

## VtoV: a perceptual cue for rhythm identification

*Massimo Pettorino, Marta Maffia, Elisa Pellegrino, Marilisa Vitale, Anna De Meo*

{mpettorino, maffia, epellegrino, vitalem, ademeo}@unior.it  
University of Naples L'Orientale, Italy

### Abstract

Current metrics for the quantification of speech rhythm take into account parameters not easily detectable by listeners. To overcome this limit, in this study we propose a new model based on a parameter that account for listeners' ability to discriminate between different rhythmic patterns.

Starting from the results of a spectro-acoustic analysis conducted on singing, we found that Perceptual Centres align close to Vowel Onset Points (VOP). To test the perceptual relevance of interval between two consecutive VOPs, that we call VtoV intervals, we analyzed a multilingual corpus of TV news, advertisings and recited speech. The signal was segmented into vocalic/consonantal portions and into VtoV intervals. The standard deviation of all parameters was calculated. The results of the analysis show that VtoV is a crucial parameter both to classify languages on a rhythmic basis and to account for intra-linguistic speech style variations.

### 1. Introduction

There is ample evidence in literature that languages differ considerably in the way they produce rhythmical contrasts. According to the way isochrony is realized languages have been traditionally classified into three main groups: syllable-timed, stress-timed and mora-timed languages (see Pike 1945; Abercrombie 1967; Ladefoged 1975).

In the first group, syllable duration tends to remain relatively constant and unstressed syllables cannot be drastically reduced. In stress-timed languages, by contrast, stressed syllables recur at equal intervals and there is a substantial degree of syllable duration variability (see Dauer 1983). In the third group the mora, a sub-syllabic unit consisting in one short vowel and any

preceding onset consonants, serves as a basic unit of rhythmical organization.

Nevertheless, over the years the attempts to find experimental evidence supporting the notion of acoustic isochrony have largely failed (for a review, see Bertinetto 1989; Kohler 2009) primarily because of the many factors that influence spoken communication, modifying its temporal organization.

Moreover, the attribution of rhythm-classes to particular languages requires very accurate syllable boundaries identification. However, syllable boundaries are hard to identify particularly when they occur within

1. a silent interval of a long stop consonant,
2. a cluster consisting of a nasal plus a voiced stop.

In this case, the signal does not contain any discontinuity for the effect of a full or partial nasalization of the voiced-stop.

A further point to be clarified on the concept of isochrony is whether the syllable duration is considered from the view point of articulatory production or from that of perception. In fact the perceptual duration of a syllable does not necessarily correspond to the duration of the articulatory gesture. For instance, in a voiceless stop CV syllable, the articulatory duration is longer than the perceptual one, because the articulatory mechanics begins before the onset of the acoustic signal.

An attempt to overcome some of these methodological limits is represented, among others, by the work of Ramus et al. (1999). Starting from many experiments on the listeners' ability to discriminate between

languages with different rhythmic patterns, they proposed a new method to assign languages to the three different rhythmic groups. To the purpose, they calculated the proportion of vocalic intervals within the sentence and the standard deviation of consonantal intervals. The results of their study have indicated that syllable, stress and mora-timed languages differ from each other in the percentage of vocalic portion (%V) and in the standard deviation of the durations of consonantal intervals ( $\Delta C$ ).

Despite the numerous methodological advantages to this procedure, the %V/ $\Delta C$  model does not seem to account for listener's ability to discriminate between languages according to their rhythmic features. Listeners are unlikely to manage to calculate, even roughly, the percentage of vocalic intervals and the standard deviation of consonantal clusters in real-time. Consequently, there should be another parameter, perceptually detectable, enabling listeners to distinguish a rhythmic pattern from another.

As rhythm is the regular succession of prominences in time (see Marotta 2011), such a parameter should be then linked to the recurrence of audible signal discontinuities.

