Chinese and Italian Speech Rhythm. Normalization and the CCI algorithm.

Chiara Bertini, Pier Marco Bertinetto, Na Zhi

Laboratorio di Linguistica, Scuola Normale Superiore, Pisa, Italy

c.bertini@sns.it, p.bertinetto@sns.it, na.zhi@sns.it

Abstract

This paper re-examines the speech rhythm of Beijing Chinese and Pisa Italian by means of the Control/Compensation Index (CCI), with a view to normalizing the speech data, in order to reduce the effect of the rate factor. Two metrics were applied: (a) $_{D}$nCCI, an adaptation to the CCI model of the nPVI normalization strategy; (b) $_{S}$nCCI, a *z*-score normalization, which takes into account the actual constitution of each V- and C-interval, by referring the individual segment's duration to the mean duration of the members of the corresponding natural phoneme class. The results indicate the advantage of the $_{S}$nCCI metrics as a normalization strategy.

Index Terms: speech rhythm, prosody, normalization

1. The CCI rationale

This study is part of an ongoing project concerning the rhythmic behavior of different languages by means of the Control/Compensation Index (CCI) algorithm, also in relation to results obtained by other methods. In previous studies [3, 4, 5], the CCI algorithm was applied to corpora of spontaneous and read Pisa Italian, and is currently being applied to spontaneous Chinese [12], German and Brazilian Portuguese. Pisa is a town in Tuscany, not far from Florence, historically (and traditionally) considered to be the Standard Italian's nest.

The basic idea of CCI consists of relativizing the PVI algorithm [8] to the number of segments composing each V- and C-interval. The duration of each interval is thus divided by the number of segments comprised in it, according to the following formula, where *m* stands for 'number of intervals' (vocalic or consonantal, as separately considered), *d* for 'duration' (in sec), *n* for 'number of segments within the relevant interval'. Except for the division by segments, the formula is exactly like the PVI one:

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

As with PVI, the CCI algorithm measures the local durational fluctuations between adjacent units of the relevant type (Vs or Cs). Due to its very conception, however, the CCI algorithm takes into account not only the speech durational behavior, but also the degree of phonotactic complexity as reflected in the number of segments composing each V- and C-interval. The CCI model aims thus at providing a more realistic representation of the rhythmic tendencies of natural languages. Indeed, it makes a big difference, in terms of phonotactics, whether a C-interval contains a single C, a geminate C, or a C cluster, and the same holds for V-intervals (with a single V, a phonologically long V, or a V sequence within the same syllabic nucleus).

In ideal situations, a perfectly "controlling" language should present tendentially identical C and V local durational fluctuations, thus falling on the bisecting line, or else it should exhibit stronger stability in the V-intervals. By contrast, "compensating" languages should fluctuate more in the V than in the C-component, for they presuppose substantial V-reduction. Fig. 1 – which modifies the initial proposal in [3] – depicts these ideal situations, obviously to be interpreted *cum grano salis*. Since the use of CCI is still in the initial phase, one should allow for a reasonable degree of approximation in the formulation of the relevant predictions.

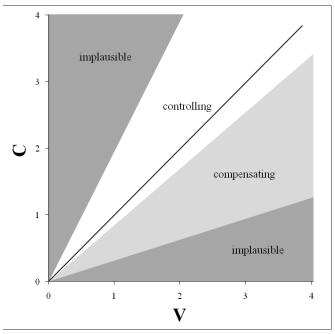


Figure 1: *Schematic representation of the major rhythmic types according to the CCI model.*

2. Previous data on Chinese and Italian

2.1. Corpus description

The Italian materials stemmed from the semi-spontaneous productions of 10 speakers of the Pisa variety of Italian, engaged in the so-called 'map-task' (corpus downloadable at: http://www.parlaritaliano.it/index.php/it/corpora/673-corpus-avip-api). This consists of a two-person dialog fostered by two slightly different maps. One speaker (Giver) provides instructions to the other (Follower), so that s/he can find the intended goal on the map. The V- and C-intervals were 2811 and 2457, respectively, extracted out of 233 utterances. The criteria for the selection of the relevant speech stretches are described in [5].

