Pier Marco Bertinetto & Chiara Bertini Scuola Normale Superiore, Pisa

Towards a unified predictive model of Speech Rhythm

(paper submitted to a journal)

To Olle Engstrand, rhythmically

1. Epistemological requirements

Research on Speech Rhythm (SR) entered a new phase around the turn of the new Millennium, when an entirely new algorithm to compare the rhythmical inclination of individual languages was proposed (see Ramus et al. 1999). The suggestion was soon followed by other scholars, suggesting revised or modified versions. The present authors will not even attempt at quoting them all. Among the revised versions, one should especially consider the Varco (Dellwo 2004) and the "semi-syllable" models (Rouas & Farinas 2004). These, like the Ramusian proposal, may be called "static" models, for the actual sequence of the relevant intervals (consonantal and vocalic) does not play a role. The results are not affected by any possible permutation of intervals; the algorithms provide an overall measure characterizing any speech stretch in its entirety, be it the standard variation of the relevant interval's duration, their mean error, the global percent value etc. Among the modified versions, it is worth mentioning the method proposed by Wagner (2007) and most notably the PVI model (Grabe & Low 2002); the latter should be characterized as "dynamic", insofar as it takes into account the local durational fluctuations between any two adjacent intervals.

Despite the merit of revitalizing the topic of SR, all these recent proposals seem to be somewhat defective on epistemological grounds. In order to grasp this, let us list the three requirements that any SR theory should fulfill, namely: (a) EXPLICITNESS, (b) PREDICTIVITY, (c) UNIFICATION. The first two are self-explaining; the third is strictly related to this particular research domain. The succinct survey that follows has no historiographic ambition; it is only meant to show that none of the models so far proposed (with one exception) fulfill all three requirements. The proposal put forth in this paper aims at remedying this fault.

Pike (1947) was a good start. The theory was perfectly explicit and predictive. It stated that languages belong to two types, each characterized by isochronicity within a specific domain: the syllable or the accentual phrase (the latter to be intended as the inter-stress interval, i.e. the stretch comprised between the onset of a stressed syllable – or, alternatively, vowel – and the next one): hence, the contrast SYLLABLE- vs. STRESS-TIMING. This theory should be praised for its explicitness. The crude linguistic facts soon falsified it (for a more recent disconfirmation, see Van Santen & Shih 2000), but one should take this as welcome result: falsified theories pave the way for better ones. There is another reason to be grateful to Pike: he pointed out the way towards the experimental testing of a prominent linguistic feature, something that still keeps people busy. As for the third requirement (unification), the Pikean theory was obviously orthogonal to it, for it postulated that languages belong to two radically alternative types. We take this to be a major flaw, for assuming the existence of mutually unrelated rhythmical types looks unattractive. One should rather start from the assumption that all natural languages share the same structural features: the differences should best be conceived of in terms of degrees along a continuum, rather than as irreconcilable.

The Pikean model's failure gave rise to a number of attempts to save its basic intuition (for a detailed survey, mirroring the situation at the end of the Eighties, cf. Bertinetto 1989). Once it was ascertained that the original formulation did not correspond to the facts, the solution was sought in other directions, among which, most notably: (i) perceptual constructs feeding impressionistic judgments (see references in Bertinetto 1989); (ii) syllabic duration compensation in the word or accentual domain (Lindblom & Rapp 1973, soon followed by others). The phonologically-oriented proposals by Bertinetto (1981) and Dauer (1983, 1987) also deserve mention: they pointed out a number of prosodic features variously feeding the rhythmical classification of languages, including – among others – the following: (a) V-reduction *vs*. full articulation in unstressed syllables; (b) complex *vs*. simple syllable structure; (c) relative flexibility *vs*. rigidity in word-stress placement; (d) tempo acceleration mainly due to compression of unstressed syllables *vs*. proportional compression. It immediately appears that the latter proposals presupposed a unified theory. Unfortunately, however, they were both wanting in explicitness and predictivity: although the features indicated (or a subset of them) are likely to have a bearing on SR, their exact contribution was not spelled out. Altogether, the intermediate post-Pikean period might be characterized as a time for rethinking: lacking a predictive theory, the main effort was put into trying to collect arguments conducive to a unified theory, based on a broad typological view of the prosodic systems of natural languages.

The most recent models, although differing in the details, share one fundamental feature with the Pikean model: they are all explicit, for they offer algorithms capable of generating the desired segregation of the alleged syllable- vs. stress-timed languages. Whether they also exhibit predictivity, is another matter. In a sense, they should be regarded as at least weakly predictive, due to their explicitness. However, they cannot be regarded as fully (or strongly) predictive, for they are reticent on unification issue. To avoid misunderstanding, one should add that the latter remark should not be read as referring to the position actually maintained by the individual models' proponents: what is meant here is that the models as such do not allow any specific inference as to whether the theory presupposes a unified design, or a bi-modal one based on radically alternative rhythmical types. Since the authors do not state what the alleged rhythmical contrast should be based on, it is impossible to shed light on the issue. Actually, considering that most scholars agree that languages cluster around two rhythmical types, one might even suppose that this should be accepted as a basic postulate. But scientific enterprises cannot merely stem from intuition. The weakness of this state of affairs is obvious. In the absence of explicit predictions at the outset, the recent SR models run a severe risk of circularity: any such algorithm is claimed to be working fine whenever it produces the correct grouping (where "correct" can only mean "consistent with the experimenter's expectations"). Thus, the interpretation can only arise *post factum*, in terms of relative positioning. The models yield a topological arrangement, whereby languages of group A vs. B (whose existence is assumed, rather than independently explained) are shown to occupy different areas on the Cartesian plane. This, however, does not tell us anything about the actual property that a language should exhibit in principle, in order to belong to the one or the other type. Consequently, none of these models can specify which language type should occupy which portion of the graphic, depending on which finely attuned structural properties. As a further consequence, the models lack an explicit metrics to effectively measure the distance between languages; hence, the "intermediate" types are merely accepted as a classification residue, rather than predicted. The recent literature on SR abounds in sentences such as: "contrary to expectations, language X clusters with stress- rather than syllable-timed languages" or "language X is intermediate between the two types". There is nothing intrinsically wrong in this, except that belonging to the one or the other type is inferred *a posteriori* from the clustering results, rather than defined on independent grounds.

