

# HOW PROSODIC VARIABILITY CAN BE HANDLED BY A DYNAMICAL SPEECH RHYTHM MODEL

Plínio A. Barbosa

Speech Prosody Studies Group/DL/IEL/State University of Campinas, Brazil

plinio@iel.unicamp.br

## ABSTRACT

This paper presents how a two-coupled-oscillator speech rhythm model can handle the variability of the durational patterns found in natural data. For doing so the model takes into account a small set of nested levels of dynamical coupling between linguistic and production-related subsystems. In this framework speech rhythm is the quasi-optimal output of the coupling between two components, a perception-oriented tendency to pattern structuring, implemented by the inter-relation between local syntactic information and a phrase stress oscillator, and a production-oriented regularity constraint, implemented by the oscillation of two components, a syllabic oscillator and a phrase stress oscillator. The syllabic oscillator's pulses are anchored at vowel onsets, implementing the carrier component of speech rhythm production, characterising the building-block of prosodic timing. The model generates complex patterns of V-to-V durations via the consequences of a phrase stress oscillator's entrainment onto the syllabic oscillator. This mechanism of generation is able to cope with intra- and inter-speaker rhythmic variability.

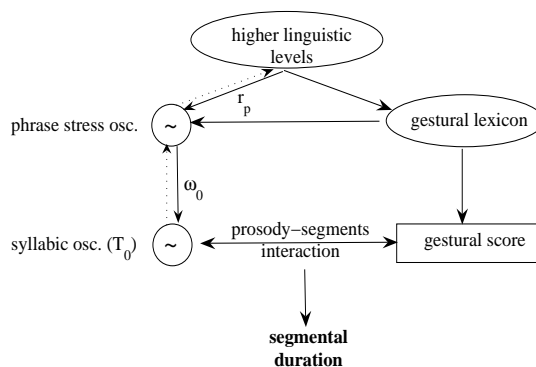
**Keywords:** Coupled oscillators; timing; dynamical systems; phonetics.

## 1. A BLUEPRINT FOR THE (IN-BEAT) SPEAKER

The dynamical speech rhythm production model (henceforth DSR model) described in this section was presented on other occasions [1, 3], where a detailed account of its dynamical features as well as its linguistic (and biomechanic) plausibility for obtaining patterns of V-to-V durations in Brazilian Portuguese (henceforth BP) was addressed. In the present paper, its ability to cope with rhythmic variability is highlighted. The model is couched in terms of Dynamical Systems Theory, more specifically the Coupled-Oscillators Theory, within which dynamical coupling is the mechanism responsible for generating complex patterns of duration at the surface. There are three sets of coupling forces acting in the

model, all three operating at distinct temporal scales [9]. A blueprint of the model can be seen in Fig. 1.

Figure 1: The DSR model.



The first set of coupling forces is implemented by the probabilistic-controlled coupling between syntax and production constraints parameterised by the coupling strength  $r_p$ , as shown in Fig. 1. Its temporal order of magnitude is the second. The second set of coupling forces operates between the phrase stress oscillator, and the syllabic oscillator shown in Fig. 1, whose coupling is parameterised by the relative coupling strength  $\omega_0$ . For the coupled-oscillators model on the left, the number and identity of phones are both unknown, which signals the abstractness of the V-to-V durations delivered by the syllabic oscillator. The temporal order of magnitude of this coupling lies in the hundreds of milliseconds. The prosody-segments interaction, which produces overt segmental durations, stands for the third set of coupling forces. A first attempt for implementing this third coupling, presented elsewhere [3], was sufficient to produce surface durations which can be perceptually evaluated by using an analysis-resynthesis technique. The temporal order of magnitude at this level lies in the tens to hundreds of milliseconds. This phoneme-sized temporal scale is not considered here.

