UNIVERSITY OF CALIFORNIA

Los Angeles

Quantifying Macro-rhythm in English and Spanish:

A Comparison of Tonal Rhythm Strength

A thesis submitted in partial satisfaction

of the requirements for the degree of Master of Arts

in Linguistics

by

Christine Prechtel

2020

ABSTRACT OF THE THESIS

Quantifying Macro-rhythm in English and Spanish: A Comparison of Tonal Rhythm Strength

by

Christine Prechtel

Master of Arts in Linguistics
University of California, Los Angeles, 2020
Professor Sun-Ah Jun, Chair

This thesis quantified macro-rhythm in English and Spanish in two speech styles. Macro-rhythm is defined as phrase-medial tonal rhythm (Jun 2014), and its strength is determined by the number of f0 alternations between peaks and valleys within a phrase, the uniformity of the rise-fall shape, and the regularity of L/H intervals. The degree of strength can be predicted based on the number of phrase-level tones in a language's tonal inventory, the most common type of phrase-medial tone, and the frequency of f0 rise per Prosodic Word. Based on these criteria, Spanish is predicted to have stronger macro-rhythm than English. Two experiments measured the variability of distance intervals between tonal targets, the variability of the slope shapes, and the number of L/H alternations per Prosodic Word per utterance in read speech (Experiment 1) and newscaster speech (Experiment 2). The results of these measures support the prediction that Spanish has stronger macro-rhythm than English.

The thesis of Christine Prechtel is approved.

Patricia Keating

Megha Sundara

Sun-Ah Jun, Committee Chair

University of California, Los Angeles

2020

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I am very grateful to the many people who have helped make this project possible. First and foremost, I want to thank my amazing committee chair and adviser, Sun-Ah Jun, for her tireless guidance and support. Her constructive feedback helped me overcome many setbacks and challenges I faced throughout this project, and I am very grateful for her patience and continued mentorship. Second, I want to thank my other committee members for their invaluable feedback. Without their input, this paper would look very different. I specifically thank Pat Keating for helping me contextualize this project within the greater speech rhythm literature, and Megha Sundara for her methodological suggestions, especially regarding prosodic annotation.

Beyond my committee, I want to thank my fellow graduate students for their insights, suggestions, and encouragement throughout this project. I am especially grateful to Adam Royer for his assistance on matters both theoretical and procedural, and to Arturo Díaz for his assistance with the Spanish stimuli in the production experiment. Special thanks to my wonderful research assistants, Samantha Gonzalez and Hua-Yu Chang, who spent many hours prosodically annotating and checking the data. I am also grateful to the UCLA statistics consultants who helped me with my data over the course of multiple visits.

Lastly, I want to thank Juan María Garrido for granting me access to a subset of the Glissando Corpus, which 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.

1. INTRODUCTION

Speech rhythm is an extensively studied phenomenon, and yet its exact acoustic and perceptual correlates remain elusive. An ever-growing body of research indicates that the perception of rhythm is complex and operates on multiple dimensions of the speech signal, with several acoustic cues interacting with one another to shape the percept of rhythmicity.

Early work focused strictly on the temporal domain, and speech rhythm was characterized by the duration intervals of linguistic units such as syllables and feet (Pike, 1945; Abercrombie, 1967). Languages were classified either as stress-timed, in which timing was coordinated between stressed syllables, or syllable-timed, in which timing was coordinated between each syllable. Although there is some evidence that infants can discriminate between languages based on this rhythm classification system (e.g., Nazzi, Bertoncini & Mehler, 1998; Nazzi, Jusczyk & Johnson, 2000; Nazzi & Ramus, 2003), and that adult listeners process speech in moras, syllables, or feet depending on the rhythm type of their native language (Cutler, Mehler, Norris & Seguí, 1986; 1992; Otake, Hatano, Cutler & Mehler, 1993; Cutler & Otake, 1994; Murty, Otake & Cutler, 2007), empirical studies have largely failed to support the predictions of isochrony across languages (Bolinger, 1968; Lehiste, 1977; Arvaniti, 2009; Arvaniti, 2012). This shifted the focus toward the timing of phonetic and phonological properties such as consonant clusters, vowel length, and vowel reduction (Dauer, 1983; Dauer, 1987), which led to the proliferation of rhythm metrics measuring the temporal properties of consonantal and vocalic intervals (e.g. Ramus, Nespor & Mehler, 1999; White & Mattys, 2007; Dellwo, 2006; Grabe & Low, 2002). However, rather than supporting the classifications under the rhythm class hypothesis, these metrics have shown substantial disagreement, in part because of different methods of data collection (Arvaniti, 2012), the syllabic properties of the stimuli

(Wiget et al., 2010; Arvaniti, 2012; Prieto et al., 2012), and inter-speaker variability (Wiget et al., 2010; Arvaniti, 2012).

Other approaches have classified speech rhythm in terms of the temporal and acoustic properties of prominence. Lee and Todd (2004) proposed that listeners segment and assign prominence to an auditory representation of the speech signal, and that stress-timed languages such as English generally exhibit greater variability in the auditory prominence of acoustic events than languages such as French. Another line of investigation on the "beat" of the syllable (Allen, 1972; Allen, 1975) and its perceptual prominence (Morton, Marcus & Frankish, 1976; Pompino-Marschall, 1989) found that speech rhythm is influenced by the onset of the amplitude envelope in the speech signal (Howell, 1988; Goswami et al., 2002). Tilsen & Johnson (2008) measured speech rhythm using spectral analysis of the amplitude envelope of filtered speech waveforms, where rhythmicity was defined as periodicity in the envelope. They found that in English, some utterances resembled stress-timed rhythm, some resembled syllable-timed rhythm, and some exhibited rhythm at the phrasal level, i.e. between pitch accents (2008:34). Tilsen & Arvaniti (2013) extended this analysis by using envelope metrics that captured periodicity at multiple timescales corresponding to higher frequency (syllable-timed) and lower frequency or supra-syllabic (stress-timed) periods, as well as their relative strengths. In a cross-linguistic comparison, they found that languages differed in supra-syllabic periodicities. Although English exhibited more low frequency periodicity than other languages (Italian, Greek, German, Korean, and Spanish), which is partially consistent with the stress-timed classification of duration interval studies, the metrics provided weak evidence for the traditional rhythm class distinctions. As the authors point out, supra-syllabic rhythms are always present, but the degree of prevalence is language-specific, so while languages like Korean do not mark prominence (e.g. stress) at the lexical level (Jun, 2005), there is still periodicity at the post-lexical level. What these amplitude studies showed is that phrasal prominence plays an important role in perceived rhythmicity. Perhaps the inconsistent isochrony findings of previous speech rhythm studies resulted from assuming a timing unit that is too narrow to capture speech rhythm within an utterance (Tilsen & Johnson, 2008). Given that languages such as English do not accent every stressed syllable (e.g. Ladd, 1996/2008), the perception of speech rhythm may instead rely on a larger timing unit, i.e. the phrasal level.

In addition to duration and amplitude, the regularity of pitch movement also plays a role in the perception of rhythmicity at the phrasal level. Studies have found that f0 is just as important of a cue to rhythm perception as duration (e.g. Barry, 1981; Andreeva, Barry & Steiner, 2007; Barry, Andreeva & Koreman, 2009). In fact, Kohler (2008) found that f0 was a stronger cue for prominence than syllable duration and overall acoustic energy, and Cumming (2011) found that f0 and duration were interdependent cues of rhythmicity. Other studies have found that repetitions of rising or falling tonal sequences affect the perceived grouping and meter of words (e.g. Thomassen, 1982; Lerdahl & Jackendoff, 1983; Handel, 1993; Dilley & Shattuck-Hufnagel, 1999; Niebuhr, 2009; Cumming, 2011). The periodicity of f0 alterations, or tonal rhythm, also plays a role in word segmentation. This has been found in languages that mark lexical stress such as English (Dilley & Shattuck-Hufnagel, 1999; Dilley & McAuley, 2008) and German (Niebuhr, 2009), as well as non-stress languages such as French (Welby, 2007), Japanese (Warner, Otake & Arai, 2010), and Korean (Kim, 2004; Kim & Cho, 2009).

The strength of tonal rhythm is language-specific and is determined by its prosodic structure. Within the prosodic hierarchy, f0 marks boundaries of linguistic units at various levels (i.e. lexical and post-lexical), and the size and structure of these units can vary widely across languages. These prosodic differences contribute to the perception that some languages sound more rhythmic than others. Jun (2005; 2014) proposed a model of prosodic typology to capture cross-linguistic differences in prosodic structure, and later to capture differences in tonal rhythm (2014). The model compares languages analyzed in the Autosegmental-Metrical (AM) framework of intonational phonology. According to the AM model, intonation marks two major properties: prominence and phrasing (e.g. Beckman, 1996; Shattuck-Hufnagel & Turk, 1996; Ladd, 1996/2008). Intonational tunes are composed of pitch accents, which are prominent pitch targets or movements that mark the head of a word (e.g. a stressed syllable), and boundary tones, which are pitch targets or movements that mark the edge of a prosodic unit. Therefore, Jun includes phrasing and prominence as parameters in the prosodic typology model.

The phrasing parameter is categorized by the prosodic units of a language (Jun 2005; 2014). At the lexical level, these units include moras, syllables, and feet, which contribute to the notion of syllable-timed and stressed-timed rhythm classifications (Pike, 1945; Abercrombie, 1967). At the post-lexical level, these units include the Accentual Phrase (AP), Intermediate Phrase (ip), and Intonational Phrase (IP).

The prominence parameter categorizes how prominence is marked at both the lexical and phrasal levels (Jun, 2005; 2014). At the lexical level, it can be marked by one or a combination of pitch accent, stress, and tone, or not marked at all (e.g. Mongolian and Seoul Korean). At the phrasal level, prominence can be marked by the head of the phrase (Head), such as a nuclear pitch accent, or by a boundary tone at the phrase edge (Edge), or by both (Head/Edge). Languages can be Head-prominent like English and Spanish, Edge-prominent like Seoul Korean, or both like Bengali and Japanese (Jun 2005; 2014). Together, the combination of prominence and phrasing at multiple levels of the prosodic hierarchy determine the f0 alternations within an utterance.

