Semi-automatic and automatic tools for generating prosodic descriptors for prosody research

Plínio A. Barbosa

Speech Prosody Studies Group, Dep. of Linguistics, State Univ. of Campinas, Brazil

pabarbosa.unicampbr@gmail.com

Abstract

This paper presents four Praat scripts which the author of this article developed for analysing speech rhythm and intonation, for helping scientific research on prosody modelling and for investigating the link between prosody production and perception. The BeatExtractor script does automatic, language-independent detection of vowel onsets; the SGdetector script does language-dependent, semi-automatic detection of syllable-sized normalised duration peaks for the study of prominence and boundary marking, the SalienceDetector script does language-independent automatic detection of syllable-sized normalised duration peaks for the study of prominence and boundary, and the ProsodyDescriptor script allows to generate 12 prosodic parameters related to duration, F_0 and spectral emphasis to the study of rhythm and intonation. All scripts are freely available and were tested in previous research since 2006. The languages tested for the first three scripts were Brazilian Portuguese, European Portuguese, German, French, English and Swedish.

Index Terms: tools, Praat, duration, F_0, speech prosody

1. Introduction

Due to volume of data, prosody research can enormously benefit from the automatization of procedures for describing prosodic functions such as prominence, boundary and discursive relations marking. Automatization is advantageous because the same procedures can be applied to analyze the entire corpus and that is useful for preparing data for the statistical analysis as well.

This paper presents four scripts running on Praat [1] which generate prosodic descriptors for prosody research. Assuming from early research that duration is a crucial parameter for signalling stress, prominence and boundary in languages such as English, Portuguese, French, German and Swedish, the scripts are able to detect prosodic boundaries and prominences, as well as to generate a 12-parameter vector of duration, intensity and F0-related descriptors for prosodic analysis. The usefulness of the scripts is discussed regarding the results obtained from their application in previous research.

2. Semi-automatic detection of acoustic salience via duration

The *SGdetector* script for Praat was implemented in 2004 and improved in 2009 and 2010 for allowing the semi-automatic detection of local peaks of smoothed, normalised syllable-sized durations. Languages that use duration to signal both stress and prosodic boundary such as Brazilian Portuguese (henceforth BP) and Swedish [2], English [3, 4], German [5, 6] and French [7] are well suited to take advantage of such a tool. Although throughly tested since 2004 with BP (see, for instance, [8, 9]), the script was used to do analyses in the other languages cited here and has potential to be applied to other languages, at least to other genetically related languages.

The input files for running the script are a TextGrid file containing a phone-sized or syllable-sized segmentation and a broad phonetic transcription of the corresponding Sound file, as well as a TableOfReal file containing a table listing the means and standard-deviations in miliseconds of the phone durations of the language under study. This latter file is delivered as part of the script and is available for BP, European Portuguese, British English, German, Swedish and French. Manual and semi-automatic segmentations and transcriptions of audio files were repeatedly tested for BP along the years (see [9]), confirming the usefulness and correction of a method for detecting prominence and boundary based on syllable-sized duration.

Syllable-sized segmentation is meant as a first step to capture prosodic-relevant duration variation along the utterances [10] and is understood here as an interval between two consecutive vowel onsets. This unit constitutes a phonetic syllable called a VV unit. Besides the crucial importance of vowel onset detection for speech signal processing [11], a clear advantage of a segmentation based on vowel onsets is its potential for automatic detection [9] even under moderately noisy conditions. Automatic detection of vowel onsets can be carried out by using a Praat script developed in 2005 [12, 9], the *BeatExtractor* script, explained in more details in section 2.1.

In the SGdetector script, detection of peaks of prosodic-relevant VV duration is carried out by serially applying a technique of normalisation followed by a smoothing technique. For normalising VV duration, the script uses the $z-score$ transformation given in equation 1, where dur is the VV duration in ms, the pair (μ_i, var_i), the reference mean and variance in ms of the phones within the corresponding VV unit. These references are found in [12, p. 489] for BP. For the other languages, they can be freely obtained from the author.

$$z = \frac{dur - \sum_i \mu_i}{\sqrt{\sum_i var_i}} \quad (1)$$

For smoothing, the script applies a 5-point moving average filtering technique given by equation 2 to the sequence of $z-scores$ (z_i).

$$z^i_{smoothed} = \frac{5.z^i + 3.z^{i-1} + 3.z^{i+1} + 1.z^{i-2} + 1.z^{i+2}}{13} \quad (2)$$

The two-step procedure described here aims at minimising the effects of intrinsic duration and number of segments in the VV unit, as well as minimising the effect of the implementation of stress irrelevant for the prosodic functions of prominence and boundary marking. Local peaks of smoothed $z-scores$ are detected by tracking the position of the VV unit for which the discrete first derivative of the corresponding smoothed $z-score$ changes from a positive to a negative value.

