On modeling the rhythm of natural languages

Pier Marco Bertinetto and Chiara Bertini

(paper presentated at Speech Prosody 2008, Campinas, Brasil)

Abstract

We describe a new model for the formal representation of the rhythmic tendencies of natural languages. The algorithm is a modification of PVI, proposed by Grabe and Low [10]. The model was tested on a fairly substantial corpus of Italian semi-spontaneous productions, and its outcome compared with that yielded by two well-known algorithms (Ramus [14] and PVI).

1. Introduction

The traditional view about rhythm in natural languages stems from the pioneering proposal by Pike [13], according to which languages differ along the divide "syllable- vs. stress-timing". Taken literally, this would imply the following, with (1a-b) depicting, respectively, an ideal syllable- vs. stress-timed language (underlining stands for stressed syllables). Isochronicity is to be found at the level of syllables in (a) and of inter-stress intervals in (b):

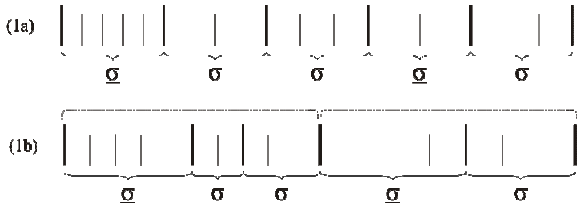

Figure 1: Schematic representation of the traditional divide: (a) syllable- vs. (b) stress-timing.

It was soon obvious that this simplified view could not possibly correspond to reality. In particular, the ideal syllable-timed type could only be approached by languages with very regular and simple syllable structure. Languages departing from this scheme would present striking deviations. This led to a number of amendments, some of which attempted to root the rhythmic typology on perception, rather than production. By the end of the Seventies, however, it became increasingly clear that the traditional view was unsatisfactory. Yet, the intuition that the basic divide contained some grain of truth persisted, leading to reinterpretation of the basic contrast in terms of an array of alternative features, mostly phonologically inspired [2, 7, 3]. Although this interpretation was probably on the right track, it was not entirely satisfactory, for it is not easy to map it onto a specific set of phonetic properties. A step in this direction was taken by [4, 5], where the original dichotomy was reinterpreted with respect to the continuum "controlling vs. compensating" languages.

The basic idea of the "control vs. compensation" (CC) hypothesis, inspired by work in articulatory phonology and earlier on by the seminal work by C.A.Fowler [9], was that languages may differ in terms of how vocalic and consonantal gestures are coupled in the articulatory flow. An ideally controlling (henceforth CTL vs. CPS) language should be conceived of as a language in which all segments receive the same amount of expenditure, i.e. articulatory effort, and (ideally) tend to have the same duration. This is obviously impossible, due to the varying points and manners of articulation; but this view acquires plausibility once we consider how languages do in fact differ in terms of the coupling of vocalic and consonantal gestures. Some languages admit a much higher segmental overlap (coarticulation) than others. This view bears resemblances with proposals recently advanced [12, 1], where the respective contribution of the syllabic and accentual oscillators are modeled by means of elegant algorithms. In fact, the latter proposals and the CC hypothesis should be viewed as converging towards a possible integration. The CC view aims at describing the intra-syllabic behavior, which in turn affects (or is possibly affected by) the overarching accentual alternation, primarily taken into account in [1, 12]. Which of these components is the dominant factor remains, for the time being, unclear.

Let us detail the theoretical consequences. The contrast "syllable- vs. stress-timing", reinterpreted in terms of CC, could be visually represented as in (2a-b):

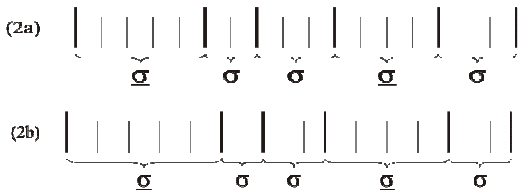

Figure 2: Schematic representation of the CC hypothesis: (a) controlling vs. (b) compensating languages.

This is a very abstract representation, but enough to grasp the essential difference vis-à-vis the traditional view. CPS languages (2b), corresponding to the hitherto hypothesized stress-timed type, paradoxically bear some resemblance with the traditional syllable-timed pattern (1a), owing to intra-syllabic duration compensation. They differ from it, however, due to the increased gestural overlap in unstressed syllables, where the segment most liable to compression / coarticulation is of course the vocalic nucleus. Needless too say, this also occurs in CTL languages: the difference is a matter of degree along a continuum. As a consequence, in CTL languages, the syllable duration is more directly related to the segments

number, due to the reduced amount of consonant/vowel overlapping (henceforth: C and V, respectively).