Since the DSR model separates stress-group-sized

and syllable-sized from phoneme-sized sources of duration specification, it exhibits similarities with both the Frame/Content [6] and the Slots/Fillers [11] theories. Nevertheless, the importance of the parametric control of dynamical behaviour, the anchoring of the syllabic oscillator at vowel onsets, and the explicit inclusion of the coupling between syntax and production constraints, are all features distinguishing both of these theories from our model.

The finite-difference equation in 1 implements the period coupling in the coupled-oscillators model, where  $\Delta T(n)$  is the current quantity to be added to the previous V-to-V duration of the syllabic oscillator,  $T(n-1)$ , to produce the current period of the syllabic oscillator,  $T(n)$ , according to the iterative equation:  $T(n) = T(n-1) + \Delta T(n)$ .

$$(1) \quad \Delta T(n) = \alpha.T(n-1).s(n).i(m) - \beta.[T(n-1) - T_0].i(m-1)$$

The real positive parameters  $\alpha$ ,  $\beta$ , and  $T_0$  are, respectively, the entrainment rate, the decay rate, and the underlying uncoupled period of the syllabic oscillator in seconds. The inverse of  $T_0$  specifies speech rate underlyingly. The index  $n$  refers to the current V-to-V unit, whereas  $m$  refers to the current stress group (also referring to the position of the phrase stress oscillator pulse along the utterance being modelled, since in BP the right edge of a stress group is the phrase stress position). The non-linear function  $s(\cdot)$  and the amplitude  $i(m)$  are, respectively, the synchronicity function (given by the set of equations in 2) and the phrase stress oscillator pulse magnitude, whose position finishes the current stress group  $m$ . The phrase stress oscillator is a train of pulses whose magnitude  $i$  and position  $m$  are both given by a probabilistic algorithm (see below).

$$(2) \quad s(n) = (1 - \omega_0).s(n-1) + \omega_0.exp(-N + n + 2), \text{ for } 0 \leq n < N - 1 \\ = \omega_0.exp(-5.81 + 0.016.T_0), \text{ for } n = N - 1$$

In the set of equations 2, defining the synchronicity function  $s(\cdot)$ ,  $N$  is the number of V-to-V units within the current stress group (dominated by the phrase stress oscillator pulse  $m$ ). The index  $n$  refers to the current V-to-V unit (which has value 0 for the first unit and  $N - 1$  for the last one in the current stress group). The relative coupling strength  $\omega_0$  is a real number in the interval [0-1], considered to vary more extremely from language to language, than from speaker to speaker within a same linguistic community [1]. This difference in cross- and intralinguistic variation needs, though, to be verified

experimentally. This coupling is related to the statistical classes of syllable-timed and stress-timed languages, considered as ideal attractors: the greater the value of  $\omega_0$ , the greater the influence of the phrase stress oscillator onto the syllabic oscillator. There is no need, though, to refer to any kind of absolute isochrony in this view of rhythm typology.

With respect to the syntax-rhythm interface, an algorithm with two coupled, probabilistic components generates the magnitude and attributes the position of duration-related phrase stress. The syntactic component is implemented by computing the conditional probability,  $p(ps/m)$ , that a phonological word bears phrase stress given a specific Dependency Grammar-marker to its right. Each marker is one of eleven syntactic markers projected onto the syntagmatic axis from dependency and strength relations between two adjacent nuclei in a dependency tree [2]. The regularity-constraint component, on the other hand, implements the production constraints on stress group size similarity by computing the conditional probability,  $p(ps/nVV)$ , that a phonological word bear phrase stress given the distance in number of V-to-V units ( $nVV$ ) from the previously assigned phrase stress. Both components are linearly combined by using the coupling strength  $r_p$ , in order to compute the likelihood of phrase stress,  $l(ps)$ , at the current phonological word (eq. 3).

