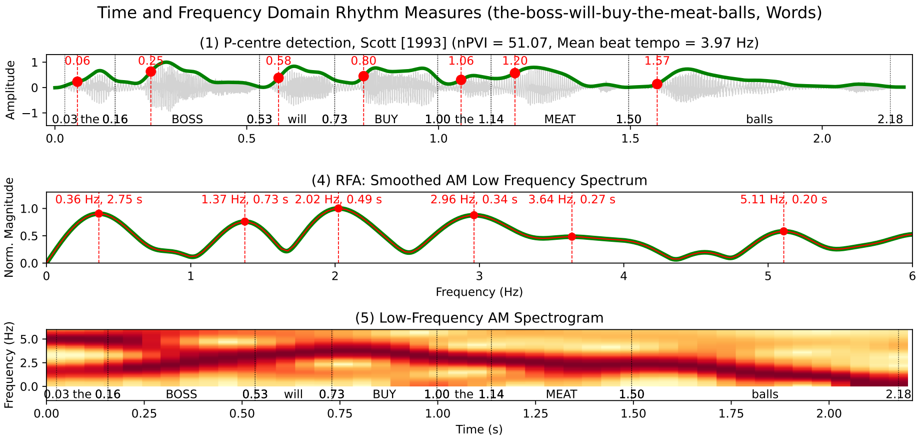

Figure 12: Time-domain and frequency-domain approaches to rhythm modelling showing stages in the development of a linear model of physical rhythm streams: (1) Scott's *p*-centre algorithm; (3) Peak detection; (3) LF spectrum with rhythm formants; (5) LF spectrogram with varying rhythm formant FM trajectory (dark line).

In signal-phonetics, on the other hand, there is a tendency to concentrate on pitch accents and intonation and less on rhythm. The models discussed above have no specific provision for speech rhythm, even though it is clear that speech models need an underlying variable 'clock' which provides the frequencies which are required for timing and synchronisation. Various models and techniques for measuring physical rhythms have been proposed, for example in cardiology with the Cosinor method (Cornelissen 2014) and in oceanography with the Empirical Mode Decomposition (EMD) method (Huang et al. 1998), and have also received attention in modulation-theoretic phonetics, mainly using FFT and wavelet analysis (Traunmüller 1994, Todd 1994 Cummins 1999, Barbosa 2002, Tilsen & Johnson 2008, Tilsen & Arvaniti 2013, Kallio et al. 2020, Gibbon 2023).

Figure 12 illustrates one time-domain and two frequency-domain methods of automatic physical rhythm analysis, using the same recorded speech data. The time-domain example in row 1 is a reconstruction of the *p*-CENTRE (perceptual centre) identification algorithm which was outlined in Scott 1993 (cf. also Cummins' related beat implementation, Cummins 1999). For the time-domain method, the *p*-centre tempo (rate, frequency) in Hz of the utterance is calculated as the inverse of the mean interval duration between the *p*-centres ($f = 1/T$). A difference measure, the $nPVI$, is used as a heuristic to assess beat-duration regularity. The results are shown in the panel titles: *p*-centre frequency of 3.97 Hz, and $nPVI$ value of 51.07. A measure based on manual annotation yielded a somewhat different $nPVI$ of 67.43, but all measures are within the expected range for English.

The frequency of time-domain beats can also be measured as peaks in the frequency domain, in the LOW-FREQUENCY SPECTRUM <6 Hz. Row 2 shows the smoothed and interpolated low-frequency spectrum with the entire input as a single FFT window, yielding formant-like RHYTHM FORMANT humps in the spectrum. The spectrum shows multiple peaks at different frequencies, not only a single frequency, representing frequencies of phrase, word, syllables, syllable parts (interval durations in parentheses): at 0.36 Hz (2.76 s), 1.38 Hz (0.73 s), 2.02 Hz (0.49 s), 2.96 Hz (0.34 s), 3.64 Hz (0.27 s), 4.56 Hz (0.22 s). Some frequencies are evidently in a near-octave relationship, an unsurprising result for language units with binary structures: if binary wholes have rhythms, the parts have rhythms with double the frequency. Tree structures representing phonetic correlates of the prosodic hierarchy can be induced from the frequencies of the rhythm formants (Gibbon 2006, Gibbon 2018).