Needless to say, the above criticism is not meant to deny the validity of the general consensus on the existence of contrasting rhythmical tendencies. This does not merely stem from intuition, but is based on objective data, two of which are worth mentioning here. One source of data is the different ease with which the various languages may be couched into musical-rhythmic frames. Although any language ultimately admits this possibility, the specific ways in which this may be obtained vary a lot. A dramatic contrast of this sort is hinted at by Cummins (2002), comparing the behavior of English speakers with that of Italian and Spanish speakers. Although a detailed comparative study of the relation of words to music has not been undertaken to date, one may surmise that it would produce exciting results. Another important source of data is the different organizational basis of traditional versification systems. Each system captures the most relevant prosodic features of the given language, turning them into a (set of) organizing principle(s), such as: inter-stress distance, syllable counting, mora or syllable quantity, tone dynamics etc., often combining more than one principle. For instance, stress-syllabic systems regulate the inter-stress distances in terms of syllable counting, using foot-measures reminiscent of the Greek and Latin tradition, although the latter implemented a quantity-syllabic system. Since metrically regulated speech is intentionally aiming at rhythmicity, one is immediately drawn to the conclusion that the rhythm organizational basis differs from language to language, for otherwise every linguistic community would have adopted the same system.

Having said this, however, one should also acknowledge that no scientific enterprise can ignore its epistemological obligations. To put it succinctly: the basic intuition should first be connected to explicit structural properties on which detailed predictions can be attached; these predictions should then be tested by appropriate tools, until they are falsified. In recent SR studies, however, the reverse happened: various tools have been devised to ascertain the initial intuition concerning rhythmic typology, without previously defining the exact structural properties on which SR rests.

The general lesson to be learned from the brief survey in this section is that, although there seems to have been a constant – albeit discontinuous – progress in SR theorizing, none of the models so far developed exhibited all three epistemological properties required by this particular research domain, as summarized in the following table:

SR models	*Unification*	*Explicitness*	*Predictivity*
Pike 1947	-	+	+
Bertinetto 1981; Dauer 1983, 1987	+	-	-
Lindblom & Rapp 1973	+	+	-
Ramus, PVI, Varco	?	+	-

Table 1. Fulfillment of the epistemological requirements by selected SR models.

It follows that the most urgent task consists in devising a unified, fully explicit and predictive theory, capable of generating the appropriate expectations as for what a language should be like (and do) in order to be assigned to a given rhythmical type.

2. Towards a new model

In a recent work (Bertinetto & Bertini 2008), the present authors presented the first outline of such a SR model. The model will be further developed here. Following the example of other scholars, the traditional terminology (syllable vs. stress-timing) will be abandoned, to avoid any misunderstanding tied to its original meaning. For simplicity's sake, this model also comprises two ideal types: CONTROLLING vs. COMPENSATING (henceforth: CNTRL vs. CMPST), except that these should be conceived of as the extremes of a continuum and thus referred to for purely descriptive reasons. The terms are borrowed from Hoeqvist (1983), although the interpretation is quite different (the same terms were also used in Bertinetto & Vékás 1991, who presented an embryonic sketch of the theory developed here).

The basic idea, inspired by work in articulatory phonology and earlier on by the seminal work of Fowler (1977), is as follows: languages may differ in terms of how vocalic and consonantal gestures are coupled in the speech flow. An ideally CNTRL language should be conceived of as a language in which all segments receive the same amount of expenditure – or articulatory effort – and tend to have the same duration. This is obviously impossible, due to the varying points and manners of articulation; yet, this view acquires plausibility as soon as one considers how languages diverge in terms of the coupling of V and C gestures. Some languages admit – or rather require – a much larger segmental overlap (i.e. co-articulation, co-production) than others. Such languages correspond to the CMPST type. Here again, the ideal maximum – whereby all adjacent C and V gestures overlap entirely – is physically impossible. It should thus be immediately clear that both extremes (CNTRL / CMPST) are artificial constructs, only used to designate two ideal cases, just as the absolute Ø temperature is a physical abstraction impossible to obtain on Earth, yet needed for reference purposes.

What one actually finds in the real world are higher or lower degrees of control / compensation. This inspires the CONTROL / COMPENSATION (**CC**) hypothesis.

The actual position along the continuum depends, to a very large extent, on the phonotactic structure of the individual language. A simple phonotactics naturally inclines towards the CNTRL setting. Note that a language consisting of just one consonant and one vowel would be perfectly rhythmical (e.g., *ba-ba-ba*). This does not follow from any musical or dancing predisposition of the human beings; it is a mere "emergent" property of gestural coordination, as task-dynamics has shown for a quite some time. If rhythmicity is indeed the simplest way to cope with complex coordination problems, then language is an obvious candidate for it, for speech production involves the fine intertwining of several articulators. Languages, however, are complex organisms based on a number of (possibly competing) structural components. Not only do their phonology involve an often fairly rich segment inventory, but word and sentence prosody interact with the segmental level in a number of ways, producing in the long run all sorts of phonological restructurings. As a result, languages often present a rich phonotactics, which forces the speaker to adopt a flexible (CMPST) articulatory setting. The natural result of this is the overlapping of C and V gestures, as Goldstein et al. (2007) have empirically shown with respect to syllabic structure: languages with a simple phonotactics have a greater chance of presenting a fairly in-phase coupling of the consonantal and vocalic oscillators. Thus, departing (more and more) from the CNTRL ideal is the automatic consequence of a (more and more) complex phonotactics. The most typical sites for gestural overlap are the unstressed syllables, where the vocalic nucleus offers itself as the privileged target for co-articulation. Needless to say, unstressed syllables reduction also occurs in CNTRL languages, but to a lower extent; conversely, and crucially, intra-syllabic durational compensation is larger in CMPST than CNTRL languages, especially (but not only) in unstressed syllables. It will not go unnoticed that this view departs radically – and somehow paradoxically – from the traditional one, despite the factual coincidence of

CMPST and so-called stress-timed languages (as well as CNTRL and syllable-timed ones). This should, however, cause no surprise, considering the empirical inadequacy of the Pikean view.