A possible test to assess this hypothesis consists in checking the effect of speech tempo. The prediction is that CTL languages should tend to reduce the segments duration in a more proportional way, whereas in CPS languages Vs should be somewhat more affected than Cs.

2. The Control/Compensation Index

In the light of the above discussion, we would like to propose a new tool for modeling the rhythmical behavior of natural languages. We shall call it 'Control/Compensation Index' (CCI). The idea consists in relativizing the PVI model [10] to the number of segments composing the V and C intervals. In practice, the duration of each interval is divided by the number of segments in it, according to the following formula, where *m* stands for 'number of intervals' (V or C, as separately considered), *d* for 'duration' (in ms), *n* for 'number of segments within the relevant interval':

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

Since we treat glides as Cs, V intervals mostly consist of a single element, except for hiatuses. It is worth noting that our formula (just as those in [10] and [14]) does not directly take the notion of syllable into account, for a C interval may begin with a coda and end with an onset. Although this is at odds with the traditional view regarding the timing of natural languages, this should not be seen as a problem. The CCI model in fig. 2 suggests that the syllable has virtually no effect, despite the traditional label of syllable-timing attached to one of the two poles of the rhythmic divide. In any case, we also checked the CCI algorithm by separately considering onset and coda intervals (see [16]). In this paper, however, we shall not report the results obtained by this further version of the algorithm. Our present goal is to compare the CCI model with a selection of the existing ones, in order to evaluate the respective virtues.

The CCI model bears resemblance with the one proposed in [15], based on so-called "pseudo-syllables", for one of the possible measures consists in the number of segments composing the C intervals. There are, however, differences. In [15], each V is treated as the nucleus of a different syllable, whereas in CCI model each V interval is divided by the number of elements composing it. This is obviously vacuous in most cases but not with hiatuses, which make up complex V intervals. More generally, CCI has the following distinctive features: (a) it takes into due account the degree of phonotactic complexity as reflected by the number of segments composing each interval (both V and C ones), a feature that is ultimately at the core of the contrast CTL vs. CPS languages; (b) it takes advantage of the PVI design, which is (in our view) superior to the models in [14] and [15] in its ability to monitor the actual durational behavior of speech, i.e. the moment-by-moment oscillations in segment durations. In a forthcoming paper [16] we provide a detailed discussion of the model in [15].

Let us describe the ideal (obviously abstract) situations, as depicted in fig. 3. A CTL language should fall in the white area, near the bisecting line, for the durational fluctuations of Cs and Vs should by and large be of the same magnitude. By contrast, a CPS language should fluctuate more in the V than in the C portions, thus falling in the bottom grey area. This would mostly be caused by the sizeable durational differences between stressed and unstressed Vs, with the latter undergoing heavier coarticulatory effects. The complementary situation (i.e. a language with larger C than V fluctuations) would by contrast undergo severe restrictions: the upper grey surface delimits an area where an actual language is unlikely to be found. Needless to say, only in models such as CCI (where each interval is divided by the number of segments composing it) may the bisecting line take such a precise interpretation; the Ramusian and PVI models do not fully allow this inference, for one does not know what is exactly at stake behind the intervals' durational variation, particularly with respect to C intervals:

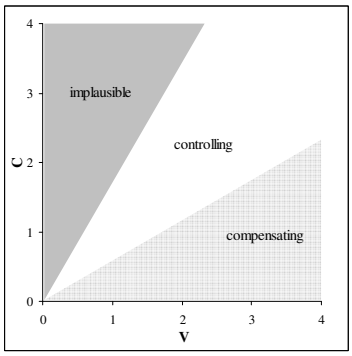

Figure 3: *Schematic representation of the major rhythmic types according to the CC hypothesis.*

Our corpus was derived from the semi-spontaneous productions of 10 speakers of the Pisa variety of Italian, engaged in the so-called 'map-task'. This consists in a two-person dialog fostered by two slightly different maps. One of the speakers (Giver) provides instructions to the other (Follower), so that s/he can find the intended goal on the map. The gradual discovery of the maps differences, of which the two speakers may or may not be aware from the outset, feeds the communicative intercourse. The V and C intervals considered were 1765 and 1755, respectively. The criteria for the selection of the relevant speech stretches from our corpus are described in [6]. In particular, we discarded the final portion of each utterance, from the last stressed syllable (inclusive) onward. This portion has an entirely different rhythmic behavior, that should best be analyzed on its own.

