10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Macro-rhythm in English and Spanish: Evidence from Radio Newscaster Speech

Christine Prechtel[1]

[1]University of California, Los Angeles, USA
cprechtel@ucla.edu

Abstract

This study quantified and compared macro-rhythm (MacR) in English and Spanish in radio newscaster speech. MacR is defined as phrase-medial tonal rhythm [1], and its relative strength is determined by three rules: 1) the presence or absence of alternating L and H tones within an IP, 2) the uniformity or similarity of the rise-fall slope shapes, and 3) the frequency of the L/H alternation intervals. The degree of MacR strength can be predicted based on the corresponding phonological criteria: the most common type of phrase-medial tone in a language's tonal inventory (rule 1), the number of phrase-level tones in the inventory (rule 2), and the frequency of f0 rise per Prosodic Word (rule 3). Based on these criteria, Spanish is predicted to have stronger MacR than English. To test this, MacR was quantified in each language by measuring the regularity of distance intervals between tonal targets, the variability of slope shapes, and the number of L/H alternations per Prosodic Word. The results provide some support for the prediction that Spanish has stronger MacR than English in this speech style and they add to previous work comparing MacR strength in English and Spanish in read speech [2, 3].

Index Terms: macro-rhythm, tonal rhythm, speech rhythm

1. Introduction

There is growing evidence that the perception of speech rhythm is at least partially based on the regularity of pitch movement within an utterance, specifically the repetition of rising and falling pitch sequences [e.g. 4, 5, 6, 7]. The periodicity of f0 alterations (tonal rhythm) has been found to play a role in word segmentation in multiple languages [e.g. 4, 8, 9, 10, 11].

Tonal rhythm is determined by language-specific prosodic structure. Within the prosodic hierarchy, f0 marks the boundaries of linguistic units at both lexical and post-lexical levels. The size and number of these units can vary widely across languages, contributing to the perception that some languages sound more rhythmic than others. [12] proposed a model of prosodic typology to capture cross-linguistic differences in phrasing and prominence-marking, which was later revised to also capture differences in the relative strength of tonal rhythm [1].

The additional parameter, macro-rhythm (MacR), is defined as phrase-medial tonal rhythm (i.e. the regularity of high and low f0 alternations) whose domain is equal to or slightly greater than a Prosodic Word (PWord) [1]; that is, a content word plus surrounding unaccented function words and/or clitics. MacR strength can differ across languages under the following three rules: the presence of alternating L and H tones (Figure 1), the uniformity of the rise-fall slope shape (Figure 2), and the frequency of the L/H intervals (Figure 3).

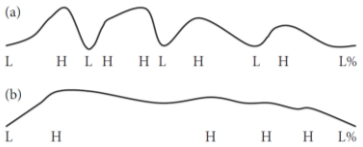

Figure 1: *Schematic pitch contours that differ in the presence of L/H alternations, adapted from [1]. The number of alternations in contour (a) is greater than contour (b), and thus (a) has stronger macro-rhythm.*

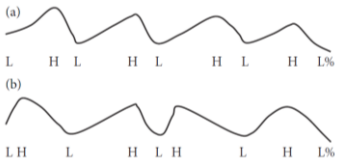

Figure 2: *Schematic pitch contours that differ in the uniformity of the rise-fall slope shape [1]. The rise-fall units in contour (a) are more regularly shaped than contour (b), and thus (a) has stronger macro-rhythm.*

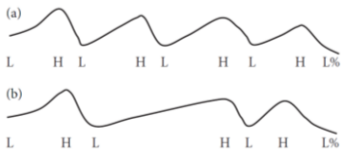

Figure 3: *Schematic pitch contours that differ in the size of the L/H interval or domain [1]. The interval size in contour (a) is more regular (i.e., has more frequent, similarly sized units) than contour (b), and thus (a) has stronger macro-rhythm.*

