A dynamical model for generating prosodic structure

Plínio A. Barbosa

Speech Prosody Studies Group & Department of Linguistics
State University of Campinas, Brazil
plinio@iel.unicamp.br

Abstract

The performance of the Monnin-Grosjean (MG) algorithm for predicting prosodic structure is compared with that of a system of dependency-grammar-based local markers (the DG system). Analyses of Brazilian Portuguese paragraphs read by five speakers reveal that the MG algorithm performs as well as the DG system when V-to-V normalised durations at word and phrase stress boundaries are used as indexes of prominence. These two procedures, however, have proved unsuccessful in dealing with individual variability. To overcome such a limitation, a dynamical model is proposed. By coupling syntactic and regularity constraints the main advantage of the model is the plausible simulation of speaker variability. Seven simulations were caried out by changing three model parameters: coupling strength, conditional probability of phrase stress placement, and V-to-V duration mean.

1. Introduction

The prosody-syntax interface has been a theme of great interest in the scientific community, as illustrated by the proposals of several algorithms that explain part of the variance of both prosodic constituency and prominence [1, 6, 8, 9]. All these algorithms have been integrated into text-to-speech synthesis systems in order to automatically generate the prosodic information necessary to produce a natural-sounding speech. Three depths of syntactic analysis prior to the obtention of prosodic structure can be identified in these models. Some use a comprehensive parser to analyse the sentences [9], others use a partial syntactic analysis [1], and the last ones use a minimal amount of syntactic information [6, 8]. All the algorithms use a set of heuristic rules to obtain prosodic constituents of similar size. This size can be measured in units of some linguistic constituent, such as the phonological phrase or the syllable. Depending on the sentence complexity, the obtained prosodic structure is often flatter and more simmetric than the syntactic one. This very fact questions the need of comprehensive parsers for prosody generation.

Within a psycholinguistic framework, the work by Gee and Grosjean [5] is often cited for the high correlation between predicted and realised structures that their ϕ algorithm produces [1, 12]. Less well known is the Monnin-Grosjean algorithm (henceforth, the MG algorithm) adapted from the ϕ one to predict the performance structure of French sentences [7].

Since the latter considers the fact that French is a right-headed language at the stress group domain, it could be used to predict prosodic structure in other right-headed languages, such as Brazilian Portuguese (henceforth BP). Here "stress group" refers to a unit delimited by two consecutive phrase stresses. Phrase stress is used here, according to a speech production criterion, as the place of one or more prominent units along the utterance. In BP, it corresponds to the position of the culmination of a quasi-monotonical increase of V-to-V (from a vowel onset to the immediately next one) durations rightwards, followed by a duration reset [3]. These peaks of duration include weak prosodic boundaries which are seldom perceived by listeners in perception tests, since perceived salience depends on others factors such as pitch accent.

Three procedures for predicting prosodic structure are presented and contrasted in the next section. Two of them are retained for comparison and the reasons for rejecting a first choice is explained. The unability of the two retained procedures to cope with prosodic individual variability guided us to propose a dynamical model of phrase stress placement and prominence generation in the third section.

2. Comparing two procedures for predicting prosodic structure

Recently, Watson and Gibson [12] have compared the performance of three algorithms for attributing intonational boundaries in English with their own, the Left/Right Constituent Boundary model (the LRB model). The LRB model predicts that the likelihood of an intonational boundary between two words increases with constituent size at both sides of the boundary.

The performance of their algorithm was compared with the other three by computing the squared correlation between the predicted weights and the probability of intonational boundary placement. The latter was estimated from the labelling of complex sentences using a subset of the ToBI break index system [10]: indexes 1, 2, and 3-4. Their model performed as well as two of the algorithms (including Gee and Grosjean's) and better than all three algorithms, for more complex sentences. Even though the sentences evaluated were produced by the interaction of 37 pairs of subjects, their model predicted a single structure for each sentence, and, therefore, it was unable to cope with individual prosodic variability. Furthermore, when applied to BP, their model is unable to predict two distinct prosodic bracketings to cope with sentences with identical syntactic structures but varied number of syllables. The following sentences illustrate such a limitation of their model (verb in bold face): [A moça **canta**] [divinamente bem] (The girl sings amazingly well), and [A afinadíssima moça] [**canta** divinamente bem] (The in-tune girl sings amazingly well). As stress groups with similar size tend to be implemented by speakers, in the first sentence - but not in the second -, the verb is included in the first stress group. Due to the strong dependency of the LRB model on syntactic constituency, it is unable to explain this common prosodic segmentation. The illustration also suggests that the verb in BP oscillates in attracting/repelling phrase stress. Prominence is expected then to be random for this grammatical category .