The CC model here described directly fulfills, due to its very conception, one of the three fundamental requirements, namely unification. The remaining two (explicitness and predictivity) need to be satisfied by appropriate computational tools. In Bertinetto & Bertini (2008) the following modified version of the PVI algorithm, called CONTROL/COMPENSATION INDEX (**CCI**), was proposed:

$$CCI = \frac{100}{m-1}\sum_{k=1}^{m-1}\left|\frac{d_k}{n_k} - \frac{d_{k+1}}{n_{k+1}}\right| \qquad (1)$$

In practice, CCI relativizes the PVI measure to the number (*n*) of segments composing each consonantal or vocalic interval. The model thus inherits the PVI's dynamic virtue, adding to it the complexity of the phonotactic structure, for it obviously makes a difference whether a given consonantal interval comprises one or several segments.

It is important to realize that CCI is a phonologically-driven model. Geminates and long vowels count as two segments (cf. Finnish), just like two vowels in synaloepha, while hyper-long segments count as three (cf. Estonian). Conversely, vowels in hiatus count as separate (monosegmental) vocalic intervals for – as detailed in § 3 – each syllable nucleus implements a vocalic oscillator's period. While applying the CCI algorithm one should thus carefully consider the phonological structure of the languages at stake, possibly adopting a double counting in delicate cases. Glides are a case in point: their treatment as either C or V segments varies from language to language. It is thus advisable to apply the algorithm in both ways, in order to ensure cross-linguistic comparison. (It can be anticipated here that the application of this double measurement strategy to the Italian data described below produced a statistically irrelevant difference. Needless to say, languages with a much larger presence of diphthongs are expected to yield a significant contrast; in the present case, the segments involved in glides were 5.1% of the total).

In Bertini & Bertinetto (in press) the criteria adopted for the coding of a semi-spontaneous Italian corpus were spelled-out. The materials consisted of excerpts of map-task dialogues, carefully segmented and labeled (the source corpus is available at: http://www.cirass.unina.it/ricerca/studi%20parlato/raccolta%20corpora/api/api.htm/). Each excerpt was at least eight (phonetically realized) syllables in length; ten speakers were involved and the intervals numbered nearly 3000 for both Cs and Vs. One detail worth mentioning is that the final portion of any sentence, from the last stressed V onward, was neglected (in addition, any C preceding the last stressed V was trimmed, on the assumption that the final lengthening phenomenon might involve at least part of that interval). The use of "trimmed sentences" is justified by the fact that the final portion has its own (language-specific) prosodic properties as a boundary signal, that should be studied on its own independently of rhythm proper.

CCI makes explicit and directly verifiable predictions, as shown in fig. 1. Languages oriented towards the CNTRL type should fall in the proximity of the bisecting line, showing that the local fluctuation of Cs and Vs tends to be of the same magnitude, whereas CMPST languages should exhibit more V than C fluctuation: [FIGURE 1 here]

The analyses carried out by Mairano & Romano (2008) on a corpus of read speech passages, produced in 8 different languages, yielded results in line with the CCI model's predictions, as shown in fig. 2: German, American and RP English (traditionally considered stress-timed) tend to be CMPST, since they present comparatively more vocalic than consonantal local variation, as a consequence of the large amount of V-reduction in unstressed syllables. Conversely, Finnish, French, Canadian French and Italian (traditionally considered syllable-timed) lie in the vicinity of the bisecting line. Note that the data in fig. 2 stem from read speech, with the exception of those indicated as CCI, corresponding to the spontaneous data analysed in Bertinetto & Bertini (2008); the latter presumably underwent some shifting towards the CMPST pole, due to hypo-articulation. [FIGURE 2 here]

Two caveat should be pointed out. First, with the exception of CCI and IC, each point on fig. 2 refers to a single speaker. However, as may be seen in the figure for FI, Ger and IT (and as is known from previous studies, e.g.: Dellwo et al. 2005, Barbosa 2006), different speakers may do quite different things, suggesting that no generalization should be drawn from small observational bases. Second, the position of Icelandic (IC) may appear to be somehow surprising, considering that it is a Germanic language like English and German, with a comparable phonotactic richness. However, Icelandic seems to exhibit a low degree of V-reduction (Mairano, p.c.), which is compatible with its position in the figure. This datum, yet to be confirmed, suggests an important theoretical consequence: a rich phonotactics is not by itself conducive to CMPST behavior, although this is the default situation. The ultimate factor is the amount of V-reduction, which may in some cases dissociate from phonotactic richness. Further support for this dissociation is provided by Singapore English, as opposed to British English (Low 1998); Western, as opposed to Eastern, Catalan (Gavaldá & Dellwo 2008); Cantonese, as opposed to Mandarin, Chinese (Mok & Dellwo 2008). This dissociation is also to be found in L2 pronunciation of CMPST languages (e.g., English as spoken by Chinese speakers, Mok 2008; see also White & Mattys 2007). The exact articulatory setting of such language varieties should be thoroughly investigated, also regardless of the SR issue.

Speech tempo variations provide a valuable test to assess the CC hypothesis. The predictions are as follows: (i) CNTRL languages should tend to reduce the segments' duration in a rather proportional way, whereas in CMPST languages Vs should be somewhat more affected than Cs; (ii) in CNTRL languages reduction should be much sharper between slow than between fast rates, whereas in CMPST languages reduction should be relatively robust even at fast rates. The latter prediction stems from the larger articulatory flexibility of CMPST languages, allowing further freedom in terms of co-production of vocalic and consonantal gestures, while CNTRL languages meet their compressibility threshold earlier (Bertinetto & Fowler 1989; cf. also Price 1980 and Davidson 2006).