These three rules correspond to phonological criteria: the most common type of phrase-medial tone in a language's tonal inventory (Figure 1), the number of phrase-level tones in the tonal inventory (Figure 2), and the frequency of f0 rise per word in a phrase (Figure 3) [1]. Languages in which the most common phrase-medial tone is rising (e.g., L+H*) or falling (e.g., H*+L) will have stronger MacR than languages whose most common tone is level (e.g., H*, L*), which corresponds to the first rule. Languages with fewer types of phrase-medial tones will have less variable f0 slope shapes and therefore stronger MacR than languages with more types of tones, corresponding to the second rule. Languages where every word is marked by a tone will have stronger MacR than languages with less or more frequent tone marking per word, corresponding to the third rule. The model can therefore predict the relative strength of MacR in any language based on its prosodic structure.

Previous studies have found phonetic evidence supporting the typological predictions of MacR strength. [13] measured

10.21437/SpeechProsody.2020-138

MacR in English and Italian spontaneous speech and found that Italian has stronger MacR than English in the regularity of f0 alternations, the magnitude of f0 excursions, and the frequency of L/H alternations per IP, supporting [1]'s hypothesis. [2, 3] quantified MacR in English and Spanish in read speech and found that Spanish had stronger MacR than English.

The current pilot study quantifies and compares MacR strength in English and Spanish, following up on [2, 3]. Although both languages have multiple pitch accent types in their respective inventories, the most common pitch accent in English is H* [14, 15] while the most common prenuclear pitch accent in Spanish is L+<H* [16, 17]. Thus, Spanish is expected to have stronger MacR than English, corresponding to Rule 1 (Figure 1). Additionally, English has frequent downstepping, so there are fewer phonological L targets between H targets than in languages with frequent bitonal pitch accents, making it less "peaky" and more step-like. Therefore, Spanish is predicted to have more L/H alternations as well as less slope shape variability than English, corresponding to Rules 1 and 2 (Figures 1 & 2). Finally, the two languages differ in the frequency at which content words (CWords) are pitch accented. With some exceptions, every CWord in Spanish is expected to have a pitch accent [18], while English frequently deaccents some types of CWords such as verbs [19, 20]. Furthermore, [21] found that Spanish places pitch accents on both new and old information, whereas English deaccents old information [22]. Therefore, Spanish is predicted to accent CWords with greater regularity and thus have stronger MacR than English, corresponding to Rule 3 (Figure 3).

The goal of this pilot study is to quantify MacR in English and Spanish in radio newscaster speech to determine if the predicted MacR differences found in [2, 3] are maintained in a different speech style. Since newscaster speech in English tends to have more bitonal pitch accents than non-newscaster speech [23], one might expect reduced differences in MacR strength between the languages.

2. Methods

2.1. Stimuli

The data in this study were taken from two speech corpora. For English, a subset of the Boston University Radio Speech Corpus [24] was used because it was prosodically annotated using ToBI conventions [25]. In addition, this corpus is the basis for the probabilistic model of American English intonation [14, 15] where H* was found to be the most common pitch accent in English, thus contributing to the prediction in [1] that English has weaker MacR than Spanish. The current study analyzed the speech of one female professional radio announcer from Boston (F1a). The total length of the analyzed data subset was 5 minutes and 23 seconds, and the recordings were composed of five news stories that had been divided into 23 total parts.

For Spanish, a subset of the Glissando corpus [26] was similarly chosen because it was prosodically annotated for intermediate phrase and Intonational Phrase-equivalent prosodic units using the SegProso tool [27]. Crucially, the part of the corpus analyzed in this study used a radio newscaster style comparable to the English data. This study analyzed the speech of one female professional reader from Valladolid, Spain (sp_f11r). The total length of the data subset was 5 minutes and 48 seconds, and the recordings were composed of 28 short news story clips.