$$(3) \quad l(ps) = \logit[p(ps/m)]r_p + (1 - r_p)\logit[p(ps/nVV)]$$

The *logit* values,  $\logit(p) = \ln(p/1-p)$ , of the probabilities were computed in order to obtain a likelihood extending outside the interval [0-1], in such a way as to correspond to the domain of normalised V-to-V duration. Observe in eq. 3 three possible sources of intra- and inter-speaker variability: the two conditional probabilities and the coupling strength between the components, all three defined within the [0-1] interval. It will be shown in sections 3 and 4 that BP speakers vary in these three dimensions when attributing phrase stress in read speech. As to  $r_p$ , if it is greater than 0.5, the syntactic component dominates the regularity component. If its value is less than 0.5, the opposite trend is simulated. This could predict the behaviour of speakers that rely more on syntax than on regularity constraints when uttering and vice-versa.

Both conditional probabilities were estimated from the Lobato corpus (an excerpt of a well-known Brazilian children's book with 110 words) at normal speech rate for four speakers, AP, AC, DP, and

PA. Since the four speakers behave in a similar way with respect to the estimation of the syntactic conditional probability, it was possible to ensure significance for additional syntactic markers by using a larger corpus read by PA, the *Pantanal* corpus (353 words). The regularity component conditional probability was implemented, on the other hand, by a lognormal distribution, since it was the best fit for all four speakers. This distribution allowed to specify this component by two parameters, log-transformed mean and st.-dev. for all speakers: (1.8, 0.4) (about 6 and 2 V-to-V units per stress group, respectively). The position of phrase stress was attributed within a rate-dependent lookahead window (measured in number of underlying syllables) for the phonological word with the greatest likelihood,  $l(ps)$ , within the window. The phrase stress (normalised) magnitude was that position is the value of its corresponding likelihood.

The DSR model allows several avenues for simulating the sources of intra- and inter-speaker rhythmic variability found in natural data. Before presenting these modes of variability, a brief summary of the importance and the methodology underlying the study of normalised V-to-V duration is presented in the next section.

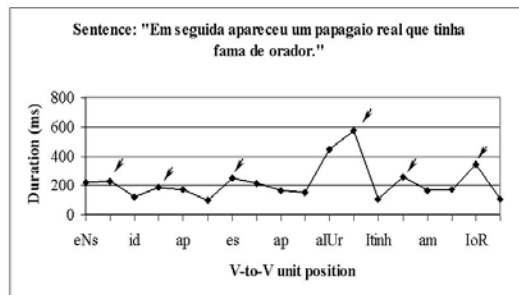
## 2. V-TO-V DURATIONAL PATTERNS IN BP

Stress groups were delimited by production criteria, namely, the position of maxima of normalised, syllable-sized duration. These positions correspond to those of produced phrase stresses whose level of prominence is signalled by the duration normalised value. Maxima of V-to-V (and not phonological syllable) duration define the right edges of stress groups in BP connected utterances, since BP is a right-headed language at the stress group domain, for which syllable-sized duration is the main correlation of both lexical and phrase stress [8]. The relevance of V-to-V units for both speech production and perception has largely been demonstrated since the 1970s in the literature on p-centers [7], as well as in the psycholinguistic literature [5], both showing the relevance of CV transition tracking for production and perception. More recently, the robustness of CV transition detection throughout the mammalian auditory pathway was demonstrated [12]. The choice of V-to-V instead of VC for the unit name is motivated for two reasons: to avoid the (possible) association of VC with tautosyllabic units only, and to remind the relevance of the vowel flow in speech production in the DSR model framework [10].