The frequency domain spectrum model is more informative than current time domain models, but neither approach fulfils the TRIM criterion of operational adequacy. A single beat or peak or *p*-centre is not a rhythm, nor are two — a rhythm consists of at least 3 beat (peak, *p*-centre) cycles which enclose two potentially isochronous periods. In the low frequency spectrum, which has no time dimension, a spectral peak can be caused by a sequence of beats, for example as a rhythm, or simply by a single beat or pair of beats, which are not rhythms.

Therefore it is necessary to reintroduce the time dimension (Table 1) in the form of a LOW-FREQUENCY SPECTROGRAM. The LF spectrogram is shown in panel 3 of Figure 12, (with a 1 s moving spectral slice FFT window, 0.05 hop ratio, with end-reflection of the signal). Evidently the many frequencies in the spectrum in the second panel are not valid for the entire signal, but are distributed at different points along the signal timeline as variable RHYTHM FORMANT-BAND time functions. A ‘rhythm formant band’ emerges, varying in frequency, becoming higher at the beginning of rhythm formant band, then showing a faster beat sequence followed by a much slower, i.e. lower, frequency, followed by a another faster rhythm, then decreasing in tempo until the end. These patterns in the spectrogram can be informally verified by comparison with the p-centre annotation. There is thus no principled reason why ‘rhythm clock’ mechanisms, modified by very slow low frequency FM, which would account for such results, should not be incorporated explicitly into models of intonation in the interest of operational adequacy.

Several applications of the rhythm formant-band spectrogram method have been made (Gibbon 2018, 2022, 2023, 2024a,b), demonstrating rhythms at different ranks (cf. Table 2) and in comparisons of languages, styles and registers. Applications are foreseen in analysis of emotion variation and of realistic rhetoric, story-telling or poetry reading, and in long-domain speech synthesis. Mel spectrograms are used in modern transformer-based generative systems, but whether there is a focus on rhythm-relevant LF components is not clear.

11 Conclusions

The TRIM (Time Type, Rank-Interpretation, Modulation) framework was developed as an integrative approach to account for and bridge differences between signal-phonetic approaches to prosody and phonology-related symbol phonetic approaches and categorial or rubber time models. On the signal-phonetic side, the Speech Frequency Scale model within Speech Modulation Theory and, on the symbol-phonetic side, the Rank-Interpretation hierarchy of structural categories, with semantic-pragmatic, prosodic and multimodal interpretations, together demonstrate the complementarity of categorial and physical methodologies.

An innovative feature of the TRIM framework is the guiding principle of operational adequacy, i.e. the specification of data structures, algorithms and their physical properties, with implementations, as an additional metatheoretical criterion alongside explanatory, descriptive and observational adequacy (Figure 1). The linear real-time properties of operational ‘working models’ of speech prosody, provide answers to complexity issues and unrealistic claims about the strong recursivity of prosody, which is conspicuously absent in spontaneous speech. Properties of symbol-phonetic prosodic models were discussed, including traditional applied phonetic models and derivational phonological models, followed by brief discussion of the linearity of the concatenative IPO and Pierrehumbert models of Dutch and English, respectively, and of the modulation-theoretic superpositional Fujisaki model. Particular attention was directed towards the Fujisaki model, which comes close to being a standard basic working model of *un système où tout se tient*.

A preliminary conclusion can be drawn, based on the architecture of the Fujisaki model with a baseline carrier signal and two superimposed information signals, the long-term phrase component and a series of short pitch accents. Given a foundation in modulation theory, the old controversy between linear and superimposed can be formally resolved: both the phrase component and the pitch accent series are simply FM: linear frequency modulations of a carrier signal. Whether the Fujisaki baseline is the carrier signal, or itself a modulation of a simpler and perceptually motivated carrier signal (Traunmüller 1994), does not change the picture: frequency modulation by partly synchronised parallel signals is still linear in terms of real time.