At the output, the script generates two text files, a new TextGrid object and an optional trace of the syllable-sized smoothed/normalised duration along the time-course of the Sound file under analysis. The first text file is a 5-column table displaying the following values for each VV unit: (1) the given transcription recovered from the TextGrid itself, e.g., "eNs", "at" (even for the case where the segmentation is made phonewise), (2) the raw duration in milliseconds, (3) the $z-score$ of the raw duration, (4) the 5-point-smoothed $z-score$ and (5) a binary value indicating if the position is a local peak of smoothed $z-score$ (value 1) or not (value 0). The second text file is a 2-column table containing (1) the raw duration in milliseconds of duration-related stress groups, delimited by two consecutive peaks of smoothed $z-scores$ and (2) the number of VV units in the corresponding stress group. This table was used a lot of times to evaluate the degree of stress-timing of a speech passage, for instance in [13].

The TextGrid generated by the script contains an interval tier delimiting the detected stress group boundaries, synchronised with the input TextGrid, which allows, when selected with the corresponding Sound file, to listen to the chunks that end with a duration-related salience. The optional feature, implemented when the option "DrawLines" is chosen in the input parameters windows, plots a trace of the smoothed $z-scores$ synchronised with the VV unit sequence: each value of smoothed $z-scores$ is plotted in the y-axis in the position of each vowel onset along the plotted original TextGrid. The advantage of this choice for integrating intonation and rhythm descriptions is discussed below.

The correspondence between smoothed $z-scores$ peaks and perceived salience, which refers to both prominence and prosodic boundary, is striking. In [9], we demonstrated an accuracy varying from 69 to 82 % between perceived and produced salience, as shown in Table 1 for the semi-automatic algorithm described here.

Table 1: it Precision, recall, and accuracy in percentage (%) for semi-automatic detected salience against perceived salience for the Lobato corpus read by a female (F) and a male (M) speaker at slow (s), normal (n) and fast (f) rates.

Sp/rate	precision	recall	accuracy
F/n	90	74	82
F/f	73	57	69
M/s	88	67	73
M/f	61	70	70

Perceived salience was determined by asking two groups of ten listeners to evaluate two readings of a passage by two BP speakers (a male and a female at two distinct rates). The listeners in both groups were lay undergraduate students in Linguistics. They were free to listen to the four readings as many times as they wanted. In the first group, each listener was given a handout with the ortographic transcription of the recording and was instructed to circle all the words s/he considered highlighted by the speaker. The second group was instructed to circle the words that preceded a boundary. In each group, the percentage of listeners that circled each word in the text for each reading was initially used to define three levels of salience, according to a one-tailed z-test of proportion. Since the smallest proportion significantly distinct from zero is about 28 % for $\alpha = 0.05$ and $N = 10$, words circled by less than 30 % of the listeners were considered non-salient. For $\alpha = 0.01$, the threshold for rejecting the null hypothesis is about 49 %. Thus, words circled by 50 % of the listeners or more were considered strongly salient. Words salient by between 30 and 50 % of the listeners were considered weakly salient. For the purpose of computing the performance measures in the table, weakly and strongly salient words were both considered as "salient".

The relatively high correspondence between perceived and produced salience allowed us to evaluate the degree of stress-timing in two different speaking styles for two varieties of Portuguese [13]. This work revealed that the speech rhythm of Portuguese speakers differs remarkably from the rhythm of Brazilian speakers when both groups narrate but nor when both groups of speakers read. This was possible to demonstrate through the linear correlation between interval durations delimited by smoothed $z-score$ peaks and number of VV units in the same interval. These two series of values were recovered from one of the tables generated by the SGdetector script.

2.1. Making the script completely automatic

For helping detecting produced salience in large corpora, the SGdetector script was modified into a *SalienceDetector* script for which phone labelling and manual vowel onset marking was made unnecessary. For this we associated a script made some time ago, *BeatExtractor* script [12], with the SGdetector script described above.