Since the MG algorithm is able to adequately predict the prosodic bracketing of the two sentences illustrated above, a comparison of this algorithm with a dependency-grammar-based system of markers (the DG-based system) is made here.

2.1. The MG algorithm

Monnin and Grosjean [7] proposed the MG algorithm for predicting the prosodic structure of French sentences. The algorithm carries out this prediction into six steps: (1) identification of the prosodic nuclei; (2) formation of basic prosodic constituents by connecting remaining words to their nuclei; (3) indexation of the basic prosodic constituents; (4) formation of higher prosodic constituents; (5) indexation of the higher prosodic constituents; (6) rhythmic adjustments to achieve constituent size equalisation, where size is measured in number of syllables. The last step explicitly states that the verb can group with a left subject depending on size constraints.

The prosodic structure predicted by the MG algorithm is assessed by computing the correlation between the predicted indexes of prosodic boundary strength between each pair of words in a set of sentences, and the duration of the final-word vowels added to a silent pause, as applicable. These durations are taken from the average duration of the reading of the sentences by eight subjects at a spontaneous speech rate. The correlation coefficient obtained for a set of nine simple sentences was 94 %, which signals the appropriateness of the MG algorithm for predicting the prosodic structure of simple sentences in French.

Since there is only one predicted prosodic structure for each sentence, and the realised prosodic structure are averaged across subjects, the algorithm is unable to deal with prosodic variability. On the other hand, the steps are relatively local by operating around prosodic nuclei.

2.2. The DG-based system

Bailly [2] proposed a set of markers based on local dependency relations of a surface tree, obtained according to the principles of Tesnière's work [11]. The set considered four kinds of relation between contiguous terms in a sentence: (1) right dependency (RD), when the dependent is to the right of the regent, as in "a casa (RD) verde" (the-house-green); (2) left dependency (LD), when the dependent is to the left of the regent, as in "a bela (LD) mulher" (the-beautiful-woman); (3) interdependency (IT), when the contiguous terms modify the same regent, as in "o belo, (IT) gentil cão" (the-beautiful-nice-dog), and (4) independency (ID), when the contiguous terms are not directly related, as well in "a quotação do dólar (ID) aumentou" (the-quotation-of-the-dollar-increased). This four-level set is combined with two levels of strength. From these principles he proposed a set of six markers, since the left and right dependencies can be strong or weak, depending whether the regent is the verb or not, respectively.

In the present work, this set is increased to eleven markers, extending the two-level strength to the other types of relation and including three markers to deal with more complex syntactic relations. The three additional markers are COORD (at coordinated clauses' boundary), DSUB (at the beggining of a subordinated clause, when the conjunction immediately follows the regent), IDSUB (at the beggining of a subordinated clause, when the regent is not contiguous to the conjunction). This extended system is called the DG-based system.

Two corpora of BP were annotated with the DG-based markers in order to assess their appropriateness in predicting prosodic structure. For doing so, the set of eleven markers was associated with natural numbers ranging from 1 to 11 in increasing order: LD, IT, RD, SRD, SLD, SIT, ID, DSUB, IDSUB, COORD, SID (where strong markers begin with 'S'). The Lobato corpus was read by five male speakers from two different dialectal regions (São Paulo and Brasília), whereas the Pantanal corpus was read by one of the speakers from Brasília. The Lobato corpus is the reading of a two-paragraph-long chunk of a story-telling text for children (110 words). The Pantanal corpus is the reading of a 353-word report on the Brazilian Pantanal.

2.3. Comparing the MG algorithm and the DG-based system

In order to compare the performance of the MG algorithm in predicting prosodic structure with that of the DG-based system, the correlation coefficient between predicted and realised prosodic strengths at each word boundary was computed. For the MG algorithm, prosodic strengths were predicted by applying the six steps above. As for the DG-based system, only the numbers associated with the local syntactic markers were used ar each phonological word boundary. A global syntactic set of indexes of strength, served as a control base of comparison. These global syntactic indexes were obtained for each phonological word boundary from a classical surface syntactic tree, according to Monnin and Grosjean's directions [7]. This is made by counting the number of non-terminal nodes dominated by the node that separates the constituents at both sides of the boundary and including the dominant node itself.