The choice of semi-spontaneous materials for our purposes deserves a note. On the positive side, the empirical basis on which we root our conclusions is considerably larger than is often the case in this sort of studies, yielding a much firmer statistical stability. On the negative side, we are aware that the use of semi-spontaneous materials is regarded as inappropriate by some authors. We would however defend our choice, for the endeavor to model linguistic rhythm should not neglect spontaneous speech. The obvious caveat consists in referring any conclusion to the given speech style.

3. Results and discussion

For the sake of homogeneity, the figures reported below invariably show the C measure on the ordinate and the V one on the abscissa. We are aware that some variants of the available models present V or C measures on both axes, but we regard this as unattractive, for both Vs and Cs contribute to the rhythmic behavior. For comparison with CCI, we chose two of the most representative models, i.e. those described in:

- [14] in two of its possible versions, with the V measure expressed as ΔV (standard deviation of the V intervals) or as %V (percentage of Vs over the global duration); the C measure is indicated as ΔC in both cases;
- [10] in its two versions, with raw (rPVI) and normalized V data (nPVI), where normalization aims at counteracting the intra-sentential tempo variations; the basic algorithm is essentially as in (1) above, except for the *n* variable;

Fig. 4 depicts the respective positioning on the Cartesian plane yielded by the five models. In order to have everything on the same graphic, the ΔV, %V and nPVI values were divided by 10. The ovals in fig. 4 hint at the spatial dispersion of the 10 speakers considered (the respective size is based on the relative error computed on Vs and Cs):

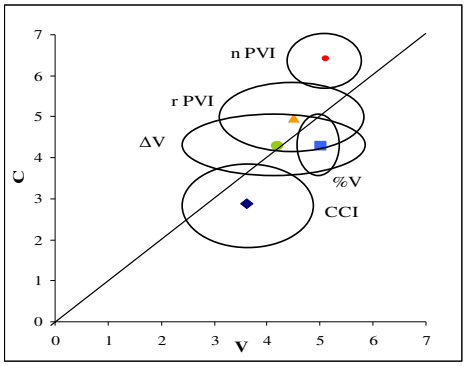

Figure 4: *Spatial representation of five rhythmic models.*

Although, according to the caveat put forth above, the fine interpretation of the bisecting line is not the same for each model, we would like to underline the following points. ΔV and (to some extent) rPVI fall in the middle of the Cartesian plane, suggesting that the V and C intervals present, altogether, a similar behavior. With nPVI, by contrast, there is more variation on the C than on the V axis, but one should note that in the latter model the measures on the two axes are of completely different magnitudes (this, as noted, also applies to %V). CCI and %V indicate the alternative tendency, whereby the durational fluctuation is larger for Vs than for Cs. Considering what is known of the phonetics of Italian, this seems to be a better representation of the actual facts. Italian presents a duration contrast between stressed and unstressed Vs (although a much weaker one than often stated, as shown by [8]), whereas no equivalent differences are to be noted among Cs. Obviously, Italian has geminate Cs, differing in duration with respect to simple Cs; however, in our model geminates are counted as two segments, as dictated by most phonological approaches.

We then examined the disaggregated data, i.e. the separate projection of the individual speakers' behavior. As shown by the ovals in fig. 4, the data are relatively scattered in each model, with the exception of %V and nPVI. This follows from the nature of the latter models: in %V, the V measure is relatively static, as compared with the other models; in nPVI, the lower dispersion is a by-product of normalization. Fig. 5a-b show the actual dispersion as represented by two selected models, namely CCI (a) and ΔV (b). The bisecting line is drawn for reference; see fig. 4 for the relative position on the Cartesian plane.

The dispersion is in part due to the different number of data points available for each speaker, as shown by the individual variances in fig. 5a-b, but it also depends on inherent individual tendencies. These may in turn depend on the different roles played by the speakers as map-task Givers or Followers (as shown by filled vs. blank symbols), but may also reflect idiosyncratic inclinations. Indeed, it can be shown that speech tempo is a better predictor of rhythmic behavior than map-task role; the two factors are strongly but not exhaustively correlated [6]. This is no wonder: it is well-known that different speakers of the same language may provide different results in terms of rhythmic behavior [1, 11]. Most models considered here exhibit a remarkable sensitivity to small individual differences, with the two exceptions pointed out above. The comparison between fig. 5a and 5b shows, however, that the relative positioning of the various speakers changes according to the algorithm employed, suggesting that the choice of the model is no innocent matter even in this respect.

