Automatic duration-related salience detection in Brazilian Portuguese read and spontaneous speech

Plínio A. Barbosa

Speech Prosody Studies Group/Dep. of Linguistics/Inst.Est. Ling., Univ. of Campinas, Brazil
pabarbosa.unicampbr@gmail.com

Abstract

This work presents an automatic prosodic salience detector algorithm which does not require the use of language-specific duration values. It is implemented in two steps: automatic detection of vowel onsets (VO) followed by the detection of normalized VO-to-VO duration peaks. The algorithm's performance is compared to that of a semi-automatic version. Perceived salience is also compared. For both fast and slower read speech, precision and accuracy of perceived word salience are between 61 and 80 %. In a larger corpus of read and storytelling speech, precision is generally higher than 70 %, whereas accuracy is higher than 80 % when the automatic version is compared with the semi-automatic one. The automatic algorithm's performance is found to be similar to that of the prominence detector reported in [12].

Index Terms: prominence detection, speech rhythm, duration

1. Introduction

When asked to associate two distinct prosodic functions such as prominence and phrasing to particular words in running speech, listeners realise both tasks with a certain level of consistency. Simple instructions on how to perform those tasks usually suffice. Listeners can be told to follow instructions such as: (1) point out the words which were highlighted (prominence) by the speaker; (2) point out the words preceding a boundary (phrasing)In general terms, it can be said that the listeners put their attention in a certain level of word salience, that is, they seem to recognise that some words emerge from a background of non-salient words, to put it in gestaltic terms. In languages that have lexical stress, such as Brazilian Portuguese (henceforth BP), the lexically stressed syllable is usually recognised as the most salient part of the word, due to both bottom-up and top-down pieces of information. The ability to distinguishing salient from non-salient syllable-sized units in running speech is an important component of the assessment of a language's rhythm. By recognising different levels of syllable salience, the listener can assess the degree of stress- and syllable-timing as a tendency between two poles, even though the influence of both poles is always present in speech [4]. A method for automatic detection of syllable-sized unit salience is an important first step to the automatic detection of rhythm types.

2. Acoustic cues signalling prominence and boundary in Brazilian Portuguese

It was pointed out by [5] that languages such as English and Swedish favour prosodic prominence marking over prosodic phrasing marking, whereas the so-called syllable-timed languages systematically signal both prosodic functions, *not often by the use of pitch accents* (our emphasis). If, on one hand, recent research on BP challenges the view signalled by the emphasised comment, because BP spontaneous speech presents a high frequency of pitch accented words [3], on the other hand, there is strong evidence that the duration of syllables and V-to-V (henceforth VV) units is a crucial parameter for signalling both prominence and phrasing in BP [1, 10]. At strong prosodic boundaries, preboundary stressed syllables are longer and higher in pitch [3] in more than 97 % of the cases. Due to the frequency that duration is used to signal both prominence and boundary, the assessment of these two functions in read and spontaneous speech also reveals that the same words fulfill these two functions. See Tab. 1 for figures.

At minor prosodic boundaries, however, duration-only or f_0-only cues can be used to signal prominence and boundary [3]. Although it is possible to signal salience only by f_0 cues, this case is not frequent in BP. This means that normalised VV duration local peaks alone can appropriately describe salience in BP, irrespective of the prosodic function. Salience is thus understood here as a general term for the functions of prominence and prosodic phrasing.

Even though listeners often associate these two functions, this does not mean that they are encoded in the same way. In BP, lexically stressed syllables are lengthened as a whole to signal emphasis, whereas the VV units are lengthened as a whole to signal a boundary. It was shown earlier that the correlation between the durations of onset consonants and the following vowel nuclei was 63 % outside preboundary position, against -31 % in phrasal preboundary position, at least for BP read isolated sentences [1]. A similar result was found in English [7].

3. Methodology

3.1. Corpora

Two corpora (Lobato's and Belém's) were used to evaluate an algorithm for detecting word salience in both read and spontaneous speech. The Lobato corpus is formed by the recording of the readings at three self-chosen nominal speaking rates (slow, normal, fast) of a 110-word excerpt of a well-known Brazilian children's book by eight São Paulo State BP speakers. For this work one male (aged 35), and one female (aged 20) were selected for analysis. Two significantly distinct rates for each subject were chosen for analysis: slow and fast for the male speaker, normal and fast for the female. Significance was determined by Kruskal-Wallis analysis of VV duration distributions with $\alpha = 0.05$.