The realised prosodic strengths were obtained by using a normalised duration, instead of Monnin-Grosjean's raw durations. The reason for that is to avoid the effect of intrinsic vowel duration. This normalised duration, called the phonological word $z - score$, is the highest smoothed $z - score$ [4] of the durations of the V-to-V units of a chunk of the phonological word preceding the boundary. This chunk extends from the lexically stressed V-to-V unit to the possible post-stressed V-to-V units. Smoothing is obtained by applying a 5-point weighted average to raw $z - scores$. Each phonological word boundary strength is then specified by a real number signalling the degree of lengthening/shortening of the boundary. Results showing correlation coefficients between the three methods of prediction and the phonological word $z - scores$ are given is table 1 for three BP male speakers from São Paulo state (AP, AC, and DP). All three speakers read the Lobato corpus at three self-chosen speech rates, but only the statistically different rates are shown. In order to evaluate the effect of restricting the computation to words bearing phrase stress, correlation coefficients for these positions only are given between parentheses. Phrase stress positions correspond to the maxima of smoothed $z - scores$ of V-to-V durations along the utterance. It is worth noting that both the MG algorithm and the DG-based system perform similarly, even though the latter one is a purely syntactic-based system. Since normalised durations of words bearing phrase stress are more reliable indexes of prosodic strength, all correlation coefficients are higher in that condition. The maximum correlation of 51 %, much smaller than the 94 % found for French [7], is explained by the use of a story-telling text that contains many complex sentences and that is pronounced in a more natural way. This correlation is also carried out with more detailed data than the perceptual indexes used in [12]. Observe also that the global syntactic system achieves correlations inferior to 30 % for two speakers, except AC, which could indicate that some speakers rely more on syntax when speaking.

Some minor differences for distinct speech rates may signal

Table 1: *Correlation coefficients between predicted prosodic structure from three methods (global syntactic indexes, GS; the DG-based system, DG; and the Monnin-Grosjean algorithm, MG), and realised prosodic structure. In all cases,* $p < 0.05$. *Values in bold face signals the highest value in each row.*

Speakers (rate)			
AP	GS	DG	MG
(slow)	28 (59)	**46 (64)**	45 (61)
(normal)	28 (47)	**45 (61)**	41 (56)
AC			
(normal)	49 (53)	**49 (54)**	42 (54)
(fast)	36 (50)	**42 (55)**	**33 (60)**
DP			
(normal)	27 (54)	26 (35)	**51 (71)**

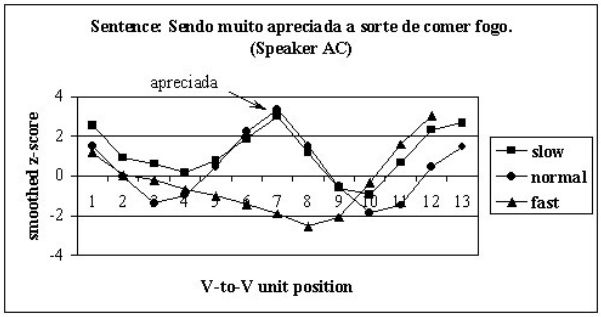

Figure 1: *Smoothed* $z-score$ *evolution (y-axis) of the utterance "Sendo muito apreciada a sorte de comer fogo." for speaker AC. Slow and normal rates are statistically indistinct.*

differences in prosodic structuring, as it is confirmed by observing the V-to-V duration patterning across rates. An illustration is given in Fig. 1, where the final sentence of the Lobato corpus, "Sendo muito apreciada a sorte de comer fogo.", is realised as two (slow and normal) or one stress group (fast), as speech rate increases. Note the absence of the medial duration peak for the fast rate. These intra- and inter-subject differences of prosodic structuring have motivated the proposal of a model able to cope with prosodic variability.

3. A dynamical model for prosodic structure generation

The model proposed here is couched on dynamical systems theory. It is part of a dynamical model of speech rhythm production that integrates two coupled oscillators able to generate durational patterns along stress groups in BP [3]. The present work completes the rhythm model by including a model for automatically attributing phrase stress position and prominence along the sentences. However, until the development of a DG-based parser is completed, it will require the manual introduction of the DG-based markers.

