A MULTILINGUAL STUDY ON INTENSITY AS A CUE FOR MARKING PROSODIC BOUNDARIES

Bogdan Ludusan, Emmanuel Dupoux

Laboratoire de Sciences Cognitives et Psycholinguistique, EHESS / ENS / CNRS, France
bogdan.ludusan@ens.fr, emmanuel.dupoux@gmail.com

ABSTRACT

Speech intensity is one of the main prosodic cues, playing a role in most of the suprasegmental phenomena. Despite this, its contribution to the signalling of prosodic hierarchy is still relatively understudied, compared to the other cues, like duration or fundamental frequency. We present here an investigation on the role of intensity in prosodic boundary detection in four different languages, by testing several intensity measures. The statistical analysis performed showed significant correlates of prosodic boundaries, for most intensity measures employed and in all languages. Our findings were further validated with a classification experiment in which the boundary/non-boundary distinction was learned in unsupervised manner, using only intensity cues. It showed that intensity range measures outperform absolute intensity measures, with the total intensity range being consistently the best feature.

Keywords: prosodic boundaries, intensity, unsupervised learning.

1. INTRODUCTION

The prosodic hierarchy is often marked by multiple prosodic cues and, among them, duration, fundamental frequency and intensity are considered to be the most important ones. The first two, duration (including the duration of silent pauses) and fundamental frequency have been extensively studied and their role in signalling the prosodic structure are well understood [17, 12, 16].

As for the role of intensity, only few studies investigated its role in marking prosodic boundaries (e.g. [6], [2], [9]). In a study of prosodic boundary detection using acoustic cues [2], the authors also perform a statistical analysis of the same cues on part of their English news corpus. They show that two of their intensity measures, end value and convexity, are significant for predicting the boundary/non-boundary distinction. A study of a corpus of conversational English [9] analysed the boundaries transcribed by a group of naive listeners, in terms of the acoustic cues signalling them. The mean intensity was successful in discriminating only part of the vowels analysed in terms of their position relative to a prosodic boundary. Other studies (e.g. [5]) have performed a more limited analysis, looking only at the intensity of syllables preceding different levels of phrase boundaries, but without analysing non-boundary syllables.

Several cues, among which the mean intensity, were investigated in the vicinity of prosodic boundaries, in a corpus of spontaneous Mandarin Chinese [6]. The authors found that intensity levels were, on average, lower in pre-boundary words as the prosodic boundary level increased, with an inverse effect being observed for post-boundary words. Further evidence on the role intensity plays in prosody organization, in Mandarin, was offered by means of a statistical model [15]. The authors showed that their intensity regression model correlated better with the original data when information about higher prosodic units was added to the model, although it did not perform as well as the other cues used (syllable duration and pause).

For a better understanding of the role of intensity in signalling prosodic structure, we have performed an investigation on four different languages: English, Japanese, Spanish and Catalan. Our investigation looked at different languages and several measures of intensity in order to be able to give a more general account of the issue at hand. For the same reason, we performed an exhaustive analysis of all the syllables in the corpora (not only pre- and post-boundary syllables) and we employed large datasets, with a combined 23 hours and more than 460,000 analysed syllables, across the four languages.

2. MATERIALS

We have chosen four typologically distinct languages for our study: English, Japanese, Spanish and Catalan and we considered as prosodic boundaries all boundaries equivalent to those of intonational and phonological phrases [10]. The two levels were then collapsed into one level, which was used as a 'gold standard' for boundaries in our experiments. We tried to use similar type of materi-

Table 1: Summary of the statistical analysis performed for English and Japanese. For each language we illustrate the mean and standard deviation of the intensity measures used, for the accented (A) and unaccented (N) syllables, as well as the boundary/non-boundary T-test value.($^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$).

Intensity measure	Cond.	English					Japanese				
		Boundary		Non-bound.		T-test	Boundary		Non-bound.		T-test
		mean	stdev	mean	stdev		mean	stdev	mean	stdev	
mean	A	67.95	4.74	68.58	4.83	-4.20(**)	68.37	4.67	68.04	4.98	0.79
	N	65.3	5.58	66.44	4.87	-3.09 (*)	65.01	6.85	65.14	6.42	-0.77
max	A	70.16	4.37	70.17	4.46	-1.47	71.60	4.89	70.03	5.03	6.27(***)
	N	67.64	5.11	67.99	4.67	-0.81	67.94	7.01	67.26	6.30	3.50(**)
min	A	62.58	6.52	64.65	6.48	-6.16(**)	59.89	6.84	62.76	6.05	-6.52(***)
	N	60.38	7.43	63.01	6.00	-4.82(**)	57.37	7.72	60.14	7.18	-9.85(***)
max-mean	A	2.21	1.54	1.59	1.55	7.12(***)	3.23	2.07	1.99	1.28	9.82(***)
	N	2.34	2.00	1.56	1.40	4.72(**)	2.93	1.92	2.12	1.57	10.9(***)
mean-min	A	5.37	3.93	3.93	3.46	5.88(**)	8.47	5.07	5.28	3.43	9.53(***)
	N	4.92	4.14	3.42	2.88	5.51(**)	7.64	5.08	5.01	3.57	15.7(***)
max-min	A	7.58	5.02	5.52	4.71	6.33(**)	11.70	6.82	7.28	4.51	9.93(***)
	N	7.27	5.74	4.98	4.05	5.26(**)	10.57	6.67	7.13	4.81	14.5(***)