An additional parameter of the prosodic typology was added in Jun (2014) to account for the similarities and differences in tonal rhythm across languages. For example, English and Greek are both Head-prominent languages and have lexical stress, but Greek has more regular phrase-medial f0 alternations than English, and this observation could not be explained by prominence and phrasing parameters alone. To capture this additional prosodic dimension, Jun proposed the macro-rhythm parameter. Macro-rhythm is defined as phrase-medial tonal rhythm, i.e. the regularity of high/low f0 alternations, whose unit is equal to or slightly greater than a Prosodic Word (PWord) (Jun 2014). It is defined by the degree of rhythmic strength in f0, which can differ across languages in the presence or absence of low/high f0 alternations (Figure 1), the uniformity of the rise-fall shape (Figure 2), and the regularity of f0 alternation intervals (Figure 3). Languages with frequent f0 alternations, similar f0 rising and falling slopes, and regular alternation intervals are said to have stronger macro-rhythm than languages with infrequent or absent alternations, variable slopes, and variable intervals.

Figure 1: *Schematic pitch contours that differ in the alternation of f0 (Jun 2014:525). The number of H and L alternations in contour (a) is greater than contour (b), thus showing stronger macro-rhythm.*

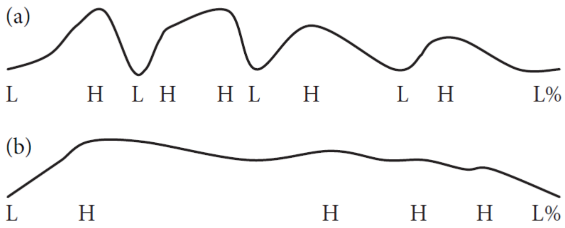


Figure 2: *Schematic pitch contours that differ in the regularity or uniformity of the shape of rise-fall slope (Jun 2014:525). The rise-fall units in contour (a) are more regularly shaped than contour (b), thus showing stronger macro-rhythm.*

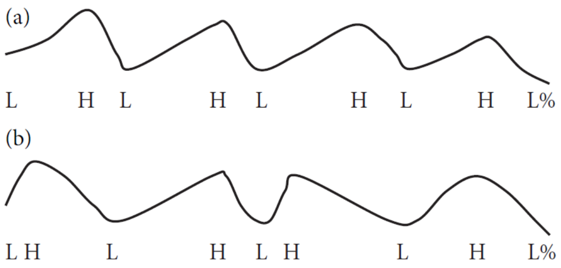
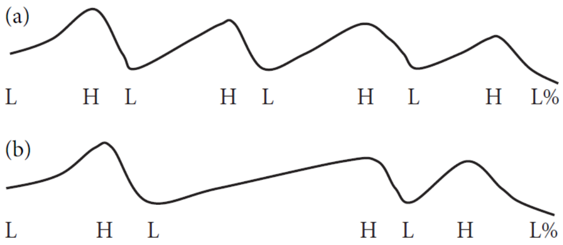


Figure 3: *Schematic pitch contours that differ in the regularity of the L/H interval or domain size (Jun 2014:525). The interval size in contour (a) is more regular (i.e., has more similarly sized units) than contour (b), and thus (a) has stronger macro-rhythm.*



These three rules are converted into the following phonological criteria: the most common type of phrase-medial tone, the number of phrase-level tones in a language's tonal inventory, and the frequency of f0 rise per word in a phrase (Jun 2014). Languages whose most common phrase-medial tone is rising (LH, L+H*) or falling (HL, H*+L) will have stronger macro-rhythm than languages whose most common tone is level (H*, L*), satisfying the rule in Figure 1. Languages with more types of phrase-medial tones such as pitch accents or AP/word tones will have more variable pitch contours and therefore weaker macro-rhythm than languages with fewer tones in the inventory, satisfying the rule in Figure 2. Languages where every word is marked by a phrasal tone will have stronger macro-rhythm than languages with less or more frequent tone marking, satisfying the rule in Figure 3. The model can therefore predict the strength of macro-rhythm in any language based on the prosodic structure as described in the AM framework.

Using these criteria, Jun (2014) divided languages into three groups of relative macro-rhythm strength: strong (e.g. Italian, Spanish, Bengali, Korean), medium (e.g. English, German, Lebanese Arabic, Chickasaw), and weak (e.g. European Portuguese, Mandarin, Cantonese). These typological classes of tonal rhythm exist on a continuum rather than having strict categorical boundaries, with some languages predicted to have stronger or weaker macro-rhythm strength relative to other languages (Jun, 2014:534).

Little previous research has been done to quantify and compare macro-rhythm strength across languages. Burdin et al. (2014) argued that macro-rhythm accounts for differences in the phonetic realization of prominence in focus-marking in English, Guaraní, Moroccan Arabic, and K'iche, but they did not quantify macro-rhythm strength for each language. More recently, however, Polyanskaya, Busà, and Ordin (2019) quantified macro-rhythm in English and Italian, two Head-prominent languages with lexical stress, and found that Italian has stronger macro-rhythm than English in the regularity of f0 alternations over time (durational variability), the magnitude of f0 excursions, and the number of tonal target points per intonational unit (frequency), providing support for Jun's (2014) hypothesis.

The goal of this paper is to phonetically quantify the macro-rhythm parameters and compare the macro-rhythm strength of English and Spanish across two different speech styles. Like Polyanskaya et al. (2019), English and Spanish were chosen for comparison because they are both Head-prominent languages with lexical stress. Although they both have multiple pitch accent types in their respective tonal inventories, the most common pitch accent in English is H* (Dainora, 2001; 2006), while the most common prenuclear pitch accent in Spanish is L+<H*

(Aguilar, de-la-Mota & Prieto, 2009; de-la-Mota, Butragueño & Prieto, 2010; Estebas-Vilaplana, 2010). Additionally, English has frequent downstepping of H*, so there are fewer low points or "sag" 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 f0 alternations and thus have stronger macro-rhythm than English.

In addition, 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 bear a pitch accent (Hualde & Prieto, 2015), while English frequently deaccents some types of CWords such as verbs (Schmerling, 1976; Ladd, 1996/2008). Although deaccenting of verbs also occurs in Spanish (Face, 2003; Ortega-Llebaria & Prieto, 2009), it varies by speech style, with spontaneous speech being more likely to deaccent than lab speech (Rao, 2009). Furthermore, Cruttenden (1993) found that Spanish places pitch accents on both new and old information, in contrast to languages such as English, where old information is deaccented (Katz & Selkirk, 2011). Therefore, Spanish is predicted to accent CWords with greater regularity and thus have stronger macro-rhythm than English.

To summarize, the goal of this paper is to test the following predictions about macro-rhythm strength between Spanish and English, based on the macro-rhythm parameters in Jun (2014):

PREDICTION A: Spanish has more f0 alternations than English

PREDICTION B: Spanish has less overall variability in contour shape than English

PREDICTION C: Spanish has more regular L/H intervals than English

To test these predictions, data from two different speech styles were collected and analyzed in two experiments. The first experiment collected read speech from multiple participants of each language, (Production Study), and the second experiment collected newscaster speech of a single speaker from corpora in each language (Corpus Study). Both experiments found acoustic differences in the regularity, variability, and frequency of f0 alternations between Spanish and English, confirming Jun's (2014) prediction that Spanish has greater macro-rhythm strength than English.

The organization of this paper is as follows: Section 2 describes the methodology, results, and analysis of the production experiment; Section 3 describes the methodology, results, and analysis of the corpus study; Section 4 compares the results of the two studies and discusses the theoretical implications; and Section 5 summarizes and concludes the findings.

2. PRODUCTION STUDY

2.1. Methods

2.1.1. Stimuli

Twenty declarative sentences containing five CWords were created for each language. The number of unstressed syllables between the stressed syllables, the interstress interval (ISI), differed so that sentences would vary in the location of pitch accents within a sentence. The sentences were designed so that the total number of different ISIs was similar between languages. This was to ensure that differences in pitch accent realizations between the two languages were not the result of differences in sentence material, especially the distance in syllable number between any two adjacent pitch accents. The number of unstressed syllables between CWords ranged from 0 to 3. The stimuli were tightly controlled for number of CWords and ISIs in order to minimize confounding variables that could contribute to macro-rhythm differences between English and Spanish. Variables like IP length, speech style, and information structure are predicted to affect macro-rhythm measures, so the starting point for investigating macro-rhythm is to constrain these factors with carefully constructed stimuli. Since all twenty sentences in each language had five CWords, it was predicted to have a maximum of five pitch accents (thus, five f0 peaks) per sentence, when read in the neutral focus condition. Therefore, each speaker could produce a maximum of 100 pitch accents. See Appendix A for the list of sentences and their ISI values for each language.

2.1.2. Participants

Participants were recruited from an undergraduate population, and they received course credit or payment for their participation. They were either monolingual native speakers of American English or Spanish-English bilingual speakers of Mexican Spanish. Eligibility was determined through a language questionnaire before the start of the experiment. The monolingual English group included data from five male and five female speakers. Participants who indicated that they had learned a language other than American English in their childhood were excluded from analysis (11 speakers). Table 1 summarizes the questionnaire data for the English group.

The Spanish-English bilingual group included four male and six female speakers. Six speakers reported that they were balanced bilinguals in speaking proficiency, three speakers reported English dominance, and one speaker reported Spanish dominance. As for reading proficiency, seven participants reported balanced proficiency, including both the participants who reported balanced speaking proficiency as well as the speaker who reported Spanish-dominant speaking proficiency. The three English-dominant speakers also reported English-dominant reading proficiency. Table 2 summarizes the questionnaire data for the Spanish-English group.

Speakers in either group who were creaky throughout the entire utterance or were disfluent readers, i.e. produced multiple IP breaks or hesitations between PWords, were also excluded (7 speakers). A total of 20 speakers were analyzed, 10 speakers for each language.