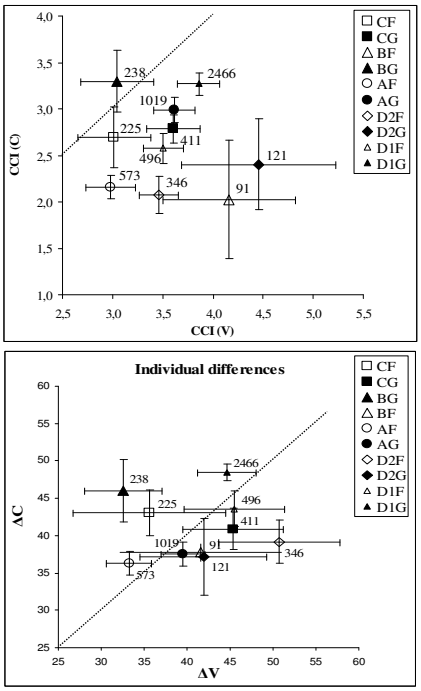

Figure 5a-b: *Individual behaviors as represented in CCI (a) and ΔV (b). The digits refer to the intervals number (V plus C).*

Finally, we examined the effect of speech tempo. For this purpose, we divided the speakers' productions in three tempo-groups, as measured in syllables per second: (I) 4.8 < 7.0 (average: 6,0); (II) ≥ 7,0 ≤ 8,1 (average: 7,6); (III) > 8,1 (average: 9,2). The groupings were obtained by evenly distributing the individual productions, so that all groups would approximately be of the same size in terms of intervals number. The results are shown in fig. 6, where again the ΔV, %V and nPVI values were divided by 10 for homogeneity. Note that, in this case, the ellipses merely connect the tempo-groups (I,II,III) within each model for ease of the reader.

The *t*-test discriminations in table 1 yield some hints. The contrast I / II is significant for both Vs and Cs in CCI, rPVI and ΔV, and for Cs alone in %V. The contrast II / III is

significant for both Vs and Cs in %V, and for Cs alone in rPVI. As for nPVI, it only discriminates I from III on the V axis. Considering the strong rhythmical effect of tempo, as also shown by [1], this calls into doubt the very essence of the latter model, based on the attempt to minimize tempo differences. %V looks somewhat extravagant, for it is the only one to emphasize the contrast II / III. This may, however, be explained: since, with the slowest tempos, the compression is very evenly distributed over Vs and Cs, the %V value remains stable. Conversely, since (according to rPVI and ΔV) Cs seem to vary more than Vs between II and III, the %V value is correspondingly higher in the latter tempo group. The model %V appears thus to be, to some extent, the mirror image of the other models, as also shown by its diverging spatial orientation. This, however, does not explain why there should be such a neat II / III contrast on the C axis (a feature shared, as noted, by rPVI and ΔV). The results yielded by CCI provide a comparatively more plausible picture. In a supposedly CTL language like Italian, tempo increases should tend to compress Vs and Cs more or less alike, and this is what we found. Besides, the compressibility threshold should be reached relatively soon, and sooner with Cs than with Vs. In our data, this limit is met within group III; and a finer separation of the tempo groups (not attempted here) might possibly show that the limit is indeed reached sooner with Cs than with Vs, as hinted at in fig. 6 and as shown by the Mann-Whitney II / III comparison, which is significant for Vs alone .

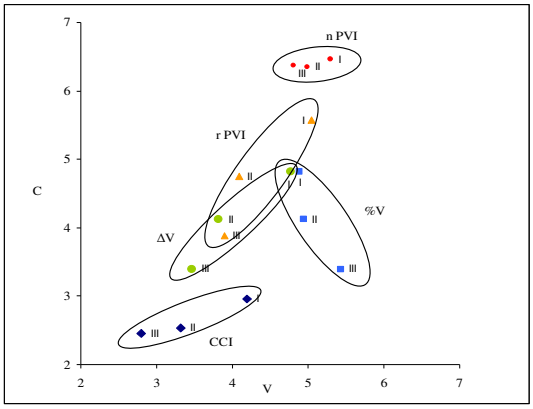

Figure 6: *Tempo groups dispersion in five models.*

		v1 vs. v2	v2 vs. v3
CCI	V	0,000	n.s.
	C	0,002	n.s.
PVI	nV	n.s.	n.s.
	nC	n.s.	n.s.
	rV	0,002	n.s.
	rC	0,000	0,000
Ramus	ΔV	0,016	n.s.
	ΔC	0,000	0,000
	%V	n.s.	0,001