The Belém corpus consists of a 1,500-word text on the origin of the Belém pastries. Although the Belḿ corpus comprises BP and European Portuguese subsets, in this article only the former subset was taken into account. Three speakers of BP,

two females and one male aged between 30 and 35 were asked to read the text. Just after their reading (reading style, RE) a second recording session was made in which the subjects were asked to tell in their own words what the text was about (storytelling, ST). The text was originally written in European Portuguese, and adapted by a native speaker to BP. With the exception of the story told by the male speaker (141 words), excerpts of circa 350 words were chosen for analysis in the other five productions (the three readings and the two stories told by the female speakers).

3.2. Perceived prominence and boundary in BP read speech

In order to highlight the importance of duration in signalling both prominence and boundary in BP, perceived and produced salience were compared at the word level for the Lobato corpus. Perceived salience/non-salience was determined by given the two readings of each speaker (slow/normal and fast speaking rates) to be evaluated by two groups of ten listeners. The listeners were lay persons and graduate students in Linguistics in both groups. In the first group, each listener was instructed to listen to the four readings as many times s/he wants, in order to circle all the words in the corresponding written passage 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 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. These three perceived levels were compared to the semi-automatic and fully-automatic versions of the algorithm described in the next two sections.

3.3. Semi-automatic and automatic detection of duration-related salience

Both the semi-automatic (henceforth SA) algorithm and the fully automatic (A) algorithm for detecting acoustic salience are entirely duration-based. Both algorithms detect local peaks of normalised VV durations. The idea underlying the procedure of salience detection came from two older scripts running on Praat, BeatExtractor and SGdetector.

3.3.1. Detecting vowel onsets: the BeatExtractor script

The BeatExtractor script [2] was implemented in Praat [6]. It implements Cummins' Beat Extractor [8] with some modifications, related to the front-end filter bandwidth and type. The script generates a grid containing intervals between consecutive vowel onsets (VOs). 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 H̃z (see step 4b) as the 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 [8, 11]; (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.

The default cut-off frequencies for the front-end filter are 1000 Hz and 2200 Hz for male speakers, and 1200 Hz and 2700 Hz for females. Note that the values used by Cummins, 700 and 1300 Hz (males), were chosen for the detection of p-centres. The band used by the algorithm proposed here allows to detect vowel onsets after BP sonorant consonants, because the frequency region containing the highest levels of energy for sonorants is within the filter rejection band. It also detects front vowels' onsets because the right cut-off frequency allows to include the mean value of F2 for [i], according to a study with 10 male and 10 female speakers of BP [9]. The effect of this choice can be seen in Fig. 1 for the female speaker AG (RE), where the first derivative of the beat wave, the VV segmentation and the spectrogram are given for the excerpt, "(en)trado par'um most**ei**r**o** há qu(ase)" (entered the monastery for almost [a year]). The pass band filter within the range $1200 - 1700$ Hz (right) allows to detect the high F2 onsets of the two last nuclei of the word "mosteiro" seen in the spectrogram (segments s17 and s18's onsets). They are missed when the p-centre-oriented pass band filter is used instead (left). In both cases the onsets were set where the derivative local maxima were higher than 10 % (0.1 in the top panels) of the global maximum.

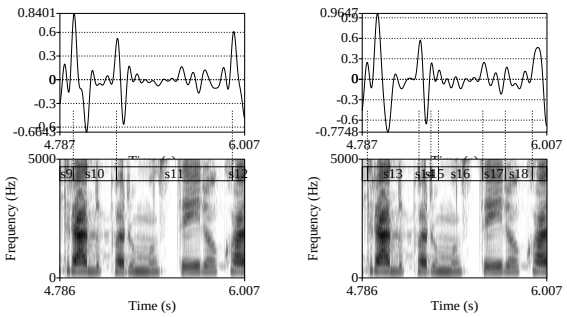

Figure 1: First derivative of the beat wave, VV segmentation and spectrogram for the excerpt "(en)trado par'um most**ei**r**o** há qu(ase)" of speaker AG in the reading condition, for two passage bands: $800 - 1500$ Hz (left), and $1200 - 2700$ Hz (right).