2.2. Annotation and analysis

Since MacR is defined as phrase-medial (within-IP) tonal rhythm, all IP-final CWords were excluded from analysis to avoid boundary tone interference. Sentences were excluded from analysis if they did not contain a minimum of three consecutive non-final CWords. The recordings were annotated in Praat [28] for f0 turning points and number of peaks per IP.

To annotate f0 turning points, the pitch tracks were schematized using the annotation process described in [29]. The annotation and schematization were done by the author and a research assistant, who annotated separately, and the data were cross-checked between annotators with a 92% agreement rate.

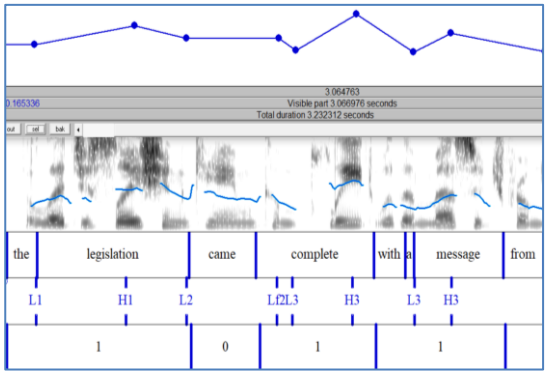

Figure 4: *Example of a schematized English sentence. The labels are numbered in the order in which each f0 point occurs in the utterance. There is no H2 label because of the plateau fall from L2 to L3.*

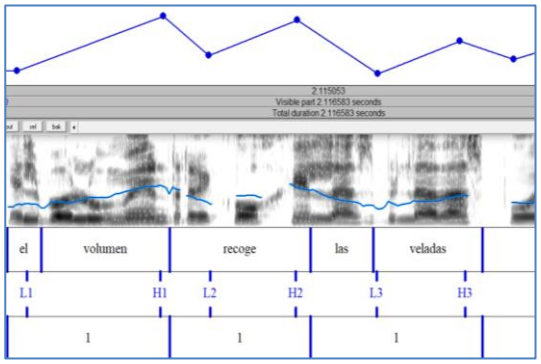

Figure 5: *Example of schematized Spanish sentence. The labels are numbered in the order in which each f0 point occurs in the utterance.*

Figures 4 and 5 show examples of the schematization and labelling for each language. Tier 2 marked f0 turning points with the following labels: L for low (valley), R for rise, H for high (peak), and Hf and Lf for a fall after a high or low f0 plateau, respectively. L marked the lowest point before the next f0 rise; R marked the end of a low plateau and the start of the rise to the next high target; H marked the highest f0 point before falling; Hf marked the end of a high plateau before falling to a low f0 point; and Lf marked the end of a low plateau before falling to even lower point. The number after the tone label indicates the order in which it occurred in the sentence. For example, in Figures 4 and 5, L1 was followed by H1, which was followed by L2, etc.

Tier 3 captured the number of peaks per PWord per sentence. The presence of a peak within the PWord interval was marked as '1'and the absence was marked as '0.' A sentence with a greater number of '1' labels is predicted to have stronger MacR than a sentence with a fewer '1' labels and more '0' labels.

2.3. MacR Measures

The time and height values of the f0 labels were extracted from each IP and used to calculate peak-to-peak distance (ms), valley-to-valley distance (ms), rising slope, and falling slope. Rising slope was calculated by taking the difference between the H target and the preceding L target or the R target if the L was followed by an f0 plateau. Similarly, falling slope was calculated by taking the difference between the L target and the preceding H (or Hf) target. Peak-to-peak distance was calculated by taking the time difference between two successive H points, and valley-to-valley distance was similarly calculated with successive L points.

