Journal of Speech Sciences (9): 105-128. 2020.
Available at: http://revistas.iel.unicamp.br/joss

MODELLING AUTOMATIC DETECTION OF PROSODIC BOUNDARIES FOR BRAZILIAN PORTUGUESE SPONTANEOUS SPEECH

RASO, Tommaso[1*]
TEIXEIRA, Bárbara[1]
BARBOSA, Plínio[2,3]

[1]Universidade Federal de Minas Gerais
[2]Universidade Estadual de Campinas
[3]CNPq

Abstract*: Speech is segmented into intonational units delimited by prosodic boundaries. This segmentation is claimed to have important consequences for syntax, information structure and cognition. This work aims both to investigate the phonetic-acoustic parameters that guide the production and perception of prosodic boundaries, and to develop models for automatic detection of prosodic boundaries in Brazilian Portuguese male monological spontaneous speech. Two samples were segmented into intonational units by two groups of trained annotators. The boundaries perceived by the annotators were tagged as either terminal or non-terminal. A script was used to extract 111 phonetic-acoustic parameters along the speech signal in both a rightward and a leftward window around the boundary of each phonological word. The extracted parameters comprise measures of (1) Speech rate and rhythm; (2) Standardized segment duration; (3) Fundamental frequency; (4) Intensity; (5) Silent pause. The script considers as prosodic boundaries positions at which at least 50% of the annotators indicated a boundary of the same type. A training of models composed by the parameters extracted by the script was developed; these models were then improved heuristically. The models were developed from the two samples considered separately and from the joined samples dataset, both using non-balanced and balanced data. A Linear Discriminant Analysis algorithm was adopted to produce the models. The models for terminal boundaries show a much higher performance than those for non-terminal ones. In this paper we: (i) show the methodological procedures; (ii) analyze the different models; (iii) discuss some strategies that could lead to an improvement of our results.*

Keywords: prosodic boundaries; automatic detection; spontaneous speech.

---

*Corresponding author: tommaso.raso@gmail.com

1 Introduction

Speech is prosodically segmented into intonation units determined by prosodic boundaries. These units can be functionally analyzed according to different theoretical perspectives, be they syntactic, pragmatic or cognitive (Cooper and Paccia Cooper, 1980; Selkirk, 2005; Halliday, 1965; Cresti, 2000; Szczepek Reed, 2012; Chafe, 1994; Croft, 1995; Bybee, 2010). However, prosodic boundaries can be also studied *per se* (Barth-Weingarten, 2016), independently of the theoretical perspective from which the units are observed, since these boundaries are clearly perceivable by listeners.

In some works, the authors investigate the opposition between presence versus absence of perceived prosodic boundary (Mo *et al.*, 2008; Park, 2002; Croft, 1995; Maschler, 2009), that is, the perceived prosodic boundaries are not differentiated and are treated equally. An alternative view suggests that prosodic boundaries vary gradiently (Byrd and Saltzman, 2003; Pijper and Sanderman, 1994; Ladd, 1988). Other works distinguish boundaries in terms of different levels of perceptual strength (Simon and Christodoulides, 2016; Reichel and Mády, 2013; Wightman *et al.*, 1992; Barbosa, 2006; Barbosa, 1994; Tabain, 2003; Tabain and Perrier, 2005; Krivokapić, 2007; Mertens and Simons, 2013). However, among the different authors who distinguish different strenghts of boundaries, there is a clear disagreement about the number of possible levels of force by which boundaries can be produced and perceived. Some authors simply distinguish between strong and weak boundaries (Simon and Christodoulides, 2016; Reichel and Mády, 2013), while others believe that it is possible to individualize more than two levels of strength (Wightman *et al.*, 1992; Barbosa, 2006; Barbosa, 1994; Tabain, 2003; Tabain and Perrier, 2005; Krivokapić, 2007).

Perceived prosodic boundaries by listeners also can be associated with the perception of conclusion or continuation of the intonation unit, showing an agreement often higher than 0.8, according to several kappa tests (Danieli *et al.*, 2004; Mello *et al.*, 2012). In general, the first type is called a terminal boundary (TB) and the second one, a non-terminal boundary (NTB). This is the perspective adopted here as a departure hypothesis.

Some of the acoustic phenomena that signal boundaries are known thanks to many studies of lab or read speech (Price *et al.*, 1991; Blaauw; 1994), and are sometimes tested in radio corpora (Ostendorf *et al.*, 1995). The main ones, commonly considered in the literature, are silent pause, pre-boundary lengthening, reset of fundamental frequency (f0) and a remarkable change in speech rate, intensity or f0 variation rate (Cruttenden, 1997; Crystal, 1969; Du Bois *et al.*, 1992; Du Bois, 2008; Kelly and Local, 1989; Amir *et al.*, 2004; Mo, 2008; Blaauw; 1994). However, other aspects are involved in the understanding of the set of acoustic correlates that signal prosodic boundaries. Among them we can cite at least the specific language and the speech style. Gender, diastratic factors and even individual variability may play a role as well (Barth-Weingarten, 2016; Barbosa and Raso, 2018; Izre'el *et al.*, forthcoming).

This work aims to investigate the acoustic-phonetic parameters that are involved in the production and guide the perception of prosodic boundaries, based on the hypothesis that they can initially be divided between two macrotypes: boundaries marking conclusion (TB) and boundaries marking continuation (NTB). It also aims to develop automatic models for detecting prosodic boundaries in Brazilian Portuguese spontaneous speech. The models shown here consider two related criteria: the acoustic-phonetic parameters automatically extracted from the sound signal and the perception of trained annotators to perceive TB and NTB. This means that human perception is assumed to be the goal that the model should reflect. So far, tools for automatically detecting TB and NTB in spoken corpora of spontaneous speech are not available.

This paper briefly presents the methodology and the results reached so far by the research. It also analyzes these results and propose possible strategies for future steps.

2 Data and data treatment

2.1 Data

The full data set comprises two samples of monological male spontaneous speech excerpts, as can be seen in Table 1. Each sample includes seven excerpts extracted from the C-ORAL-BRASIL I corpus (Raso and Mello, 2012) and from two sections of C-ORAL-BRASIL II (Raso *et al.*, forthcoming), with on average 190 words. The excerpt taken from the C-ORAL-BRASIL I represents natural informal monological spontaneous speech. The other two excerpts are taken from the sections *media* and *formal speech in natural context*[1] of C-ORAL-BRASIL II.

Table 1: Sample description.

Context	Sample	Text[2]	Time	Words
Natural informal	I	bfammn11	01’11’’	189
		bfammn24	00’58’’	151
	II	bpubmn12	01’26’’	198
		bpubmn13	01’00’’	180
Media	I	bmidmasc01	01’23’’	212
		bmidmasc02	01’21’’	238
		bmidmasc03	01’07’’	183
	II	bmedsp03_1a	01’02’’	206
		bmedsp03_1b	01’07’’	200
		bmedts10_1	01’11’’	180
Natural formal	I	bnatmasc01	01’30’’	205
		bnatmasc02	01’09’’	161
	II	bnatco03	01’00’’	202
		bnatpr05	01’43’’	181

2.2 Data treatment

The excerpts were segmented into intonation units by two groups of trained annotators. The annotators were previously trained to perceive and annotate prosodic boundaries. The first group, who annotated sample I, includes 14 annotators; the second one, who annotated sample II, includes 19 annotators[3].

Each annotator received an audio file with the excerpts and their orthographic transcription without any further annotation; their task was to annotate the two main types of boundaries following their perception using a simple slash symbol (/) to indicate a NTB and a double slash to indicate a TB (//). Disfluencies were marked with (+), but they were excluded, since they were considered non-planned boundaries. The agreement among the annotators,

[1] Formal speech in natural context comprises a set of natural contexts that all the C-ORAL corpora (Cresti and Moneglia, 2005; Raso *et al.*, forthcoming) partake, such as preaching, political speech and debate, professional explanation, teaching, conference and law.