Table 1: *Language and demographic data for monolingual English speakers.*

Speaker	Age	Gender	Birthplace
02en	19	M	CA
04en	19	F	CA
05en	21	F	CA
10en	20	M	CA
12en	21	M	IL
14en	25	M	CA
18en	21	F	CA
21en	20	F	CA
27en	19	M	CA
28en	21	F	CA

Table 2: *Language and demographic data for bilingual Mexican Spanish speakers. Speaking and reading proficiencies were self-reported.*

Speaker	Age	Gender	Birthplace	Speaking Proficiency	Reading Proficiency
01sp	20	M	CA	Balanced	Balanced
02sp	22	F	Mexico	Balanced	Balanced
03sp	20	M	TX	English-dominant	Balanced
04sp	21	F	CA	English-dominant	English-dominant
08sp	21	F	CA	English-dominant	Balanced
10sp	19	F	Mexico	Balanced	Balanced
11sp	22	F	Mexico	Spanish-dominant	English-dominant
13sp	19	M	TX	Balanced	Balanced
14sp	17	F	CA	Balanced	Balanced
17sp	19	M	Mexico	Balanced	Balanced

2.1.3. Procedure

To reduce the likelihood of disfluencies, participants were first given the list of sentences to read silently to themselves. They were then presented with each sentence one at a time on a computer screen and instructed to read the sentence aloud fluently and without any pauses. Each sentence appeared twice, and two filler sentences were shown at the beginning of the experiment to familiarize the participants with the reading task. All recordings were made in a sound-attenuated room at a sampling rate of 44.1 kHz (32 bit) using SM10A ShureTM microphone and headset.

2.1.4. Annotation

Each recording was segmented into IPs and the first repetition of each sentence was chosen for analysis. If the first sentence was disfluent, then the second repetition was analyzed instead. All IP-final CWords were excluded from analysis because of utterance-final creak and boundary tone interference. Sentences were excluded from analysis if they did not contain a minimum of three consecutive non-disfluent CWords. The recordings were annotated in Praat (Boersma & Weenink, 2019) for words, syllables, f0 turning points, and the number of peaks per IP.

To annotate f0 turning points, the pitch tracks were schematized using the annotation process described in Mennen, Schaeffler, and Docherty (2012). The purpose of schematization was to create a simplified representation of the f0 contour without fluctuations caused by pitch estimation errors such as octave shifts and micro-prosody caused by nearby consonants. First, the sound object was selected in Praat, and a manipulation object was created. Next, all original f0 points were deleted and the sentence received an initial and final point (marking the start and end of the f0 contour of an IP). Points were added for each f0 minimum and maximum within an IP, excluding perturbations due to micro-prosody. Additional points were added wherever the interpolation between existing points differed substantially from the original contour, e.g. when the f0 plateaued before a rise or fall. The data collected from the pilot study had previously been hand-annotated, so the labels were compared to the schematized labels and corrections were made as needed. 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. An example of a schematized pitch track is shown in Figure 4. The definition of each label is given in Table 3. The numbering of the labels is based on the order of each sequential f0 target label and assumes alternations between L and H points. As such, H labels match in number with the preceding L/R label. For example, the first low target would be labeled L1, followed by the first H target, labelled H1.

Figure 4: *Example of a schematized pitch track, shown on top of the original pitch track. Labels are shown on the "f0" tier (H=peak, L= valley, R=rise). The labels are numbered in the order in which each f0 point occurs in the utterance. The "freq" tier marks the number of peaks per PWord per IP, with '1' indicating the presence of a peak.*

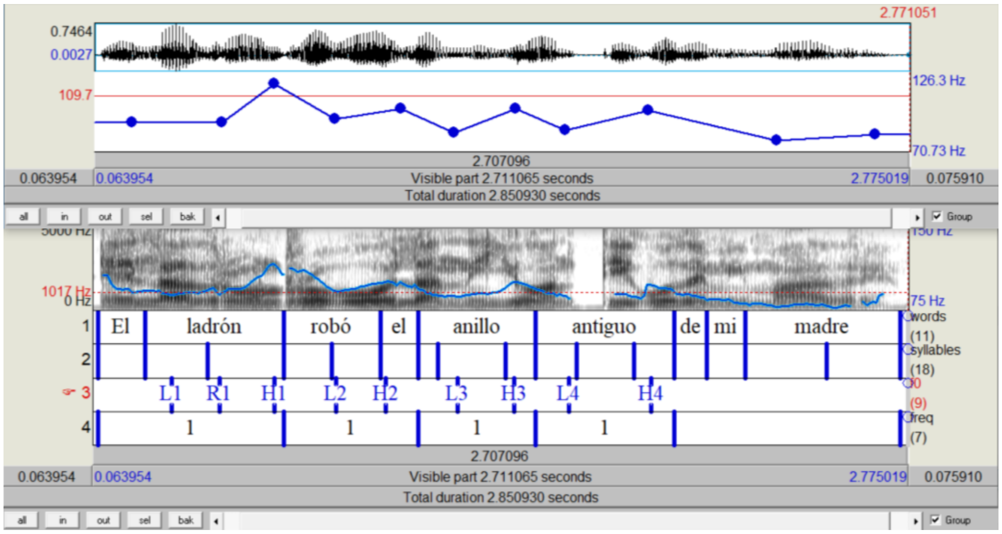


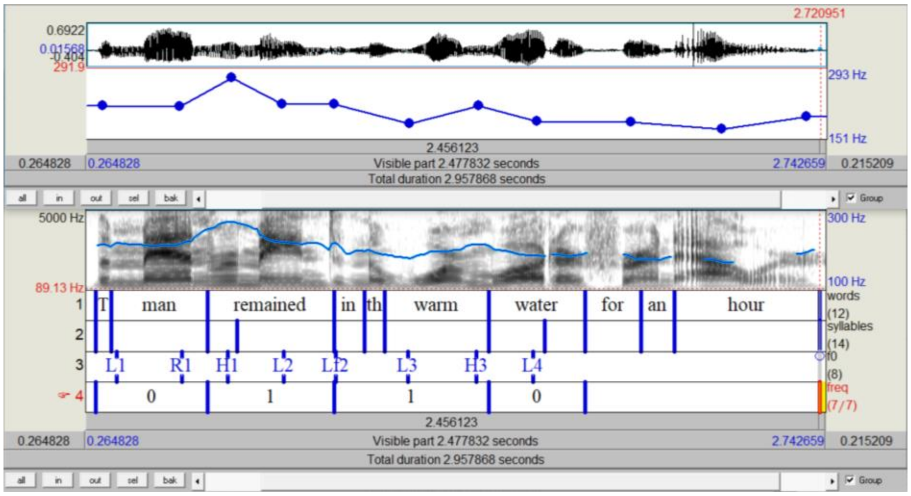
Table 3: *F0 labeling conventions.*

Label	Description
L	Marks the lowest F0 point before the next F0 rise
R	Marks the beginning of the F0 rise after a low plateau
H	Marks the highest F0 point in a rise (peak)
Hf	Marks the beginning of a fall after a high plateau; must follow a H label
Lf	Marks the beginning of a fall after a low plateau; must follow a L label

Polyanskaya et al. (2019) used the MOMEL (Modeling Melody) algorithm (Hirst & Espesser, 1993; Hirst, DiCristo & Espesser, 2000; Hirst, 2007) to automatically interpolate the f0 contour and detect f0 turning points. The reason for using a manual schematization process instead of an automatic algorithm in the current paper was to annotate the data with more phonetic detail and capture f0 events such as plateaus between peaks and valleys.

Figure 5 shows an example where the f0 plateaus after the first peak, marked by L2 and Lf2, before falling to a lower target labelled L3. Because macro-rhythm is defined as phrase-medial tonal rhythm, sentences were only labelled up to the final H peak (or Hf) to avoid influence from the IP-boundary tone. If there was no final H, which was common in the English data, the last L point was labelled before the f0 dropped again. The absence of a peak within the PWord interval was marked with a '0.' Sentences with a greater number of '1' labels are predicted to have stronger macro-rhythm than sentences with fewer number of '1' labels (PREDICTION A).

Figure 5: *Example sentence read by a female English speaker (05en). H=peak, L= valley, R=rise, Lf=fall at the end of a plateau to an even lower f0. The labels are numbered in the order in which they occur in the utterance. There is no H2 label because of the plateau fall from L2 to L3, and the H target number corresponds to the preceding L target. The '0' label in the "freq" tier indicates no peak within the PWord interval.*



2.1.5. Macro-rhythm Measures

A script was used to extract the time and height values of the f0 labels, which were 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.

To phonetically quantify the differences in L/H alternations between the two languages (PREDICTION A), I measured the variability in distance intervals between f0 peaks (H targets) valleys (L targets) using nPVI (Normalized Pairwise Variability Index) (Grabe, 2002). This measure has traditionally been used to quantify speech rhythm in terms of duration by calculating pairwise variability in consonantal and vocalic segment durations in speech. Polyanskaya et al. (2019) used nPVI to calculate variability in the distance intervals between f0 peaks and valleys. nPVI, shown in (1), was used to calculate 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 k^{th} measurement.

$$(1) \qquad nPVI = 100 \times \left[\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| / (m-1) \right]$$

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. Adapting Polyanskaya et al.'s method, I calculated nPVI values based on the distance between H targets (nPVI-H), the distance between L targets (nPVI-

L), and the distance between alternating H and L targets (nPVI-all). Since the original authors' nPVI measurements were based on the labeling conventions with the MOMEL algorithm, which only marked H and L points, I excluded the R, Lf, and Hf labels from the calculations.

To quantify slope shape variability (PREDICTION B), I used the method proposed by Jun (2014) called the Macro-rhythm Variation Index (MacR_Var), which is the sum of the standard deviations of the rising slope (rSD), falling slope (fSD), peak-to-peak distance (pSD), and valley-to-valley distance (vSD), summarized in (2).

(2) $$\text{MacR\_Var} = \text{rSD} + \text{fSD} + \text{pSD} + \text{vSD}$$

A high MacR_Var value indicates weaker macro-rhythm 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 variability, than Spanish (PREDICTION B).