The utterances, intonationally neutral, consisted of at least 8-syllables. They did not present disrupting phenomena, such as speakers' overlapping, laughters, background noise, etc. Moreover, each utterance-final portion was trimmed from the last stressed syllable (inclusive) onward, in order to minimize the durational distortion due to final lengthening. The trimming procedure was also applied to sentence initial syllable-onsets consisting of an unvoiced plosive or affricate, due to impossibility to measure the first C-interval's duration.

The Chinese materials stemmed from the *Chinese Spontaneous Conversation Corpus* [9, 10], distributed by the Chinese Academy of Social Sciences, Beijing. It consists of 12 units of daily conversations between native Beijing speakers. Each unit is a 1-hour dialogue between two speakers. For this study, 607 utterances produced by 7 speakers (4 females and 3 males) were selected, segmented and manually labeled by one of the authors (a native speaker). The V- and C-intervals were, respectively, 6648 and 5609. The utterances were selected as much as possible according to the same criteria used for the Italian corpus. In particular, trimming was applied to each utterance-final syllable (note that, in contrast to Italian, for Chi-

10.21437/Interspeech.2011-504

Copyright © 2011 ISCA

nese it would be much harder to look for the last syllable bearing sentence stress). See [3, 12] for further details.

The latter paper also describes the strategy adopted in order to deal with segment deletions. The generally high rate of spontaneous speech brings about a number of unintended deletions. Since, however, the relevant segments should be considered part of the speaker's articulatory plan, they should best be computed in the CCI metrics, which aims at measuring the language phonotactics. This is inspired by the general orientation of the CCI model, intrinsically phonological rather than phonetic. However, not all deletions are irregular: some of them have become conventionalized in the given dialect, and should thus be accepted as a regular feature.

The Italian and Chinese data in the corpus received a parallel treatment, whereby the discrimination between regular and irregular deletions was finely tuned to the individual language variety considered (Beijing Mandarin, Pisa Italian). The computations were, however, performed according to two alternative strategies, in order to have full control of the results: phonological (with irregularly deleted segments computed as part of the articulatory input) and phonetic (whereby they were ignored, in accordance with the actual output). This paper, for simplicity, only reports the results of the phonological analysis, which best mirrors the CCI rationale.

2.2. CCI computations

Fig. 2 presents the results of the CCI metrics as applied to the Pisa Italian and Beijing Chinese data. The empty diamond and square indicate the overall means of the two languages. The Italian mean projection is shown in two versions, depending on glides' treatment. With the empty square, the glides are treated as part of the (vocalic) syllable nucleus; with the cross, they are treated as part of the (consonantal) onset or coda, as appropriate. The latter treatment is more consistent with the phonology of Italian. Since, however, the Chinese phonology is usually assumed to dictate the alternative treatment with all vocalic segments regarded as part of the nucleus – although the phonological status of 'pre-nucleus' glides is still under debate [1, 7, 11] – the alternative treatment was extended to Italian for comparison. Considering that the two computations yielded fairly similar results for Italian (cf. fig. 2), in the following the data will only be presented according to the first strategy.

Considering the mean projections, it appears that Italian lies towards the low margin of the conceivable controlling area, while Chinese definitely falls within it. This was indeed predicted, for Chinese has a very simple phonotactics. The only type of coda consists of the nasals /n ŋ/, whereas Italian presents a relatively richer phonotactics. Fig. 2 also indicates the projections for three speech-rate groups in each language. As is well-known, speech rate is a powerful rhythm predictor, whatever the metrics used (see, e.g., [6]). This also turned out to be the case with the Italian and Chinese materials used in this study. The three rate groups were obtained by dividing the utterances of each language corpus into three subsets, as shown in table 1. The overall speed of the Chinese speakers was only slightly higher, as measured in segment/sec, than that of the Italian speakers. The comparison is thus sufficiently fair. The difference is supposedly due to the eliciting procedure used in the Italian corpus: in a map-task, Followers tend to have a more hesitant behavior than Givers.