Finally, an open issue, the paucity of treatments of physical speech rhythm was discussed, and methods of signal-phonetic rhythm analysis were sketched as a step to fulfilling the TRIM framework criteria, showing that the path is open for explicitly incorporating underlying 'rhythm clock' models with FM variation into intonation models with a modulation-theoretic foundation.

It might be objected that current transformer models, as applied to prodigiously 'big data' in large language models to produce AI-powered speech input and output, render discussions of the kind followed in this study irrelevant. Not so. It has sometimes been suggested that LLMs have been likened to finite state automata with large but finite working memory (though the working memory is tiny in comparison to the background knowledge network). In this sense an LLM is linear, and close to the TRIM framework and to neural models of prosody and cognition (cf. contributions to Meyer & Strauß 2026) than to informal categorial models. That they are real-time capable, when trained, seems evident and their behaviour resembles human behaviour, while categorial models exclude such issues.

But such models require empirical token information, whether from annotations and LLM-compatible acoustic codes or other kinds of signal analysis (Latif et al. 2023, Xie et al. 2024) for realistic results, for example in the development of therapeutic and didactic tools. Mathematical models with similarities to the Fujisaki model, with mel spectrograms and HMM techniques, are already used in front ends to transformer models, with code-chunk pretraining, and could benefit from tokenising discourse-rank rhythm-formant-band AM and FM modulations of time types and metalocutionary rank-interpretation mappings, such as those outlined in the present context of the TRIM framework.

References

Allen, James F. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM* 26(11). 832–843. DOI: 10.1145/182.358434.

Armstrong, Lilias Eveline & Ida Caroline Ward. 1926. *Handbook of English intonation*. Berlin: B. G. Teubner.

Barbosa, Plinio. 2002. Explaining cross-linguistic rhythmic variability via a coupled-oscillator model for rhythm production. In Bernard Bel & Isabel Marlien (eds.), *Proceedings of speech prosody 2002*, 163–166. Aix-en-Provence: Laboratoire Parole et Langage.

Beckman, Mary E. & John Kingston. 1990. Introduction. In John Kingston & Mary E. Beckman (eds.), *Papers in Laboratory Phonology I: Between the grammar and physics of speech*, 1–16. Cambridge: Cambridge University Press.

Beesley, Kenneth & Lauri Karttunen. 2003. *Finite State Morphology*. Stanford, CA: CSLI Publications.

Bird, Steven. 1995. *Computational Phonology: A constraint-based approach* (Studies in Natural Language Processing). Cambridge, UK: Cambridge University Press.

Bird, Steven & Ewan Klein. 1990. Phonological events. *Journal of Linguistics* 26(1). 33–56.

Browman, Catherine P. & Louis Goldstein. 1986. Towards an articulatory phonology. *Phonology Yearbook* 3. 219–252.

Campbell, W. Nicholas. 1992. *Multilevel speech timing control*. Completed September 1992; degree conferred June 1993; supervisor: Dr. S. D. Isard. Sussex, UK: University of Sussex, Department of Experimental Psychology. (Doctoral dissertation).

Carson-Berndsen, Julie. 1998. *Finite state models and event logics in speech recognition*. Dordrecht & Boston: Kluwer Academic Publishers.

Cassimjee, Farida & Charles W. Kisseberth. 1998. Optimality Domains Theory and Bantu tonology: A case study from isiXhosa and Shingazidja. In Larry M. Hyman & Charles W. Kisseberth (eds.), *Theoretical aspects of Bantu tone*, 33–132. Stanford, CA: CSLI Publications.

Chao, Yuen Ren. 1968. *Language and symbolic systems*. Cambridge: Cambridge University Press.

Chien, Jen-Tzung & Sadaoki Furui. 2004. F_0 modeling in HMMbased speech synthesis using the Fujisaki model. In *ICASSP 2004 – IEEE International Conference on Acoustics, Speech and Signal Processing*, I341–I344.