To quantify the regularity of the L/H alternation intervals (PREDICTION C), Jun (2014) proposed a method of counting the frequency of low/high alternations in a phrase, known as the Frequency Index (MacR_Freq). The domain of these alternations should roughly correspond to the size of a Prosodic Word (PWord), i.e., a Cword plus surrounding unaccented function words and/or clitics. MacR_Freq is calculated by dividing the number of f0 peaks per sentence by the number of PWords in the sentence (3). A language with stronger macro-rhythm will have a MacR_Freq value close to 1, meaning each PWord will have one f0 peak. Spanish is predicted to have a MacR_Freq value closer to 1 than English.

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

2.2. Results

The results in Table 4 show the total number of IPs and PWords in each language, as well as the average number of IPs and PWords per speaker. Some of the speakers were more disfluent readers than others and did not have a minimum of three CWords within an IP, so not every speaker contributed the maximum 80 CWords (5 Cwords x 20 sentences minus the 20 excluded IP-final CWords). The contributions of individual speakers are listed in Appendix B. Spanish speakers were more likely to have excluded PWords than English speakers. However, the mean number of IPs per speaker was comparable, as well as the mean number of PWords per language.

Table 4: *Total and mean number of IPs and PWords analyzed for each language. Standard deviations are included in parentheses.*

	English	Spanish
Total number of IPs	193	180
Total number of PWords	790	721
Mean number of IPs included per speaker	19.3 (1.6)	18 (1.9)
Mean number of PWords per language	79 (6.4)	72.1 (8.9)

Table 5 compares the average number of words, syllables, and PWords within an IP for each language. On average, English had more content and function words than Spanish ($t(372) = 11.6$, $p < 0.01$), but both languages had similar numbers of syllables and PWords. This indicates that Spanish CWords tended to have more syllables than English ones, and that the English sentences tended to have more unstressed monosyllabic words (e.g. function words like ‘the’).

Table 5: *Average syllable count, words, and PWords per IP for each language. Standard deviations are included in parentheses.*

	English	Spanish
Mean number of words per IP	7.6 (1.6)	6.0 (1.0)
Mean number of syllables per IP	11.6 (1.5)	11.7 (1.4)
Mean number of Pwords per IP	4.1 (3.6)	4 (0.3)

2.2.1. nPVI

Though individual speakers across languages varied on the number of peaks per IP, the general trend for speakers in each language is represented by the examples in Figures 6 and 7. The English example (Figure 6) has fewer peaks than Spanish example (Figure 7), which occur regularly and within each PWord.

Figure 6: *Example of a sentence read by a male English speaker (27en). The small dip in f0 after L2 is the result of micro-prosody.*

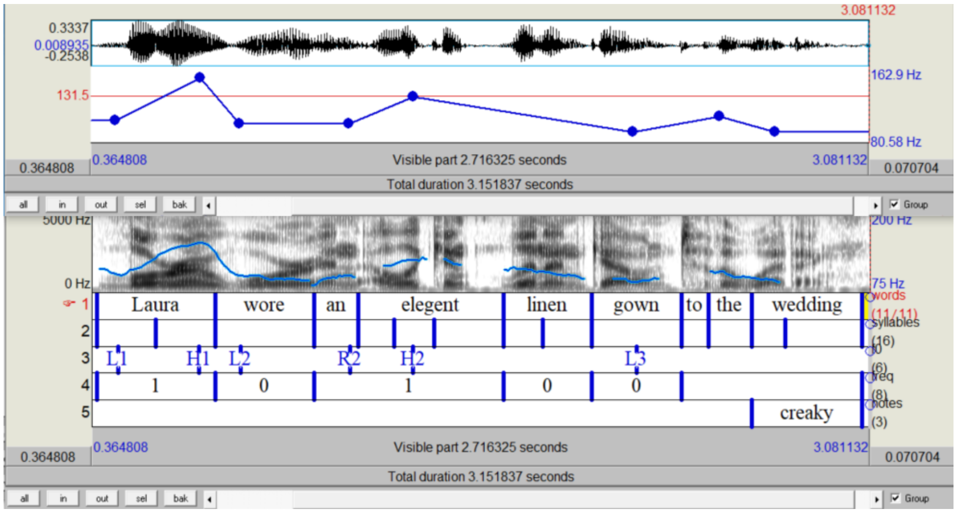
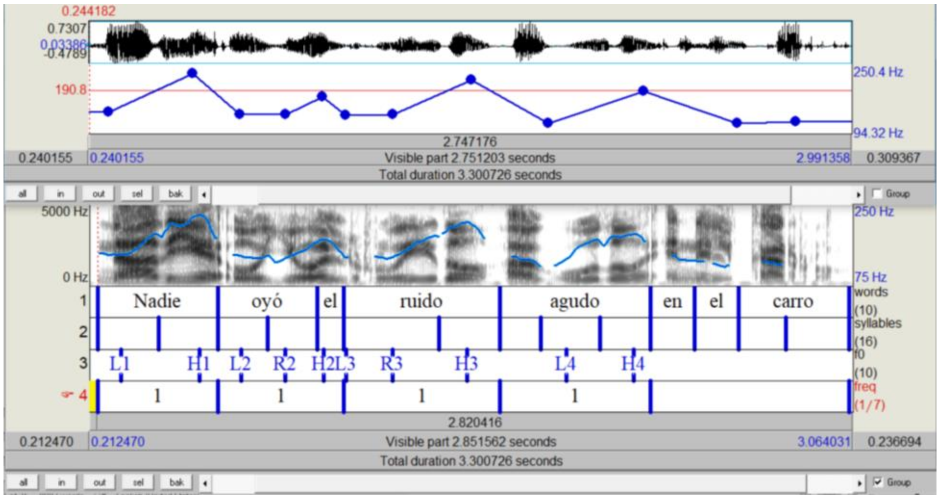


Figure 7: *Example of a sentence read by a male Spanish speaker (03sp).*



The nPVI values were calculated for peak-peak distance intervals (nPVI-H), valley-valley distance intervals (nPVI-L), and L and H intervals (nPVI-LH), per Polyanskaya et al. (2019). Linear mixed effects models were run on each measure with group as the predictor and speaker as random intercept, and the results showed that only nPVI-H was significant (Table 6), providing support for PREDICTION A.

Table 6: *Linear mixed effects model results for nPVI values. The cells with a significant p-value are highlighted in gray.*

	β	*SE*	*t stat.*	*p*
nPVI-L	-0.01	0.05	-0.29	0.77
nPVI-H	-0.14	0.04	-3.29	0.004
nPVI-LH	0.04	0.03	-1.49	0.15

2.2.2. MacR_Var

To calculate the MacR_Var index, the standard deviations were taken for rising slope, falling slope, peak-to-peak distance, and valley-to-valley distance. The raw data were then transformed into z-scores and added together. To determine if English had greater variability than Spanish, a linear mixed effects model was run with MacR_Var index values as the dependent variable, language group as the predictor, and speaker as the random intercept. The results showed a significant difference for group, with Spanish having less overall variability compared to English, supporting PREDICTION B. Linear mixed effects models were also run on each individual measure, and the results for all five measures are summarized in Table 7. There was no significant difference in rising slope or falling slope between languages. Peak-to-peak distance was marginally significant, suggesting that Spanish has slightly shorter distance intervals between peaks than English, while valley-to-valley distance was not significant.

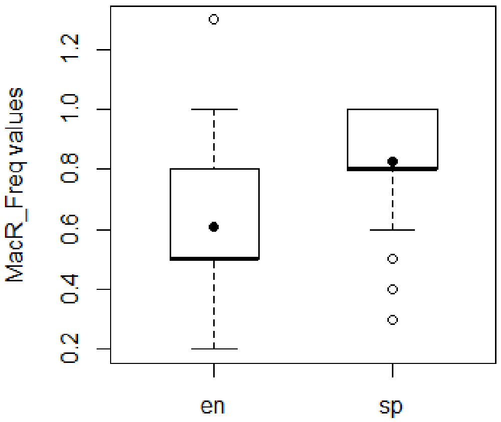
Table 7: *Linear mixed effects model results for MacR_Var and related measures. Rslope = rising slope, Fslope = falling slope, Pdist = peak-to-peak distance, and Vdist = valley-to-valley distance. The cells showing a significant p-value are highlighted in gray. A marginally significant p-value is highlighted in lighter gray.*

	β	*SE*	*t stat.*	*p*
MaR_Var	-1.59	0.52	-3.06	0.006
RSlope	-0.02	0.03	-0.66	0.52
Fslope	-0.05	0.03	-1.79	0.09
Pdist	-107.14	52.19	-2.05	0.055
Vdist	-23.02	32.67	-0.71	0.49

2.2.3. MacR_Freq

To calculate the MacR_Freq index, the number of H targets was divided by the number of PWords (maximum of 4) to get the ratio. A generalized linear mixed effects model was run with language group as the predictor and speaker as a random intercept. It modeled the number of peaks per PWord by creating a ratio using the number of peaks and number of PWords. Sentence was not included as a random effect because there was not enough variability for the model to converge. The results showed that language group was a significant predictor (*β* = 0.31, SE = 0.08, z = 3.95, *p* < 0.01), indicating that Spanish speakers had higher MacR_Freq values than English speakers, consistent with PREDICTION C. This means that Spanish tends to have one peak per PWord and therefore greater regularity of peaks than English, which also provides some support for PREDICTION A. Figure 8 shows the differences in the distribution of MacR_Freq values between English and Spanish.

Figure 8: *Distribution of the MacR_Freq values by language (en= English and sp=Spanish). The means are represented by black dots. The outlier in the "en" column is greater than 1 because an English speaker (14en) produced 5 peaks over 4 PWords in a sentence.*



2.3. Discussion

The results of the nPVI values showed that the only cross-language difference was the variability in time between H targets, with Spanish having less inter-peak variability than English, consistent with the peak-to-peak distance measure. This provides some support for PREDICTION A that Spanish has greater regularity of peak-to-peak distance intervals than English. However, there was no difference for L targets or alternating L/H targets. The results for nPVI-L are consistent with the results of the valley-to-valley measurements. Polyanskaya et al.'s (2019) study found the exact opposite nPVI results for Italian and English: nPVI-L and nPVI-LH were significant while nPVI-H was not significant, with Italian having lower variability and therefore greater regularity of the distribution of f0 targets than English. They did not report individual peak-to-peak or valley-to-valley measurements in their study.