[2] The excerpts with a number followed by underscore and another number are different parts of the same recording.

[3] The agreement data will be presented in Table 15.

including disfluencies, evaluated through the Fleiss kappa coefficient (Fleiss, 1971), was 0.80 for TB and 0.75 for NTB in the first sample, and 0.73 for TB and 0.72 for NTB in the second one.

The audio files were annotated into six Praat TextGrid tiers (Boersma and Weenink, 2014) as follows:

1) vowel-to-vowel (V-V)[4] interval tier with a broad phonetic transcription (Albano and Moreira, 1996);
2) point tier with points at every phonological word boundary. In each point tier, it was informed how many annotators signalled the focused upon phonological word boundary as a NTB;
3) point tier with points at every phonological word boundary. In each point tier, it was informed how many annotators signalled the focused upon phonological word boundary as a TB;
4) point tier with points at every phonological word boundary. In each point tier, it was informed how many annotators signalled the focused upon phonological word boundary as a disfluency;
5) interval tier delimiting silent pauses;
6) text tier with the textual transcription of utterances.

Table 2: Summary of extracted acoustic parameters[5].

Class	Type	Measurement
Speech rate and rhythm	Global	Rate of V-V unit normalized duration per second (right window context, left window context and difference)
		Rate of non-salient V-V units per second
Standardized VV duration	Local	Mean of smoothed z-score (adjacent right context, adjacent left context and difference)
	Global	Mean of smoothed z-score (right window context, left window context and difference)
		Standard deviation of smoothed z-score (right window context, left window context and difference)
		Skewness of smoothed z-score (right window context, left window context and difference)
		Peak rate of smoothed z-score (right window context, left window context and difference)
Fundamental frequency	Local	F0 median for each V-V (left and right V-Vs in window and difference at window center) in semitones re 1 Hz
		First derivative of F0 median for each V-V unit (left and right V-Vs in window and difference at window center) in semitones re 1 Hz/s
	Global	Mean of F0 medians (right window context, left window context and difference) in semitones re 1 Hz
		Standard deviation of F0 medians (right window context, left window context and difference) in semitones re 1 Hz
		Skewness of F0 medians (right window context, left window context and difference)

[4] About V-V units, see Barbosa (2006).

[5] See Appendixes in the metadata section for the role of these parameters for the models.

		Mean of F0 median first derivative (right window context, left window context and difference) in semitones re 1 Hz/s Standard deviation of F0 median first derivative (right window context, left window context and difference) in semitones re 1 Hz/s Peak rate of smoothed F0 peaks per second (right window context, left window context and difference)
Intensity	Local	Mean spectral emphasis[6] for V-V unit at window center in dB
	Global	Mean spectral emphasis (right window context, left window context and difference) in dB
Pause	Local	Pause presence (0 = absence or 1 = presence) Pause duration in seconds

The Praat script BreakDescriptor (Barbosa, 2016-2018) was used to extract 111 phonetic-acoustic measurements along the speech signal for all the V-V units in a window centered at all the boundaries between phonological words[7]. The windows scanned by the BreakDescriptor scan a maximum of V-V units that includes the target V-V unit plus ten V-V units to the left and ten V-V units to the right of each analyzed V-V unit. The extracted parameters comprise measures of: 1) Speech rate and rhythm (6 global measurements, see below); 2) Normalized duration (34 measurements – 12 global and 22 local, see below); 3) Fundamental frequency (65 measurements – 21 global and 44 local); 4) Intensity (4 measurements – 3 global and 1 local); 5) Silent pause (presence/absence and duration). Positions at which at least 50% of the annotators indicated a boundary of the same type were considered as a boundary. BreakDescriptor allows reducing the size of the scanned window if required.

Table 2 shows a summary of the measurements extracted for prosodic analysis, divided into global and local. Global measurements are calculated considering the values in the whole left and right windows, plus the difference between those values. Local values are calculated for every single V-V unit of the left and right windows plus the target V-V position.

Below, Figure 1 shows the windows scanned by BreakDescriptor. Starting from the top: wave form, broad-band spectrogram, and all tiers in a Praat TextGrid. The position of NTB used here constitutes the central point of the analyzed window; the windows scanned by BreakDescriptor are highlighted in yellow.

[6] "Spectral emphasis may be described as an acoustic feature reflecting the relative intensity in the higher frequency bands" (Heldner, 2001). See Traunmüller and Eriksson (2000).

[7] See Teixeira (2018) for the complete list of measurements.

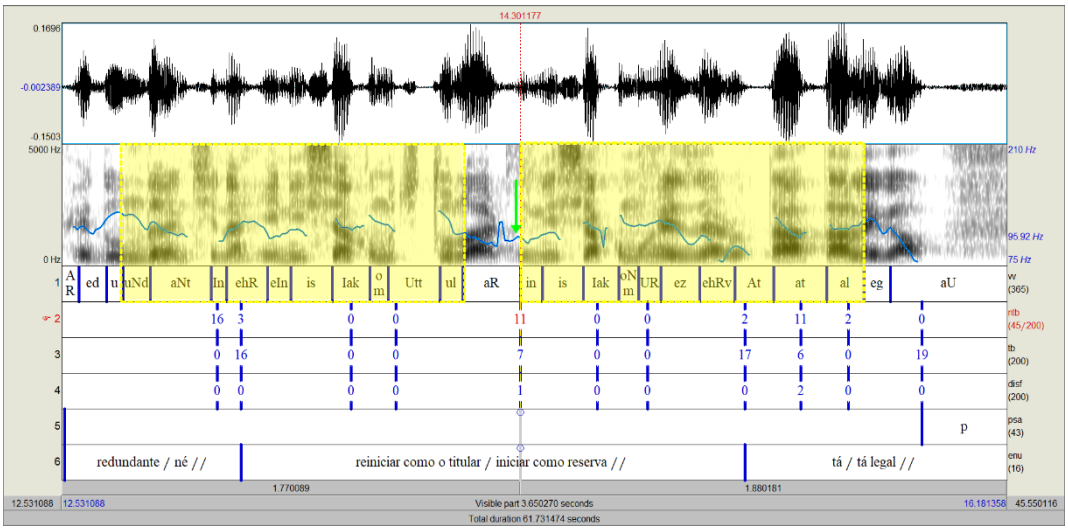

Figure 1: Windows scanned by BreakDescriptor.

In Figure 1, the target V-V unit is positioned between the right and left windows scanned by the script. In this case, 11 annotators marked the target position as a NTB, 7 annotators marked it as TB and 1 of them marked it as a disfluency position. For the position highlighted with the green arrow, the acoustic-phonetic parameters are calculated in the 10 previous V-Vs units (left shaded area of the spectrogram), in the unit that marks the boundary position and in the 10 V-Vs units after the target position (right shaded area of the spectrogram). The target position under analysis is taken as NTB by BreakDescriptor because at least 50% of the annotators considered it as a NTB. Table 3 shows the total number of perceived boundaries in samples I and II.

Table 3: Analyzed positions.

Tag	Total	%	Sample	Frequency
Terminal	116	4.8	I	70
			II	46
Non-terminal	534	22.3	I	242
			II	292
Non-boundary	1744	72.8	I	985
			II	759

Table 4 shows (silent) pause distribution for TB tags, while Table 5 shows the same data for NTB tags. This information is anticipated here since, as we will see, this constitutes a very important aspect for the analysis of the models and the main point for future research.

Table 4: Pause distribution in terminal boundaries.

Sample	TB	TB with pause	TB without pause
I	70	80%	20%
II	46	76%	24%

Table 5: Pause distribution in non-terminal boundaries.