als, with the English, Spanish and Catalan data containing news recordings, while for Japanese we had academic speech. Also in terms of speakers, we balanced between male and female speakers. Further details are given in the following sections.

2.1. English

The Boston University radio news corpus [11] was chosen for English. The corpus was partly annotated for prosody using the ToBI standard for American English [13] and we employed in our studies level 3 and level 4 breaks, roughly corresponding to intermediate and intonational phrase boundaries. From the entire corpus we have chosen all the recordings having both segmental annotation and level 3 and level 4 break annotations, which gave us about 3 hours of data. It contained speech from 6 speakers, 3 males and 3 females and a total of 49,419 syllables (out of which 8,516 boundary syllables).

2.2. Japanese

The Japanese data used is an 8 hours subset from the core part of the Corpus of Spontaneous Japanese [7] and includes recordings from 12 females and 14 males. The prosody was annotated using the X-JToBI standard [8] and we have chosen all level 2 and level 3 breaks (equivalent to accentual and intonational phrase boundaries). Similar to the English, the material chosen has both prosodic and segmental information. In terms of number of syllables analyzed, we had 33,932 and 126,891 boundary and non-boundary syllables, respectively.

2.3. Spanish

For Spanish, we employed the news part of the GLISSANDO corpus [3]. The corpus was aligned at the segmental level and it also has annotations for minor and major prosodic boundaries. The Spanish subset used consists of 8 speakers (4 males, 4 females) and a total of just over 6 hours of recordings. From a total of 131,015 syllables, 18,368 are found at boundary positions.

2.4. Catalan

As in the previous section, the Catalan data also belongs to the news subset of the GLISSANDO corpus. It is similar to the Spanish data, containing 6 hours of recordings from 8 speakers (4 males, 4 females). It has a total of 21,157 boundary syllables and 98,594 non-boundary syllables.

3. METHODS

The intensity of all the recordings was first computed using Praat [1]. Then, we extracted, for each syllable nucleus, several acoustic measures related to the intensity. The measures contained the following: the average (*mean*), minimum (*min*) and maximum (*max*) intensity over the nucleus, respectively. We have also extracted several range measures (computed as differences in decibel scale), the idea being that such measures should be less affected than absolute measures by variations due to speaking style or recording conditions. These were, respectively, the difference between maximum and av-

Table 2: Summary of the statistical analysis performed for Spanish and Catalan. For each language we illustrate the mean and standard deviation of the intensity measures used, for the accented (A) and unaccented (N) syllables, as well as the boundary/non-boundary T-test value.($^{*}p < .05$; $^{**}p < .01$; $^{***}p < .001$).

Intensity measure	Cond.	Spanish					Catalan				
		Boundary		Non-bound.		T-test	Boundary		Non-bound.		T-test
		mean	stdev	mean	stdev		mean	stdev	mean	stdev	
mean	A	71.70	4.36	72.05	3.70	-0.64	71.16	5.18	73.10	3.48	-4.90(**)
	N	67.81	6.45	70.60	4.08	-3.85(**)	67.40	7.34	71.38	3.88	-7.10(***)
max	A	74.17	3.81	73.71	3.57	1.25	74.06	3.81	74.69	3.27	-3.30(*)
	N	71.20	5.52	72.33	3.80	-2.00	71.29	5.27	72.99	3.63	-7.76(***)
min	A	65.95	7.63	68.15	5.48	-4.21(**)	64.64	9.96	69.13	5.52	-6.81(***)
	N	59.93	10.14	66.58	6.18	-9.60(***)	59.67	12.62	67.58	6.05	-11.6(***)
max-mean	A	2.47	2.18	1.66	1.36	6.27(***)	2.89	3.35	1.59	1.45	4.99(**)
	N	3.39	2.96	1.74	1.56	8.72(***)	3.89	4.47	1.61	1.56	5.80(***)
mean-min	A	5.75	5.12	3.90	3.46	8.76(***)	6.52	6.34	3.97	3.53	8.20(***)
	N	7.88	5.80	4.02	3.62	17.0(***)	7.73	6.92	3.79	3.65	20.2(***)
max-min	A	8.22	6.97	5.57	4.62	8.76(***)	9.42	9.26	5.56	4.77	6.93(***)
	N	11.27	8.34	5.75	5.02	18.5(***)	11.62	10.87	5.40	5.05	12.3(***)

erage intensity (*max* − *mean*), average and minimum intensity (*mean* − *min*), and maximum and minimum intensity (*max* − *min*).