Two measurements were used to quantify differences in the presence of L/H alternations between the two languages. The first measurement was the peak-to-peak and valley-to-valley distance intervals, which were normalized into z-scores for each utterance. The second measurement was nPVI (Normalized Pairwise Variability Index) [30], which was adopted by [13] to calculate variability in the distance intervals between f0 peaks and valleys. nPVI, shown in (1) calculated pairwise variability in the distribution of f0 targets, where *m* is the number of adjacent tonal intervals in an utterance and *d* is the score of the *kth* measurement.

$$100 \times \left\lfloor \sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| / (m-1) \right\rceil \quad (1)$$

This measurement calculates the difference in duration between each pair of successive intervals, takes the absolute value of the difference, and divides it by the mean duration of the pair to get the normalization factor for speech rate. [13] calculated nPVI values based on the distance between H targets (nPVI-H), between L targets (nPVI-L), and between alternating H and L targets (nPVI-all). The current study adopted the same three nPVI measurements, although the R, Lf, and Hf labels were excluded from the calculations.

To quantify slope shape variability, [1] proposed a metric called MacR Variation Index (MacR_Var), which is the sum of the standard deviations of rising slope (rSD), falling slope (fSD), peak-to-peak distance (pSD), and valley-to-valley distance (vSD) per IP, summarized in (2).

$$\text{rSD} + \text{fSD} + \text{pSD} + \text{vSD} \quad (2)$$

A high MacR_Var value indicates weaker MacR because greater variability suggests irregularly shaped peaks and/or variable distance intervals between peaks. English is predicted to have a higher MacR_Var value and thus greater slope shape variability than Spanish because of its common use of H* and downstepping, as well as its less frequent pitch accent marking of CWords.

To quantify the frequency of L/H alternations (i.e., size of the L/H interval), [1] proposed a metric called the MacR Frequency Index (MacR_Freq). The interval or domain of L/H alternations should roughly correspond to the size of a PWord. MacR_Freq is calculated by dividing the number of f0 peaks per sentence by the number of PWords in the sentence, as summarized in (3).

$$\frac{\text{Number of f0 peaks per sentence}}{\text{Number of PWords per sentence}} \quad (3)$$

A language with stronger MacR will have a MacR_Freq value closer to 1, meaning each PWord will have one f0 peak. Spanish is predicted to have a MacR_Freq value closer to 1 than English because it tends to accent every CWord.

3. Results

Table 1 shows the total number of IPs and PWords for each language. The Spanish IPs tended to be shorter than the English ones, so more Spanish IPs were included to have a comparable number of PWords with English. Table 2 shows the average number of words (both CWords and function words), syllables, and PWords within an IP for each language. Although the number of words and PWords was similar between languages, there was a significant difference in average syllable number, with Spanish having more syllables on average than English ($t(263) = 6.91$, $p < 0.001$), indicating that the Spanish words tended to be longer and contain more syllables than the English words.

Table 1: *Total number of IPs and PWords per language.*

	English	**Spanish**
Number of IPs	122	143
Number of PWords	578	573

Table 2: *Average number of syllables, words, and PWords per IP. Standard deviations are in parentheses.*

	English	**Spanish**
Words per IP	6.2 (2.2)	6.8 (2.4)
Syllables per IP	9.7 (4.2)	13.7 (5.2)
PWords per IP	4.7 (1.6)	4 (1.2)

3.1. nPVI

The nPVI values were calculated for peak-peak distance intervals (nPVI-H), valley-valley intervals (nPVI-L), and L/H intervals (nPVI-LH), and linear mixed effects models were run with speaker ("language group") as the predictor and news story clip (i.e., the story that individual IPs came from) as random intercept. The results show that none of the nPVI values were significantly different between Spanish and English, as summarized in Table 3. The negative coefficient for nPVI-H suggests that Spanish trended toward less inter-peak variability than English.

Table 3: *Results of the linear mixed effects models for nPVI values.*

	nPVI-L	**nPVI-H**	**nPVI-LH**
β	0.04	-0.05	0.001
SE	0.07	0.1	0.06
t-stat.	0.64	-0.57	0.02
p	0.52	0.57	0.99

3.2. MacR_Var

A linear mixed effects model was run with MacR_Var index as the dependent variable, speaker ("language group") as the predictor, and news story clip as random intercept. The results show that the MacR_Var index was marginally significant, with Spanish having slightly more overall variability than English, contrary to the prediction. Neither of the slope measures were significantly different between the languages, suggesting that slope shape was equally variable between Spanish and English. However, both peak-to-peak and valley-to-valley distance were significantly different, with Spanish having shorter distance intervals between H targets and between L targets compared to English. The results of all five models are shown in Table 4.