The BeatExtractor script implements Cummins' Beat Extractor [14] with some modifications. It generates a TextGrid containing intervals between consecutive vowel onsets. It runs according to five steps: (1) the speech signal is filtered by a default second-order Butterworth (or Hanning) filter; (2) the filtered signal is then rectified; (3) the rectified signal is low-pass filtered using 20 Hz (see step 4a) or 40 H̆z (see step 4b) as cut-off frequencies. This signal is normalised by dividing all points by the maximum value. This normalised, band-specific amplitude envelope is called the beat wave, a technique also applied by [14, 15]; (4) a vowel onset is set either (a) at a point where the amplitude of the beat wave local rising is higher than a certain threshold, or (b) at a local maximum of the normalised first derivative of the beat wave, provided this maximum is higher than a certain threshold; (5) a Praat TextGrid is generated that contains all vowel onsets as interval boundaries. More details in [9].

After obtaining the vowel onset positions, the SalienceDetector script proceeds by cumputing duration $z-scores$ by using fixed values for the reference mean ($Refmean = 193$ ms) and standard-deviation ($RefSD = 47$ ms) duration according to equation 3, where m estimates the actual number of VV units between each interval generated by the BeatExtractor algorithm, which may miss vowel onsets (up to 20 % from all vowels effectively present in the Sound file).

$$z = \frac{\frac{\sqrt{m}}{m}.dur - \sqrt{m}.Refmean}{RefSD} \quad (3)$$

Smoothed $z - scores$ are determined in the same way as before, by using the 5-point moving average filter. The output files are the same of the semi-automatic SGdetector script. The performance of this algorithm is a little lesser than the semi-automatic algorithm, as it can be seen in Table 2, for which accuracy varies from 53 to 80 %.

Table 2: *Precision, recall, and accuracy in percentage (%) for automatic detected salience against perceived salience for the Lobato corpus read by a female (F) and a male (M) speaker at slow (s), normal (n) and fast (f) rates.*

Sp/rate	precision	recall	accuracy
F/n	80	69	74
F/f	61	53	61
M/s	75	57	62
M/f	78	67	79

Its performance can be enhanced by manually changing the input parameters or by using a gradient-descent technique to find the input parameters that achieve the better performance in a limited set of utterances of a particular language, since this script is not language-dependent. Its usefulness depends essentially on the relevance of syllable-sized duration to signal both boundary and prominence. As an additional feature, the SalienceDetector script also indicates the occurence of silent pauses in the corresponding TextGrid interval.

3. Describing the relations between F0 trace and syllable-sized duration trace

The normalised syllable-sized duration trace obtained with the "DrawLines" option of the SGDetector script was conceived in such a way as to give the value of normalised duration along the vowel onsets of the utterance. This feature allows the possibility of plotting the F0 contour of the utterance against the evolution of normalised duration and examining the VV units for which pitch accents and boundary tones coincide with normalised duration peaks. This was presented in [8].

Table 3 presents results of such coincidences in terms of a priori and conditional probabilities for both read paragraphs (two male subjects) and spontaneous speech (a male and a female subject). A priori probabilities are the proportion of pitch-accents, p(F_0), and normalised duration peaks, p(dur), considering the total number of phonological words. Conditional probabilities consider the co-occurrence between a duration peak with a pitch accent over the total number of duration peaks, p(f_0/dur), or the total number of pitch accents, p(dur/F_0). A significant difference, computed from a test of proportions with $\alpha = 0.02$, between a priori and conditional probabilities signals a dependence between pitch accent and duration peak.

The table shows that there is a dependence between duration peak and pitch accent for the female speaker in spontaneous speech, as well as for speaker AC in read speech: for the latter, a pitch accent implies 76 % of chance of a duration peak. For the female speaker both are inter-related. This inter-relation is confirmed when the analysis is restricted to major prosodic boundaries in read speech (utterance boundaries, clause and subject-predicate boundaries): 98 % (speaker AP), and 100 % (AC) of the time, both pitch accent and duration peak occur in the same lexical item, usually in the stressed vowel for pitch accents, and in the stressed or pre-pausal VV unit for duration peaks. Fig. 1 illustrates how both traces can be vizualised. This was possible with the use of the "DrawLines" option of the SGDetector script. In this figure, the labels "sg1" and "sg2" signal the first two stress groups. The first rising contour during "sg1" signals a prominence not accompanied by a duration peak. The two low boundary tones inside the stress groups ending in "ano" and "viver" occur during a VV unit with a duration peak.