The MacR_Var index was also able to capture differences in overall variation across languages, and the results support PREDICTION B that Spanish has less overall variability in f0 shape than English. Regarding slope measures, neither rising nor falling slope values were significantly different, although Spanish falling slope values tended to be shallower than English, as indicated by the negative coefficient in Table 7. This is surprising, given that Figures 6 and 7 show steeper falling slopes for Spanish than English. The lack of significance, and the trend toward shallower slopes in Spanish could be the result of the labelling conventions. Spanish sentences tended to have shorter f0 plateaus than English, and the labelling may have marked L in the middle of a small plateau, thus making the slope values shallower. As for the distance interval measures, peak-to-peak distance was marginally significant while valley-to-valley distance was not significant, with Spanish having slightly smaller distance intervals between peaks than English. The shorter peak distance intervals may also provide support for PREDICTION C because Spanish is more likely to have one peak per PWord within an IP than English. While it is not necessarily the case that shorter peak-to-peak intervals mean that the distance is more regular, we would expect a correlation between peak frequency and distance, with Spanish having shorter distance intervals between peaks than English. While one might also expect a correlation between peak distance and valley distance (i.e. shorter intervals between peaks would also mean shorter intervals between valleys), the prevalence of L plateaus between two H tones means that peak-to-peak distance is easier to measure and represent than valley-to-valley distance, which is marked by two points (L and R). Therefore, it is unclear whether valley-to-valley distance intervals in Spanish truly have more variability than peak-to-peak distance intervals.

The MacR_Freq index was able to capture differences in f0 alternations between languages. As expected, Spanish had higher MacR_Freq values than English, which is consistent with PREDICTION C. Because English has frequent downstepping of H* pitch accents, there are fewer L/H f0 alternations, which means fewer peaks and weaker macro-rhythm strength compared to languages like Spanish, which have more bitonal pitch accents with clear L and H targets. The greater regularity of peaks in Spanish compared to English also provides indirect support for PREDICTION A.

3. CORPUS STUDY

3.1. Methods

In addition to a production experiment, which analyzed read speech, a corpus experiment was also conducted to analyze a different speech style. The stimuli in the production experiment were tightly controlled for utterance length, ISI, and number of CWords. While this allowed for comparison between languages that minimized confounding variables, it also limited the number of sentences analyzed and the number of tokens per sentence, which does not reflect other speech styles. Indeed, reading prosody differs from other styles such as spontaneous speech (Howell & Kadi-Hanifi, 1991). In more naturalistic speech settings, utterance length can differ considerably, as well as the number of CWords within an IP. The goal of the corpus experiment is to quantify macro-rhythm between English and Spanish radio newscaster speech, which, while scripted, is closer to naturalistic speech than read speech (Ostendorf, Price & Shattuck-Hufnagel, 1996), having greater variation in IP length, ISI, and number of CWords. A secondary goal of this experiment is to determine if the macro-rhythm differences found in read speech of the previous experiment are maintained in a speech style closer to spontaneous speech. Although newscaster speech tends to have more L+H* pitch accents than non-newscaster speech in English (Gasser, Ahn, Napoli & Zhou, 2019), Spanish is still expected to have stronger macro-rhythm than English, although the differences between the languages may be reduced compared to read speech.

In addition, the production study compared monolingual speakers of English group to bilingual speakers of Spanish group, which introduces the potential interaction of language, i.e., English influencing Spanish intonation patterns. The corpus study, in contrast, compared one monolingual American English speaker and one monolingual Castilian Spanish speaker.

3.1.1. Stimuli

A subset of the Boston University Radio Speech Corpus (BU corpus, Ostendorf et al., 1996) was chosen for English for two reasons. First, it was prosodically annotated using ToBI conventions (Beckman & Ayers, 1997), so the IP breaks were already labelled. Second, Dainora (2001; 2006) based her probabilistic model of American English intonation on this corpus, where she found H* to be the most common pitch accent in English. This finding is one of the reasons why Jun (2014) predicted that English is less macro-rhythmic than Spanish. The current study analyzed data from one speaker, F1a, a female professional radio announcer from Boston. The total length of the data subset analyzed in this study was 5 minutes and 23 seconds. The recordings were composed of five news stories that had been divided into multiple parts (23 parts total).

For Spanish, the Glissando corpus (Garrido et al., 2014) was chosen for similar reasons. First, the corpus was annotated for intermediate phrase and Intonational Phrase-equivalent prosodic units using the SegProso tool in Praat (Garrido, 2013). Second, the part of the corpus analyzed in this study used a newscaster style similar to the BU corpus. This study analyzed the speech of one speaker, sp_f11r, a female professional reader from Valladolid, Spain. The total length of the data subset was 5 minutes and 48 seconds. The recordings were composed of 28 short news story clips.

As with the production study, IPs were excluded from analysis if they did not contain a minimum of three consecutive non-disfluent CWords or were not part of a declarative utterance, as the macro-rhythm typology is based on the intonation of declaratives (Jun, 2014).

Additionally, IPs with a focused CWord were also excluded from analysis because it changes the default prosody in both languages.

3.1.2. Annotation and Analysis

The same annotation method from the production study was used for the corpus study. The data were segmented into IPs, schematized, and annotated for f0 turning points and number of peaks. All IP-final CWords were excluded from analysis because of utterance-final creak and possible boundary tone interference. The same quantification measures (nPVI, MacR_Var, and MacR_Freq) were also calculated and compared between the two corpora.

3.2. Results

Table 8 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 9 shows the average number of words, syllables, and PWords within an IP for each language. Although the number of words and PWords was similar between the languages, there was a significant difference in average syllable number between the languages, 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 8: *Total number of IPs and PWords per language group.*

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

Table 9: *Average number of syllables, words, and PWords per IP. Standard deviations are in parentheses.*

	English	Spanish
Mean number of words per IP	6.2 (2.2)	6.8 (2.4)
Mean number of syllables per IP	9.7 (4.2)	13.7 (5.2)
Mean number of PWords per IP	4.7 (1.6)	4 (1.2)

3.2.1. nPVI

The nPVI values were calculated for peak-peak distance intervals (nPVI-H), valley-valley distance intervals (nPVI-L), and L/H intervals (nPVI-LH) and linear mixed effects models were run with speaker (language) as the predictor and news story file as random intercept. The results showed that none of the nPVI values were significantly different between Spanish and English, as summarized in Table 10.

Table 10: *Results of the linear mixed effects models for nPVI values.*

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

3.2.2. MacR_Var

A linear mixed effects model was run with MacR_Var index as the dependent variable, language group as the predictor, and news story recording clips as the random intercept. The results showed The MacR_Var index was marginally significant, with Spanish having slightly more overall variability than English, contrary to PREDICTION B. Neither of the slope measures were significantly different between the languages, suggesting that slope steepness 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 11.

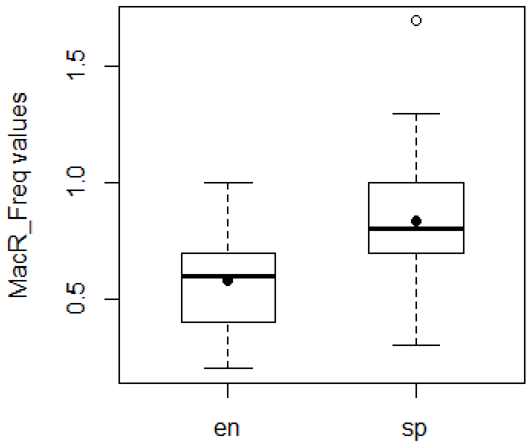
Table 11: *Linear mixed effects model results for MacR_Var, rising slope, falling slope, peak-to-peak distance, and valley-to-valley distance. Cells with a significant p-value are highlighted in gray. A marginally significant p-value is highlighted in lighter gray.*

	β	*SE*	*t stat.*	*p*
MaR_Var	0.39	0.2	1.98	0.056
RSlope	-0.01	0.01	-0.59	0.60
Fslope	-0.01	0.01	-0.51	0.62
Pdist	-54.80	22.36	-2.45	0.02
Vdist	-50.88	20.12	-2.53	0.02

3.2.3. 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 language group as the predictor. A generalized linear mixed effects model was also run with news story file clips as the random intercept, but the model was overfitted. The results show that group was a significant predictor (β = 0.37, SE = 0.07, z = 5.11, *p* < 0.01), indicating that the Spanish sentences had higher MacR_Freq values than the English sentences, supporting PREDICTION C. Since Spanish has greater regularity of peaks than English, this also provides indirect support for PREDICTION A. Figure 9 shows the difference in distribution of MacR_Freq values between the two speakers.

Figure 9: *Distribution of MacR_Freq values by language. Means are represented by black dots.*



3.3. Discussion

Unlike the production study, none of the nPVI measures were significant. Given the results of the distance interval measures and MacR_Freq index, one would expect the nPVI results to show that Spanish has less pairwise variability than English in at least one measure. The lack of significance may be due to the labelling conventions, as the R and Lf/Hf turning points were excluded from the calculations. Whenever there was an f0 plateau, the beginning would be marked with H or L and the end would be marked with R, Lf, or Hf. To calculate pairwise variability, only the L and H turning points were compared, which leaves out information about the duration of low and high plateaus. However, these results still differ from the production data, which used the same labeling conventions and found a significant difference in the pairwise variability for nPVI-H, so this could indicate a speech style difference.

The MacR_Var index was able to capture differences in variability between Spanish and English, although marginally, and in the opposite direction of the prediction. The size of the effect is not necessarily surprising, given the prediction that the greater number of bitonal pitch accents in the English data would result in smaller differences between the languages. Neither slope measure captured differences in variability, suggesting that slope shape did not differ drastically between languages. However, both peak-to-peak and valley-to-valley distance differed between languages, indicating shorter distance intervals in Spanish than English. As with the first experiment, these results do not necessarily indicate greater regularity of distance intervals, but given the greater regularity of peak frequency in Spanish compared to English, this may provide some support for PREDICTION C.