These predictions were tested against the afore-mentioned spontaneous Italian corpus (Bertini & Bertinetto, in press). The speech materials were divided into 3 naturalistically obtained tempo groups (T1, T2, T3). The assignment of each utterance to a given group was done *a posteriori*, rather than directly elicited from the speakers: this avoids any possible distortion induced by the conscious effort to comply with the experimenter's demand. The rate measures used were segments or syllables x sec.; the rate groups were obtained by evenly distributing the V and C intervals, yielding the following classes:

Segments x sec.: T1 ≤ 15,6 / av. 14.2; 15.6 < T2 ≤ 17.65 / av. 16.6; T3 > 17.65 / av. 19.2

Syllables x sec.: T1 ≤ 6.75 / av. 6.1; 6.75 < T2 ≤ 7.75 / av. 7.3; T3 > 7.75 / av. 8.9

The data reported below refine those of Bertinetto & Bertini 2008 (where rate was measured in syllables x sec. and the groups were equalized with respect to the number of utterances). Interestingly, the general trend of all models is strictly linear, with the exception of %V, nPVI(C) and Varco(C):

T1 // T2 // T3	CCI(V), Ramus(V+C), rPVI(V+C), Varco(V), RF(V+C)
T1 <> T2 <> T3	nPVI(C), Varco(C)
T1 <> T2 // T3	nPVI(V)
T1 // T2 <> T3	CCI(C), %V

Table 2a.

T1 // T2 // T3	CCI(V), Ramus(C), rPVI(C), RF(C)
T1 <> T2 <> T3	nPVI(V+C), Varco(V+C)
T1 <> T2 // T3	CCI(C), %V
T1 // T2 <> T3	Ramus(V), rPVI(V), RF(V)

Table 2b.

Statistical analysis, based on three rate classes, according to alternative rhythm models: CCI, Ramus, PVI (raw and normalized), Varco, Rouas & Farinas (RF). Speed measures: Table 2a = segments x second; Table 2b = syllables x second. The diacritics <> and // stand, respectively, for statistically 'not-separable' vs. 'separable' according to pairwise t-tests carried out on T1 vs. T2, and T2 vs. T3.

The first row in table 2a-b indicates that the relevant model is very sensitive to the rate differences as considered here; the second row, on the contrary, indicates that no difference is detected. The third row presents a rather implausible situation, whereby the contrast appears to be sharp only at fast rates; the last row, instead, suggests that the incompressibility threshold is reached between T2 and T3, which is definitely more reasonable. As it happens, the results depend heavily on the rate measure used. On the whole, the one used in table 2a (segm. x sec.) seems to work better, as shown by the fact that the third row is less populated. In addition to this, all the models mentioned in the forth row of table 2b refer to V measures only, whereas the corresponding C measures appear in the first row. This suggests, rather implausibly, that Cs are more compressible than Vs at fast rates. Considering then the results of table 2a, it appears that CCI is among the most sensitive models and, above all, it is the only one to capture the plausible propensity of Cs to attain incompressibility before Vs. With respect to the predictions spelled-out above, the picture emerging from the statistical analysis based on CCI suggests that Italian does not conform entirely to the idealized CNTRL type. Indeed, the acceleration's effects are not strictly proportional for Vs and Cs, for only the latter reach threshold in the T2 vs. T3 comparison.

Although this cannot be regarded as the last word on the matter, the results, together with the ones reported in fig. 2, look promising. One aspect of the model is especially worth highlighting here: namely, its predictive character. This enables the researcher to put forth meaningful predictions with respect to tempo variations within a single language. Inter-language comparison is a useful – and typology-wise unavoidable – perspective, but is not necessary to validate the model. This solves the circularity problem referred to in sect. 2.

3. A bi-level model of SR

The above sketch of a SR model was devised to capture the rhythmical consequences of phonotactic structure. This, however, does not exhaust the picture, for languages are based on a complex architecture. Over and above the segments' concatenation, they present overarching levels, among which ACCENTUAL PHRASES are especially relevant to the present concern. The model should thus be extended in the direction of a bi-level architecture, conceived of as two pairs of coupled oscillators, comprising:

- Level-I (PHONOTACTIC), based on the coupling of the vocalic and consonantal oscillators, along the lines suggested by Goldstein et al. (2007);

- Level-II (PHRASAL), based on the coupling of the accentual and syllabic oscillators, adopting suggestions by O'Dell & Nieminen (1999).

Pluri-level conceptions of SR have already been advanced in the literature (e.g., Barbosa 2007; O'Dell et al. 2007). The major claim for originality of the model proposed here, apart from its specific shape, lies in the possible divergence of the two levels, as detailed below.

Since, in the present model, syllable and accent are no longer regarded as the source of two alternative rhythmical tendencies, they should be regarded as basic prosodic features necessarily shared by all languages. In particular, although the phonological role of word stress differs from language to language, one may assume that phrase accents are universally present as rhythm regulators, whatever their language-specific phonetic implementation. The latter is no doubt the product of several intermixed acoustic components, as emphasized by Kohler 2008: one should, for example, note that dynamic tones – especially descending ones – yield an impression of longer duration, as opposed to static tones, adding further complications at the perceptual level. In stress languages, like English or Italian, there is an interplay between word-stresses (including secondary ones) and phrase accents: at slow rates, the latter tend to coincide with the former, whereas at faster rates only the most salient stresses are preserved. As a result, the average number of syllables per accentual phrase increases along with the tempo, preserving some sort of durational regularity.

Of paramount importance is the contrast 'rigid' vs. 'mobile' word stress. In Italian, word stress may be downplayed or even (at faster rates or in stress clashes) deleted, but – with very few exceptions – it cannot be shifted. In English, on the contrary, a large part of the lexicon may undergo optional stress shift, as in words like *coronal*, *exponent*, *contribute*, *subsidence*, *exquisite*, *satiety* etc. (also depending on specific sociolects). This may be regarded as the Level-II equivalent of the Level-I CC-divide: the more mobile the stress is, the more flexible (CMPST) the accentual structure, for the speaker may then have a larger degree of freedom in regulating the inter-accentual distances. Once again, one finds a gradient between two extremes, in accordance with the unification requirement. The underlying assumption is that speakers follow their spontaneous inclination towards rythmicity as long as the language intricacies do not constrain their behavior. Word stress rigidity is such a constraint at Level-II, just as rich phonotactics is at Level-I.