The dynamical model has two coupled probabilistic components, a syntactic component and a regularity-constraint component (see eq. 1). The syntactic component is realised by computing the conditional probability that a phonological word bears the phrase stress, given a specific DG-marker, to its right, $p(ps/m)$. The regularity component implements the production constraints on stress group size similarity by computing the conditional probability that a phonological word bears the phrase stress, given n V-to-V units (nVV) following the assignment of a previous phrase stress, $p(ps/nVV)$). Both components are linearly combined by using a coupling strength factor, r_p, in order to compute the likelihood of phrase stress, $l(ps)$, at phonological word boundary.

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

The $logit$ values ($logit(p) = ln(p/1-p)$) of the probabilities are computed in order to obtain a likelihood extending outside the interval [0-1]. This is necessary because the smoothed V-to-V duration $z-scores$ are not restricted to that interval. In the model equation there are three sources of intra- and inter-speaker variability: the two conditional probabilities and the coupling strength between the components. As to the latter ($0 \leqslant r_p \leqslant 1$), if r_p is greater than 0.5, the syntactic component dominates the regularity one. If its value is less than 0.5, the opposite trend is realised. This can simulate speakers that rely more on syntax than on regularity constraints when speaking, as speaker AC.

Both conditional probabilities were estimated from the Lobato corpus in a pilot experiment for four speakers (AP, AC, DP, and PA). As significance was achieved for a few markers, in the case of the syntactic conditional probability, the computation was achieved by using the Pantanal corpus with speaker PA from Brasília. In order to assess his compatibility with the other speakers, the conditional probabilities from both corpora were compared for this speaker and revealed that the same trends are found for the four speakers. The regularity component probability was implemented by a lognormal distribution, since it was the best fit for all four speakers. The distribution allowed to specify this component by two parameters, log-transformed mean and standard-deviation, for all speakers: $(1.9, 0.4)$ (approximately 8 and 3 V-to-V units per stress group, respectively).

The conditional probabilities for the syntactic component are given in table 2 for five markers. The values for the other six were used in the simulations with the model, but were not significant (DSUB, IDSUB, and COORD gave the value of 1, for instance). Significance in table 2 refers to statistical differences between conditional probability and a priori probability of phrase stress (the latter corresponds to the random case). This is done in order to evaluate the attraction or repulsion of specific markers to/from phrase stress. The repulsion of the RD marker from phrase stress (the conditional probability is lesser than the a priori probability) indicates that this speaker (and the other speakers, since the results were identical in this respect) rejects to put a phrase stress between a noun and its complemetnt in a nominal phrase, for instance. The stronger markers SID and ID, on the other hand, attract phrase stress (0.86 and 0.76 are statistically greater than 0.48), which means that sentence boundary

Table 2: *Conditional probability of phrase stress given a specific DG-based marker, computed from the corpus Pantanal for speaker PA. The* $p-values$ *refer to significance from a priori probability of phrase stress* $p(ps) = 0.48$.

marker	n	$p(ps/m)$	$p <$
RD	66	0.26	0.002
SRD	24	0.30	ns
SID	34	0.86	0.0002
ID	17	0.76	0.03
IT	32	0.52	ns

and the boundary between a complement and a verb, for instance, are preferred places for assigning phrase stress. The random behaviour of the SRD marker (between a verb and a right dependent) suggests that verbs behave ramdomly with respect to phrase stress attraction. This is confirmed by computing the conditional probability of phrase stress given a grammatical category in the same Pantanal corpus. This analysis revealed that verbs are in fact indifferent to phrase stress (probability statistically indistinct from a priori one: 0.35), whereas nouns and adjectives attract them (probabilities 0.53 and 0.60, respectively).

The dynamical model proceeds as following: (1) four input parameters are introduced: a V-to-V mean duration specifying speech rate, μVV; the two parameters specifying the conditional probability of phrase stress given the extension from the last assigned phrase stress (or the beginning of the sentence), (μnVV,σnVV); the coupling strength r_p; and the conditional probabilities for the 11 DG-based markers; (2) a lookahead window in phonological words is computed from the ratio between the mean value of stress group duration across the four speakers ($1300\ ms$), and the ratio between μVV and the averaged number of V-to-V units per phonological word (3.5); (3) the likelihood of phrase stress assignment at each lexically stressed V-to-V unit within the current lookahead window is computed from equation 1; (4) a phrase stress is assigned to the position with the higher likelihood (this value is retained for computing the correlation between simulated and realised prosodic structures). The model proceeds to step 3 until the sentence ends.