As with the production study, the MacR_Freq index was able to capture language differences in the frequency domain. The number of peaks, and therefore the MacR_Freq index values, were higher in Spanish than English, supporting PREDICTION C and providing indirect support for PREDICTION A.

4. GENERAL DISCUSSION

The results of both experiments provide some support for all three predictions regarding macro-rhythm strength in English and Spanish. In other words, Spanish has stronger macro-rhythm than English in the number of L/H f0 alternations (PREDICTION A), slightly stronger macro-rhythm in the uniformity of slope shapes than English (PREDICTION B), and greater regularity of L/H intervals than English (PREDICTION C). These results add to a growing body of literature supporting cross-linguistic differences in the frequency, variability, and regularity of f0 alternations predicted by the prosodic typology model (Jun, 2014).

Overall, each type of measure was able to quantify an aspect of macro-rhythm strength in the two languages, although some were marginal. Regarding distance interval measures (Pdist and Vdist), peak-to-peak distance was significant in the corpus data, indicating shorter distance intervals in Spanish than in English, but only marginally so in the production data. Similarly, Spanish had significantly shorter valley-to-valley distance intervals than English in the corpus data, but there was no significant difference in the production data. The differing results between the two experiments may be due to the differences in sentence materials. The production data were tightly controlled for the number of CWords and the ISI between prominent syllables, so perhaps this constrained the distance variability differences between the two languages. The corpus data, in contrast, were less constrained by number of CWords and entirely uncontrolled for ISI between prominent syllables. Despite the greater frequency of L+H* pitch accents in the English corpus data, English still showed longer distance intervals for L targets than Spanish, supporting the prediction that macro-rhythm strength is consistent across speech styles.

The nPVI results were surprising, as only the production study showed a significant difference between Spanish and English, and the difference was only the variability between peaks (nPVI-H), with Spanish having less variability than English. This is consistent with Spanish having the smaller MacR_Var index number, the larger MacR_Freq index number (i.e. greater number of peaks), and the marginally smaller peak-to-peak distance intervals compared to English. In contrast, the corpus study found no significant differences between Spanish and English in any of the nPVI measures. This contradicts the results of the peak-to-peak and valley-to-valley distance interval measures, and the MacR_Freq index, which indicate that Spanish has more frequent peaks and shorter inter-peak distance intervals than English. The results could partially be attributed to the labelling conventions; the length of the low plateaus can vary (i.e. the distance between the L label and the R label), especially when ISIs contained multiple unstressed syllables. Since the nPVI-L and nPVI-LH measures only compared pairwise variability using L labels and did not take R labels into account, the results may not reflect the variability of Spanish compared to English.

In the production experiment, the MacR_Var index showed that Spanish had less overall variability than English, supporting PREDICTION B. In the corpus study, however, the difference was marginally significant in the opposite direction. One would expect a smaller effect size given the greater frequency of bitonal pitch accents in English newscaster speech, but it is unclear why Spanish has greater variability than English in newscaster speech. Given the conflicting results of the individual distance and slope measures, the MacR_Var index provides only weak evidence of variability.

None of the slope measures were significant in either study, which is surprising given the visual difference in slope steepness between Figures 6 and 7. The lack of significant differences could partially be the result of the labelling and the placement of the L and R labels. Future work could compare the absolute values of rising and falling slope within each language and then compare these values between languages to see if there are any differences. Spanish is expected to show less variability in slope shape compared to English, although the difference may be reduced in speech styles such as newscaster speech.

The MacR_Freq index was significant in both experiments. As expected, Spanish had higher values overall, reflecting the greater number of phonological low/high alternations than English, and confirming PREDICTION C. This index seems to be the most robust measure for macro-rhythm quantification; the results of a pilot study comparing a subset of the production data found that this was the only measure that captured the differences between the two languages. MacR_Freq also provides indirect support for PREDICTION A because Spanish tends to have a L/H alternation for every PWord, and thus has more alternations than English.

The results of both experiments differed from Polyanskaya et al.'s (2019) results for Italian and English. They found that Italian had more regular distribution (i.e. less variability) of both L targets and f0 turning points (nPVI-LH), but not H targets. They argue that the difference in the distribution of L targets can be explained by the greater frequency of phonological L tones in Italian than in English. Since one of the most common pitch accents in Italian is L+H*, speakers need to plan f0 valleys that are associated with the prominent syllable. The most common pitch accent in English is H*, so the L tones are not phonological, but rather unplanned "sagging" between a sequence of H tones, and thus are more variable than Italian. While the results of the current study may differ for language-specific or speech style reasons, the interpretation that lower L target variability is the result of planning for the phonological L target does not explain the Spanish results. Like Italian, the most common pitch accent in Spanish is bitonal; specifically, the most common prenuclear pitch accent is L+<H*. Based on their analysis, Spanish should behave like Italian and show less variability of the L target. One difference between the most common (prenuclear) bitonal pitch accents in Spanish and Italian is that the peak is delayed in Spanish, meaning that the f0 maximum is realized on a syllable after the prominent one (e.g. Face & Prieto, 2007). Perhaps these differences in Italian and Spanish effect slope shape or even distance intervals. However, it would be surprising if the L targets of the delayed peaks exhibit more variability than non-delayed peaks because the L tone is still associated with the prominent syllable.

A more likely explanation for the differing results is the difference in labelling conventions. The current study annotated f0 movements with greater phonetic detail than Polyanskaya et al. (2019). For sentences where there was an f0 plateau between peaks, the L label marked the beginning of the plateau and the R label marked the beginning of the rise to the next peak. Therefore, the nPVI measure may not have been as useful for capturing the variability of phonological f0 targets, particularly L targets. In contrast, the MOMEL algorithm used by Polyanskaya et al. (2019) calculated and labeled L and H points, and it is unclear how the authors treated f0 plateaus in the data. If the only points included for analysis were L and H alternations, then their slope and distance measures may not faithfully reflect the f0 contours of each language. For example, if the algorithm marked a L tone in the middle of a low plateau, the falling slope of the preceding peak and the rising slope of the following peak will be shallower than if the end of the fall and beginning of the rise were marked separately. If L targets were consistently marked in the middle of a plateau, then the distances between L targets may be more regular and therefore exhibit less variability in distance. The authors emphasize that their analysis was based on phonological f0 targets, which is perhaps why only L and H labels were used. Their results may also be partially attributed to speech style. Their experiment elicited spontaneous speech by having participants read a short story, watch a cartoon of the story, and then recount it themselves. In contrast, the current study analyzed two types of read speech, one of which is closer to spontaneous speech style.

There were a few limitations and potential confounds in both experiments. For the production experiment, monolingual English speakers were compared to Spanish-English bilingual speakers. Although more than half reported that they were balanced bilinguals, there is likely an interaction of language. Indeed, four of the ten Spanish-English bilingual speakers reported English-dominance in either speaking or reading, so this may have affected the variability measures in particular. Even though the bilingual group's Spanish still had stronger macro-rhythm than the English group, future work should compare monolingual English speakers with monolingual Spanish speakers to avoid the confound. The corpus experiment did compare monolingual speakers of Spanish and English, but the comparison was between a single speaker of each language, so the data likely also captured speaker-specific variation. Perhaps the results of the variability and distance measures would differ if more speakers were added. For both experiments, the f0 annotation procedure, while phonetically detailed, made the analysis of distance intervals less straightforward because of the f0 plateaus. In both experiments, only the L and H points were used to calculate distance intervals, but this only tells us where the end of the f0 fall is, which can differ depending on the number of syllables in an ISI. Polyanskaya et al. (2019) treated L points as phonological low targets, and conducted their analysis accordingly, while the current study treated L points as the end point of the falling slope.

The results of the current study, as well as the results of Polyanskaya et al. (2019) have proposed a number of metrics for phonetically quantifying macro-rhythm. However, these are not the only measures that could be used to capture tonal rhythm, and future work should continue to explore other methods of quantifying macro-rhythm strength. One possibility is to use autocorrelation to extract the f0 contour and determine the unit of repetition in the signal. In other words, one could examine how periodic the pitch track contour is, i.e., create an autocorrelation or cepstral analysis of a pitch track and determine the periodicity of the f0 alternations. There are a variety of approaches to this methodology, although it would require a large enough dataset to calculate the f0 periodicity, and the specific technique may require longer sentences. The advantage of a mathematical approach like this is that it would avoid the problem of human decision in locating a specific H and L points on an f0 contour, whether that be phonological or phonetic, and check a large dataset of each language. This method would be robust, reproducible, and independent of language.

Overall, macro-rhythm quantification is a promising approach to the study of speech rhythm because it captures the periodicity of f0 movement as predicted by a language's intonational phonology. Since tonal rhythm plays an important role in word segmentation and in marking word prominence (Jun, 2014), one would expect it to be a salient correlate of rhythmicity. Previous work on isochrony that measured segmental or syllabic durations was able to capture some aspects of speech rhythm, but the results were inconsistent because the perception of rhythm is not based only on these duration measures. Speech rhythm perception is likely the result of multiple acoustic correlates operating at multiple levels of prosodic structure. Tilsen and Arvaniti (2013) found evidence of rhythmicity differences in prominence between languages at various timescales. The rhythmicity of f0 movement captures both the temporal domain of rhythm (i.e. distance intervals between tonal targets) and the frequency domain (i.e. the number of f0 alternations within a phrase), which combines previous approaches and applies them to the phrasal level.

While the results of this study provide support for acoustic differences in measures of tonal rhythmicity, they do not make claims about the perceptibility of tonal rhythm across languages. The next step is to test the perceptibility of macro-rhythm differences between languages. Previous work on speech perception has shown that a listener's native language and linguistic experience determines the weighting of the acoustic cues associated with speech rhythm (Cumming, 2011), affects rhythmic grouping and segmentation (Tyler & Cutler, 2009; Bhatara et al., 2013; Molnar, Carreiras & Gervain, 2016; Ordin, Polyanskaya, Laka & Nespor, 2017), and causes speakers to use different means of producing rhythm (Niebuhr, 2009; Cumming, 2011; Mori, Hori & Erickson, 2014). In terms of language processing, speech rhythm determines how listeners segment the speech signal (e.g. Cutler et al., 1986; Dilley & Shattuck-

Hufnagel, 1999; Dilley & McAuley, 2008), which plays a role in language acquisition (Nazzi et al., 2006). Given these findings, one would predict that listeners can perceive differences in macro-rhythm strength. Specifically, one would predict that listeners perceive Spanish as more tonally rhythmic than English. This would provide further evidence that phrase-level tonal rhythm plays an important role in the perception of speech rhythm.