Table 3: *A priori probability of pitch accent p(F_0) and duration peak p(dur) in percentage (%) of number of phonological words. Speaker and speaking style are indicated. Stars signal significant differences between a priori and conditional probabilities ($\alpha = 0.02$).*

sp (sp.sty)	p(F_0)	p(F_0/dur)	p(dur)	p(dur/F_0)
F (spont.)	63 *	79 *	49 *	63 *
M (spont.)	73	80	48	56
AC (read)	54	66	56 *	76 *
AP (read)	70	83	65	74

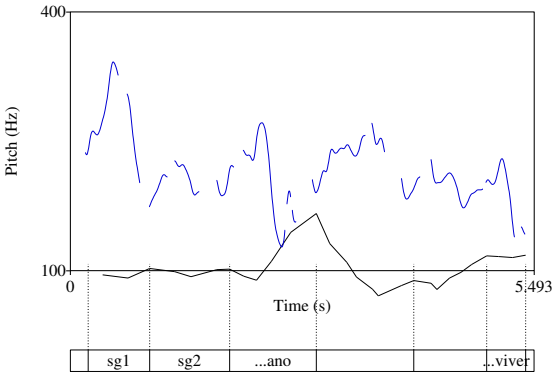

Figure 1: *F0 contour superposed on the VV normalised duration contour of read utterance "Manuel tinha entrado para o mosteiro há quase um ano, mas ainda não se adaptara àquela maneira de viver."*

4. Semi-automatic extraction of global prosodic parameters

The *ProsodyExtractor* script delivers 12 prosodic descriptors for whole utterances or chunks of the same utterance in order to allow research on the link between prosody production and perception. This script has as input parameters the names of the Sound and corresponding TextGrid files. The TextGrid file must be composed of two interval tiers, one with the labelling and segmentation of the VV units (VV tier), and the other with the delimitation of the chunks of the audio file for analysis (Chunk tier). The number of intervals in the Chunk tier can vary from one to any number of units corresponding to any kind of phras-

ing needed for the intended analysis (e.g., syntactic phrases, prosodic constituents like stress groups, content-based chunks, among others). F_0 contour is also computed, thus, it is necessary, as for the Pitch buttons in Praat, to inform minimum and maximum pitch range.

For each chunk in the corresponding interval tier, the algorithm generates (a) 6 duration-related measures computed from the metadata obtained by using the algorithm of the previously described SGdetector script, (b) 5 descriptors obtained from the Pitch object computed by the script and (c) a measure of spectral emphasis as defined by [16]. The six duration-related measures computed in each chunk are: speech rate in VV units per second (sr), maximum of smoothed VV duration $z - score$, mean of smoothed VV duration $z - score$, standard-deviation of smoothed VV duration $z - score$, rate of smoothed VV duration $z - score$ local peaks (pr), and rate of non-salient VV units. The five F_0 descriptors are F_0 median, range, maximum, minimum, as well as F_0 peak rate. For computing the latter measure a smoothing function (with cut-off frequency of 1.5 Hz) followed by a quadratic interpolation function are applied before the F_0 peak rate computation.

The 12 measures generated per chunk can be used both to study the evolution of these prosodic parameters throughout a speech signal, as well as to correlate prosody production and perception. As regards the latter, we used the difference of these values between paired utterances as predictors of the degree of discrepancy between perceived manner of speaking [17]. The experimental design consisted in instructing 10 listeners to evaluate two subsets of 44 audio pairs combining 3 different speakers of BP and two speaking styles, storytelling and reading. The instruction was "Evaluate each pair of excerpts as to how they differ according to the manner of speaking given a scale from 1 (same manner of speaking) to 5 (very different manner of speaking)". After testing more than 50 models of multiple linear regression, results showed that the best model was the one which explained 71 % of the variance of the listeners responses (lr), as given in equation 4 with $p - value$ of at least 0.009 for all coefficients ($F_{3,11} = 12.4$, $p < 0.0008$).

$$lr = -1.5 + 10.4pr + 2.65sr - 10.75pr * sr \qquad (4)$$

This reveals that the significant production parameters that explain the listeners' performance are speech rate in VV units/s and normalised duration peak rate, which can be associated with the syllable succession and salient syllable succession.