As fig. 2 shows, the impact of speech rate is quite dramatic in both languages, although with slightly different effects. At the highest speed, the Italian speakers tend to converge towards the bisecting line, showing an increased controlling behavior. This is not surprising, considering that the faster one speaks, the less room there is for local durational fluctuations. In comparison, Chinese speakers seem to present a substantially linear behavior, equally distributing the speed impact between Vs and Cs.

The substantial effect of speed suggests that the use of a normalization procedure might contribute to gather a deeper view of the rhythmic inclinations of the two languages, by reducing the perturbations produced by the local speed fluctuations. The next section will detail the analysis.

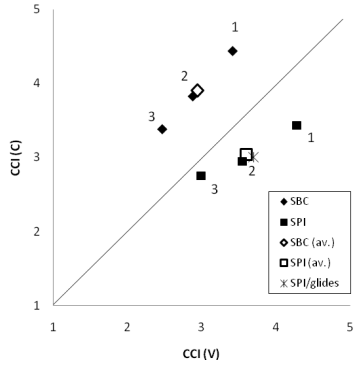


Figure 2: *Global rhythmic tendencies of Spontaneous Beijing Chinese (SBC; empty diamond) and Spontaneous Pisa Italian, with glides assigned to V- (SPI; empty square) or C-intervals (SPI/glides; cross). Filled markers: slow (1), medium (2) and fast (3) speech-rate projections.*

language	group	speech rate (*segm/sec*)	utterances n°
SBC	1	low ≦ 16.1 (average: 14.3)	205
	2	medium >16.1, < 18.8 (average: 17.4)	203
	3	hight ≧ 18.8 (average: 20.5)	199
SPI	1	low (≦ 14,7) (average: 13.4)	78
	2	medium >14.7, < 16.8 (average: 15.7)	74
	3	high ≧ 16.8 (average: 18.7)	81

Table 1. *Spontaneous Beijing Chinese (SBC) and Spontaneous Pisa Italian (SPI) data as divided into speech rate groups: slow (1), medium (2) and fast (3).*

3. Data Normalization

3.1. Normalization: first procedure

A normalization procedure has been applied to the PVI metrics [8]. Each difference between subsequent intervals (V or, respectively, C) was calculated as a proportion of the mean difference value between them, and the result was multiplied by 100 to yield a whole digit, as in (2), where *d* stands for "interval duration":

$$nPVI = \frac{100}{m-1}\sum_{k=1}^{m-1}\left|\left(d_k - d_{k+1}\right) \Big/ \frac{1}{2}\left(d_k + d_{k+1}\right)\right| \quad (2)$$

In the CCI metrics, each interval's duration is divided by the number of segments comprised in it; hence, calculating the corresponding proportion of the mean difference value amounts to normalize with respect to the local mean segment duration (once again, separately considered for V and C). See the $_{D}$nCCI formula in (3) (subscript *D* stands for "difference"):

$$ {}_{D}nCCI = \frac{100}{m-1}\sum_{k=1}^{m-1}\left|\left(\frac{d_k}{n_k}-\frac{d_{k+1}}{n_{k+1}}\right)\Big/\frac{1}{2}\left(\frac{d_k}{n_k}+\frac{d_{k+1}}{n_{k+1}}\right)\right| \qquad (3) $$

The results of the computation are shown in fig. 3, with both formulas (2) and (3) applied. As it happens, the formulas yield fairly comparable results, particularly so for the Chinese data. This is not surprising, for many V-intervals consist of just one V, and most C-intervals of just one C. By contrast, the Italian C-intervals often consist of two Cs. With respect to Italian, it is also to be noted that fig. 3 indicates a more markedly controlling behavior than fig. 2 (no such conclusion should be drawn from nPVI, due to its conception).