Table 4: *Linear mixed effects model results for MacR_Var, rising slope (RS), falling slope (FS), peak-to-peak distance (PD), and valley-to-valley distance (VD). Cells with a significant p-value are highlighted in gray. A marginally significant p-value is highlighted in lighter gray.*

	MacR_Var	**RS**	**FS**	**PD**	**VD**
β	0.39	-0.01	-0.01	-54.80	-50.88
SE	0.2	0.01	0.01	22.36	20.12
t-stat.	1.98	-0.59	-0.51	-2.45	-2.53
p	0.056	0.60	0.62	0.02	0.02

3.3. MacR_Freq

To compare MacR_Freq values, a generalized linear model was run with the number of peaks per PWord as the dependent variable and speaker (language) as the predictor. A generalized linear mixed effects model was also run with news story clips as random intercept, but the model was overfitted. The results show that group was a significant predictor ($\beta = 0.37$, SE = 0.07, z = 5.11, $p < 0.01$), indicating that the Spanish sentences had higher MacR_Freq values and therefore greater regularity of peaks per PWord than the English sentences. Figure 6 shows the difference in distribution of MacR_Freq values between the two speakers.

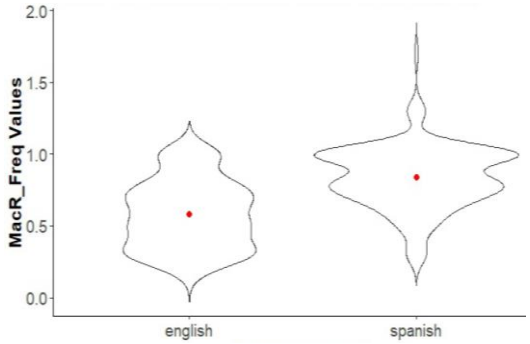

Figure 6: *Distribution of the MacR_Freq values by language group. The means are represented by dots.*

4. Discussion

Overall, the results of this pilot provide preliminary, though mixed, support for the prediction that Spanish has stronger MacR than English in radio newscaster speech. Although none of the nPVI measures were significant, the peak-to-peak and valley-to-valley distance intervals did differ between the two languages, suggesting shorter distance intervals and therefore more L/H alternations in Spanish than English. The lack of significance for the nPVI metric could partially result from including only H and L points in the calculations, which left out information about the duration of f0 plateaus, as the beginning of a plateau was marked with L or H while end was marked with R, Lf, or Hf. A planned follow-up will include this information in nPVI calculations.

The metric for quantifying slope shape variability, MacR_Var index, captured marginal differences between Spanish and English, but in the opposite direction of the prediction, suggesting that the Spanish sentences had more overall variability than English. Neither slope measure captured differences in variability, indicating that slope shape did not differ between languages. Taken together with the small effect size, this could indicate that the English data have a greater number of bitonal pitch accents than other speech styles, resulting in smaller differences between the languages. With more data, slope shape variability may be equal between languages, suggesting that the similarity of slope shape is less important to relative rhythmicity than the number of L/H alternations and the size of the L/H interval.

The metric for quantifying the size and frequency of L/H intervals, MacR_Freq index, captured significant differences in the number of peaks per PWord between the languages, with Spanish having more peaks per PWord than English. This indicates that while English radio newscaster speech may have more L/H alternations due to the greater frequency of bitonal pitch accents in this speech style, it does not mark CWords with the same regularity as Spanish, and therefore has weaker MacR.