It is important to realize, however, that the CC parameter does not necessarily converge at both levels. The interaction may be complex, due to the vagaries of linguistic typology. The examples in the table below may not be the most prototypical ones, but will suffice for the present purpose:

TYPE	*LEVEL-I*	*LEVEL-II*	*EXAMPLE*
1	CNTRL	CNTRL	*Italian*: relatively simple phonotactics, fairly rigid word stress pattern
2	CMPST	CMPST	*English*: fairly complex phonotactics, fairly mobile word stress pattern, density of secondary stresses yielding further prominence sites
3	CMPST	CNTRL	*Polish*: very complex phonotactics (Bertinetto et al. 2007), rigid word stress pattern
4	CNTRL	CMPST	*Japanese*? *Chinese*? (see the text for comments)

Table 3. Interplay of Level-I and Level-II with respect to the CC contrast.

To avoid confusion, one should speak of CNTRL-1, CMPST-2 etc., with integers referring to the appropriate level. Needless to say, several other components may cooperate to yield the final result, most notably word structure. For instance, a language with many polysyllables and rigid stress pattern (cf. again Polish, with fixed penultimate stress) offers little ease to the speaker to produce regularly recurring prominences. Conversely, a language whose lexicon mainly consists of mono- or disyllables has a much greater chance of presenting regular inter-accentual distances. What is especially relevant is that a bi-level model of SR seems to justify the often vague intuitions that people have, with respect to the rhythmical inclinations of the languages. The suggestion underlying the model proposed here is that no single measure can assess the actual behavior of any language: both Level-I and Level-II should be taken into account. Their possible divergence justifies the sometimes elusive character of rhythm judgments, including scholarly judgments. This may, for instance, explain why Polish is alternatively assigned, impressionistically, to syllable- or stress-timing, depending on the perceiver.

Two important caveat should be put forward here. The first concerns the lack of objective criteria for locating the phrasal prominences. Individual speakers may or may not detect as prominent a given syllable in a speech chain, and even one and the same speaker may hesitate and provide different judgments in different moments. Apart from very salient phrasal prominences, there is a grey zone of ambiguity often to be found in spontaneous speech. Indeed, the "news reading" style, some version of which seem to be practiced in most language communities, sounds so peculiar precisely because of the constantly emphatically realized prominences. One should thus be aware that the individuation of phrasal prominences is not a straightforward process. It is advisable to adopt multiple measures, e.g., limited to the most prominent peaks (MEASURE α) or including the intermediate ones (MEASURE β) .

The second caveat is even trickier. As it happens, dynamic stress – as conceived of for English, Italian, Polish etc. – is not a feature of every language. For instance, it does not play a role in Chinese, Japanese, Korean, Tamil and Mongol (Akamatsu 1997, Nolan 2008), although even there polysyllables normally have a prominent syllable – or rather, as in Japanese, a mora – presenting distinctive tonal features; indeed, even monosyllables may be a site for tonal prominence. (Incidentally, the literature often hints at the notion "mora-timing", as applying to languages such as Japanese, Korean, Sinhalese, Tamil, Hindi; in the view of the present authors, however, mora-timing is not regarded as an autonomous type, but rather as the most extreme form of CNTRL behavior.) To avoid confusion, in this paper the term "stress" will not be employed with reference to the word-level prominence in Chinese and Japanese, although the general use is – as it often happens in linguistic terminology – far from converging. This may or may not be a problem for the proposed view and should be regarded as a matter for further research. Table 2 is based on the assumption that phrasal accent, however realized, is a universal trait as rhythm regulator. Every language is assumed to present phrasal prominences whose more or less regular distribution accounts for a great deal of rhythm perception. Their presence is normally tied to word-prominence locations, although the relation is not one-to-one, for phrase accents only exist beyond the word, at the intonational level. Their function is purely communicative-pragmatic: they partition the speech chain into conveniently sized chunks, providing anchoring points that help the hearer to process the intended meaning. Interestingly, this sort of chunking seems to matter with respect to memory processes (Boucher 2006), and at the lowest production level it possibly subserves the respiratory activity. One might thus want to consider this as a kind of expansion into the cognitive domain of the task-dynamics proposal, concerning the emergent nature of rhythmical behaviors: supposedly, the rhythmical organization of speech at the phrasal level is exploited by both speaker and hearer for the sake of thought coordination. If this is true (at least in part), then Chinese and Japanese – plus any phonotactically simple language where dynamic word-stress does not play a role – are good candidates for type (4) above. Should this not be the case, then one should limit the role of Level-II to a subset of the languages, reducing somehow the scope of the bi-level model presented here. There is, in any case, little doubt as to the extremely simple phonotactics of languages such as Japanese and Chinese (Bauer 1995, Akamatsu 1997). This proposes them as very likely candidates as CNTRL-1 languages.

Be it as it may, what one needs in order to validate the Level-II hypothesis is, once again, a convenient algorithm. The one proposed by O'Dell & Nieminen (1999), exploiting the "Averaged Phase Difference" theory (APD), is a viable option. It has the following shape: $I = a + bn$, where I stands for 'duration of inter-stress intervals', n for 'number of syllables', while a and b are coefficients. More specifically, with r indicating the relative strength parameter:

$$T(n) = \frac{r}{r\omega_1 + \omega_2} + \frac{1}{r\omega_1 + \omega_2} n \qquad (2)$$

In practice, the formula relates the accentual phrase's duration to the number of syllables composing it. If r is greater than 1, then the overarching (accentual) oscillator predominates; if r is less than 1, the subordinated (syllabic) oscillator prevails.