Sample	NTB	NTB with pause	NTB without pause
I	242	39%	61%
II	292	42%	58%

3 Statistical analysis and results

The Linear Discriminant Analysis (LDA) algorithm was used to develop models composed by multiple parameters designed for the automatic identification of boundaries.[8] Different models were used to tackle the problem of automatic boundary detection. For TB, positions of TB and absence of TB (NTB and non-boundary) were used. For NTB, only positions of NTB and non-boundary (NB) were used, since LDA presents too many false alarms due to confusion between TB and NTB, mainly caused by the effect of pause-related parameters[9].

All models independent variables were selected heuristically, starting from the output of the BreakDescriptor software and trying to reach the best recognition with smaller numbers of measurements and false alarms. Of course, this means that we had to decide which model attained the best balance among these three goals. These were our steps so far:

1. we developed models for detecting TB and NTB using non-balanced data[10] extracted from sample I;
2. we validated these models on the data of sample II and on the full data (sample I plus sample II);
3. we developed models from the non-balanced full data;
4. we developed models with balanced data from each sample and from the full data, and applied them to the different samples and to the full data.

These procedures yield a lot of information. The fact that we have different samples allows us to better understand the impact of the data on the models results. In fact, the different samples present different characteristics. This is important not only to evaluate the capacity of

[8] Statistical analysis of data was performed using the environment for statistical computing R (R Core Team, 2019)

[9] Throughout this paper we will discuss the problems caused by the overestimation of pause by both TB and NTB models, which lead to the greatest amount of false alarms in all models. For a tool that gives a relevant weight to pause, see Avanzi *et al*. (2008).

[10] The balancing process consists in using the data that should be captured by the model and the same amount of randomly chosen data that the model should not capture. For instance, if we want a model that recognizes all TB in a sample with 40 TB marked by the annotators, the balancing needs 40 randomly chosen positions that are not TB (they can be NTB or NB). In fact, Machine Learning models generalize taking into account the amount of data. If data are non-balanced, the model tends to privilege those kinds of data that are present in larger number (Wei and Dunbrack Jr., 2013). As for non-balanced data, we mean all the data of the excerpts, irrespectively of the fact that they are NB, NTB or TB positions.

each model for generalization, but also to better understand the reasons for false alarms, and therefore to consider heuristics that could solve the problems highlighted by them. As already mentioned, the presence of a silent pause seems to be an important reason for false alarms and at the same time a problem easy to solve following perceptual criteria, as we will see later.

3.1 Models developed for TB

3.1.1 Models built with non-balanced data

We started looking for the best model using the data of the first sample. As a first step, we used 70% of the data to train the model and 30% for test. The second step was to select and annotate sample II and to use the whole sample I for training and sample II for test. At this point, we did not foresee a balancing of the data; so, we called it TB-nb (i.e. non-balanced). The TB-nb model developed from sample I gave us a good result (80%), with a high performance of recognition and a relatively small amount of false alarms. The TB-nb model trained for the entire sample I can be seen in Appendix 1[11]. When we tested it on the second sample, we found that the recognition decayed (-10%) and a growth of false alarms was observed (+3.5%). The test with the full data (TB-Full) produced an intermediate result (75.6% with 7.4% of false alarms). The results of the application to the different samples are shown in Table 6, where the results of the TB model extracted from the whole non-balanced data are also shown. We will come back to this later in this section. *Main parameters* on the Tables means that the model is formed basically by parameters that involve a certain type of measurements, considering their number and their weight. This column should be interpreted as an extreme synthesis on what can be found during the discussion of each model and in the appendixes.

Table 6: TB models developed with non-balanced data

Boundary	Model	Main parameters	Data set	Sample	Performance (%)	False alarms (%)
TB	TB-nb	Pause and f0	training	I	80	5
			test	II	70	8.5
			test	Full data	75.6	7.4
TB	TB-full	Pause and f0	training	Full data	65.5	2.9

Comparing the two models, we realized that the main reason for the loss of explanatory potential was due to the significantly different number of TB in the two samples, and essentially the different number and distribution of pauses (see table 4). This is confirmed by analyzing the false alarms of the two samples and observing that the model seems to overestimate the relevance of a pause as a feature for detecting TB. Besides these two reasons, many problems with false alarms related to pause emerge also in the NTB models, as we will see later. It is known, in fact, that silent pauses, while always being seen as a sufficient condition for a boundary, do not seem a relevant one to distinguish between the two different kinds of boundaries we were looking for. Pauses can mark both TB and NTB, and, according to research conducted on spontaneous speech corpus data (Raso *et al.*, 2015), its duration cannot be seen as a strong correlate of the nature of the boundary. The TB-nb model is made up by 20 different

[11] All the appendixes can be seen in the metadata section.

measurements. The two most important parameters are *pause duration* and *pause presence*. Their weight is much higher than the weight of the other 18 measurements.

Besides this, it is interesting for the discussion to notice that eleven parameters point to the importance of f0 descriptors. Moreover, among these eleven f0 measurements, four of them have the highest weight after the pause measurements. The most important of them was f0 reset (F0 median difference across boundary) and the second one was F0 median slope in the first V-V unit of the right window. The third one was F0 median slope mean within the left window. This might be correlated with declination. Duration at a first sight did not seem to play an important role, as happens in all the TB models, but this deserves some more consideration and we will come back to it later. The first duration-related measurement appears only in 7th position as an LDA load and concerns the difference in rate of duration-related peak across boundary. The load of the V-V normalized duration in 1st V-V unit on the left window appears only in 13th position, and four other measurements related to duration and rhythm appear with smaller loads just before one measurement related to intensity at the last position.

Let us now make some observations about the TB-full model, developed with the non-balanced data of the two samples together. This model has a lower performance because our goal was to ensure a lower number of false alarms. It can be seen in details in Appendix 2. This model is the only one for which pause duration clearly has a much higher load than any other parameter. Also, the presence of pause (which usually is the most relevant measurement for explaining a TB), even occupying the second position in the hierarchy of the model, does not differ much in load from the other measurements. The model features seventeen measurements: fourteen are related to f0 and only the last one to duration. What we can observe is that the burden caused by a model that gives too much weight to pause duration compared to other measurements seems to lead to a much lower performance, probably leaving aside many positions followed by short pauses. This confirms the impression that pause duration, if considered the main feature to indicate terminality, does not necessarily work well.

3.1.2 Models built with balanced data

Our next step was to develop models from balanced data, which we call TB-b. The model developed from balanced data from sample I (see Appendix 3), when tested with the non-balanced data of each sample and with the full data, shows an increase of performance, but also an increase of false alarms. Considering these two aspects, it would not be easy to choose between the models developed from balanced or non-balanced data, as Table 7 shows.

Table 7: TB models developed with balanced data of sample I.

Boundary	Model	Main parameters	Data set	Sample	Performance (%)	False alarms (%)
TB	TB-b1	Pause and f0	training	I-b	84.2 (+4.2)	7 (+2)
			test	II-nb	76.3 (+6.3)	12 (+3.5)
			test	Full data-nb	79 (+3.5)	9.7 (+2.3)
			test	I-nb	84.2 (+4.2)	7.6 (+2.6)

The reason that probably renders the model extracted from balanced data preferable is the fact that it is based on only eight parameters instead of twenty. These parameters substantially confirm the analysis made based on the non-balanced data model. The first parameter, with a much higher load over the others, is the presence of a pause. Pause duration is the third parameter, but with a much lesser load. The second parameter is the general changes in the

intonational contour of f0 on the left window. Then we have one duration parameter and four f0-related ones. The duration-related one is the change of articulation rate. The f0-related ones are, in decreasing order of load, the change of rate of f0 maxima (which is related to the rate of pitch accent), the change of f0 between the V-V at the boundary and the first V-V on its right, the reset of f0 after the boundary, and the median f0 slope on the last V-V unit on the left window.