The reason for extending the passage band in step (3) above to 20 Hz, for choice 4a, or 40 Hz, for choice 4b ([8] used 10 Hz) is related to the need of detecting vowel onsets after fast beat wave amplitude changes such as those produced by the tap in intervocalic position (e.g., "xícara", cup). If criterion 4a above is chosen for vowel onset detection, the default threshold value is 0.15. This constraint allows the algorithm to ignore steep risings associated with very small amplitudes. If criterion 4b is used instead, the default threshold value is 0.12.

It is important to signal that, as regards glides, the procedure is so that offglides are included in the VV unit containing the vowel leftwards, whereas onglides are included in the VV unit containing the vowel rightwards. Silent pauses form a VV interval with the preceding sound stretch.

In the SA algorithm, all VOs were obtained automatically using the BeatExtractor script, then manually corrected by the author (up to 20 % of boundaries' displacements and inclusions/exclusions). In the A algorithm, on the other hand, the automatically detected vowel onsets were uncorrected. The vowel

onsets of both corpora were entirely determined in both ways.

3.3.2. Detecting salient VV durations: the SGdetector script

Also implemented in Praat, a second script, SGdetector, detects local peaks of normalised and smoothed VV durations from a grid containing VV intervals. In the SA algorithm the detection is carried out by serially applying two techniques for normalising the VV durations: a $z - score$ transform ($z = \frac{dur - \sum_i \mu_i}{\sqrt{\sum_i var_i}}$, 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 [2, p. 489]), followed by a 5-point moving average filtering ($z^i_{smoothed} = \frac{5.z^i + 3.z^{i-1} + 3.z^{i+1} + 1.z^{i-2} + 1.z^{i+2}}{13}$). The labeling of the phoneme-sized segments within each VV interval is a necessary step, and this was manually done for the two corpora. The two-step normalisation aims at minimising the effects of intrinsic duration and number of segments on the VV raw duration.

In the A algorithm, on the other hand, no phoneme labelling is necessary: $z - scores$ are computed by using fixed values for reference mean ($Refmean = 193$ ms) and standard-deviation ($RefSD = 47$ ms) duration: $z = \frac{dur - \sqrt{ratio}.Refmean}{RefSD}$. The reference values were those of a canonical VV, calculated on the basis of durations measured in a reference subject [2, p. 489]. This unit was defined to be the sequence of the most frequent vowel in BP and a plosive. The value of $Refmean$ is the sum of the /a/ mean duration, and the estimation of the duration for the voiceless plosives (the mean of the three BP voiceless plosives average durations for the reference subject). The value of $RefSD$ is the square root of the sum of the variances of the duration for the same segments. To correct the mean value for the cases where there is more than one VV unit inside the automatically defined interval, an estimation of the number of VV units is given by $\sqrt{ratio}$, where $ratio$ is $\frac{dur}{meandur}$ if $thresh._1 < dur < thresh._2$, and 1 otherwise. The thresholds, set to $thresh._1 = meandur + SDdur$, and $thresh._2 = 1.5 \times meandur$, were defined in such a way as to capture the region of the distribution that probably contains more than one VV unit but does not contain a silent pause. The values were determined heuristically. The values $meandur, SDdur$ are the mean and standard-deviation of the VV duration distribution in the analysed sound file. Smoothed $z - scores$ are determined in the same way as before, by using the 5-point moving average filter. The A algorithm was implemented as a single script combining the two scripts just described.