In this regard, a sizeable body of research has demonstrated the existence of prominent instants in the speech signal that are perceptually more salient than others. These instants, called Perceptual Centres or P-Centres (see Morton et al. 1976) correspond to a particular point within the syllable that perceptually corresponds to its "moment of occurrence" (see Marcus 1981).

Additionally, sequences of P-Centres are thought to underlie the perception and production of rhythm in perceptually regular speech sequences. However, physical correlates of P-Centres have not been firmly established (see Villing 2003) and their exact location is a current matter of experimental verification (see Villing 2010 for reviews).

## 2. The study

### 2.1. First experiment

The existence of P-Centres is particularly evident in singing. In this case, the tempo of the music requires the singer to produce each syllable at precise time points. But how can a syllable, which corresponds to a time interval, be synchronous with an instant? There should exist within that interval a perceptually prominent point which allows for such synchronization.

In order to answer this question, we asked a professional singer to record an Italian song going in time with the beats of a metronome (92 bpm). The spectro-acoustic analysis of the corpus, carried out by means of Praat, has shown that all beats align with the vowel onsets, thus confirming some data present in the literature (see Tuller and Fowler 1980). 74% of beats occurs within 0.005 s from the vowel onset, while 26% shifts on average 0.034 s ( $\sigma = 0.008$  s) (fig. 1).

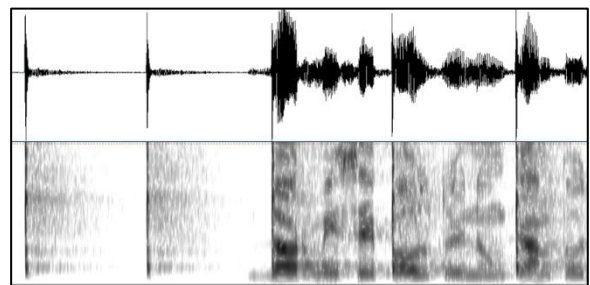


Figure 1

In order to collect further evidence that P-Centres tend to align close to the Vowel Onset Point (VOP), we examined spectro-acoustically some commercial songs played by very well-known artists. It is to be underlined that, unlike the song performed on the beats of a metronome, for songs with a complex instrumental accompaniment, it is not always possible to check on the spectrogram the synchronization between lyrics and music. However, where this synchronization was verifiable, the analyses have confirmed that P-Centres were located close to VOPs. It is therefore possible to

conclude that the VOPs represent those audible signal discontinuities that would guide listeners in the perception of rhythm. As a consequence, the interval between two consecutive vowel onset points (henceforth called VtoV interval) seems to be the cue enabling listeners to identify the rhythmic pattern of a language.

2.2. Second experiment

To test the role of VtoV in rhythm perception, we collected a multilingual corpus of about 15 minutes. The corpus was composed of TV news readings, speech taken from drug advertisements and recited speech. The languages were representative of the three rhythmic groups: Italian, French, English and Japanese.

As for the TV news, the speech samples were taken from RAI, RTF, BBC and NHK channels. Drug ad samples were drawn from the end of pharmaceutical television commercials when the voiceover recites the contraindications and side effects. Here, the speech is deliberately accelerated, sometimes through the use of signal manipulation, in order to hinder the full understanding of the message. As for the recited speech, the corpus consisted of verses from: 1) Shakespeare's 20th sonnet "A woman's face", 2) Montale's "Le quattro stagioni" and 3) Prévert's "Cet amour".

The entire corpus was segmented into vocalic/consonantal portions and into VtoV intervals on two separate tiers. The segmentation of glides followed the rules adopted by Ramus et al. (1999): [w] and [j] were treated as consonants and the boundary was placed between the approximant and the vowel. Falling diphthongs were segmented in one or two vowels intervals depending on the spectro-acoustic characteristics of the tract. If both vowels presented a specific steady-state formant pattern, the diphthong was divided into two VtoV intervals; otherwise, it was treated as a single interval.