Since intensity may be influenced by stress in three of the languages used in this study, and its role in marking Japanese pitch accent is not generally agreed upon [14], we have decided to perform separate analyses for stressed (English, Spanish and Catalan)/pitch accented (Japanese) syllables (thereafter called condition A) versus unstressed/no pitch accented syllables (thereafter called condition N), for both boundary and non-boundary cases.

In a first step, descriptive statistics on the boundary versus non-boundary syllables were computed and two-tailed paired t-tests were applied to test the significance of the difference between the two cases. The analysis was done separately for each language, for each of the six intensity measures computed and for the two conditions we considered (A and N).

Next, we performed a boundary/non-boundary classification task run on one intensity measure (one syllable) at a time. This was done by first training in unsupervised manner a Gaussian binary classifier, obtained after fitting two Gaussians by means of the expectation maximization (EM) algorithm. The implementation used for the EM algorithm was the one offered by the Weka toolbox [4]. Two types of experiments were performed: for the first one, similar to the statistical analysis, we classified separately the syllables belonging to the A and N conditions. This would be equivalent to applying before the classification a stress/pitch accent detector having an accuracy of 100%. In the second case instead, we combined all the syllables together and classified them. Thus, we will be able to see which is the performance cost for not having any knowledge of stress/pitch accent.

4. RESULTS

4.1. Statistical analysis

The results of the statistical analyses performed are illustrated in Table 1, for English and Japanese, and Table 2, for Spanish and Catalan. It appears that *mean* and *max* intensity are non-reliable indicators of prosodic boundaries, their corresponding differences having been found non-significant in at least two languages. Interestingly, the variation of the *max* intensity goes even in the opposite direction for several conditions (Japanese A and N, Spanish A), with respect to the other absolute measures. The remaining absolute measure, *min*, seem to be highly significant in all the four languages tested. Also the three range measures were found to be highly significant in all the languages, although the variation direction was different from that of the absolute measures (a larger range for boundary compared to non-boundary syllables).

4.2. Unsupervised classification

We present the classification performance of each of the six measures in Table 3. For the evaluation of the results we used the area under the receiver operating characteristic (ROC) curve (AUC). The ROC curve is obtained by varying the threshold for the class probability estimates and plotting

Table 3: Area under the ROC curve results obtained for the classification of boundary/non-boundary syllables, for different languages and intensity measures. An EM-based classifier was employed and three cases were considered: only accented syllables (A), only non-accented syllables (N), or all the syllables (All) were clustered.

Intensity	English			Japanese			Spanish			Catalan		
measure	A	N	All	A	N	All	A	N	All	A	N	All
mean	.531	.565	.543	.520	.506	.510	.514	.615	.593	.596	.643	.589
max	.513	.528	.523	.566	.539	.525	.512	.545	.531	.530	.564	.522
min	.605	.610	.599	.624	.607	.620	.566	.689	.660	**.626**	**.672**	.630
max-mean	.616	**.634**	**.628**	.698	.641	.646	**.599**	.671	.654	.623	.652	.633
mean-min	.615	.613	.609	.689	.655	.654	.589	**.696**	.670	.608	.659	.634
max-min	**.627**	.629	.625	**.700**	**.657**	**.658**	.598	.695	**.672**	.618	.659	**.636**

for each value the resulting true positive rate versus false positive rate. The AUC can be interpreted as the probability of making a correct choice in a forced choice task where one is given a random pair of tokens, one instantiating a boundary, and the other a non-boundary. Since the chance level for the AUC is 0.5, the measure is especially useful when comparing databases of different sizes and distributions of boundaries and non-boundaries, as in our case.

The AUC results show that both *mean* and *max* are close to chance level, for at least two languages each, while the other measures perform better. The range measures perform better, with the best performance obtained (0.7) for the Japanese, $max - min$, in the A condition. The results obtained when no knowledge about stress/pitch accent is available seem to be similar to the average of the performance between the A and N conditions.