It seems that pause continues to be an overestimated parameter, since the almost totality of false alarms coincide with a position where there is a pause. The load of pause presence is clearly much higher than the load of all the other parameters. Lower in the hierarchy, f0-related parameters continue to be relevant. An f0 measurement is in the second position in the load hierarchy, and comes before pause duration parameter, which has a much lesser load in comparison with that of the non-balanced data model. The change in articulation rate is the only duration-related parameter present in the model. It is difficult to understand why the change in peak rate of smoothed F0 peaks per second appears in 5th place, but it is easy to understand the importance of the other f0 measurements. No intensity-related parameters are relevant.

On the basis of early work on the matter, we tried to insert other duration-related parameters in the model, mainly parameters related to lenghtening in the target V-V just before the boundary and at the first V-V unit on his left and on his right, which yields an interesting result: the insertion of these parameters did not change performance. We will come back to this point later.

We also developed other models for TB. One from the balanced data extracted from sample II and one extracted from the balanced data of the full data (sample I plus sample II). For the entire composition of these models, see Appendix 4 and 5.

The model extracted from the balanced data of sample II was called TB-b2 and the results of its applications can be seen in Table 8:

Table 8: model developed with balanced data extracted from sample II.

Boundary	Model	Main parameters	Data set	Sample	Performance (%)	False alarms (%)
TB	TB-b2	Pause and f0	training	II-b	82.6	12.3
			test	I-nb	80.8	7.8
			test	Full data-nb	81.5	10
			test	II-nb	82.6	12.3

This model shows more similar performances than the previous one when applied to different data, but this does not seem to make it immediately better or worse than the model extracted from sample I. What is confirmed is that the different data sets of the two samples have an impact on the performance, and this must be considered. The relevance of certain parameters is confirmed, with a few differences. Pause presence remains the main parameter with a much higher load with respect to the others. Now pause duration is the 5th and last parameter. This means that one advantage of this model is that it is simpler in terms of number of measurements. The other three measurements are related to f0 and reflect previous knowledge on boundary-related parameters: reset of f0, f0 slope median difference across boundary, and median f0 slope in the boundary unit itself. Additionally in this case, inserting duration-related measurements does not change the results, just increases the number of measurements.

A last TB model was developed using balanced data from the two samples together (which we call TB-bFull) and it was applied to the three groups of non-balanced data. The results can be observed in Table 9. The whole model can be seen in Appendix 5.

Table 9: Model depeloved with data from the two samples together.

Boundary	Model	Main parameters	Data set	Sample	Performance (%)	False alarms (%)
TB	TB-bfull	Pause and f0	training	Full data-b	81.5	12.6
			test	I-nb	83.5	7.3
			test	II-nb	78.2	11.4
			test	Full data-nb	81.5	9.2

Again, we do not see clear evidence that this model is either better or worse than the other two, if we just look at its performance. It seems to work better on sample I than on sample II data, which shows how the results are sensitive to the characteristics of each sample. If we investigate the parameters of the model, we must first observe that it presents six measurements, being therefore simpler than TB-b from sample I and presenting only one more measurement than TB-b from sample II. Once again, the two measurements related to pause occupy the first two positions in terms of load, coherently with the fact that the model performs better with sample I than with sample II. In general, presence of pause seems more important than pause duration, except for the TB-nb model extracted from sample II and for the TB-nbFull. It seems, therefore, that an important effect of data balance is the reduction of the importance of pause duration. Presence of pause also presents a clearly higher load compared to the other measurements. Then, we have three measurements related to f0, and a last measurement related to duration. The three f0-related measurements are, in decreasing order of load: f0 reset, median f0 slope in the last V-V unit on the left of the boundary position V-V, and F0 median slope in V-V unit on the left window immediately before the boundary. The duration measurement concerns the first V-V unit on the left of the boundary unit.

If we compare the three balanced models and their composition, we observe many elements of coherence among them, as well as between them and what is usually said in the literature (Cruttenden, 1997; Wagner and Watson, 2010; Amir *et al.*, 2004; Blaauw, 1994; Mo, 2008). In terms of number of parameters, there is no evident difference, despite the fact that TB-b2 features only 5 measurements and TB-b1 8 measurements. TB-bFull features 6 measurements. In terms of results (considering both correct identification and false alarms) model TB-b1 reaches the best result, but only on sample I, decaying in correct identification especially in sample II. Model TB-b2 reaches the most similarity when applied to the different samples, but it also features an increase of false alarms; as expected, model TB-bFull shows an intermediate situation, and seems more appropriate for the data of sample I than for those of sample II.

What seems especially interesting is the coherence in terms of measurements: presence of pause is clearly the most important one in the three models, especially in the model extracted from the two samples in isolation. Its importance in the model extracted from the full data seems less different from that of other parameters, but still occupies the first place. Duration of pause is present in all the models, but while it is the second most important parameter in TB-bFull and the third one in TB-b from sample I, it occupies the last position in TB-b trained on sample II. This may be due to the limited presence of TB and especially those followed by a

pause. However, also in TB-b from sample I, duration of pause does not seem to have a significant difference in load compared to the other parameters. This poses a question: why TB-bFull gives to pause duration more importance than the two samples from which it is built? It does not seem easy to answer this question, but it shows that, once more, pause is a parameter that, despite its importance, generates particular effects on the different models, their performance and the false alarms.

All the TB models show that f0 measurements are very important: f0 reset is the second most relevant factor in TB-b2 and the third one in TB-bFull, while it occupies the 7th position in TB-bI; the change in f0 between the V-V unit at a boundary and the first unit on the right is the fourth most important measurement in TB-bFull and the 6th one in TB-b1; f0 slope median difference across boundaries is the fourth measurement in TB-bII; f0 median slope in first V-V unit of the right window and f0 median slope in the immediately leftward of the boundary unit is the 8th one in TB-b1.

All these measurements deal with f0 changes in the three V-V units that comprise the boundary unit and the adjacent one on the left and/or the right. They might refer, at least partially, to the same phenomena: mainly f0 reset or shift and a clear change of the movement at the boundary point or immediately before or after it. Two other measurements related to f0 appear only in TB-b1: f0 median slope mean within the left window appears in the 2nd position; one possible hypothesis is that this parameter refers to declination. The alternative hypothesis is variability in using pitch accent. Both hypotheses are, at least partially, consistent with terminality, since terminality signals the end of the utterance. Each utterance in fact is characterized by declination and by possible variability in its main pitch accent, which is crucial to signal the illocutionary value of the utterance. However, we may have (especially in monological speech) long terminated sequences with more than one illocution, each one followed by a NTB. Change in F0 peak rate appears in 5th position. It is related to change in pitch accent rate, which concerns mainly expressivity, and, to a certain degree, is associated with the semantic/pragmatic (illocutionary) value of the utterance. In any case, recall that TB-b1 features more measurements than the other two models.

On the other hand, V-V duration-related measurements play a marginal role. No such a measure appears in TB-b2 model, and just one appears in the other two: in TB-b1 change in articulation rate appears in 4th position, and duration of the first V-V unit on the left window appears in 6th position of TB-bFull. No intensity measurement is present in any model.

We can therefore say that the models are largely coherent among themselves. With respect to what is usually said in the literature, the only surprise is the lower relevance of durational measurements. This will be discussed later.

3.2. Models developed for NTB

For the NTB models, we used only positions that the majority of the annotators marked as NTB or NB (no boundary). In fact, the identification of NTB positions seems a much more difficult task. Our steps were as follows:

1. Firstly, we built a model from the non-balanced data of sample I. This model is called NTB-1.
2. Since we did not reach a satisfactory result, what we did was to withdraw the positions correctly identified by this first model and to develop a second model with the remaining positions (NTB-2); we again withdrew the positions correctly identified by this second model and developed a third model for the remaining positions (NTB-3). At the end of this process, 98% of NTB were correctly identified. This result can be seen in

Table 9, by summing 68% (NTB-1 training), 25% (NTB-2 training) and 5% (NTB-3 training). These models were tested on sample II and to full dataset, as shown in Table 9. These three models can be seen respectively in Appendix 6, 7 and 8.