There were a few potential confounding factors in this study. Only one speaker per language was analyzed, so there is likely an effect of speaker-specific variation. In addition, only a small subset of each corpus was analyzed, so future work should include more speakers and IPs from each corpus, or potentially other radio news corpora in the target languages.

This study provides some phonetic evidence for MacR differences between languages as predicted by their respective prosodic structure. In addition, these results add to the growing body of work of MacR quantification as a promising approach to the study of speech rhythm. Since tonal rhythm is an important cue for word segmentation [e.g. 4, 8, 9, 10, 11] and marking prominence within a phrase [1], one would expect it to be a perceptually salient correlate of rhythmicity. These results do not directly address the perceptibility of MacR across languages, so the next step is to investigate whether listeners perceive these acoustic differences in tonal rhythm strength. Future work should also investigate the role of linguistic experience on the perception of tonal rhythm. L1 has been found to influence other aspects of rhythmic perception, grouping, and segmentation [e.g. 31, 32, 33], which suggests that language experience also influences how listeners weight acoustic cues associated with tonal rhythm, similarly to other speech rhythm cues [6].

5. Acknowledgements

The Glissando Corpus [26] is licensed under the Creative Commons Attribution-ShareAlike 4.0 International License http://creativecommons.org/licenses/by-sa/4.0/. The data used in this study have been modified from the original.

6. References

[1] S-A. Jun, “Prosodic typology: by prominence type, word prosody, and macro-rhythm,” in S-A. Jun (ed), *Prosodic Typology II*. Oxford University Press, 2014, pp. 520-539.

[2] C. Prechtel, “Quantifying Macro-rhythm in English and Spanish,” in *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia*, pp. 2896-2900, 2019.

[3] C. Prechtel, “Quantifying Macro-rhythm in English and Spanish: A Comparison of Tonal Rhythm Strength,” MA Thesis, University of California, Los Angeles, 2019.

[4] O. Niebuhr, "F0-based rhythm effects on the perception of local syllable prominence," *Phonetica*, vol. 66, pp. 95–112, 2009.

[5] W. J. Barry, B. Andreeva, and J. Koreman, "Do rhythm measures reflect perceived rhythm?" *Phonetica*, vol. 66, no. 1-2, pp. 78-94, 2009.

[6] R. Cumming, "The language-specific interdependence of tonal and durational cues in perceived rhythmicality," *Phonetica*, vol. 68, pp. 1–25, 2011.

[7] R. Cumming, "Perceptually informed quantification of speech rhythm in pairwise variability indices," *Phonetica*, vol. 68, pp. 256–277, 2011.

[8] L. Dilley and S. Shattuck-Hufnagel, "Effects of repeated intonation patterns on perceived word-level organization," *Proceedings of the 14th International Congress of Phonetic Sciences, San Francisco*, pp. 1487-1490, 1999.

[9] L. Dilley and J. D. McAuley, "Distal prosodic context affects word segmentation and lexical processing," *Journal of Memory and Language*, vol. 59, pp. 294-311, 2008.

[10] P. Welby, "The role of early fundamental frequency rises and elbows in French word segmentation," *Speech Communication*, vol. 49, pp. 28-48, 2007.

[11] S. Kim and T. Cho, "The use of phrase-level prosodic information in lexical segmentation: Evidence from word-spotting experiments in Korean," *Journal of the Acoustical Society of America*, vol. 125, no. 5, pp. 3373-3386, 2009.

[12] S-A. Jun, “Prosodic Typology,” in S-A. Jun (ed), *Prosodic Typology*. Oxford University Press, pp. 430-453, 2005.

[13] L. Polyanskaya, M. G. Busà, and M. Ordin, "Capturing cross-linguistic differences in macro-rhythm: the case of Italian and English," *Language and Speech*, March 2019, https://doi.org/10.1177/0023830919835849.

[14] A. Dainora, "An Empirically Based Probabilistic Model of Intonation in American English," Ph.D. dissertation, University of Chicago, 2001.