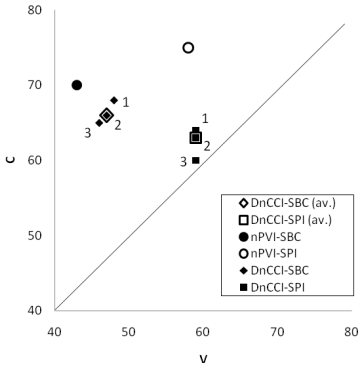


Figure 3: *Application of nPVI (circles) and* $_{D}$*nCCI (diamond and square) to the SBC and SPI data. Empty diamond and square = average; small filled markers = slow (1), medium (2) and fast (3) speech-rate projections.*

3.2. Normalization: second procedure

The nPVI and $_{D}$nCCI metrics present an obvious drawback. Suppose that, among two syllables, one is stressed: by normalizing over this sequence, one disruptively wipes out the stress effect. This weakness was pointed out by [2], who suggested larger normalization windows. The CCI model offers, however, another possibility for, as explained in § 1, its logic rests on the internal constitution of the V- and C-intervals. This allows for a more refined normalization procedure to be adopted. This is done in two steps.

First, the Italian and Chinese phonemes are grouped into natural classes, and the average duration of the elements in each class is calculated (table 2). Actually, not all classes in table 2 are "natural", for the ones consisting of diphthongs and triphthongs are a rough simplification. It should however be considered that any further subdivision might introduce idiosyncratic perturbations, due to the small digits in some cells. It should also be observed that, among the Chinese segments, the nasal /n/ was distinguished according to its syllabic status (onset vs. coda). A similar subdivision would also have been welcome for the Italian sonorants, of which Italian codas mostly consist. However, given the exploratory nature of this attempt, this improvement was for the time being neglected.

The second step consists in calculating the *z*-score normalization for the individual segments, according to the $_{S}$nCCI formula in (4) (with subscript *S* = "segment" and *z* = (d-m)/st.Dev, where *m* = mean duration of the relevant natural class's unit). The results are shown in fig. 4.

$$ {}_{S}nCCI = \frac{100}{m-1}\sum_{k=1}^{m-1}\left|\frac{z_k}{n_k}-\frac{z_{k+1}}{n_{k+1}}\right| \qquad (4) $$

2a Italian phoneme class	n°	m. duration
High vowels: *i u*	689	58.38
Mid-high vowels: *o e*	919	60.33
Mid-low vowels: *ɛ ɔ*	192	97.12
Low vowel: *a*	847	77.03
On-gliding diphthongs: *jV wV*	72	80.49
Off-gliding diphthongs: *Vj Vw*	56	105.76
Triphthongs: *wjV wVj*	2	98.20
Plosives : *p b t d k g*	759	70.18
Affricates : *ts tʃ dz dʒ*	90	106.70
Fricatives: *f v s z ʃ ʒ*	578	88.59
Liquids : *l r ʎ*	725	41.96
Nasals: *m n ɲ*	480	50.67
2b Chinese phoneme class	n°	m. duration
High vowels: *i u y*	1635	80.20
Mid-high vowels: *o ɤ*	1167	70.48
Mid vowel: *ɚ*	55	93.78
Low vowel: *a*	748	81.40
Diphthongs: *ai aɚ au ei ia iɚ ie ou ua uə uo ya ye yu əɚ*	2360	101.17
Triphthongs: *iaɚ iau iou uai uaɚ uei uəɚ yaɚ*	626	112.65
Aspirated plosives: t^h p^h k^h	440	90.28
Non-aspir plosives: *t p k*	1152	54.73
Aspirated affricates: $tʂ^h$ $tɕ^h$ ts^h	308	105.44
Non-aspir affricates : *tʂ tɕ ts*	903	61.44
Fricatives: *s f ʂ ʐ ɕ x v*	1253	78.00
Liquid: *l*	305	41.24
Nasals (onset): *m n*	1318	52.42
Nasals (coda): *n ŋ*	901	45.57

Table 2. *Mean segment duration (in ms) of the Italian and Chinese phonemic classes.*