3. We built a model using balanced data from sample II and a model using balanced data extracted from the Full data (sample I plus sample II), and applied them to all the non-balanced samples.

3.2.1. Models developed from non-balanced data

The performance of the NTB models built with non-balanced data extracted from sample I and applied to the other samples can be seen in Table 10:

Table 10: NTB models developed with non-balanced data.

	Model	Main parameters	Data set	Sample	Performance (%)	False alarms (%)
NTB	NTB-1	Standardized V-V duration and pause	training	I	68	22
			test	II	66	28
			test	Full data	66	25.6
NTB	NTB-2	Speech rate and f0	training	I	25	20
			test	II	19	47
			test	Full data	17.1	38.8
NTB	NTB-3	Standardized V-V duration and f0	training	I	5	12
			test	II	3.4	13.6
			test	Full data	4.7	16.4

In Table 10, NTB-1 refers to the first model, extracted from the whole data (after the exclusion of the TB positions marked by the majority of the annotators) of sample I. This model reached an agreement of 68% with the annotators. This performance is slightly lower when the model is applied to sample II and to the full data (sample I plus sample II), but still very close, which signals generalization was achieved. However, the number of false alarms is high. NTB-2 refers to the model built on the data that NTB-1 did not correctly identify, after having withdrawn the data recognized by NTB-1. NTB-2 comprises the false alarms of NTB-1. In this case, what seems to be relevant to point out is the increase of the number of false alarms when the model is applied to sample II or the Full data. Correct identification also decays more than in the case of NTB-1. The same happens with NTB-3 (obtained with the rest of data of sample I after the withdrawal of the already correctly identified positions but including all the false alarms). Because NTB-3 was trained to a restricted amount of data, its results must be taken with reserve. The fact that NTB-2 and NTB-3 decay so much is expected, since the number of positions is much lower and therefore more dependent on the characteristic of the specific data.

It is important to say that there is a significant number of NTB positions that can be captured with more than one model: some of them can be captured by the three models, some by two of them, and only a remaining part is captured by one model alone. This confirms what we have already seen with TB models: sample I and sample II present different characteristics that have an impact on the performance of the models based on one sample. However, it is interesting to observe that the NTB-1 model shows less difference in performance than the other two models. This suggests that it would be interesting to analyze in more detail the

characteristics of data recognized by the different models. We analyzed the false alarms, and again a large number of them is related to pause presence. Having pause presence as a boundary predictor leads to the fact that both TB and NTB models frequently signal the same positions.

NTB-1 comprises nine measurements; NTB-2 comprises ten measurements and NTB-3 eight measurements. It is interesting to observe the composition of the three models, especially the first and more accurate one, and compare them with the composition of the TB models, which, as we have already seen, are very similar to each other.

NTB-1 features six duration-related measurements, the two measurements for pause and just one f0-related measurement with a very low load in the penultimate position among the nine measurements of the model. Looking at the load of the measurements in this model, we can clearly divide them in three groups: the first 3 measurements, all duration-related, present a load between 4.5 and 4.2; the following two measurements present a load of 2.6 and 2.3 and are presence and duration of pause; the other ones have 0.3 and 0.2 as loads.

The duration-related measurements with the highest loads are: normalized duration of the V-V unit at the boundary position; normalized duration of the first V-V unit on the right window; normalized duration change between the first V-V unit on the right window and the V-V unit at the boundary point. The other durational measurements are: change in articulation rate, change in speech rate, and normalized duration of the first V-V unit immediately leftwards. The only f0-related measurement, with a very low load, is change in f0 slope, which is related to f0 variability.

These observations, if compared to those made for the TB models, show that duration is very important for the correct identification of NTB, while it was absent or almost absent in the TB models. On the other hand, in this NTB model, f0 is almost absent, while it was the most important factor, together with pause, in TB models. If we turn to pause, we can see that this parameter plays an important role both for TB and NTB, even if it seems more important for TB than for NTB. This is another important argument to explain why the great majority of false alarms, both in TB and in NTB models, are related to positions followed by pause.

If we compare the composition of the three NTB models, we can also say that, while NTB-1 basically comprises V-V duration-related and pause measurements, NTB-2 comprises mainly f0-related measurements, and NTB-3 comprises a mix of duration-related and f0-related measurements. However, while NTB-1 seems to separate the load of the measurements in three clear groups, as we already said, the NTB-2 model does not present a clear difference in terms of load among the parameters. In the case of NTB-3, the measurements related to V-V duration present a much higher load than the measurements related to f0 (which, on the other side, are in larger number), but we cannot give much importance to a model extracted from very little data. However, the main general impression is that NTB cannot be seen as just one type of boundary, while this seems more likely for TB.

We did not develop a model from non-balanced data from sample II, and directly developed a model from non-balanced data from the two samples combined together, which we call NTB-nbFull, and whose results are shown in Table 11 (see Appendix 9 for the details about the model):

Table 11: Results of the model NTB-nbFull.

Boundary	**Model**	**Main parameters**	**Stat. analysis**	**Sample**	**Performance (%)**	**False alarms (%)**
NTB	NTB-nbFull	Pause and duration	Development	Full data-nb	40.6	0.6

This model recognizes a reduced number of positions, compared to the NTB-1, but has the advantages of presenting very few false alarms and only five measurements. Among them, presence of pause has a much higher load compared to all the others. The other measurements do not present a relevant difference in load among themselves, and are all related to V-V duration. Difference in V-V normalized duration mean between first unit of the right window and the boundary unit; V-V normalized duration in the V-V unit immediately leftward the boundary; V-V normalized duration of V-V unit at boundary point; change in articulation rate. Therefore, the importance of duration-related parameters and the fact that pause is a parameter that the two main kinds of boundaries have in common is confirmed.

3.2.2. Models built from balanced data

Before getting in a general discussion and proposing some strategies to improve correct identification and reduce the false alarms, we still need to show the models obtained from the balanced data. The model extracted from balanced data of sample I gave the results shown in Table 12 (see Appendix 10 for the details):

Table 12: Results of model NTB-b1.

Boundary	Model	Main parameters	Stat. analysis	Sample	Performance (%)	False alarms (%)
NTB	NTB-b1	Pause, articulation rate and stand. segment dur.	Development	I-b	72	23.5
			Validation	II-nb	71.2	38.2
			Validation	Full data-nb	71.9	33.4
			Validation	I-nb	72	29.8

This model reached much more satisfactory results than NTB-1 and its recognition power remains stable in all the applications to the different samples of non-balanced data, but, at the same time, it presents a great amount of false alarms, especially when applied to sample II. Again, the false alarms involve principally positions followed by pause. This model comprises eleven measurements. Once again presence and duration of pause are the first ones, with a clearly higher weight with respect to all the other measurements. At the same time, we have the confirmation that duration-related parameters are decisive for the correct identification of NTB. They occupy the hierarchical positions of the model from the 3rd to the 6th ones besides the 8th and the 9th ones. The other measurements involve f0.

In order of load, the durational parameters are: change in articulation rate; Difference in V-V normalized duration mean between first unit of the right window and the boundary unit; V-V normalized duration in the V-V unit immediately leftward the boundary unit; V-V normalized duration of V-V unit at boundary point; change in normalized duration variability and change in speech rate. Among the f0 parameters, the most important is the f0 mean slope in boundary unit, followed by the f0 change between the boundary unit and the two adjacent ones.