This allows to put forth exact predictions as for Level-II, again with respect to speech rate variations: (i) At slow rates, the accentual oscillator should predominate in all languages, following the universal tendency towards rhythmicity alluded to above; (ii) At faster rates, the syllabic oscillator should prevail; however, its dominance is expected to emerge earlier, and more emphatically, with CNTRL-2 languages. The rationale is as follows: CMPST-2 languages present a relatively flexible structure, allowing the speaker more freedom to adjust the inter-accentual distances. In a language like English, this may be obtained by downgrading some of the word prominences and possibly promoting some of the secondary ones, and above all by shifting the word prominences as the case requires. In a stress-less language, like Chinese or Japanese, this may supposedly be achieved by appropriately redistributing the phrasal prominences, assuming that they are exceedingly flexible due to the non-dynamic character of the word prominences. By contrast, since none of these possibilities is available to CNTRL-2 languages, the dominance of the syllabic oscillator should tend to emerge as soon as the speech rate begins to increase, although some restructuring is available to the speaker, mostly by way of accent deletions.

These predictions were tested against the same Italian corpus exploited in sect. 2. Note that, in this case, the C interval preceding the last stressed vowel was not trimmed, in order to preserve the integrity of the accentual phrase. Besides, since the observations number was higher than in the CCI calculus, it was possible to partition the materials not only into 3, but into 5 rate classes, with segments x sec. as criterion. Table 4a-b presents the results according to MEASURE α and β, respectively:

Tempo	*segments x second*	*N*	*r*	*Tempo*	*segments x second*	*N*	*r*
T1	< 15,7 / av. 14.2	264	1.15	t1	< 14.9 / av. 13.6	166	1.05
T2	< 17.8 / av. 16.6	277	1.03	t2	< 16.1 / av. 15.4	157	1.30
T3	> 17.7 / av. 19.2	275	0.71	t3	< 17.2 / av. 16.6	161	0.84
				t4	< 18.9 / av. 17.9	167	0.91
				t5	> 18.8 / av. 20,0	165	0.57

Table 4a. Output of the APD algorithm as applied to rate classes naturalistically extracted from a spontaneous Italian corpus. MEASURE α: *limited to the most prominent peaks. Coupled oscillators: accentual vs. syllabic.*

Tempo	*segments x second*	*N*	*r*	*Tempo*	*segments x second*	*N*	*r*

T1	< 15,7 / av. 14.2	264	1.29	t1	< 14.9 / av. 13.6	166	1.20
T2	< 17.8 / av. 16.6	277	1.06	t2	< 16.1 / av. 15.4	157	1.36
T3	> 17.7 / av. 19.2	275	0.54	t3	< 17.2 / av. 16.6	161	1.02
				t4	< 18.9 / av. 17.9	167	0.76
				t5	> 18.8 / av. 19.9	165	0.46

Table 4b. Output of the APD algorithm as applied to rate classes naturalistically extracted from a spontaneous Italian corpus. MEASURE β*: including the intermediate peaks. Coupled oscillators: accentual vs. syllabic.*

The results show that, at slow rates, the accentual oscillator does indeed predominate; however, as rate increases, the syllabic oscillator definitely prevails. This tendency is emphasized by MEASURE *b*, whereby the intermediate-level accentual peaks are included: at the slow tempos, the intermediate peaks contribute to regularize the inter-accentual distances, whereas at faster tempos they obtain the contrary effect. In the present case this occurred despite the relative rarity (4.3%) of intermediate peaks vis-à-vis the most salient ones. Interestingly, when 5 rate classes are considered, the above tendency turns out to be non-monotonic, showing that tempo variation is accompanied by some restructuring in the implementation of accentual prominences (i.e., deletions or insertions). For instance, t2 allows a more regular accent distribution, yielding a sharper dominance of the accentual oscillator. Apart from this detail, Italian appears to behave as a CNTRL-2 language.

This result was expected, but what really matters is that it was autonomously derived: it stems from behavioral measures mirroring relevant structural properties. As noted above, this avoids the risk of circularity implicit in basing one’s interpretation on the mere contrastive distribution on the Cartesian plane of allegedly prototypical languages. At the present stage of our knowledge, no language can really be considered prototypical.

4. Expanding the model

The CCI algorithm described in sect. 2 aims at capturing the intra-syllabic rhythmical behavior, which in turn affects (and is possibly affected by) the overarching accentual oscillations, as described in sect. 3. Which of these components is the dominant factor remains – for the time being – unclear, although one may want to assign this role to Level-I due to the pervasive nature of phonotactics. What one can emphatically assert, in any case, is that the inter-level relation is not deterministic, for the two levels may diverge along the CC continuum. This, however, does not imply that no attempt should be undertaken to combine the two levels into a single design. It is very tempting to reduce the two pairs of coupled oscillators described above to a cascade of hierarchically ordered oscillators: accentual > vocalic > consonantal. Indeed, the Level-I vocalic oscillator, implementing the syllabic nucleus, may be conflated with the Level-II syllabic oscillator. As for the consonantal oscillator, it clearly acts upon the vocalic one very much in the same way as the syllabic oscillator acts upon the accentual one at Level-II.

As a first step, the O'Dell & Nieminen formula was applied to the Level-I oscillators, relating the duration of inter-V-onset intervals – from one V-onset to the next – to the number of intervening Cs (the relevance of the inter-V-onset interval as a rhythmic unity is underlined, e.g., by Keller & Port 2007). Once again, *r* greater than or less than 1 indicates whether the overarching (vocalic) or the subordinated (consonantal) oscillator prevails.

The predictions are as follows: (i) In general, the consonantal oscillator should emerge as the dominant factor along with tempo increases, for the Cs comprised between two vocalic gestures cannot be compressed beyond a certain threshold, whereas Vs allow for more compression; (ii) In CNTRL languages, however, due to the relative incompressibility of unstressed Vs, the vocalic oscillator should partly compensate the previous effect.

The computation's results, again referred to 3 or, alternatively, 5 rate classes, appeared to be compatible with both expectations. As table 5 shows, the dominance of the consonantal oscillator increases from T/t1 to T/t2, but then begins to decrease towards the fastest rates. Needless to say, these predictions should be checked against other languages, particularly those expected to follow the CMPST pattern. The present authors are currently engaged in such a task.