Seven simulations of the model for the Lobato corpus are presented in table 3. They consider changes in all input parameters except the syntactic conditional probabilities. Correlation coefficients are computed between realised phonological word $z-scores$ for four speakers, and the likelihoods estimated by the model. Since the model predicts phrase stress position and prominence, the values in table 3 should be compared to those between parentheses in table 1, ranging from 54 to 71 %. The highest correlation among the four speakers is given in table 3, with the correspoonding speech rate (correlations for all speakers and rates were greater than 45 %). Observe that, as the simulated speech rate increases (by decreasing μVV), the extension of the lookahead window increases, which ensures a smaller number of assigned phrase stresses. The coupling strengths in table 3 vary from 0.5 to 0.9, which means that a regularity constraint is necessary for predicting the prosodic structure. Even though the correlations are statistically similar to those in tables 1 and 3, the dynamical model is able to cope with prosodic variability by changing the degree of the syntactic component influence, as well as by decreasing the number of attributed phrase stresses as speech rate changes (achieved by the change of the lookahead window size). The upper limit of about 70 % for all correlations shown here strongly suggests that part of the remaining variance is certainly due to semantic factors.

Table 3: *Simulations of phrase stress assignment with the dynamical model. Lookahead (LA) and correlation coefficients (R) between simulated and realised prosodic structures of specific speakers are given. See text for additional information.*

μVV	(μnVV,σnVV)	r_p	*LA*	R
300	(1.9,0.4)	0.7	2	58 (AP, slow)
300	(1.6,0.4)	0.8	2	58 (AP, slow)
300	(2.6,0.4)	0.5	2	67 (PA, normal)
225	(1.85,0.4)	0.6	3	65 (AP, slow)
225	(2.1,0.4)	0.6	3	55 (AP, slow)
150	(1.9,0.4)	0.9	4	49 (AP, slow)
150	(2.8,0.4)	0.9	4	46 (AP, slow)

4. Summary

After assessing the performance of two algorithms for predicting prosodic structure with data from BP text reading data, a dynamical model for assigning phrase stress is proposed. The model is able to deal with at least two sources of speaker variability: changes in phrase stress prominence and place due to speech rate, and changes triggered by syntactic structuring.

5. Acknowledgements

This work is part of the FAPESP project *Dynamical analysis and modelling of speech prosody* (05/02525-7). I thank J. M. Vieira's help with the segmentation and labelling of the Pantanal corpus, and S. Madureira for suggestions. Additional information available at <http://www.unicamp.br/iel/site/docentes/plinio/index.htm>.

6. References

[1] Bachenko, J.; Fitzpatrick, E., 1990. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16(3), 155-170.

[2] Bailly, G. 1986. Un modèle de congruence relationnel pour la synthèse de la parole du français. *Actes des 15es Journées d'Etude sur la Parole*. Aix-en-Provence, France, 75-78.

[3] Barbosa, P.A., 2002. Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production. *Proceedings of the first International Conference on Speech Prosody*. Aix-en-Provence, France, 163-166.

[4] Campbell, W.N.; Isard, S.D., 1991. Segment durations in a syllable frame. *Journal of Phonetics*, 19, 37-47.

[5] Gee, J.P.; Grosjean, F.E., 1983. Performance structures: a psycholinguistic and linguistic appraisal. *Cognitive Psychology*, 15, 411-458.

[6] Martin, P., 1987. Prosodic and rhythmic structures in French. *Linguistics*, 25, 925-949.

[7] Monnin, P.; Grosjean, F., 1993. Les structures de performance en français : caractérisation et prédiction. *L'Année Psychologique*, 93, 9-30.

[8] Pasdeloup, V., 1992. A prosodic model for French text-to-speech synthesis: a psycholinguistic approach. In *Talking Machines: Theories, Models and Designs*, G. Bailly; C. Benoît (eds.). Amsterdam: Elsevier, 335-348.

[9] Schweitzer, A.; Braunschweiler,N.; Morais, E., 2002. Prosody generation in the SmartKom project. *Proceedings of the first International Conference on Speech Prosody*. Aix-en-Provence, France, 639-642.

[10] Silverman, K. et al., 1992. ToBI: a Standard for Labeling English Prosody. *Proceedings of the 2^{n}d ICSLP*. Banff, Canada, v. 2, 867-870.

[11] Tesnière, L., 1967. Eléments de syntaxe structurale. Paris: Klincksieck.

[12] Watson, D.; Gibson, E., 2004. The relationship between intonational phrasing and syntactic structure in language production. *Language and Cognitive Processes*, 19(6), 713-755.