This model, obtained with the balanced data, seems to make a good synthesis of the NTB1 and the NTB2. The fact that the two measurements of pause are the first ones could be an important indication for future strategies, as we will see.

The model extracted from the balanced data of sample II presents the results shown in Table 13 (see Appendix 11 for details):

Table 13: Results of model NTB-b2.

Boundary	Model	Main parameters	Stat. analysis	Sample	Performance (%)	False alarms (%)
NTB	NTB-b2	F0, articulation rate and stand. segment dur.	Development	II-b	69	17.4
			Validation	I-nb	75.6	40.9
			Validation	Full data-nb	71.7	35.4
			Validation	II-nb	73.9	31.2

Comparing the potential for correct identification and the false alarms of this model with the model extracted from sample I and presented in Table 12, we mainly observe a small gain in terms of correct identification and an increase of false alarms, especially if we compare the results when applied to the whole data. We also need to observe that the NTB-b2 model shows a lower performance compared with NTB-b1 and at the same time a lower number of false alarms.

This model is even more complex than the previous one, since it comprises thirteen measurements. However, it is interesting to compare their composition: seven of the eleven measurements of NTB-b1 are also present in NTB-b2. Six of them are duration-related measurements and only one is related to f0 (f0 slope mean difference across boundary). A few other measurements seem to be related to similar phonetic phenomena. In NTB-b1 the f0 mean slope appears in boundary unit, while in NTB-b2 the f0 median slope shows up in the V-V unit immediately leftward of the boundary. In NTB-b1 (in 11th position), the f0 in the V-V unit occurs immediately leftward the boundary unit, while in NTB-b2 the f0 median slope appears in first V-V unit of the right window. All these measurements are related to something that changes in the f0 at the boundary unit or/and the adjacent ones. In NTB-b2 the f0 reset also shows up as the most important parameter. This parameter can be related to the same group of phenomena just mentioned. The main difference between the two models is due to the presence of pause and pause duration as the two most important parameters in NTB-b1, while no pause measurement appears in TB-b2. Once again, we observe that pause represents an important predictor, not only for most of the model composition, but also for its different impact on the data of the two samples; and once again, we observe that a significant part of the false alarms is related to positions followed by pause.

Finally, let us see what happens with the NTB model extracted from the balanced data of the two samples together. The results of the model NTB-bFull can be seen in Table 14 and the model details are in Appendix 12.

Table 14: Results of model NTB-bFull.

Boundary	**Model**	**Main parameters**	**Stat. analysis**	**Sample**	**Performance (%)**	**False alarms (%)**
NTB	NTB-bfull	Pause, f0 and speech rate	Development	Full data-b	54.5	6
			Validation	I-nb	51	11
			Validation	II-nb	57.5	13.4
			Validation	Full data-nb	54.5	12

The model extracted from the full balanced data and tested with the other sample is still under development; so far it offers an agreement with humans less interesting than the other two NTB balanced models, but at the same time it presents much fewer false alarms, which should not be underestimated. This model, at this stage, presents sixteen measurements. Once again the two measurements related to pause are at the top of the hierarchy: presence of pause has a higher weight than pause duration (as usually happens), but the two pause measurements have a higher weight than all the other measurements, whose weight diminishes slowly. The model presents nine f0 measurements and five duration ones. Both types of measurements confirm that what happens in the three units around the boundary point is essential, while global measurements seem to have much less weight, with the exception of articulation rate, and local measurements besides the three central V-V units do not seem to have any weight. In all the models, we have seen so far, only in one case, in NTB-1, a local measurement related with the penultimate V-V unit appears with a very low weight.

At this point, we are ready to make some final considerations and propose strategies for the future of this research.

4 Final considerations and future strategies

Analyzing the models developed so far, we found eight recurring aspects that can be put together, in order to elaborate future strategies to reach better models. They are listed in 1 to 8 below.

1. We observed that the data sets have an impact on the models: training on sample I and sample II leads to different models. However, we can identify some parameters that constantly appear in the models built from the two samples. At the same time, the results indicate that testing on a different data set does not lead to radically different performances. Looking from this perspective, we can say that we have a good departure point and we should look now for strategies that are flexible with respect to specific aspects of the data. Another potential issue is whether the amount of data we annotated and analyzed can be considered sufficient to produce models that can be generalized. This aspect goes beyond the scope of this paper.

2. An aspect that often emerges is the relative role of pause parameters in almost all the models, and at the same time the fact that these parameters seem the main reason for false alarms and the distinction in performance between the data sets. Actually, false alarms in positions not followed by pauses are very rare. Pause is, of course, a necessary parameter for boundary identification, since it is the only parameter that always generates a boundary automatically, but it also is responsible for a confusion between the two types of boundaries. Therefore, we need a strategy to face this problem. The fact that presence of pause is a highly perceivable parameter by human annotators may facilitate the task. In fact, as we will propose later, we need more models, that could be applied on the data, arranged into a hierarchy. This means that we need to ask the annotators to separate data according to certain salient parameters. Presence/absence of pause seems a very good candidate to separate the data for specific models: they are a very relevant parameter for our task, the principal cause of false alarms, and at the same time very easily perceived by annotators.

3. We observed, in all the models, that local measurements seem to have a greater importance and that only the three central positions of the window taken in consideration by the BreakDescriptor look really important. This is a relevant point, since it might suggest a strong reduction of window extension.

4. It might also be useful to investigate the possible negative effect, from a statistical point of view, that different measurements have on the evaluation of the same phenomenon.

This means that there might exist colinearity or nested variables among different measurements. For instance, pause duration overlaps with pause presence, and it might be useful to exclude presence of pause, since pause duration, of course, already implies presence of pause. This kind of situation might cause the overestimation of the load attributed to one phenomenon, since it is considered by more than one measurement. This might happen also with duration-related and f0-related measurements, which should be carefully considered.

5. Intensity-related parameters do not seem relevant in any model.

6. Special attention should be paid to the effect of duration-related parameters, especially if we compare TB and NTB. We noticed that: (i) the models for TB do not really profit from the insertion of V-V duration-related parameters; (ii) durational parameters seem necessary in all the NTB models; (iii) if we add some durational parameters to the TB models, we do not change the performance and just add more parameters to the model. This seems to suggest that V-V duration-related parameters are present in any type of boundary, but do not play a special role in distinguishing between TB and NTB. This is something that resembles what happens with pause, but while pause is something very easily perceived by annotators, V-V duration effects are not (they are related to segment lengthening and shortening).

7. As for f0-related measurements, they seem very important to capture TB, but they are much less important for NTB. In the main non-balanced models (those that capture more than 2/3 of the positions), only one f0-related measurement appears in NTB-1, with a very low load, and no f0 measurement appears in the NTB-Full. This picture changes partially if we consider models extracted from balanced data and if we consider NTB-2. Let us see first what happens in NTB-2. In this model, f0-related parameters seem very important, but we should keep in mind that this model identified 25% of boundary position, since we have already withdrawn the data recognized by NTB-1, and that NTB-2 does not feature any measurement for pause. Now, let us see what happens in the balanced models. While the NTB-b built from sample I still gives a very reduced importance to f0-related measurements, the NTB-b built from sample II shows a sort of balance between f0-related and duration-related measurements. This should be analyzed together with the fact that NTB-b II is the only model that does not feature any pause measurement. In fact, in the NTB-bFull model the pause-related measurements are the two most important ones, before mixing f0 and duration-related parameters, with more emphasis to the former. Therefore, it seems that the importance of f0 is related to the behavior of the two samples with respect to pause, presented above in Table 4. Sample I features 56 TB with pause and 14 TB without pause; sample II features 35 TB with pause and 11 without pause. On the other hand, sample I features 91 NTB with pause and 142 without pause, while sample II features 120 NTB with pause and 170 without pause. The balance of data with and without pause with respect to TB and to NTB is very different. This might be one of the main reasons, if not the main one, for the different models inferred from the two samples.