Tempo	*segm.s x sec.*	*N*	*r*	*Tempo*	*segm.s x sec.*	*N*	*r*
T1	< 15,7 / av. 14.2	913	0.97	t1	< 14.9 / av. 13.6	561	1.01
T2	< 17.8 / av. 16.6	947	0.72	t2	< 16.1 / av. 15.4	573	0.74
T3	> 17.7 / av. 19.2	951	0.84	t3	< 17.2 / av. 16.6	549	0.78

t4	< 18.9 / av. 17.9	556	0.81
t5	> 18.8 / av. 20.0	572	0.84

Table 5. Output of the APD algorithm as applied to the rate classes naturalistically extracted from a spontaneous Italian corpus. Coupled oscillators: vocalic vs. consonantal.

A further attempt was made by the present authors to model the combined effect of Level-I and -II, by extending the algorithm in (2) to a system of three cascaded oscillators. It is useful to reiterate that the vocalic oscillator is common to both levels, although at Level-II it is more appropriately called "syllabic" oscillator. In the case at stake, the formula yields two indexes: *r1* referring to the relation between the accentual and the vocalic (= syllabic) oscillators, *r2* referring to the relation between the vocalic and the consonantal oscillators. As it happens, the joint consideration of the two oscillator pairs brings about the emphatic dominance of the vocalic oscillator, since *r1* is often below 1 and *r2* constantly much above. The interpretation of the results (not reported here) is, however, far from easy. Apparently, the dramatic oscillation of Cs duration has strong repercussions on the dominant oscillator, yielding fairly high *r2* values. Moreover, one should consider that the actual durational behavior of speech does not only depend on the interaction of the three oscillators considered here, as shown by the explained variance, that never exceeds 69%. Indeed, over and above the interplay of the three oscillators considered, other factors play an important role, like the unpredictable pragmatic behavior of the speaker (emphasis, hesitation etc.), that may cause major local disturbances in the rhythmic flow.

5. Conclusion

The main attempt carried out in this paper was to propose a unified and predictive SR theory. Inevitably, the hypothesis presented here will in due time – perhaps very soon – be disconfirmed, but the present authors will not be upset about that: any theory's crisis, or even death, should be viewed as a step forward, paving the way to improved conceptions. It remains to be seen whether this sketch of a theory will be globally disconfirmed or only with respect to some of its predictions. Should the latter be the case, there would be room for reformulation of the details; alternatively, an entirely new hypothesis should be devised. Whatever the case, future theories will necessarily presuppose the spelling-out of explicit language features, from which specific rhythmical consequences can be derived. Returning somehow to the original spirit of the Pikean proposal, one should realize that SR is the observable consequence of precise – albeit so-far poorly understood – structural properties, rather than a sort of phonetic primitive. The ultimate goal is to isolate and define those basic structural properties.

As noted above, the first results obtained should be checked against other linguistic materials. These should be selected out of conveniently sized corpora, for it is now clear that no meaningful conclusion can be drawn from scanty data. One should thus compare the present Italian data, stemming from spontaneous speech, both with read speech from the same language, and with speech from other languages, both read and spontaneous (as for style variability, see e.g. Wiget et al. 2008). It is however important to note that, once the topic is addressed within a sound epistemological perspective, cross-linguistic comparison becomes a useful – indeed necessary – tool for theory testing, rather than being the precondition for the results' assessment. The latter should rather follow from the constant interplay between predictions and results, progressively extended to a larger array of data.

It is equally important to observe that whoever engages in this research domain should be aware that this is a cumulative scientific enterprise. Whatever new insight one develops will rest on previous successes and failures, just as the model presented in this paper exploits a number of ideas developed by other scholars, whose inspiration is gratefully aknowledged by the authors. Hopefully, by joining the efforts, a better understanding of this fascinating language aspect will be achieved.

References

Akamatsu, T.: Japanese Phonetics. Theory and Practice. (LINCOM, München / Newcastle 1997).

Arvaniti, A.: Rhythm, timing and the timing of rhythm. Paper presented at the conf. on Empirical Approaches to Speech Rhythm. (Univ. College London 2008).

Barbosa, P.: Incursões em torno do ritmo da fala. (Pontes, Campinas 2006).

Barbosa, P.: From syntax to acoustic duration: A dynamical model of speech rhythm production. Speech Communication **49**: 725-742 (2007).

Bauer, R.S.: Syllable and word in Cantonese. Journal of Asian Pacific Communication **6**: 245-306 (1995).

Bertinetto, P.M.: Strutture prosodiche dell'italiano. Accento, quantità, sillaba, giuntura, fondamenti metrici. (Accademia della Crusca, Firenze 1981).

Bertinetto, P.M.; Bertini, C.: On modeling the rhythm of natural languages. Proc. of the 4th Speech Prosody Conf. (Campinas 2008)

Bertinetto, P.M.; Scheuer, S.; Dziubalska-Kolaczyk, K.; Agonigi, M.: Intersegmental cohesion and syllable division in Polish. Proc. of the 16th Int. Cong. of Phonetic Sciences: 1953-1956 (Saarbrücken 2007).

Bertinetto, P.M.; Fowler, C.A.: On sensitivity to durational modifications by Italian and English speakers. Rivista di Linguistica **1**: 69-94 (1989).

Boucher, V.J.: On the function of stress rhythms in speech: Evidence of a link with grouping effects on serial memory. Language and Speech **49**: 495-519 (2006).

Cummins, F.: Speech rhythm and rhythmic taxonomy. Proc. of the 2nd Speech Prosody Conf.: 121-126 (Aix en Provence 2002).

Dauer, R.M.: Stress-timing and syllable-timing reanalyzed. J. of Phonetics **11**: 51-62 (1983).

Dauer, R.M.: Phonetic and phonological components of language rhythm. Proc. of the 11th Int. Cong. of Phonetic Sciences: vol.5, 447-50 (Tallinn, URSS 1987).