For analysis, the sequence of phrase stress posi-

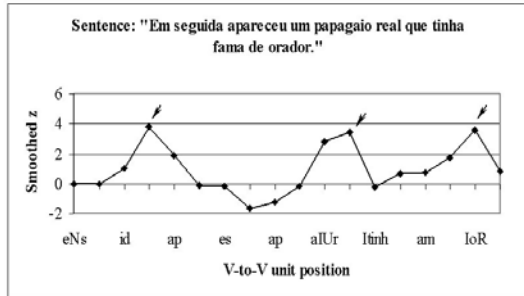
tions in the Lobato corpus, read at three self-chosen speech rates, was automatically tracked by serially applying two techniques for normalising raw duration: a  $z$ -score transform procedure, and a 5-point moving average filtering procedure. The two-step normalisation technique, and the detection of duration-related phrase stress boundaries were implemented in a Praat [4] script written by the author. If the first technique is classical in statistical analysis, the second is not. The  $z$ -score smoothing attenuates local duration decreases attributed to the implementation of BP lexical stress in penultimate stress (e.g. *casa*, house) and antepenultimate stress words (e.g. *xícara*, cup). The effect of smoothing is to ensure that mostly prosodically prominent V-to-V durations will remain. The application of the two normalising techniques produces the V-to-V duration patterns shown in Fig. 3 from the raw durations in Fig. 2. The sentence “Em seguida apareceu um papagaio real que tinha fama de orador” from the Lobato corpus illustrates the procedure. The three duration-related phrase stress boundaries in Fig. 3 delimit two stress groups, excluding the anacrusis “Em seguida”: [apareceu um papagaio real **qu-**] and [-e tinha fama de orador] (bold face signals phrase stress position). The DSR model is able to gener-

**Figure 2:** V-to-V raw duration evolution in ms for the sent. “Em seguida apareceu um papagaio real que tinha fama de orador.” read by speaker AP.



ate these patterns of normalised duration. For doing so, the initial model parameters were estimated from BP data with a set of 108 isolated utterances read by a fifth speaker. The values of  $\omega_0$ ,  $\alpha$  and  $\beta$  were optimised in such a way as minimising the error between the values  $T(n)$  of the entrained period of the syllabic oscillator generated by the model, and the averaged V-to-V durations in corresponding position  $n$ , for stress groups of the same size in the corpus. This classical technique of Control Theory gave the approximative values of  $\omega_0 = 0.8$ , and  $(\alpha, \beta) = (0.4, 1.0)$ . The value of  $T_0$  was estimated by averaging all raw V-to-V durations not corresponding to phrase stress position and one unit to its left

**Figure 3:** V-to-V smoothed  $z$  – scores for the sentence “Em seguida apareceu um papagaio real que tinha fama de orador.” read by speaker AP. The three local peaks are taken as phrase stresses.



position. Otherwise referred to, all simulations presented here used the following input parameter values:  $\alpha = 0.4$ ,  $\beta = 1.0$ ,  $\omega_0 = 0.8$ , syntactic conditional probability for speaker AP (very similar to the others), log-transformed mean and st.-dev. of stress group extension (in number of V-to-V units) of [1.8 0.4], mean and st.-dev. of stress group extension (in ms) of [1300, 600] (this latter to estimate the extension of the lookahead), and  $r_p = 0.6$ .

By optimising the magnitude  $i$  of the period coupling in eq. 1 in such a way as obtaining a normalised, abstract V-to-V duration at the right end of the stress group identical to the likelihood of phrase stress, it was possible to obtain (4), which allowed the generation of phrase stress oscillator pulse magnitude from the probabilistic estimation of phrase stress magnitude.

$$(4) \quad i = 1.11 + 0.72l(p \ s).T_0 + 0.5T_0$$

The DSR model parameters can be changed in order to cope with at least two main modes of rhythmic variability, intra- and inter-speaker variability.

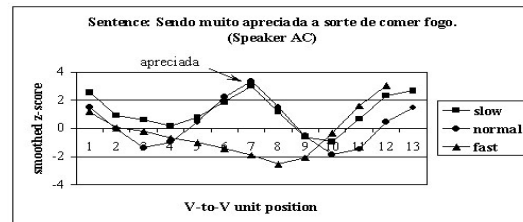
### 3. INTRA-SPEAKER RHYTHMIC VARIABILITY

A single speaker can modify or perturb the way s/he controls the three prosodic functions, namely establishing phrasing, marking prominence, and signalling discourse-related function, depending not only on uncontrolled factors such as tiredness, emotion, disease, but crucially depending on volitional, situational-dependent demands that allows her/him to change speaking style, speech rate, and discursive function. These kinds of effects have immediate consequence on the formation of patterns of V-to-V duration, as is the case of speech rate changes.