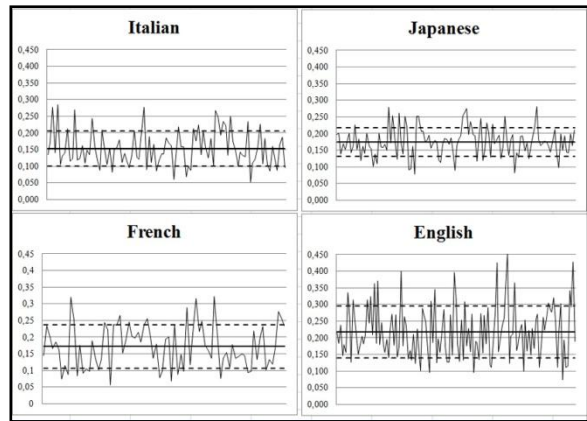


Figure 2

For each speech sample we calculated the average value of VtoV intervals, their standard deviation ( $\Delta VtoV$ ), the percentage of vocalic (%V) and consonantal (%C) portions and their standard deviations ( $\Delta V$  and  $\Delta C$ ).

Figure 2 represents the sequence of VtoV intervals on *x* axis, and on *y* axis their duration (s). The data refer to TV news in Italian, French, Japanese and English. The continuous horizontal line indicates the average value of VtoV, while the dashed lines indicate the  $\Delta VtoV$ . As the figure shows, English differs from the other languages, both for higher VtoV and for larger  $\Delta VtoV$ . These results reflect the fact that English, compared to Italian, French and Japanese, is characterized by wider variety of syllable structure type, more complex consonant clusters, higher frequency of closed syllables and drastic reduction of unstressed vowels.

Tables 1 and 2 present VtoV and  $\Delta VtoV$  for recited speech in Italian, French and English, in comparison with TV news. VtoV increases by about 40% in the three languages, and  $\Delta VtoV$  undergoes an increase of 40% in English, of 14% in French, and of 55% in Italian.

	TV news (a)	Recited speech (b)	Difference (b-a)	%
Italian	0.153	0.213	+ 0.060	+39
French	0.172	0.250	+ 0.078	+45
English	0.215	0.299	+ 0.084	+39

Table 1

	TV news (a)	Recited speech (b)	Difference (b-a)	%
Italian	0.053	0.082	0.029	+55
French	0.065	0.074	0.009	+14
English	0.089	0.125	0.036	+40

Table 2

An opposite trend to recited speech was observed in the drug ads. The VtoV decreases by 38% for Italian and by 35% for French (Tab. 3). In both languages  $\Delta VtoV$ , instead, decreases by 55% (Tab. 4).

	TV news (a)	Drug ad. (b)	Difference (b-a)	%
Italian	0.153	0.095	-0.058	-38
French	0.172	0.112	-0.060	-35

Table 3

	TV news (a)	Drug ad. (b)	Difference (b-a)	%
Italian	0.053	0.024	-0.029	-55
French	0.065	0.029	-0.036	-55

Table 4

Figure 3 shows the relationship between VtoV and  $\Delta VtoV$ , analyzed *per* different languages and speech styles. Data indicate quite evidently that there is a direct relationship between the two variables: the higher the VtoV the larger  $\Delta VtoV$ . Additionally, regardless of language and speech style, wider intervocalic intervals undergo greater variations. On the contrary the closer the vowels are, the more constant the intervocalic intervals.

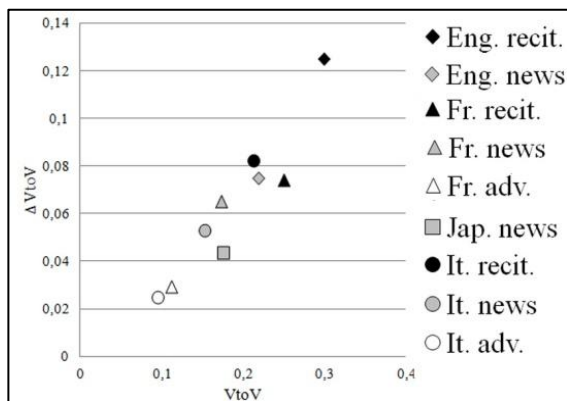


Figure 3

### 2.3. Third experiment

Another experiment was conducted to better investigate the relationship between VtoV variation and vocalic and consonantal variability. To the purpose, we plotted VtoV with  $\Delta V$  and  $\Delta C$ , and then  $\Delta C$  with  $\Delta V$  (fig. 4). From the three graphs it is possible to infer that VtoV variations are more greatly determined by  $\Delta C$  rather than by the  $\Delta V$ .