For both the SA and the A algorithms, smoothed $z - score$ maxima are taken as degrees of salience for the word containing the corresponding local VV duration peak. For instance, if the word sequence *palmas acolheu* (claps welcomed), labelled as [**paʊ**mɐs akoʎe**ʊ**] (lexically stressed syllables in bold), has a local duration peak in the post-stressed VV [ɐs], the [ɐs] duration smoothed $z - score$ is taken as the salience degree of the word *palmas*. Even though lexically stressed syllables are potential places for higher values of acoustic cues signalling both prominence and boundary, these local peaks not necessarily coincide with a lexically stressed VV. Two main reasons for that are: (1) prepausal lengthening, which affects post-stressed VVs, making them often longer than lexically stressed VVs, and (2) the presence of silent pauses irrespective of lengthening. In fact, silent pauses after prosodic boundaries are included in the previous sound stretch, and then, VV intervals containing these pauses are usually local maxima. This is desirable, because the words preceding them signal strong prosodic boundaries.

In the following section two accounts of the A algorithm's performance are given. First, its precision, recall and accuracy are compared to the same signal-detection indexes of the SA algorithm, taking as a reference the perceived salient words in the Lobato corpus. Then, its performance is described by the same three indexes considering as references, the salient words detected by the SA algorithm in the Belém corpus. It is shown that the performance of the A algorithm does not justify the need of manual intervention.

4. Results

Tab. 1 shows precision, recall and accuracy in percentage for the SA and the A algorithms for the two speakers (and speaking rates) of the Lobato corpus. For this comparison, the three levels of perceived salience (see section 3.2) were reduced to two: both weakly and strongly salient words were classified as salient, against the non-salient words. The accuracy of the A algorithm, taking as reference the salient words detected by the SA algorithm, is also given (SA vs A). The table also shows the percentage of perceived prominent words perceived as preboundary (B/P), and vice-versa (P/B) in each reading.

Table 1: Precision, recall, and accuracy in percentage for the SA and A algorithms in the Lobato corpus. Proportion of prominent words perceived as preboundary (B/P), and vice-versa (P/B) are also given. For female (F) and male (M) speakers, and the speaking rates slow (s), normal (n) and fast (f). VO detection at 5 % according to criterion 4a. The accuracy of the A algorithm as referred to the SA algorithm is also given (SA vs A).

Sp/rate	precision SA/A	recall SA/A	accuracy SA/A (SA vs A)	B/P	P/B
F/n	90/80	74/69	82/74(80)	64	72
F/f	73/61	57/53	69/61(77)	41	55
M/s	88/75	67/57	73/62(72)	45	86
M/f	61/78	70/67	70/79(76)	57	67

From this table it can be seen that the SA algorithm is slightly closer to perceived salience than the A algorithm, with the exception of accuracy and precision for the fast reading of the male speaker. For the male speaker, the percentages of precision and recall are similar to those in [12] using syllable duration as a cue of prominence (respectively 64.7 and 65.7 %). The authors compared the performance of their algorithm with a dialogue corpus whose most prominent words were labelled by three listeners. It is important to remind that the listeners use additional acoustic cues to take their decisions, which partly explain the algorithms' performances, especially for the female speaker, who often associate f_0 and duration to signal salience.

Given its crucial communicative function, if perceived strong salience is taken for assessing the A algorithm performance in terms of undetected words, the following picture emerges. For the male speaker: at fast speech, 3 missed words (against 1 for the SA algorithm), and at slow speech, 4 missed words (against 3 for the SA algorithm). For the female speaker: at fast speech, 8 missed word (the same for the SA algorithm), and at normal speech, 4 missed words (against 3 for the SA algorithm). It can be seen that there are only slight differences between the two algorithms.

In order to test the A algorithm in more critical conditions, an evaluation of its performance was done in spontaneous speech. The Belém corpus contains instances of storytelling, considered here as a sub-class of spontaneous speech. The crucial step for ensuring an acceptable performance of the A algorithm is that vowel onset detection be appropriate. The technique of detection and respective threshold for each subject and speaking style that reached the best performances in vowel onset detection are given in Tab. 2 for the Belém corpus. The best performances were chosen by simply checking the positions of vowel onsets according to the vicinity of F2 onset in vowels given by the spectrogram.

Tab. 2 shows precision, recall, accuracy in percentage for the A algorithm as referred to the salient words detected by the SA algorithm in the Belém corpus. The table also gives the number (out of the total) of strongly salient words missed by the A algorithm, but detected by the SA algorithm. Strong salience was determined by using a k-means clustering technique to split the distribution of all smoothed $z - score$ maxima determined by the SA algorithm into two clusters. All words containing VV units whose smoothed $z - score$ maxima are included in the cluster with the highest mean were considered strongly salient.