Dellwo, V.: Rhythm and speech rate: A variation coefficient for ΔC. In: Karnowski, P.; Szigeti, I. (eds.) Language and language-processing: 231-241 (Peter Lang, Frankfurt am Main 2004).

Dellwo, V.; Steiner, I.; Aschenberner, B.; Dankovicova, J.; Wagner, P.: Bonn-Tempo Corpus and Bonn-Tempo Tools: A database for the study of speech rhythm and rate. Proc of INTERSPEECH 2004 – 8th Int. Conf. of Spoken Language Processing. (Jeju Island, Korea 2005).

Fowler, C.A.: Timing Control in Speech Production. (Indiana University Linguistics Club 1977).

Gavaldá, N.; Dellwo, V.: Vowel reduction and Catalan speech rhythm. Poster presented at the conf. on Empirical Approaches to Speech Rhythm. (Univ. College London 2008).

Goldstein, L.; Chitoran, I.; Selkirk, E.: Syllable structure as coupled oscillator modes: Evidence from Georgian vs. Tashlhiyt Berber. Proc. of the 16th Int. Cong. of Phonetic Sciences: 241-244 (Saarbrücken 2007).

Grabe, E.; Low, E.L.: Durational variability in speech and the rhythm class hypothesis. Papers in Laboratory Phonology 7: 515-546 (Berlin, Mouton de Gruyter 2002).

Keller, E.; Port, R.: Speech timing: Approaches to speech rhythm. Proc. of the 16th Int. Cong. of Phonetic Sciences: 327-329 (Saarbrücken 2007).

Kohler, K.: Prominence and rhythm rivisited perceptually. Paper presented at the conf. on Empirical Approaches to Speech Rhythm. (Univ. College London 2008).

Lindblom, B.; Rapp, K.: Some Temporal Regularities of Spoken Swedish. Papers of the Inst. of Linguistics n. 21. (Stockholm 1973).

Low, E.L.: Prosodic Prominence in Singapore English. (Doctoral dissertation, Univ. of Cambridge 1998).

Mairano, P.; Romano, A.: A comparison of four rhythm metrics for six languages. Poster presented at the conf. on Empirical Approaches to Speech Rhythm. (Univ. College London 2008).

Mok, P.P.K.: Using durational measures with non-native speech rhythm. Poster presented at the conf. on Empirical Approaches to Speech Rhythm. (Univ. College London 2008).

Mok, P.P.K. & Dellwo, V.: Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English. Proc. of the 4th Speech Prosody Conf. (Campinas 2008).

Nolan, F.: The peevy eye: A critical look at a rhythm measure. Paper presented at the conf. on Empirical Approaches to Speech Rhythm. (Univ. College London 2008).

O'Dell, M.; Nieminen, T.: Coupled oscillator model of speech rhythm. Proceedings 14° Int. Cong. Phonetic Sciences 2, 1075-1078 (Berkeley 1999).

O'Dell, M.; Lennes, M.; Werner, S.; Nieminen, T.: Looking for rhythms in conversational speech. Proc. of the 16th Int. Cong. of Phonetic Sciences: 1201-1204 (Saarbrücken 2007).

Pike, K.L.: The Intonation of American English. (Ann Arbor 1945).

Price, P.J.: Sonority and syllabicity: acoustic correlates of perception. Phonetica **37**: 327-43 (1980).

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

Rouas, J.L.; Farinas, J.: Comparaison de méthodes de caractérisation du rythme des langues. Workshop MIDL. (Paris 2004).

Van Santen, J.P.H. & Shih, C.: Suprasegmental and segmental timing models in Mandarin and American English. J. of the Acoustical Soc. of America **107**: 1012-1026 (2000).

Wagner, P.: Visualizing levels of rhythmic organization. Proc. of the 16th Int. Cong. of Phonetic Sciences: 1113-1116 (Saarbrücken 2007).

White, L.; Mattys, S.: Calibrating rhythm: First language and second language studies. J. of Phonetics **35**: 501-522 (2007).

Wiget, L.; White, L.; Mattys, S.: Poster presented at the conf. on Empirical Approaches to Speech Rhythm. (Univ. College London 2008).

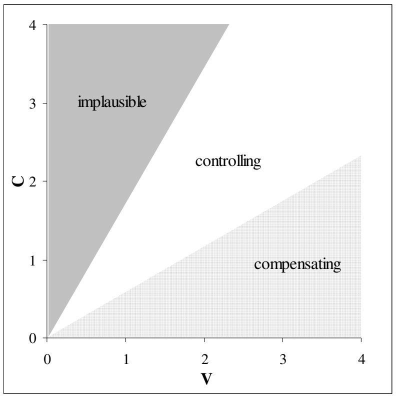

Figure 1: *Schematic representation of the two ideal rhythmic types according to the CC model.*

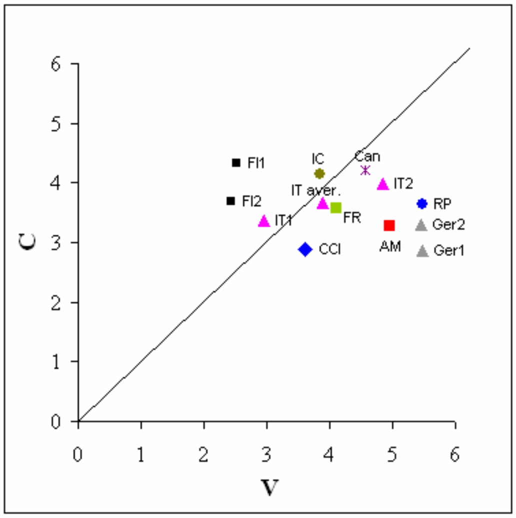

Figure 2: *Application of CCI by Mairano & Romano (in prep.):* AM = *Amer. Eng.,* Can = *Can. Fr.,* FI(1,2) = *Finnish,* FR = *French,* Ger(1,2) = *German,* IC = *Icelandic (average of 10 speakers),* IT(1,2) = *Ital. (+* IT aver.*),* RP = *Eng. RP*. CCI = *spontaneous Ital. corpus as analyzed in Bertinetto & Bertini (2008).*