Chomsky, Noam. 1957. *Syntactic structures*. The Hague: Mouton & Co.

Chomsky, Noam. 1964. *Current issues in linguistic theory*, vol. 38 (Janua linguarum, Studia memoriae Nicolai van Wijk dedicata, Series minor). Den Haag: Mouton.

Chomsky, Noam. 1965. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, Noam & Morris Halle. 1968. *The Sound Pattern of English*. Cambridge, MA: MIT Press.

Chomsky, Noam & Marcel P. Schützenberger. 1963. The algebraic theory of contextfree languages. In P. Braffort & D. Hirschberg (eds.), *Computer programming and formal systems*, 118–161. Amsterdam: North Holland.

Church, Kenneth W. 1980. *On memory limitations in Natural Language Processing*. M.S. thesis, MIT Department of Electrical Engineering and Computer Science. Cambridge, MA: Massachusetts Institute of Technology. (MA thesis). https://groups.csail.mit.edu/medg/ftp/church/ken-church-ms.pdf.

Collier, René P. G. & Johan 't Hart. 1971. A grammar of pitch movements in Dutch intonation. *IPO Annual Progress Report* 6. 17–21.

Connell, Bruce. 2001. Downdrift, downstep, and declination. In Ulrike Gut & Dafydd Gibbon (eds.), *Typology of African Prosodic Systems Workshop (TAPS)*. Bielefeld: Applied Phonetics, Bielefeld University.

Cornelissen, Germaine. 2014. Cosinor-based rhythmometry. *Theoretical Biology and Medical Modelling* 11(16). 1–24.

Couper-Kuhlen, Elizabeth & Margret Selting. 2018. *Interactional Linguistics: Studying language in social interaction*. Cambridge: Cambridge University Press.

Crystal, David. 1969. *Prosodic systems and intonation in English*. Cambridge: Cambridge University Press.

Cummins, Fred. 1999. *Instructions for estimating the location of beats in a soundfile*. https://cspeech.ucd.ie/Fred/beatExtraction.php.

Cutler, Anne. 2012. *Native listening: Language experience and the recognition of spoken words*. Cambridge, MA: MIT Press.

Dilley, Laura C. 2005. *The phonetics and phonology of tonal systems*. Massachusetts Institute of Technology, Dept. of Linguistics & Philosophy. (Doctoral dissertation). http://dspace.mit.edu/handle/1721.1/30274.

Everett, Daniel L. 2005. Cultural constraints on grammar and cognition in Pirahã. *Current Anthropology* 46(4). 621–646. DOI: 10.1086/431525. https://www.journals.uchicago.edu/doi/10.1086/431525.

Firth, John Rupert. 1948. Sounds and prosodies. *Transactions of the Philological Society*. 127–152.

Fujisaki, Hiroya & Keikichi Hirose. 1984. Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan (E)* 5(4). 233–241.

Fujisaki, Hiroya, Keikichi Hirose & Kazuhiko Ohta. 1979. Acoustic features of the fundamental frequency contours of declarative sentences in Japanese. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics*. 163–173.

Fujisaki, Hiroya & Shigeo Nagashima. 1969. A model for the synthesis of pitch contours of connected speech. *Annual Report of the Engineering Research Institute* 28. 53–60.

Gibbon, Dafydd. 1976. *Perspectives of intonation analysis*. Doctoral thesis, Göttingen University. Berne, Frankfurt/M, Munich: Herbert Lang, Peter Lang.

Gibbon, Dafydd. 1981. Metalocutions, structural types and functional variation in English and German. *Papers and Studies in Contrastive Linguistics* 12. 17–39.

Gibbon, Dafydd. 1987. Finite state processing of tone systems. In *Proceedings of the Third Conference of the European Association for Computational Linguistics (EACL)*, 291–297. European Association for Computational Linguistics.

Gibbon, Dafydd. 1992a. ILEX: A linguistic approach to computational lexica. *Zeitschrift für Dialektologie und Linguistik, Beiheft Computation Linguae* 73. 32–53.