5. CONCLUSIONS

We have performed here an investigation into the role of intensity in marking prosodic boundaries in four languages. The main finding of this paper is that intensity correlates with the presence of boundaries across the analysed languages. Furthermore, the unsupervised classification experiment showed that intensity can be used as a cue for unsupervised boundary detection with a modest, but better than chance, performance. Averaging across the four languages, intensity range measures outperform the absolute measures, with the total range ($max - min$) being the best feature for both accented, unaccented, and pooled syllables. Since, in the latter case, we do not need to take into account the stress value of the syllable, in order to exploit the intensity range cues, these measures seem to be particularly suitable to be used in a bottom-up approach.

There are several directions which we can take to build upon the current study. First, we would like to explore whether there are differences between the different levels of phrase boundaries in terms of intensity, as previous studies have shown mixed results for this [6, 9]. At the same time, we envisage using new measures of intensity, spanning neighbouring syllables, in order to capture some local context. This appears to be important in light of the findings by Liu and Li [6] that pre-boundary syllables tend to have decreasing intensity with the increase in boundary strength, and the post-boundary syllables exhibiting the inverse trend.

Finally, in this work we investigated the intensity as a cue of prosodic boundaries in isolation. While this is a good initial step, allowing us to discover its potential in marking prosodic boundaries, such boundaries are usually marked by more than one cue. We would like to expand out study to take into account also the interaction between intensity and other correlates, like presence of pause or f0 reset. We are interested to see whether the information carried by intensity is complementary to the one given by the other cues, or whether it would become redundant when in combination.

ACKNOWLEDGEMENTS

The research leading to these results was funded by the European Research Council (ERC-2011-AdG-295810 BOOTPHON). It was also supported by the Agence Nationale pour la Recherche (ANR-10-LABX-0087 IEC, ANR-10-IDEX-0001-02 PSL*), the Fondation de France, the École des Neurosciences de Paris, and the Région Île-de-France (DIM cerveau et pensée).

6. REFERENCES

[1] Boersma, P., Weenink, D. Praat: doing phonetics by computer [Computer program]. http://www.praat.org. version 5.4.01, retrieved Nov-14.

[2] Choi, J.-Y., Hasegawa-Johnson, M., Cole, J. 2005. Finding intonational boundaries using acoustic cues related to the voice source. *The Journal of the*

Acoustical Society of America 118(4), 2579–2587.

[3] Garrido, J., Escudero, D., Aguilar, L., Cardenoso, V., Rodero, E., de-la Mota, C., Gonzalez, C., Vivaracho, C., Rustullet, S., Larrea, O., Laplaza, Y., Vizcaino, F., Estebas, E., Cabrera, M., Bonafonte, A. 2013. Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan. *Language Resources and Evaluation* 47(4), 945–971.

[4] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1).

[5] Kim, H., Yoon, T.-J., Cole, J., Hasegawa-Johnson, M. 2006. Acoustic differentiation of L- and LL% in switchboard and radio news speech. *Proc. Speech Prosody 2006* Dresden. 214–217.

[6] Liu, Y., Li, A. 2003. Cues of prosodic boundaries in Chinese spontaneous speech. *Proc. ICPhS 2003* Barcelona. 1269–1272.

[7] Maekawa, K. 2003. Corpus of Spontaneous Japanese: Its design and evaluation. *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition* Tokyo.

[8] Maekawa, K., Kikuchi, H., Igarashi, Y., Venditti, J. 2002. X-JToBI: an extended J-ToBI for spontaneous speech. *Proc. Interspeech 2002* Denver. 1545–1548.

[9] Mo, Y. 2008. Duration and intensity as perceptual cues for naive listeners' prominence and boundary perception. *Proc. Speech Prosody 2008* Campinas. 739–742.

[10] Nespor, M., Vogel, I. 2007. *Prosodic phonology* volume 28. Berlin: Walter de Gruyter.

[11] Ostendorf, M., Price, P., Shattuck-Hufnagel, S. 1995. The Boston University radio news corpus. *Linguistic Data Consortium* 1–19.

[12] de Pijper, J., Sanderman, A. 1994. On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *The Journal of the Acoustical Society of America* 96(4), 2037–2047.

[13] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J. 1992. TOBI: a standard for labeling English prosody. *Proc. ICSLP 1992* Banff. 867–870.

[14] Sugiyama, Y. 2012. *The production and perception of Japanese pitch accent*. Newcastle upon Tyne: Cambridge Scholars.

[15] Tseng, C.-Y., Fu, B.-L. 2005. Duration, intensity and pause predictions in relation to prosody organization. *Proc. Interspeech 2005* Lisbon. 1504–1407.

[16] Vaissière, J. 2005. Perception of intonation. In: Pisoni, D., Remez, R., (eds), *The Handbook of speech perception*. Oxford: Blackwell 236–263.

[17] Wightman, C., Shattuck-Hufnagel, S., Ostendorf, M., Price, P. 1992. Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America* 91(3), 1707–1717.