Some differences of normalised V-to-V duration patterning for distinct speech rates may signal differences in prosodic phrasing, as it is confirmed by

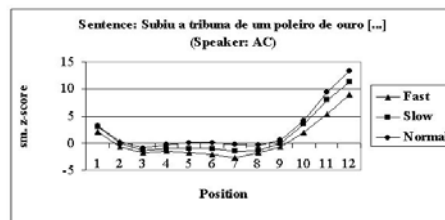
observing duration patterns across rates. An illustration is given in Fig. 4, where the final sentence of the Lobato corpus, “Sendo muito apreciada a sorte de comer fogo.” (with the fire eating trick being the most appreciated), is realised by speaker AC with two stress groups at slow and normal self-chosen rates, and one stress group at the fast rate (note the absence of the medial duration peak for the fast rate).

**Figure 4:** Smoothed  $z$  – scores for the sentence “Sendo muito apreciada a sorte de comer fogo.” read by speaker AC at three speech rates. Slow and normal rates are indistinct.



This change in durational patterning when constrained by a faster speech preserves, however, the major boundaries (in this case both edges of the utterance). This does not mean that this speaker will exhibit the same behaviour every time there is a minor boundary in the middle of the utterance. Realising an intermediate peak at a particular, usually slower rate, depends on a decision which affects his performance as a speaker communicating a particular message to its potential audience. In the passage “Subiu a tribuna de um poleiro de ouro.” (He stepped up onto a platform of a golden perch) AC produces a similar durational pattern at all three speech rates, with no minor boundary in the middle of the utterance (Fig. 5).

**Figure 5:** Smoothed  $z$  – scores for the sentence “Subiu a tribuna de um poleiro de ouro.” read by speaker AC at three speech rates.



Although the same general configuration (two major boundaries separated by a minor one) of Fig. 4 is realised at the beginning of the paragraph reading, in the passage “Em seguida apareceu um papagaio real que tinha fama de orador.” (After that a royal parrot entered, famous for his rhetorical qualities),

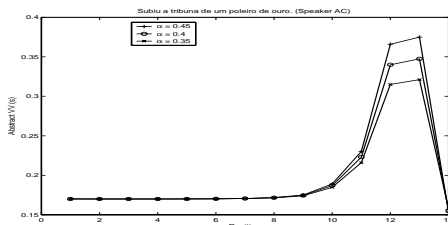
speaker AC only attenuates the level of the prominence for the minor boundary at the fast rate (Fig. 6).

**Figure 6:** Smoothed  $z$  – scores for the sentence “Em seguida apareceu um papagaio real que tinha fama de orador” read by speaker AC at three speech rates.



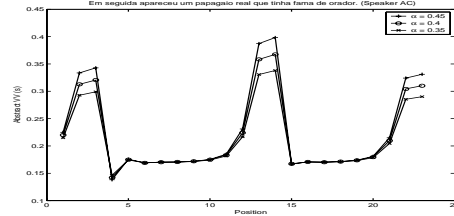
These intra-speaker changes in prominence level and phrasing can be simulated in the DSR model by modifying the input parameters which indirectly control prominence (entrainment rate, coupling strength between syntax and regularity constraints), and stress group extension (both in seconds and in number of V-to-V units). The attenuation of the middle prominence in the speaker AC speech can be obtained (a) by preserving the underlying speech rate for simulating the three nominal rates he produced (it can be seen above that there are no crucial differences for V-to-V durations in non-prominent positions) with  $T_0 = 0.170$  s, and (b) by using the change of the entrainment rate  $\alpha$  (from 0.35 to 0.45) to finely controlling duration at prominent position. These results can be seen in Figs. 7 and 8, where the greater  $\alpha$ , the greater the V-to-V duration. The other parameters changed were  $\beta = 0.7$ , and log-transformed stress group extension descriptor [2.6, 0.4]. In order to simulate the non-

**Figure 7:** Abstract V-to-V duration along the sentence “Subiu a tribuna de um poleiro de ouro” for three sim. degrees of prominence (cf. Fig. 5).