[15] A. Dainora, "Modelling Intonation in English," in L. Goldstein, D. H. Whalen, and C. T. Best (eds), *Laboratory Phonology 8*. Berlin: Mouton de Gruyter, pp. 107-132, 2006.

[16] L. Aguilar, C. de la Mota, and P. Prieto, “SP_ToBI training materials,” 2009, http://prosodia.upf.edu/sp_tobi/en/index.php

[17] E. Estebas-Vilaplana and P. Prieto, "Castilian Spanish intonation," in P. Prieto and P. Rosano (eds), *Transcription of Intonation of the Spanish Language*. Munich: Lincom Europa, pp. 17-48, 2010.

[18] J. I. Hualde and P. Prieto, "Intonational variation in Spanish: European and American varieties," in S. Frota and P. Prieto (eds), *Intonation in Romance*. Oxford: Oxford University Press, pp. 350-391, 2015.

[19] S. F. Schmerling, *Aspects of English sentence stress*. Austin: University of Texas Press, 1976.

[20] D. R. Ladd, *Intonational Phonology*. Cambridge: Cambridge University Press, 1996/2008.

[21] A. Cruttenden, "The de-accenting and re-accenting of repeated lexical items," in *Proceedings of the ESCA Workshop on Prosody, Lund, 16-19*, 1993.

[22] J. Katz and E. Selkirk, "Contrastive focus vs. discourse-new: evidence from phonetic prominence in English," *Language*, vol. 87, no. 4, pp. 771-816, 2011.

[23] E. Gasser, B. Ahn, D. J. Napoli, and Z.L. Zhou, "Production, perception, and communicative goals of American newscaster speech," *Language and Society*, vol. 48, no. 2, pp. 233-259, 2019.

[24] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, Boston University Radio Speech Corpus LDC96S36. DVD. Philadelphia: Linguistic Data Consortium, 1996.

[25] M. E. Beckman and G. Ayers, "Guidelines for ToBI labelling ver. 3," *The Ohio State University Research Foundation*, 1997.

[26] J. M. Garrido, D. Escudero, L. Aguilar, V. Cardeñoso, E. Rodero, C. de-la-Mota, C. González, S. Rustullet, O. Larrea, Y. Laplaza, F. Vizcaíno, M. Cabrera, and A. Bonafonte, "Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan," *Language Resources and Evaluation*, vol, 47, no. 4, pp. 945-971, 2013.

[27] J. M. Garrido, "SegProso: A Praat-Based Tool for the Automatic Detection and Annotation of Prosodic Boundaries in Speech Corpora," *Proceedings of TRASP, Aix-en-Provence*, pp. 74-77, 2013.

[28] P. Boersma and D. Weenink, *Praat*, Version 6.0.31, 2018.

[29] I. Mennen, F. Schaeffler, and G. Docherty, "Cross-language differences in fundamental frequency range: a comparison of English and German," Journal of the Acoustical Society of America, vol 131, no. 3, pp. 2249–2260, 2012.

[30] E. Grabe and E. L. Low, "Durational variability in speech and the rhythm class hypothesis," in C. Gussenhoven and N. Warner (eds), *Laboratory Phonology*. Berlin: Mouton de Gruyter, pp. 515-546, 2012.

[31] M. Tyler and A. Cutler, "Cross-language differences in cue use for speech segmentation," Journal of the Acoustical Society of America, vol. 126, pp. 367–376, 2009.

[32] A. Bhatara, N. Boll-Avetisyan, A. Unger, T. Nazzi, and B. Höhle, "Native language affects rhythmic grouping of speech," *Journal of the Acoustical Society of America*, vol. 134, pp. 3828–3843, 2013.

[33] M. Ordin, L. Polyanskaya, I. Laka, and M. Nespor, "Cross-linguistic differences in the use of durational cues for the segmentation of a novel language," *Memory and Cognition*, vol. 45, pp. 863–876, 2017.