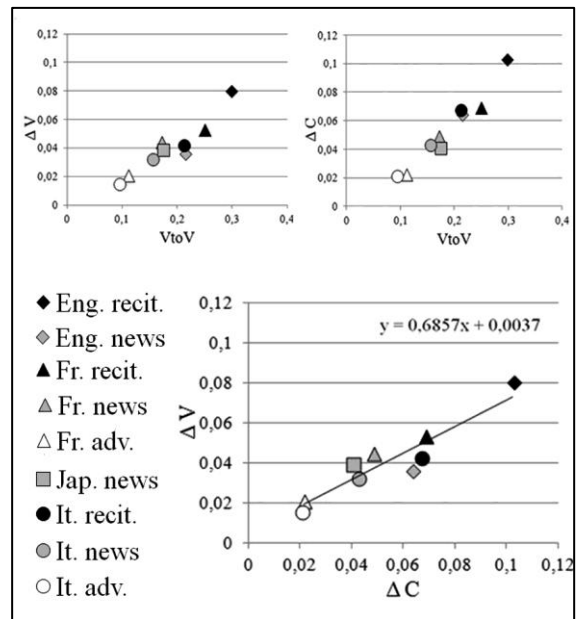


Figure 4

Given this direct relationship between VtoV and  $\Delta C$ , we propose to revisit the %V/ $\Delta C$  model by substituting  $\Delta C$  with VtoV, the only perceptually salient and detectable parameter for listener. Therefore, we analyzed our multilingual corpus according to both models (%V/ $\Delta C$  and %V/VtoV). The figures 5 and 6 show the results of both analyses. In the two graphs languages of different rhythmic groups are distributed along the x axis. From left to right there is English, stress-timed language, then Italian and French, syllable-timed languages, and then Japanese, isomoraic language. The distribution of languages along y axis indicates intra-language differences due to the diverse speech styles. In fact, going from the bottom to the top, the

rate of the speech samples moves from very fast to very slow.

To confirm whether differences in speech rate were recognized also on a perceptual level, 80 Italian listeners, aged between 18 and 23, were involved in a perception test. They were asked to evaluate the speech rate of 9 speech samples on a three-point scale (slow, medium and fast). To eliminate the message component, the samples were manipulated through lowpass filtering technique (cut-off frequency 400 Hz). The results show that, regardless of language and speech style, the excerpts judged as “slow” are those with a VtoV higher than 0.250 s; those recognized as “fast” corresponded to VtoV of about 0.1 s; those considered as “medium” were between 0.1s and 0.2 s.