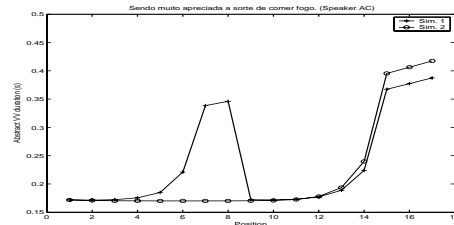
realisation of the prominence at the minor boundary for sentence ‘Sendo muito apreciada a sorte de comer fogo’, changes in stress group extension descriptor and syntax-regularity coupling strength are necessary. In the two simulations in Fig. 9, these parameters are [1800, 600] ms, [2.0, 0.6],  $r_p = 0.7$  (Sim. 1), and [2300, 600] ms, [2.8, 0.6],  $r_p = 0.9$

**Figure 8:** Abstract V-to-V duration along the sentence “Em seguida apareceu um papagaio real que tinha fama de orador” for two sim. degrees of prominence (cf. Fig. 6).



(Sim. 2), with  $\alpha = 0.5$ , and  $\beta = 0.4$  in both cases.

**Figure 9:** Simulated V-to-V duration along the sentence “Sendo muito apreciada a sorte de comer fogo” for three sim. degrees of prominence.



The change in the model parameters, illustrated here, suggests that a single subject can change the inter-relation s/he usually establishes between syntactic and production constraints, as well as how far s/he looks ahead to reach her/him communicative goals. As far as the data and simulation are concerned, this change can operate along the reading of a short text as the one read by the speakers studied here. The same input parameters can be changed for simulating other subjects, reflecting the way different subjects utter the same material.

#### 4. INTER-SPEAKER RHYTHMIC VARIABILITY

Although different speakers have at their disposal the same resources for signalling prosodic function, they can produce distinct durational patterns when asked to utter at different, self-chosen speech rates. Observe for the natural normalised duration, Fig. 10, that speaker AP attenuated almost completely the first prominence at the fast rate, while speaker AC (Fig. 6) didn’t.

Observe also that speaker AP realises the three nominal rates by trying not only to change the levels of prominence, as AC seems to do (see previous section), but to produce distinct durations along the entire utterance. This is apparent in the sentence “Subiu a tribuna de um poleiro de our(o)” shown in

**Figure 10:** Smoothed  $z$  – scores for the sentence “Em seguida apareceu um papagaio real que tinha fama de orador” read by speaker AP at three speech rates. Normal and fast rates are indistinct.

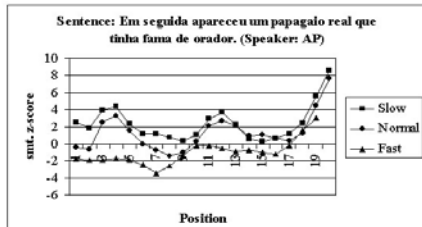
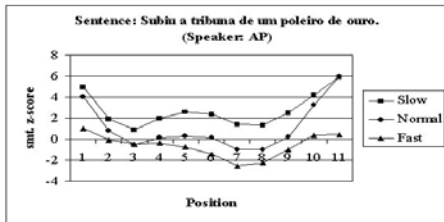


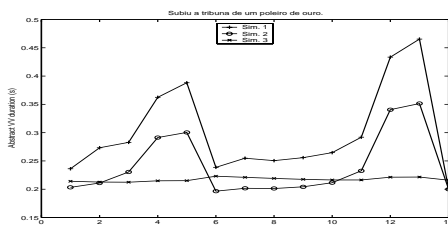
Fig. 11. Note also that, contrary to AC, speaker AP preserves the minor middle prominence.