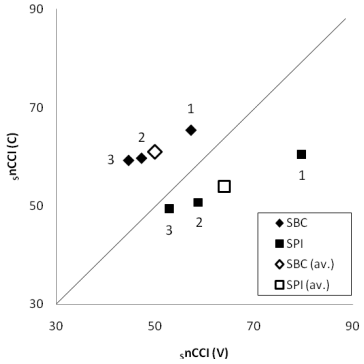


Figure 4. *Application of* $_{S}$*nCCI to SBC and SPI. Empty markers = average; filled markers = rate groups.*

Normalization had a different impact on Chinese and Italian. In order to clarify this point, table 3 was built so that the mean values of each rate-group show up as percentage values, with medium rate as the anchor. This highlights a number of facts. First, in the CCI data of table 3a, the acceleration effect is roughly the same in both languages. The slow-rate V-values are 24% (for SBC) and 23% (for SPI) larger than the medium rate values, while fast rate differs from the latter by -15% and -14%, respectively. The picture for the C-values is again fairly similar for the two languages. After $_{S}$normalization, however, things change significantly. In SBC, the V-values distance between medium and fast rate reduces considerably, and the same happens for the C-values distance between slow and me-

dium and, even more, medium and fast. As for SPI, the major change concerns the V-values distance between slow and medium, which increases substantially. To the extent that the reference corpus is sufficiently representative, these results suggest that the interaction between speed compression and the language-specific phonotactic constraints creates in both languages (particularly so for Italian) larger local durational fluctuations within the V- than the C-component. Moreover, in both languages the C-compressibility threshold seems to be tendentially reached already at medium rate.

3a		CCI/V	CCI/C
SBC	slow (≦16,1)	124	117
	medium	100	100
	fast (≧ 18,8)	85	87
SPI	slow (≦14,7)	123	115
	medium	100	100
	fast (≧16,8)	86	92
3b		$_{S}$nCCI/V	$_{S}$nCCI/C
SBC	slow (≦16,1)	121	109
	medium	100	100
	fast (≧18,8)	94	99
SPI	slow (≦14,7)	136	119
	medium	100	100
	fast (≧16,8)	90	97

Table 3. *Proportional effect of speech rate (in percentage with respect to the medium rate value) on the CCI (3a) and $_{S}$nCCI (3b) measures for Chinese and Italian.*

Italian	N	*r*		**Chinese**	N	*r*	
CCI/V	233	-0,509	**	CCI/V	607	-0,394	**
CCI/C	233	-0,278	**	CCI/C	607	-0,326	**
$_{D}$nCCI/V	223	-0,071		$_{D}$nCCI/V	587	0,051	
$_{D}$nCCI/C	223	-0.003		$_{D}$nCCI/C	587	0,062	
$_{S}$nCCI/V	233	-0,450	**	$_{S}$nCCI/V	607	-0,403	**
$_{S}$nCCI/C	233	-0,222	**	$_{S}$nCCI/C	607	-0,278	**
rPVI/V	233	-0,385	**	rPVI/V	607	-0,281	**
rPVI/C	233	-0,488	**	rPVI/C	607	-0,358	**
nPVI/V	233	-0,161	*	nPVI/V	607	-0,071	
nPVI/C	233	0,019		nPVI/C	607	0,072	

Table 4. *Spearman's correlation coefficient of the various algorithms relative to the speech rate factor (** = p≤0.01; * = p≤0.05).*

In order to compare the impact of the normalization procedures, one can compute the correlation between the outputs of the three CCI formulas (plus PVI for comparison) with the speed factor as measured in segment/sec (see table 4). The statistical measure used was the Spearman coefficient, since part of the data (even within one and the same rhythm measure) had a non-normal distribution. Altogether, the correlation is not high: the highest *r* value occurs for Italian CCI/V. In fact, the percentage of the variance explained by the linear model was below 30%; evidently, other factors were involved beyond speed (e.g., stress). The partial correlations separately computed for Vs and Cs, assuming the other component (Cs and Vs, respectively) as a dependent factor, were also calculated, but no substantial difference emerged.