8. We clearly observed that TB can be identified much more easily than NTB. Any single model already reaches very good results for TB, while no NTB model reaches satisfying results, either because of low correct identification or because of the great amount of false alarms, and often for both reasons. This seems to have two consequences: (i) it confirms that TB and NTB should be treated separately, reinforcing the hypothesis that we perceive TB as something different from just a boundary irrespectively of its nature; (ii) it shows that NTB cannot be treated as just one category; we need to deal with different kinds of NTB.

Together with the eight above listed considerations, we need to recall that this research has two goals: one is of course to reach models that can be applied to spontaneous speech corpora and perform an automatic segmentation as trustable as possible. The other goal is to understand better how we signal and perceive boundaries in natural speech, and to investigate if

we can distinguish among boundaries of different nature. This means that any model needs to interact with perceptual cues. It should somehow help us to better understand what we perceive when we judge that a particular position in the speech chain is a boundary and when we judge that a boundary has a specific function.

However, our findings not necessarily reflect exactly the physical nature of prosodic boundaries or parameters that human cognition uses to perceive a boundary. The models try to reproduce the results achieved by the annotators through their perception, but we cannot state that the models capture precisely the perceptual parameters of humans, or that the parameters weight and combinations reflect exactly human perception. Nevertheless, our findings can be considered an attempt to better understand the physical cues behind human perception and allow to make good assumption on the importance of features (or combinations of features) that might lead us to a more advanced knowledge of the relationship between human perception and physical cues with respect to these different kinds of boundaries

So far, we are working with the hypothesis that there are two different kinds of boundaries from a functional point of view. However, our results suggest the reevaluation of this hypothesis for at least NTB, which seems to be associated with different functions. Building on this work, our next steps will be based on the following considerations:

1. It seems to be more important to have models with a very low rate of false alarms. This is motivated by the fact that we can easily apply more models in a hierarchical way to achieve higher correct identification rates than deal with false alarms.

2. It is crucial to differentiate the data in a way that we can create different models with specialized functions.

3. Since the performance of the model is measured with respect to human performance, we need to differentiate the data using a highly perceivable parameter. The most important one in this vein is pause, which also seems to be responsible for the majority of false alarms

4. Our plan is to ask the annotators to also inform whether they perceive a pause or not from the majority of TB and NTB (the ideal segmentation based on the agreement among annotators). For this task, we will invite only the annotators that better respect the following criteria: (i) a higher inter-rater agreement reached during the annotation (done more or less two years ago); and (ii) a high intra-annotator agreement, i.e. the agreement with themselves, repeating the same task two years later. The first criterium seems more important, but the second one could lead to a different decision in some cases. If we keep the annotators with higher degree of agreement, we observe that this agreement is very high. The inter-rater agreement among the 19 annotators is presented in Table 15, taking as reference the ideal segmentation (i.e. the segmentation used to build the model, which is the result of the decision of the majority of the annotators). The intra-annotator agreement computation was already done, among ten annotators, and the results are shown in Table 16.

Table 15. Inter-annotator agreement, taking as reference the segmentation applied to build the models.

Annotator	General	TB	Disfluency	NTB	NB
To.	0.92	0.86	0.87	0.92	0.94
Lui.	0.91	0.80	0.94	0.89	0.94
Áb.	0.91	0.87	0.96	0.88	0.93
Gi.	0.91	0.83	0.95	0.88	0.93
Fr.	0.89	0.82	0.91	0.88	0.92
Sa.	0.89	0.88	0.91	0.87	0.92
Ba.	0.89	0.88	0.92	0.86	0.91
Ad.	0.88	0.87	0.91	0.86	0.89
Th.	0.88	0.83	0.96	0.83	0.90
Al.	0.87	0.77	0.96	0.83	0.91
He.	0.87	0.68	0.99	0.83	0.90
Mr.	087	0.68	0.99	0.83	0.90
Ol.	0.87	0.82	0.92	0.83	0.89
Ma.	0.85	0.80	0.81	0.82	0.89
Luc.	0.84	0.75	0.95	0.79	0.88
Ta.	0.80	0.80	0.84	0.76	0.81
Br.	0.78	0.87	0.34	0.73	0.88
Is.	0.76	0.80	0.30	0.70	0.89
Ca.	0.75	0.71	0.87	0.67	0.79

Table 16: Intra-annotator agreement.

Annotator	General	TB	Disfluency	NTB	NB
To.	0,92	0,90	0,90	0,94	0,93
Gi.	0,90	0,86	0,96	0,83	0,94
Sa.	0,90	0,91	0,90	0,87	0,93
Fr.	0,88	0,79	0,99	0,83	0,91
Lui.	0,88	0,85	0,93	0,83	0,90
Th.	0,87	0,71	1,00	0,76	0,92
Al.	0,87	0,80	0,97	0,78	0,90
Luc.	0,86	0,80	0,88	0,78	0,91
Ba.	0,84	0,82	0,88	0,76	0,89
Br.	0,79	0,90	0,54	0,71	0,91
Ca.	0,76	0,56	0,93	0,63	0,83

The intra-annotator test was satisfying, since it ranges from 0.92 to 0.76, and for eight annotators it ranges from 0.92 to 0.84 (Fleiss's Kappa).

5. Once the annotators have performed this new task, we can separate positions with and without pause and develop two different models. What we hope is that these two models will avoid the confusion caused by the co-presence of boundaries with and without pause, since pause is a very relevant parameter to recognize the presence of boundary, but at the same time seems to be the main parameter yielding a confusion between the two types of boundaries. Once

we have different models for data with and without pause, we can look for the best hierarchy by applying these models, progressively withdrawing the data recognized by the previous model. The hierarchy will be guided by a simple criterium: the model exhibiting the best performance will be applied first, followed by the models with lower performance. Subsequent models will, therefore, need to deal with fewer and more coherent data, because the elimination of data recognized by previous models will automatically make easier to categorize the remaining data.

6. In parallel, we will reduce the window size scanned by the BreakDescriptor, in order to verify if this strategy (as suggested by the data analysis) reduces the noise produced by so many measurements. Our idea is to reduce the windows from 21 V-V units to 7 and 3 V-V units. This means that we will try two different strategies of reduction of the window, in order to verify the impact that this reduction may have on the global parameters. At the same time, we will try to understand better if some measurements could be overestimated because they partially overlap with each other in capturing the same phenomenon. We also need to consider that the phonetic phenomenon that leads to boundary perception may happen in a region not coinciding with the phonological boundary; then it is our expertise that decides the exact phonological position where to place the boundary, which in most cases coincides with the end of a phonological word.

ACKNOWLEDGEMENTS

We thank the text annotators, all members of the Lab of Empirical and Experimental Linguistic Studies (LEEL) at UFMG. We also thank Capes, CNPq and Fapemig for funding the research.

REFERENCES

1. Albano E C, Moreira A A. *Archisegment-based letter-to-phone conversion for concatenative speech synthesis in Portuguese.* Proceedings of the ICSLP'96, October 3-6, v.3, 1996, 1708–1711.

2. Amir N, Silver-Varod V, Izre'el S. *Characteristics of intonation unit boundaries in spontaneous spoken Hebrew – perception and acoustic correlates*. Speech Prosody. Nara, 2004.

3. Avanzi M, Lacheret-Dujour A, Victorri B. ANALOR: *A Tool for Semi-Automatic Annotation of French Prosodic Structure: ANALOR*, Campinas, 2008, 119–122.

4. Barbosa P. *Automatic Duration-Related Salience Detection in Brazilian Portuguese Read and Spontaneous Speech*. Speech prosody international conference ISCA, Chicago, 2010.

5. Barbosa P. *Caractérisation et génération automatique de la structuration rythmique du français*, PhD thesis, Institut National Polytecnique de Grenoble, 1994.