Future work on the perception of speech rhythm should specifically focus on the timing of f0 and prominence. F0 is an acoustic correlate of prominence at both the lexical and phrasal levels, and there is evidence that the periodicity of another acoustic correlate, amplitude, operates on multiple levels of prosodic structure (Tilsen & Arvaniti, 2013). In addition, Jun (2014:536) notes that one of the primary functions of intonation is to mark word prominence, so tonal rhythm must be the result of prominence marking within a phrase. She further observes an inverse correlation between macro-rhythm strength and the phonetic realization of stress, where languages that mark stress with strong amplitude and long duration (e.g. English) tend to have weaker macro-rhythm than languages with phonetically weak stress (e.g. Bengali). Future research should explore how this language-specific relationship between macro-rhythm and what Jun (2014) calls ‘micro-rhythm’, the traditional speech rhythm metrics such as stress-timed and syllable-timed intervals), informs the perception and production of speech rhythm and word segmentation.

5. CONCLUSION

The goals of this paper were to quantify macro-rhythm in English and Spanish across two speech styles and test the predictions of macro-rhythm strength with phonetic data. The results provide preliminary evidence that Spanish has stronger macro-rhythm than English in the number of L/H f0 alternations, the overall variability of slope shapes, and the regularity of L/H intervals. These differences can be quantified using MacR_Freq index, and, with mixed success, MacR_Var index and nPVI measures. Overall, the results support the predictions about acoustic differences in tonal rhythm strength between English and Spanish, and they provide new potential metrics for capturing and comparing speech rhythm. Future work should test whether these acoustic differences in tonal rhythm are perceptible to listeners, and how these measures contribute to the perception of speech rhythm more broadly.

APPENDIX A: PRODUCTION STIMULI

The bolded syllables are predicted to bear a pitch accent. The numbers to the right of the sentence indicate the ISI between each stressed syllable.

English stimuli

1. **Mi**lo and **A**my ran through **ru**ral Al**ber**ta in the **rain**. 2 3 2 3
2. **Lau**ra wore an **e**legant **lin**en **gown** to the **wed**ding. 3 2 1 2
3. **Dan**ny **mar**ried a re**li**able and **or**derly **wo**man. 1 3 3 2
4. **Ma**lory re**mem**bered the **grim** and a**larm**ing **night**mares. 3 2 2 1
5. **E**mily **wa**tered the **beau**tiful **li**ly on the **win**dowsill. 3 2 2 3
6. My **mom** went to **Ma**ryland and **Maine** in the **mid**dle of the **year**. 2 3 2 3
7. The **brave** and **ho**norable **knight** won **all** of the **duels**. 1 3 1 2
8. **Me**lanie de**light**ed in **knit**ting the **co**lorful **mit**tens. 3 2 2 2
9. The **wa**ter in the **well** will be **gone** by the **end** of the **month**. 3 2 2 2
10. **Mol**ly and **Gre**gory **yelled** at the **land**lord for an **ho**ur. 2 2 2 3
11. The **man** re**mained** in the **warm** **wa**ter for an **ho**ur. 1 2 0 3
12. **Ga**ry led the **wo**man and her **dog** into the **nar**row **al**ley. 3 3 3 1
13. The **mall** had **grim**y **win**dows and a **faul**ty **el**evator. 1 1 3 1
14. My **neigh**bor and my **mom** like **le**mon and **or**ange **mar**malade. 3 1 2 1
15. The **lone**ly **mail**man had **lain** under the **wil**low to **nap**. 1 2 3 2
16. He be**lieved** that the **meet**ing with **ar**rogant **law**yer went **well**. 2 2 2 2
17. **Al**len hid **wine** in the **yel**low **row**boat near the **lake**. 2 2 1 3
18. Their **new** **ar**mor was **light**er than their **new** **weights**. 0 2 3 0
19. **O**liver **wad**ed through the **mud**dy and **tur**bulent **ri**ver. 2 3 2 2
20. The **moo**dy and **ir**ritable **room**mate had a **bad** **a**libi. 2 3 3 0

Table A1: *Number of Inter-Stress Intervals (ISI) of different lengths in English. 0 indicates that there are two consecutive stressed syllables, while 3 indicates three unstressed syllables between stressed ones.*

Length of ISI	Number of Occurrences
0 syllables	4
1 syllable	14
2 syllables	36
3 syllables	26

Spanish stimuli

1. El **ni**ño con ca**be**llo **ru**bio me **dio** un re**ga**lo. 3 1 2 2
2. **Fá**tima vol**vió** a **Mé**rida du**ran**te el ve**ra**no. 3 1 3 3
3. El o**lor** en la ne**ve**ra prove**ní**a de **car**ne po**dri**da. 3 3 2 2
4. **Na**die o**yó** el **rui**do a**gu**do en el **ca**rro. 2 1 2 3
5. El ve**na**do pe**que**ño hu**yó** del **lo**bo ham**brien**to. 2 2 1 2
6. Ma**rí**a llega**rá** ma**ña**na a **Vi**ña del **Mar**. 3 1 2 2
7. El **rui**do del le**ón** le **da**ba **mie**do al **hom**bre. 3 1 1 2
8. Mi a**bue**la le **da** de co**mer** la a**ve**na a Ma**rí**a. 2 2 2 3
9. El gue**rre**ro hono**ra**ble y **no**ble pe**lea**ba con vi**gor**. 3 2 2 3
10. Ra**úl** volve**rá** a Uru**guay** el pri**me**ro de **ma**yo. 2 3 2 2
11. Mi **ma**dre me lla**ma**ba por te**lé**fono **ca**da **lu**nes. 3 3 2 1
12. No te**ní**a i**de**a **dón**de pon**er** la **lla**ve. 2 1 2 1
13. Mi**guel ri**ña al **pe**rro que a**rrui**na la no**ve**la. 0 2 3 3
14. E**le**na llo**ró** de ale**grí**a du**ran**te la **bo**da. 2 3 2 2
15. Ma**nuel Lu**na mi**ró** una a**ra**ña en el **ba**ño. 0 2 3 3
16. El la**drón** ro**bó** el a**ni**llo an**ti**guo de mi **ma**dre. 1 2 2 3
17. **Al**ma devo**ró** el po**bla**no re**lle**no con a**rroz**. 3 2 2 3
18. No vo**lé** por avi**ón** du**ran**te **nue**ve **a**ños. 2 1 1 1
19. **E**va mar**có** el pri**mer nú**mero muy **rá**pido. 2 2 0 3
20. El **ni**ño de Ra**món Ál**varo **bai**la con Ma**ri**na. 3 0 2 3

Table A2: *Number of Inter-Stress Intervals (ISI) of different lengths in Spanish. 0 indicates that there are two consecutive stressed syllables, while 3 indicates three unstressed syllables between stressed ones.*

Length of ISI	Number of Occurrences
0 syllables	4
1 syllable	14
2 syllables	36
3 syllables	26

APPENDIX B: INDIVIDUAL SPEAKERS

Table B1: *Total number of IPs and PWords included for each English speaker*.

Speaker	Number of IPs	Number of PWords
02en	20	82
05en	20	82
10en	20	81
13en	19	77
14en	20	84
18en	20	82
21en	20	82
26en	19	77
27en	15	81
28en	20	82

Table B2: *Total number of IPs and PWords included for each Spanish speaker*.

Speaker	Number of IPs	Number of PWords
01sp	15	59
02sp	17	70
03sp	19	77
04sp	16	63
08sp	20	82
10sp	20	80
11sp	16	63
13sp	20	79
14sp	19	76
17sp	18	72

REFERENCES

Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh University Press.

Aguilar, L., de-la-Mota, C., Prieto, P. (2009). SP_ToBI training materials. http://prosodia.upf.edu/sp_tobi/en/index.php

Allen, G. D. (1972). The location of rhythmic stress beats in English: An experimental study, parts I and II, *Language and Speech* 15, 72–100, 179–195.

Allen, G. D. (1975). Speech rhythm: Its relation to performance and articulatory timing. *Journal of Phonetics* 3, 75–86.

Andreeva, B., Barry, W. J., & Steiner, I. (2007). Producing phrasal prominence in German. *Proceedings of the 16th International Congress of Phonetic Sciences*, Saarbrücken, Germany, 1209-1212.

Arvaniti, A. (2009). Rhythm, timing, and the timing of rhythm. *Phonetica* 66(1-2):46-63.

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, 40(3):351–373.

Barry, W. J. (1981). Prosodic functions revisited again. *Phonetica* 38, 320-340.

Barry, W. J., Andreeva, B., & Koreman, J. (2009). Do rhythm measures reflect perceived rhythm? *Phonetica* 66 (1-2), 78-94.

Beckman, M. E. (1996). The Parsing of Prosody. *Language and Cognitive Processes* 11, 17-67.

Beckman, M.E. & Ayers, G. (1997). Guidelines for ToBI labelling ver. 3. *The Ohio State University Research Foundation*.

Bhatara, A., Boll-Avetisyan, N., Unger, A., Nazzi, T., Höhle, B. (2013). Native language affects rhythmic grouping of speech. *Journal of the Acoustical Society of America*, 134, 3828–3843.

Boersma, P., Weenink, D. (2018). Praat [Computer program]. Version 6.0.31.

Bolinger, D. (1968). *Aspects of Language*. New York: Harcourt, Brace & World, Inc.

Burdin, R., Phillips-Bourass, S., Turnbull, R., Yasavul, M., Clopper, C., & Tonhauser, J. (2014). Variation in the prosody of focus in head- and head/edge-prominence languages, *Lingua 165*, 254-276.

Cruttenden, A. (1993). The de-accenting and re-accenting of repeated lexical items. In *Proceedings of the ESCA workshop on Prosody*, Lund, 16-19.