Gibbon, Dafydd. 1992b. Prosody, Time Types, and Linguistic Design Factors in spoken language system architectures. In Günther Görz (ed.), *Konvens 92, 1. Konferenz „Verarbeitung natürlicher Sprache“ Nürnberg, 7.–9. Oktober 1992*, 90–99. Best Paper Award.

Gibbon, Dafydd. 2001. Finite state prosodic analysis of African corpus resources. In Paul Dalsgaard et al. (eds.), *EUROSPEECH 2001, 7th European Conference on Speech Communication and Technology, 2nd INTERSPEECH*, 83–86. Aalborg.

Gibbon, Dafydd. 2006. Time Types and Time Trees: Prosodic mining and alignment of temporally annotated data. In Stefan Sudhoff, Denisa Lenertova, Roland Meyer, Sandra Pappert, Petra Augurzky, Ina Mleinek, Nicole Richter & Johannes Schließer (eds.), *Methods in empirical prosody research*, 209–281. Berlin: Walter de Gruyter.

Gibbon, Dafydd. 2018. The future of prosody: It’s about time. In Katarzyna Klessa, Jolanta Bachan, Agnieszka Wagner & Maciej Karpinski (eds.), *Proceedings of Speech Prosody 2018*, 1–9.

Gibbon, Dafydd. 2022. Speech rhythms: Learning to discriminate speech styles. In *Proceedings of Speech Prosody 2022*, 302–306. SProSIG.

Gibbon, Dafydd. 2023. The rhythms of rhythm. *Journal of the International Phonetic Association* 53(1). 233–265.

Gibbon, Dafydd. 2024a. Cohesive rhythms: Choral narrative in Ega. In Edward Gibson & Moshe Poliak (eds.), *From fieldwork to linguistic theory: A tribute to Dan Everett* (Empirically Oriented Theoretical Morphology and Syntax), 85–110. Berlin: Language Science Press. DOI: 10.5281/zenodo.12665911.

Gibbon, Dafydd. 2024b. Time, cohesion, style: Rhythm formants in oral narrative. In Lars Meyer & Antje Strauß (eds.), *Rhythms of speech and language: Physiology, cognition, culture*. Cambridge: Cambridge University Press.

Gibbon, Dafydd & Flaviane Romani Fernandes. 2005. Annotation-mining for rhythm model comparison in Brazilian Portuguese. In *Proceedings of Interspeech 2005*, 3289–3292. Lisboa, Portugal. DOI: 10.21437/Interspeech.2005-8600.

Gibbon, Dafydd & Sascha Griffiths. 2017. Multilinear Grammar: Ranks and Interpretations. *Open Linguistics* 3(1). 265–307.

Gibbon, Dafydd, Inge Mertins & Roger K. Moore. 2000. *Handbook of Multimodal and Spoken Dialogue Systems. Resources, Terminology and Product Evaluation* (The Kluwer International Series in Engineering and Computer Science). Dordrecht & New York: Kluwer & Springer.

Gibbon, Dafydd, Roger Moore & Richard Winski (eds.). 1997. *Handbook of Standards and Resources for Spoken Language Systems*. Berlin: Mouton de Gruyter.

Gibbon, Dafydd & Helmut Richter. 1984. *Intonation, accent and rhythm: Studies in Discourse Phonology*. Berlin: Walter de Gruyter.

Goldsmith, John A. 1976. *Autosegmental Phonology*. Massachusetts Institute of Technology. (Doctoral dissertation). http://dspace.mit.edu/handle/1721.1/16388.

Halliday, Michael A. K. 1967. *Intonation and grammar in British English*. The Hague: Mouton.

Hao, Sophie Yiding. 2019. Finite-state Optimality Theory: Non-rationality of Harmonic Serialism. *Journal of Language Modelling* 7(2). 49–99.

Hao, Sophie Yiding. 2024. Universal generation for Optimality Theory is PSPACE-Complete. *Computational Linguistics* 50(1). 83–117.

’t Hart, Johan, René Collier & Antonie Cohen. 1990. *A perceptual study of intonation: An experimental-phonetic approach to speech melody*. Cambridge: Cambridge University Press.