Table 1: *Tempo groups discrimination* (t-*test).*

4. Conclusion

We compared four well-known rhythm modeling algorithms with the CCI model first proposed here. The results obtained were promising. The CCI algorithm appeared: (i) to allow neatly interpretable hypotheses in terms of positioning on the Cartesian plane (fig. 4); (ii) to be as fine-tuned as its best competitors in capturing the individual speakers' behavior (fig. 5); (ii) to provide very plausible tempo-groups discrimination (fig. 6). This suggests that CCI might turn out to be a viable representation of the rhythmic tendencies of natural languages.

Needless to say, the ultimate proof must rest on the ability to discriminate between different languages, providing consistent groupings among them. Our next goal will thus be to apply this method to different languages, obviously with comparable speech materials. We also intend to compare spontaneous and read speech, to check for style variations.

Our data were so-far based on spontaneous productions. This is not often the case in speech rhythm studies, although any realistic modeling of natural languages should include this sort of data. At any rate, the size of the corpus analyzed provided robust statistical reliability to our conclusions. One major result was that speech tempo has a striking impact on the rhythmical behavior. This is not new, but it turns out to be too often overlooked when different languages are compared. Meaningful comparisons should only involve homogeneous data.

5. References

[1] Barbosa, P., 2006. *Incursões em torno do ritmo da fala.* Campinas: Pontes.

[2] Bertinetto, P.M., 1981. *Strutture prosodiche dell'italiano. Accento, quantità, sillaba, giuntura, fondamenti metrici.* Firenze: Accademia della Crusca.

[3] Bertinetto, P.M., 1989. Reflections on the dichotomy 'stress- vs. syllable-timing'. *Revue de Phonétique Appliquée* 91/93. 99-130.

[4] Bertinetto, P.M.; Fowler, C.A., 1989. On sensitivity to durational modifications in Italian and English. *Rivista di Linguistica 1*. 69-94.

[5] Bertinetto, P.M.; Vékás, D., 1991. Controllo vs. compensazione: sui due tipi di isocronia. In *L'interfaccia tra fonologia e fonetica*, E. Magno Caldognetto; P. Benincà (eds.). Padova: Unipress, 155-162.

[6] Bertini, C.; Bertinetto, P.M., to appear. Prospezioni sulla struttura ritmica dell'italiano basate sul corpus semi-spontaneo AVIP/API. *Atti del 4° Conv. AISV 2007*. Cosenza.

[7] Dauer, R.M., 1983. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics* 11. 51-62.

[8] Dell'Aglio, M.; Bertinetto, P.M.; Agonici, M., 2002. Le durate dei foni vocalici in rapporto al contesto nel parlato di locutori pisani. Primi risultati. In *La 2 fonetica acustica come strumento di analisi della variazione linguistica in Italia*, A. Regnicoli (cur.). Roma: Il Calamo, 53-58.

[9] Fowler, C.A., 1977. *Timing Control in Speech Production*. Indiana University Linguistics Club.

[10] Grabe, E.; Low, E.L., 2002. Durational variability in speech and the rhythm class hypothesis. *Pap.s in Laborat. Phonology 7*. Berlin: Mouton de Gruyter, 515-546.

[11] Mairano, P.; Romano, A., in press. Lingue isosillabiche e isoaccentuali: misurazioni strumentali su campioni di it., fr., ing. e ted. *Atti del 3° Convegno AISV 2006*. Trento.

[12] O'Dell, M.; Nieminen, T., 1999. Coupled oscillator model of speech rhythm. *Proceedings 14° Int. Cong. Phon. Sciences* 2, 1075-1078.

[13] Pike K.L., 1947 [1945], *The Intonation of American English*. Ann Arbor.
[14] Ramus, F. ; Nespor, M., Mehler, J., 1999. Correlates of ling. rhythm in the speech signal. *Cognition* 73. 265-292.
[15] Rouas, J.L.; Farinas, J., 2004. Comparaison de méthodes de caractérisation du rythme des langues. *Workshop MIDL*. Paris.
[16] Bertinetto, P.M.; Bertini, C., 2008. Modelización del ritmo y estructura silábica, con aplicación a l'italiano. In Sanchez-Miret, F. (ed.), *Romanística sin complejos: Homenaje a Carmen Pensado*. Bern.