With most metrics, Vs unsurprisingly turned out to be more speed-correlated than Cs. The fact, however, that Chinese altogether emerged as less speed-dependent than Italian could not be predicted. rPVI is the odd-man out, for it yielded higher values for Cs than for Vs. This output is ostensibly due to the very logic of this metrics, since V- and C-intervals are viewed as indivisible units, without considering the number of elements included. This looks like a fundamental drawback. More generally, CCI, $_{S}$nCCI and rPVI are negatively, but highly significantly, correlated with speed, whereas $_{D}$nCCI and to a large extent nPVI are non-correlated (Italian nPVI/V is marginally significant).

These results seem to suggest that $_{D}$nCCI and, in part, nPVI depart more radically from any possible effect of rate. Note, however, that sterilizing the rate effect is not a virtue in and by itself, for this result might be a by-product of an invasive data transformation. nPVI might be one such case, since it is based – in contrast to $_{D}$nCCI – on the average difference between adjacent intervals, rather than segments. This may have disruptive results on languages with large enough differences in the segmental constitution of the intervals. While further research is needed, the following indications seem to emerge: (i) rPVI ostensibly exaggerates the role of Cs w.r.t. to Vs; (ii) since $_{S}$nCCI does not substantially differ from CCI, it is probably inadequate for the normalization purpose. Nevertheless, the comparison between fig. 2 and 4 (as well as the data in tables 3 and 4) shows that the application of $_{S}$nCCI is not without consequences, for it has a different impact on Chinese as opposed to Italian.

Our laboratory is now examining further theoretical options, which will be the object of future communications.

4. References

[1] Bao, Z. "The asymmetry of the medial glides in middle Chinese", Proc. 7th International and 19th National Conferences on Chinese Phonology, 7-27, Taipei, 2001.

[2] Benton, M. "A preliminary analysis of the relationship of speech rate to speech-timing metrics as applied to large corpora of non-laboratory speech in English and Chinese broadcast news", Proc. 4th International Conference on Speech Prosody, 11-14, Chicago, 2010.

[3] Bertinetto, P.M. and Bertini, C. "On modelling the rhythm of natural languages", Proc. 4th International Conference on Speech Prosody, 427-430, Campinas, 2008.

[4] Bertinetto, P.M. and Bertini, C. "Towards a unified predictive model of natural language rhythm", In: Russo, M. [ed.], Prosodic Universals. Comparative Studies in Rhythmic Modeling and Rhythm Typology, 43-77, Naples: Aracne, 2010.

[5] Bertini, C. and Bertinetto, P.M. "Propezioni sulla struttura ritmica dell' italiano basate sul corpus semispotaneo AVIP/API", In: Romito, L., Galatà, V. and Lio, R. (eds.), La fonetica sperimentale: metodo e applicazioni-Proc. 4o Convegno Nazionale AISV, EDK Eitore, 2007.

[6] Dellwo, V. "Rhythm and speech rate: a variation coefficient for deltaC", In: Karnowski, P. and Szigeti, I. [eds.], Language and language processing, 231-241, Frankfurt: Peter Lang, 2006.

[7] Duanmu, S. The Phonology of Standard Chinese 2nd edition, New York: Oxford University Press Inc, 2007.

[8] Grabe, E. and Low, E.L. "Durational variability in speech and the rhythm class hypothesis", In: Gussenhoven, C. and Warner, N. [eds.], Papers in Laboratory Phonology 7, 515-546, Berlin: Mouton de Gruyter, 2002.

[9] Li, A.J. "Chinese Prosody and Prosodic Labeling of Spontaneous Speech", Proc. 1st International Conference on Speech Prosody, Aix-en-Provence, 2002.

[10] Li, A.J., Yin, Z.G., Wang, M.L., Xu, B. and Zong, C.Q. "A spontaneous Conversation Corpus CADCC", Oriental COCOSDA Workshop, South Korea, 2001.

[11] Wan, L.P. "Alignments of Prenuclear Glides in Mandarin", Taipei: Crane Publishing Co. LTD, 2002.

[12] Zhi, N., Bertinetto, P.M., and Bertini, C. "Modelling the speech rhythm of Beijing Chinese in the CCI framework", submitted to Proc. 17th ICPhS, HongKong, 2011.