Table 2: Precision, recall and accuracy in percentage for the A algorithm in the Belém corpus referred to the salient words given by the SA algorithm. Speaker (LL, AG, and FA), sex (F/M), speaking style (RE/ST)), and technique (**D**erivative/**A**mplitude), with corresponding threshold (th), for detecting vowel onsets are given. Number of missed items is also given (msed stB) among the strongly salient words detected by the SA algorithm.

spk./spk. st.	tech. (th.)	prec.	rec.	accur.	msed stB
LLF/RE	D (0.12)	97	63	91	$4/34$
LLF/ST	D (0.08)	64	72	85	$0/13$
AGF/RE	D (0.10)	76	68	87	$0/26$
AGF/ST	D (0.08)	74	74	86	$3/24$
FAM/RE	A (0.15)	76	68	87	$1/17$
FAM/ST	D (0.08)	77	51	82	$5/13$

Observe that the percentage of accuracy and precision are very satisfactory, considering the results of Tab. 1, for which a comparison with perceived salience in the Lobato corpus is given. As regards strong salience detection, the A algorithm performance drops for the male speaker in the ST style. In the five cases, the A algorithm failed to detect a local duration peak, according to the SA algorithm, because it failed to detect one vowel onset, delaying the detected duration peak to the next, postpausal VV unit. In the RE style of this speaker and for the other speakers, the A algorithm detected a salient word just before or after the word detected by the SA algorithm, usually in cases of hesitation involving silent pauses and vowel elongation. Another important aspect of the comparison concerns the position of the VV unit detected by both algorithms. There is coincidence of the VV unit considered as salient for between 67 and 79 % among all words detected in the Belém corpus. For those cases, in which an algorithm signals the lexically stress VV unit as salient, the other signals an unstressed VV unit immediatelly before or after the one detected by the other algorithm. These differences arise from distinct ways of peak detection and normalisation, as presented in section 3.3.2.

5. Conclusion

As just pointed out, the A algorithm has a performance comparable to that of [12] for detecting word salience. This is done at the level of syllable-sized units using only the detection of local peaks of VV normalised duration. The algorithm works similarly well for read and for spontaneous speech. Since the algorithm does not require the use of language-specific duration values, it can be tested in other languages. It can also be used as a front-end to automatic rhythm type detection, if a procedure relating number of syllable-sized units and stress group duration is chosen to assess rhythm in speech [4].

6. Acknowledgments

A grant from CNPq (300371/2008-0). The Belém corpus was recorded in the context of the FCT project PTDC/PLP/72404/2006, for which INESC-ID Lisboa had support from the POSI Program of the "Quadro Comunitário de Apoio III". M. Céu Viana, I. Trancoso, and S. Madureira for helpful discussion and suggestions. A. C. Constantini, L. Lucente, S. Merlo, and W. Silva for help with the perception test.

7. References

[1] 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, 85-88, 1996.

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

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

[4] Barbosa, P. A., "Measuring Speech Rhythm Variation in a Model-Based Framework", Proc. Interspeech 2009, Brighton, 1527-1530, 2009.

[5] Beckman, M. E., "Evidence for speech rhythms across languages" in Tohkura, Y. et al. [Eds], Speech perception, Production and linguistic structure, 457–463, IOS Press, 1992.

[6] Boersma, P., Weenink, D., "Praat: doing phonetics by computer" (Version 5.1.08) [Computer program], Online: http://www.praat.org, accessed in 2009.

[7] Campbell, W. N., "Automatic detection of prosodic boundaries in speech", Speech Communication, 13:343-354, 1993.

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

[9] Escudero, P. et al., "A cross-dialect acoustic description of vowels: Brazilian and European Portuguese", J. Acoust. Soc. Am., 126(3):1379-1393, 2009.

[10] Massini, G., A duração no estudo do acento e do ritmo em português., Master's thesis. Univ. of Campinas, 1991.

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

[12] Wang, D., Narayanan, S., "An acoustic measure for word prominence in spontaneous speech", IEEE Trans. ASLP, 15(2):690-701, 2007.