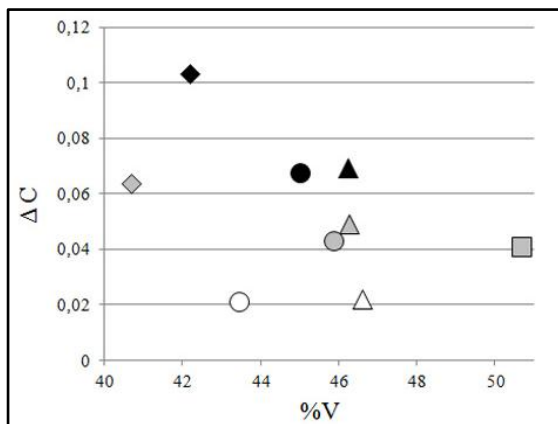


Figure 5

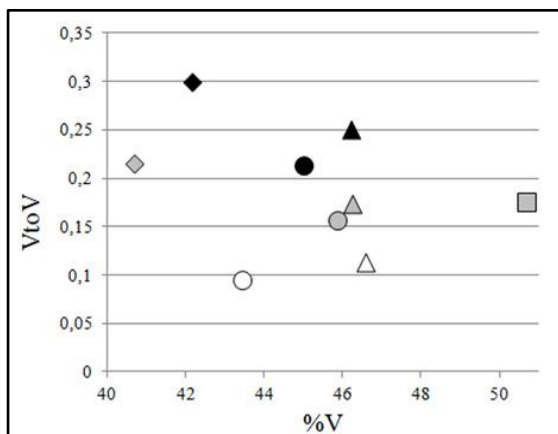


Figure 6

Nevertheless, it is important to underline that English presents VtoV values that are higher than those in French and Italian both for recited speech and news reading. This proves that VtoV is a parameter that does not only depend on speech rate but also on the rhythmic characteristics of the language. To test this hypothesis we will extend our analysis to other languages belonging to the three different rhythmic groups.

### 3. Conclusions

This study, performed on languages with different rhythmic characteristics and on different speech styles, shows that VtoV interval represents a relevant cue in the perception of rhythm. Under this perspective, the consonantal intervals can be considered as the interruptions or attenuations in the speech signal that determine those discontinuities underlying the perception of rhythm. These discontinuities, indeed, consist in the periodic recurrence of fully resonant vowel sounds. The %V/VtoV model is therefore very effective to represent the different rhythmic patterns of languages, providing a very articulated framework of the possible combinations among different languages and different types of speech. Data from our study seem to show that, speech style being equal, there is an inverse relationship between the two parameters: the higher %V, the lower VtoV.

In further steps of our research we will investigate whether the perception of speech rate, that is proved to be linked to VtoV variations, depends on the different rhythmic groups of languages. To the purpose, we will administer perception tests based on natural speech to native speakers of the target languages.

### References

Abercrombie, D. (1967). *Elements of general phonetics*. Aldine, Chicago.

- Bertinetto, P. M. (1989). Reflections on the dichotomy «stress» vs «syllable timing». *Révue de Phonétique Appliquée* 91, pp. 99-129.
- Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11, pp. 51-62.
- Kohler, K. J. (2009). Rhythm in speech and language. A new research paradigm. *Phonetica* 66, pp. 29-45.
- Ladefoged, P. (1975). *A course in phonetics*. Harcourt Brace Jovanovich, New York.
- Marotta, G. (2011). Ritmo, voce dell'*Enciclopedia dell'Italiano Treccani*, vol. II, Istituto dell'Enciclopedia Italiana, Simone Raffaele, 1262, 2011.
- Marcus, S. M. (1981). Acoustic determinants of Perceptual-centre (P-Centre). *Perception and Psychophysics*, 30, pp. 247-256.
- Morton, J., S. Marcus, & C. Frankish (1976). Perceptual Centers (P-centers). *Psychological Review*, 83:5, pp. 405-8.
- Pike, K. L. (1945). *The intonation of American English*. Ann Arbor, Michigan: University of Michigan Press.
- Ramus, F., M. Nespors, & J. Mehler (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 72, pp. 1-28.
- Tuller, B. & C. A. Fowler (1980). Some articulatory correlates of perceptual isochrony. *Perception and Psychophysics*, 27: 4, pp. 277-283.
- Villing, R., T. Ward & J. Timoney (2003). P-Centre Extraction from Speech: the need for a more reliable measure *Proceedings of ISSC 2003*, Limerick.
- Villing, R. (2010) Hearing the Moment: Measures and Models of the Perceptual Centre [http://eprints.nuim.ie/2284/1/Villing\\_2010\\_-\\_PhD\\_Thesis.pdf](http://eprints.nuim.ie/2284/1/Villing_2010_-_PhD_Thesis.pdf)