**Figure 11:** Smoothed  $z$  – score evolution for the sentence “Subiu a tribuna de um poleiro de ouro(o)” read by speaker AP at three speech rates.



A simulation of the kind of change exhibited by speaker AP when speaking faster, for the sentence in Fig. 11, can be obtained by changing both the underlying speech rate ( $T_0$ ), and a more syntactic-oriented phrasing. Sim. 1 (the highest values of duration) was obtained by using  $T_0 = 0.25$  s, while sim. 2 uses  $T_0 = 0.2$  s. All the other parameters took the initial values presented in the first section. For simulating the fastest rate (the line with the smallest durations), the following parameters were modified:  $\alpha = 0.02$ ,  $\beta = 0.04$ ,  $T_0 = 0.18$  s, and  $r_p = 0.9$ .

**Figure 12:** Abstract V-to-V duration along the sentence “Subiu a tribuna de um poleiro de ouro” for three simulated underlying speech rates.



## 5. SUMMARY

These simulations and data suggest that a synergy of dynamical parameters is necessary in order to pro-

duce a change in speech rate. Shorter (or longer) syllable-sized units are realised by changes in both the syntax-rhythm interface and by how far we look ahead when speaking. Prominence degree and number of phrase prominences can also be modified to realise the functional task of uttering slower or faster than at a comfortable rate.

## 6. ACKNOWLEDGEMENTS

This work is part of the FAPESP project n. 05/02525-7, and the CNPq project n. 300296/2005-3. The paper is part of the Special Session *Speech Timing: Approaches to Speech Rhythm*, coordinated by E. Keller and R. Port, having as co-participants, M. Brad, N. Campbell, Y. Hirata, and myself.

## 7. REFERENCES

- [1] Barbosa, P.A., 2002. Explaining Brazilian Portuguese resistance to stress shift with a coupled-oscillator model of speech rhythm production. *Cadernos de Estudos Linguísticos* 43, 71-92.
- [2] Barbosa, P.A., 2006. A Dynamical Model For Generating Prosodic Structure. *Proc. of the Speech Prosody 2006 Conf.* Dresden, Germany, 366-369.
- [3] Barbosa, P.A., to appear. From syntax to acoustic duration: a dynamical model of speech rhythm production. *Speech Communication*.
- [4] Boersma, P., Weenink, D., 2005. Praat: doing phonetics by computer. Accessible in <<http://www.praat.org/>>, version 4.4.
- [5] Dogil, G., Braun, G. 1988. *The PIVOT model of speech parsing*. Wien: Verlag.
- [6] MacNeilage, P.F. 1998. The frame/content theory of evolution of speech production. *Behavioral and Brain Sciences* 21, 499-511.
- [7] Marcus, S.M., 1981. Acoustic determinants of Perceptual-center (p-center) location. *Perception and Psychophysics* 30 (3), 247-256.
- [8] Massini, G., 1991. A duração no estudo do acento e do ritmo em português. Master’s thesis. Univ. of Campinas, Brazil.
- [9] Mauk, M.D., Buonomano, D.V., 2004. The neural basis of temporal processing. *Annu. Rev. Neurosci* 27, 307-340.
- [10] Öhman, E.G., 1966. Coarticulation in VCV utterances: spectrographic measurements. *J. Acoust. Soc. Am.* 39 (1), 151-168.
- [11] Shattuck-Hufnagel, S., Klatt, D. 1979. The limited use of distinctive features and markedness in speech production: evidence from speech error data. *J. of Verb. Learn. and Verb. Behav.* 18, 41-55.
- [12] Wong, S.W., Schreiner, C.E., 2003. Representation of CV-sounds in cat primary auditory cortex: intensity dependence. *Spe. Com.* 41, 93-106.