5. Summary and availability of the tools

The tools presented here were used to conduct research on speech rhythm analysis and modelling either in a single language or crosslinguistically, on the relation between intonation and rhythm both stricto sensu, as well as on the link between speech rhythm production and perception. They were tested in French, German, Brazilian and European Portuguese, Swedish and English, the latter two less systematically. All scripts are available freely from the author, including a Praat Script not considered in this article but which might be of interest to those who investigate speech expressivity, the ExpressionEvaluator script, which extracts five classes of acoustic parameters and four statistical descriptors, producing 12 acoustic parameters.

All scripts are available freely from the author, are licensed under the terms of the GNU General Public License as published by the Free Software Foundation; version 2 of the License. They were tested in French, German, Brazilian and European Portuguese, Swedish and English, the latter two less systematically.

6. Acknowledgment

The author thanks a research grant from CNPq (301387/2011-7) and Sandra Madureira for revising the manuscript.

7. References

[1] Boersma, P., Weenink, D., "Praat: doing phonetics by computer" [Computer program], Online: http://www.praat.org.

[2] Barbosa, P.A., Eriksson, A. and Åkesson, J., "Cross-linguistic similarities and differences of lexical stress realisation in Swedish and Brazilian Portuguese", in E.L. Asu and P. Lippus [Eds], Nordic prosody. Proceedings from the XIth conference, Tartu 2012 (pp. 97-106). Frankfurt am Main: Peter Lang, 2013.

[3] Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M. and Price, P.J., "Segmental durations in the vicinity of prosodic boundaries", Journal of the Acoustical Society of America, 91(3),:1707–1717, 1992.

[4] Fry, D. B., "Experiments in the perception of stress", Language and Speech, 1,:126–152, 1958.

[5] Dogil, G.,"Phonetic correlates of word stress", in Van der Hulst, [Ed], Word Prosodic System of European Languages, 371-376, De Gruyter, Berlin, 1995.

[6] Sluijter, A. M.C., "Phonetic Correlates of Stress and Accent", Ph.D. Thesis, Holland Institute of Generative Linguistics, Leiden, 1995.

[7] Barbosa, P.A., "Caractérisation et génération automatique de la structuration rythmique du français". PhD thesis, ICP/Institut National Polytechnique de Grenoble, France, 1994.

[8] Barbosa, P. A., "Prominence- and boundary-related acoustic correlations in Brazilian Portuguese read and spontaneous speech", Proc. Speech Prosody 2008, Campinas (pp. 257-260), 2008.

[9] Barbosa, P. A., "Automatic duration-related salience detection in Brazilian Portuguese read and spontaneous speech", Proc. Speech Prosody 2010, Chicago (100067:1-4), 2010. Online: "http://www.speechprosody2010.illinois.edu/papers/100067.pdf."

[10] Barbosa, P.A., "At least two macrorhythmic units are necessary for modeling Brazilian Portuguese duration", Proc. of the 1^{st} ETRW on Speech Production Modeling, Autrans, (pp. 85-88), 1996.

[11] Dogil, G., Braun, G., The PIVOT model of speech parsing, Verlag, Wien, 1988.

[12] Barbosa, P. A., "Incursões em torno do ritmo da fala", Campinas: RG/Fapesp, 2006.

[13] Barbosa, P. A., Viana, M. C. and Trancoso, I., "Cross-variety Rhythm Typology in Portuguese", Proc. of Interspeech 2009 - Speech and Intelligence. Brighton, UK (pp. 1011 1014). London: Causal Productions, 2009.

[14] Cummins, F., Port, R., "Rhythmic constraints on stress timing in English", J. Phon., 26:145–171, 1998.

[15] Tilsen, S., Johnson, K., "Low-frequency Fourier analysis of speech rhythm", JASA Express Letters, 124(2), EL34, 2008.

[16] Traunmüller, H. and Eriksson, A., "The frequency range of the voice fundamental in the speech of male and female adults", Unpublished Manuscript. Online: http://www.ling.su.se/staff/hartmut/aktupub.htm.

[17] Barbosa, P. A. and da Silva, W., "A New Methodology for Comparing Speech Rhythm Structure between Utterances: Beyond Typological Approaches", in H. Caseli et al. [Eds], PROPOR 2012, LNAI 7243 (pp. 329-337). Springer, Heidelberg, 2012.