Hirose, Keikichi, Hiroya Hashimoto, Daisuke Saito & Nobuaki Minematsu. 2016. Superpositional modeling of fundamental frequency contours for HMM-based speech synthesis. In *Proceedings of Speech Prosody 2016*, 771–775. Boston, MA, USA. DOI: 10.21437/SpeechProsody.2016-158.

Hirst, Daniel. 2024. *Speech prosody: From acoustics to interpretation*. Berlin: Springer.

Huang, Norden E., Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Qiang Zheng, Nai-Chyuan Yen, Chi Chao Tung & Henry H. Liu. 1998. The Empirical Mode Decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* 454(1971). 903–995. DOI: 10.1098/rspa.1998.0193.

Hudson, Roxanne F., Holly Lane & Paige C. Pullen. 2005. Reading fluency assessment and instruction: What, Why and How? *The Reading Teacher* 58(8). 702–714.

Jakobson, Roman, Gunnar Fant & Morris Halle. 1952. *Preliminaries to speech analysis: The distinctive features and their correlates*. Cambridge, MA: MIT Press.

Johnson, Douglas. 1972. *Formal aspects of phonological description*. Den Haag: Mouton.

Kallio, Heini, Antti Suni, Juraj Šimko & Martti Vainio. 2020. Analyzing second language proficiency using wavelet-based prominence estimates. *Journal of Phonetics* 80. 1–12.

Kaplan, Ronald M. & Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics* 20(3). 331–378.

Karttunen, Lauri. 1998. On the proper treatment of optimality in computational phonology. In *Proceedings of the International Workshop on Finite-State Methods in Natural Language Processing (FSMNLP)*, 1–12.

Kay, Martin. 1987. Nonconcatenative Finite-State Morphology. In *Third Conference of the European Chapter of the Association for Computational Linguistics*, 2–10. Copenhagen: Association for Computational Linguistics. https://aclanthology.org/E87-1002/.

Keating, Patricia A. 1990. Phonetic representations in a generative grammar. *Journal of Phonetics* 18(3). 321–334.

Kingdon, Roger. 1958. *The groundwork of English intonation*. London, New York, Toronto: Longmans, Green & Co.

Kornai, Andras. 1985. Natural languages and the Chomsky Hierarchy. In Maghi King (ed.), *Proceedings of the 2nd European Conference of the Association for Computational Linguistics*, 1–7.

Kornai, Andras. 1987. Finite State Semantics. In Ursula Klenk, Peter Scherber & Manfred Thaller (eds.), *Computerlinguistik und philologische Datenverarbeitung*, 59–70. Hildesheim: Georg Olms Verlag.

Latif, Siddique, Aun Zaidi, Heriberto Cuayahuitl, Fahad Shamshad, Moazzam Shoukat, Muhammad Usama & Junaid Qadir. 2023. Transformers in speech processing: A survey. https://arxiv.org/abs/2303.11607.

Leben, Will. 1973. *Suprasegmental Phonology*. Cambridge, MA: Massachusetts Institute of Technology. (PhD dissertation).

Lehiste, Ilse. 1970. *Suprasegmentals*. Cambridge, MA: M.I.T. Press. viii, 194.

Levelt, Willem J. M. 1983. Monitoring and self-repair in speech. *Cognition* 14(1). 41–104. DOI: 10.1016/0010-0277(83)90026-4.

Liberman, Mark & Janet Pierrehumbert. 1984. Intonational invariance under changes in pitch range and length. In Mark Aronoff & Richard Thomas Oehrle (eds.), *Language sound structure*, 157–233. Cambridge, MA: MIT Press.

Liberman, Mark & Alan Prince. 1977. On stress and linguistic rhythm. *Linguistic Inquiry* 8(2). 249–336.

Lin, Xuewei & Dafydd Gibbon. 2023. Distant rhythms: Computing fluency. In *Proceedings of the International Congress of Phonetic Sciences*, 4219–4223. Prague: Charles University.