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

7. Barbosa P. *BreakDescriptor (2.0)*. Available with the author, 2019.

8. Barth-Weingarten D. *Intonation Units Revised: Cesuras in talk-in-interaction*. Philadelphia: John Benjamins Publishing Company, 2016.

9. Blaauw E. *The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech*. Speech communication 14, Elsevier Science Publishers, 1994, 359–375.

10. Boersma P, Weenink D. *Praat: doing phonetics by computer*, 2015.

11. Bybee J. *Language, Usage and Cognition*. Cambridge: CUP, 2010.

12. Byrd D, Saltzman E. *The elastic phrase: Modeling the dynamics of boundary adjacent lengthening*. Journal of Phonetics 31, 2003, 149–180.

12. Chafe W. *The Deployment of Consciousness in the production of a Narrative.* In: Chae W. (ed.). The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. Norwood: Ablex, 1980, 9–50.

14. Cooper W, Paccia Cooper J. *Syntax and Speech*. Cambridge: Harvard Universty Press, 1980.

15. Cresti E. *Corpus di Italiano parlato*. v. 1. Firenze: Accademia della Crusca, 2000.

16. Cresti E, Moneglia M. (eds.). *C-ORAL-ROM. Integrated Reference Corpus for Spoken Romance Languages*. Amsterdam: John Benjamins, 2005.

17. Croft W. *Intonation Units and Grammatical Structure*. Linguistics 33 (5), 1995, 839–882.

18. Cruttenden A. *Intonation*. Cambridge: CUP, 1997.

19. Crystal D. *Prosodic Systems and Intonation in English*. Cambridge: CUP, 1969.

20. Du Bois J. *Rhythm and Tunes: The notation Unit in the Structure of Dialogic Engagement*. Conference Prosody and Interaction, University of Potsdam, 2008.

21. Du Bois J, Chafe W, Meyer Ch, Thompson S, Englebretson R, Martey N. *Discourse Transcription.* Santa Barbara Papers in Linguistics 4. Santa Barbara: Department of Linguistics, University of California, 1992.

22. Fleiss J. *Measuring nominal scale agreement among many raters*. Psychological Bulletin 76(5), 1971, 378–382.

23. Halliday M. *Speech and Situation*. Londres: University College, 1965.

24. Heldner M. S*pectral emphasis as an additional source of information in accent detection*. Prosody 2001: ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding, 2001.

25. Izre'el S, Mello H, Panunzi A, Raso T. *In search of a baisc unit of spoken language: Segmenting speech.* In Izre'el S, Mello H, Panunzi A, Raso T (eds). In search of basic units of spoken language: A corpus-driven approach. Amsterdam: John Benjamins, forthcoming.

26. Krivokapić J. *The planning, production and perception of prosodic structure*, PhD thesis, University of Southern California, 2007.

27. Ladd R. *Declination reset and the hierarchical organization of utterances.* Journal of the Acoustical Society of America 84, 1988, 530–544.

28. Kelly J, Local J. *On the Use of General Phonetic Techniques in Handling Conversational Material.* In Roger D, Bull P. Conversation: An Interdisciplinary Perspective. Clevedon: Multilingual Matters, 1989.

29. Maschler Y. *Metalanguage in Interaction: Hebrew Discourse Markers*. Amsterdam: John Benjamins, 2009.

30. Mertens P, Simon A. *Towards Automatic Detection of Prosodic Boundaries in Spoken French*. In: Mertens P, Simon A. Proceedings of the Discourse-Prosody Interface Conference (IDP 2013), Leuven: University of Leuven, 2013, 81–87.

31. Mello H, Raso T, Mittmann M, Vale H, Côrtes P. *Transcrição e segmentação prosódica do corpus C-ORAL-BRASIL: critérios de implementação e validação.* In Raso T, Mello H. C-ORAL-BRASIL: corpus de referência do português brasileiro falado informal (I). Editora UFMG, 2012, 125–176.

32. Mo Y. *Duration and intensity as perceptual cues for naïve listeners' prominence and boundary perception.* In Barbosa P, Madureira S, Reis C. Proceedings of Speech Prosody. Campinas: ISCA, 2008, 39–742.

33. Mo Y, Cole J, Lee E. *Naïve listeners' prominence and boundary perception*. In Barbosa P, Madureira S, Reis C. *Speech Prosody.* Campinas: ISCA, 2008, 739–742.

34. Moneglia M, Fabbri M, Quazza S, Panizza A, Danieli M, Garrido J, Swerts M. *Evaluation of Consensus on the Annotation of Terminal and Non-Terminal Prosodic Breaks in the C-ORAL-ROM corpus.* In Cresti E, Moneglia M. (eds.). C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages. Amsterdam: John Benjamins, 2005, 257–276.

35. Ostendorf M, Price P, Shattuck-Hufnagel S. The Boston University Radio News Corpus, Boston University Technical Report, No. ECS-95-001, 1995.

36. Park J. *Cognitive and interactional motivations for the intonation unit.* Studies in Language 26(3), 2002, 637–680.

37. Pijper J, Sanderman A. *On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues.* Journal of the Acoustical Society of America 96, 1994, 2037–2047.

38. Price P, Ostendorf M, Shattuck-Hufnagel S, Fong C. *The use of prosody in syntactic disambiguation*. Journal of the Acoustical Society of America 90(6), 1991, 2956–2970.

39. Raso T, Mello H. *C-ORAL-BRASIL: corpus de referência do português brasileiro falado informal (I)*. Editora UFMG, 2012.

40. Raso T, Mittmann M, Oliveira A. *O papel da pausa na segmentação prosódica de corpora de fala.* Revista de Estudos da Linguagem (23), 2015, 883–922.

41. Raso T, Mello H, Ferrari L. *C-ORAL-BRASIL II: corpus de referência do português brasileiro falado formal*. Forthcoming.

42. Reichel U, Mady K. *Parameterization of F0 register and discontinuity to predict prosodic boundary strength in Hungarian spontaneous speech*, Elektronische Sprachsignalverarbeitung ESSV 26, 2013, 223–230.

43. Selkirk E. *Comments on Intonational Phrasing in English*. In: Frota S, Vigário M, Freitas, M (eds.) Prosodies. Berlim: Mouton de Gruyter, 2005, 11–58.

44. Simon A, Christodoulides G. *Perception of Prosodic Boundaries by Naïve Listeners in French*. In Proc. of the 8th Speech Prosody Conference, Boston, USA, 2016

45. Szczepek Reed B. *Prosody, Syntax and Action Formation: Intonation Phrases and Action Components.* In Bergmann P et al. (eds.), Prosody and Embodiment in Interactional Grammar. Berlin: Mouton de Gruyter, 2012, 142–169.

46. Tabain M. *Effects of prosodic boundary on /aC/ sequences: acoustic results*. Journal of the Acoustical Society of America 113, 2003, 516–531.

47. Tabain M, Perrier P. *Articulation and acoustics of /i/ in pre-boundary position in French*. Journal of Phonetics 33, 2005, 77–100.

48. Teixeira B. *Correlatos fonético-acústicos de fronteiras prosódicas na fala espontânea*, Master Thesis, Federal University of Minas Gerais, 2018.

49. Traunmüller H, Eriksson A. *Acoustic effects of variation in vocal effort bt men, women, and children.* Journal of the Acoustical Society of America 107, 2000, 3438–3451.

50. Wei Q, Dunbrack Jr R L. *The role of balanced training and testing data sets for binary classifiers in bioinformatics.* PloS one 8(7), e67863, 2013.

51. Wightman C, Shattuck-Hufnagel S, Ostendorf M, Price P. *Segmental durations in the vicinity of prosodic phrase boundaries.* Journal of the Acoustical Society of America 91, 1992, 1707–1717.