Cumming, R. (2011). The language-specific interdependence of tonal and durational cues in perceived rhythmicality. *Phonetica* 68, 1–25.

Cutler, A., Mehler, J., Norris, D. G., Seguí, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language* 25, 385–400.

Cutler, A., Mehler, J., Norris, D., Seguí, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology* 24, 381–410.

Cutler, A., Otake, T. (1994). Mora or phoneme? Further evidence for language-specific listening. *Journal of Memory and Language* 33, 824–844.

Dainora, A. (2001). An Empirically Based Probabilistic Model of Intonation in American English. Ph.D. dissertation, University of Chicago.

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

Dauer, R.M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics 11*, 51-62.

Dauer, R.M. (1987). Phonetic and phonological components of language rhythm. *Proceedings of the International Congress of Phonetic Sciences*, Tallinn, 447–449.

de-la-Mota, C., Butragueño, P.M., & Prieto, P. (2010). Mexican Spanish Intonation. In: Prieto, P., Roseano, P. (eds.), *Transcription of Intonation of the Spanish Language*. Munich: Lincom, 319-350.

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for delta C. In Karnowski, P. & Szigeti, I. (eds), *Language and Language Processing: Proceedings of the 38th Linguistics Colloquium, Piliscsaba 2003*. Frankfurt am Main, Germany: Peter Lang Publishing Group, 231-241.

Dilley, L. & Shattuck-Hufnagel, S. (1999). Effects of repeated intonation patterns on perceived word-level organization. *Proceedings of the 14th International Congress of Phonetic Sciences*, San Francisco, 1487-1490.

Dilley, L. & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language* 59, 294-311.

Estebas-Vilaplana, Prieto, P. (2010). Castilian Spanish intonation. In P. Prieto, P. Rosano (eds), *Transcription of Intonation of the Spanish Language*. Munich: Lincom Europa, 17-48.

Face, T.L. (2003). Intonation in Spanish declaratives: differences between lab speech and spontaneous speech. *Catalan Journal of Linguistics* 2: 115-131.

Face, T. & Prieto, P. (2007). Rising accents in Castilian Spanish: a revision of Sp_ToBI. *Journal of Portuguese Linguistics* 6.1, 117-146.

Gasser, E., Ahn, B., Napoli, D.J., & Zhou, Z.L. (2019). Production, perception, and communicative goals of American newscaster speech. *Language and Society* 48(2): 233-259.

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

Garrido, J. M., Escudero, D., Aguilar, L., Cardeñoso, V., Rodero, E., de-la-Mota, C., González, C., Rustullet, S., Larrea, O., Laplaza, Y., Vizcaíno, F., Cabrera, M., & Bonafonte, A. (2013). Glissando: a corpus for multidisciplinary prosodic studies in Spanish and Catalan. *Language Resources and Evaluation* 47(4): 945-971.

Goswami, U., Thomson, J., Richardson, U., Stainthorp, R., Hughes, D., & Scott, S. K. (2002). Amplitude envelope onsets and developmental dyslexia: A new hypothesis. *Proceedings of the National Academy of Sciences* 99(16), 10911–10916.

Grabe, E. & Low, E.L. (2002). Durational variability in speech and the rhythm class hypothesis. In: C. Gussenhoven & N. Warner (eds.), *Laboratory Phonology*. Berlin: Mouton de Gruyter, 515-546.

Handel, S. (1993). The effect of tempo and tone duration on rhythm discrimination. *Perception and Psychophysics 54*(3), 370-382.

Hirst, D., Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. Travaux de I'Institut de Phonetique d'Aix 15, 71–85.

Hirst, D., Di Cristo, A., Espesser, R. (2000). Levels of representation and levels of analysis for intonation. In: Horne, M. (Ed.), *Prosody: Theory and Experiment*. Dordrecht, NL: Kluwer Academic Publishers, 51-87.

Hirst, D. (2007). A Praat plugin for MOMEL and INTSINT with improved algorithms for modelling and coding intonation. *Proceedings of the International Congress of Phonetic Sciences*, Saarbruecken, 1233–1236.

Howell, P. (1988). Prediction of P-center location from the distribution of energy in the amplitude envelope. *Perception and Psychophysics* 43(1), 90–93.

Howell, P. & Kadi-Hanifi, K. (1991). Comparison of prosodic properties between read and spontaneous speech material. *Speech Communication* 10: 163-169.

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

Jun, S-A. (2005). Prosodic Typology. In: Jun, S-A. (ed), *Prosodic Typology*. Oxford: Oxford University Press, 430-453.

Jun, S-A. (2014). Prosodic typology: by prominence type, word prosody, and macro-rhythm. In: Jun, S-A. (ed), *Prosodic Typology II*. Oxford University Press, 520-539.

Katz, J., Selkirk, E. (2011). Contrastive focus vs. discourse-new: evidence from phonetic prominence in English. *Language* 87(4), 771-816.

Kim, S. (2004). The role of prosodic phrasing in Korean word segmentation. Ph.D. dissertation, UCLA.

Kim, S. & Cho, T. (2009). The use of phrase-level prosodic information in lexical segmentation: Evidence from word-spotting experiments in Korean. *Journal of the Acoustical Society of America* 125(5), 3373-3386.

Kohler, K. (2008). The perception of prominence patterns. *Phonetica* 65, 257-269.

Ladd, D. R. (1996/2008). *Intonational Phonology*. Cambridge: Cambridge University Press.

Lee, C. S., and Todd, N. P. M. (2004). Towards an auditory account of speech rhythm: Application of a model of the auditory 'primal sketch' to two multi-language corpora. *Cognition* 93, 225–254.

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics* 5(3), 253–263.

Lerdahl, F. & Jackendoff, R. (1983). *A Generative Theory of Tonal Music*. Cambridge, MA: MIT Press.

Mennen, I., Schaeffler, F., & Docherty, G. (2012). Cross-language differences in fundamental frequency range: a comparison of English and German. *J. Acoust. Soc. Am.* 131(3): 2249–2260.

Molnar, M., Carreiras, M., Gervain, J. (2016). Language dominance shapes non-linguistic rhythmic grouping in bilinguals. *Cognition* 152, 150–159.

Mori, Y., Hori, T., Erickson, D. (2014). Acoustic correlates of English rhythmic patterns for American versus Japanese speakers. *Phonetica* 71, 83–108.

Morton, J., Marcus, S., and Frankish, C. (1976). Perceptual Centers (P-centers). *Psychological Review* 83(5), 405–408.

Murty, L., Otake, T., Cutler, A. (2007). Perceptual tests of rhythmic similarity. I. Mora rhythm. *Language and Speech* 50(1): 77–99.

Nazzi, T., Bertoncini, J. & Mehler, J. (1998). Language discrimination by newborns: Towards an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance* 24, 756-766.

Nazzi, T., Jusczyk, P.W., Johnson, E.K. (2000). Language discrimination by English-learning 5-month-olds: effects of rhythm and familiarity. Journal of Memory and Language 43, 1–19.

Nazzi, T., Ramus, F. (2003). Perception and acquisition of linguistic rhythm by Infants. Speech Communication 41, 233–243.

Nazzi, T., Iakimova, G., Bertoncini, J., Frédonie, S., Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences. *Journal of Memory and Language* 54, 283–299.

Niebuhr, O. (2009). F0-based rhythm effects on the perception of local syllable prominence. *Phonetica* 66, 95–112.

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

Ortega-Llebaria, M., Prieto, P. (2009). Perception of word stress in Castilian Spanish: the effects of sentence intonation and vowel type. In M. Vigário, S. Frota, M.J. Freitas (eds), *Phonetics and Phonology: Interactions and Interrelations*. Amsterdam: Benjamins, 35-50.

Ostendorf, M., Price, P., & Shattuck-Hufnagel, S. (1996). Boston University Radio Speech Corpus LDC96S36. DVD. Philadelphia: Linguistic Data Consortium.

Otake, T., Hatano, G., Cutler, A., Mehler, J. (1993). Mora or syllable? Speech segmentation in Japanese. *Journal of Memory and Language* 32, 358–378.

Pike, K. (1946). *The Intonation of American English*. Ann Arbor: University of Michigan Press.

Polyanskaya, L., Busà, M.G., & Ordin, Mikhail. (2019). Capturing cross-linguistic differences in macro-rhythm: the case of Italian and English. *Language and Speech* https://doi.org/10.1177/0023830919835849.

Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-center phenomenon. *Journal of Phonetics* 17, 175–192.

Prieto, P., Vanrell, M., Astruc, L., Payne, E., and Post, B. (2012). Phonotactic and phrasal properties of speech rhythm: Evidence from Catalan, English, and Spanish. *Speech Communication* 54, 681–702.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition 73*, 265-292.

Rao, R. (2009). De-accenting in spontaneous speech in Barcelona Spanish. *Studies in Hispanic and Lusophone Linguistics* 2(1): 31-75.

Schmerling, S.F. (1976). *Aspects of English sentence stress*. Austin: University of Texas Press.

Shattuck-Hufnagel, S., Turk, A. (1996). A Prosody Tutorial for Investigators of Auditory Sentence Processing. *Journal of Psycholinguistic Research*, 25, 193-247.

Thomassen, J. M. (1982). Melodic accent: experiments and a tentative model. *Journal of the Acoustical Society of America 71*(6), 1596-1603.

Tilsen, S. & Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. *Journal of the Acoustical Society of America* 124(2): EL34-39.

Tilsen, S. & Arvaniti, A. (2013). Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages. *Journal of the Acoustical Society of America* 134(1): 628-639.

Tyler, M., Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *Journal of the Acoustical Society of America* 126, 367–376.

Warner, N., Otake, T., & Arai, T. (2010). Intonational structure as a word boundary cue in Japanese. *Language and Speech* 53, 107-131.

Welby, P. (2007). The role of early fundamental frequency rises and elbows in French word segmentation. *Speech Communication* 49, 28-48.

White, L. & Mattys, S.L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics* 35, 501-522.

Wiget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., and Mattys, S. L. (2010). How stable are acoustic metrics of contrastive speech rhythm? *Journal of the Acoustical Society of America* 127, 1559–1569.