Meyer, Lars & Antje Strauß (eds.). 2026. *Rhythms of speech and language: Physiology, cognition, culture*. Cambridge: Cambridge University Press.

Mixdorff, Hansjörg. 2000. A novel approach to the fully automatic extraction of Fujisaki model parameters. In *Proceedings of ICASSP 2000*, vol. 3, 1281–1284. Istanbul, Turkey.

Möbius, Bernd. 1993. *Ein quantitatives Modell der deutschen Intonation: Analyse und Synthese von Grundfrequenzverläufen*. Universität Bonn. (Doctoral dissertation).

O'Connor, Joseph Desmond & Gordon Frederick Arnold. 1961. *Intonation of Colloquial English*. London: Longman.

Ohala, John J. 1990. There is no interface between phonology and phonetics: A personal view. *Journal of Phonetics* 18(2). 153–171.

Pierrehumbert, Janet Breckenridge. 1980. *The Phonology and Phonetics of English Intonation*. Massachusetts Institute of Technology. (Doctoral dissertation). http://dspace.mit.edu/handle/1721.1/16065.

Pike, Kenneth L. 1947. *Phonemics: A technique for reducing languages to writing*, vol. 3 (University of Michigan Publications Linguistics). Ann Arbor: University of Michigan Press.

Pike, Kenneth Lee. 1945. *The intonation of American English*. Ann Arbor: University of Michigan Press.

Reich, Peter. 1969. The finiteness of natural language. *Language* 45(4). 831–843.

Scott, Sophie K. 1993. *P-centers in speech: An acoustic analysis*. University College London. (Doctoral dissertation).

Selkirk, Elisabeth O. 1984. *Phonology and syntax. the relation between sound and structure*. Cambridge, MA: MIT Press.

Tillmann, Hans-G. & Phil Mansell. 1980. *Lautsprachliche Zeichen, Sprachsignale und lautsprachlicher Kommunikationsprozeß*. Stuttgart: Klett-Cotta.

Tilsen, Sam & Amalia Arvaniti. 2013. Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *Journal of the Acoustical Society of America* 134(1). 628–639.

Tilsen, Sam & Keith Johnson. 2008. Low-frequency Fourier analysis of speech rhythm. *Journal of the Acoustical Society of America* 124(2). 34–39.

Todd, Neil P. McAngus. 1994. The auditory "Primal Sketch": A multiscale model of rhythmic grouping. *Journal of New Music Research* 23(1). 25–70. DOI: 10.1080/09298219408570647.

Trager, George Leonard & Henry Lee Smith. 1951. *An outline of English structure*. Washington, D.C.: American Council of Learned Societies.

Traunmüller, Hartmut. 1994. Conventional, biological, and environmental factors in speech communication: A modulation theory. *Phonetica* 51(1-3). 170–183.

Trim, John L. M. 1959. Major and minor tone groups in English. *Le Maître Phonétique* 112. 26–29.

Trubetzkoy, Nikolai Sergejewitsch. 1939. *Grundzüge der Phonologie*. 1st edn. Prague: Travaux du Cercle Linguistique de Prague.

Wang, Bojia & Dafydd Gibbon. 2024. Duration and declination in L2 reading. In *Proceedings of the 5th International Symposium on Applied Phonetics (ISAPh 2024)*, 125–130. DOI: 10.21437/ISAPh.2024-24.

Whorf, Benjamin Lee. 1940. Linguistics as an exact science. *The Technology Review (Massachusetts Institute of Technology)* 4(3). (FSA model), 61–83.

Xie, Tianxin, Yan Rong, Pengfei Zhang, Wenwu Wang & Li Liu. 2024. Towards controllable speech synthesis in the era of large language models: A survey. Equal contribution: Xie, Rong, Zhang. https://arxiv.org/abs/2412.06602.

Yu, Jue, Dafydd Gibbon & Katarzyna Klessa. 2014. Computational annotation-mining of syllable durations in speech varieties. In *Proceedings of Speech Prosody 2014*, 443–447. DOI: 10.21437/SpeechProsody.2014-76.