Title: Capturing cross-linguistic differences in macro-rhythm: the case of Italian and English

Running title: Cross-language differences in macro-rhythm

Leona Polyanskaya[1*]

Maria Grazia Busà[2]

Mikhail Ordin[1,3]

[1]BCBL – Basque Centre on Cognition, Brain and Language
Mikeletegi 69,
20009 Donostia, Spain

[2]Università degli Studi di Padova
Dipartimento di Studi Linguistici e Letterari (DiSLL)
Via Pellegrino, 26
35137 Padova, Italy

[3]Ikerbasque – Basque Foundation for Science
Maria Diaz de Haro 3,
48013 Bilbao, Spain

*corresponding author

**ABSTRACT**

We tested the hypothesis that languages can be classified by their degree of tonal rhythm (Jun, 2014). The tonal rhythms of English and Italian were quantified using the following parameters: (1) regularity of tonal alternations in time, measured as durational variability in peak-to-peak and valley-to-valley intervals; (2) magnitude of F0 excursions, measured as the range of frequencies covered by the speaker between consecutive F0 maxima and minima; (3) number of tonal target points per intonational unit; and (4) similarity of F0 rising and falling contours within intonational units. The results show that, as predicted by Jun's prosodic typology (2014), Italian has a stronger tonal rhythm than English, expressed by higher regularity in the distribution of F0 minima turning points, larger F0 excursions, and more frequent tonal targets, indicating alternating phonological H and L tones. This cross-language difference can be explained by the relative load of F0 and durational ratios on the perception and production of speech rhythm and prominence. We suggest that research on the role of speech rhythm in speech processing and language acquisition should not be restricted to syllabic rhythm, but should also examine the role of cross-language differences in tonal rhythm.

## *INTRODUCTION*

Rhythmicity is an intrinsic property of the physical environment and social behaviour. Rhythm is therefore an integral part of both nature and nurture (see Strogatz, 2004 for an overview of studies on rhythmic alternations in nature). When people talk about rhythm, they usually mean the systematic recurrence of recognizable patterns at regular or quasi-regular intervals. Rhythm in speech is a remarkable exception. In speech, people often perceive rhythms, yet no uncontroversial evidence has been found that recognizable events or patterns of events recur at regular intervals. Some researchers have even said that speech is arrhythmic by nature (see, e.g., Nolan & Jeon, 2014), and that rhythm is imposed on the acoustic signal by the listener in the course of perception (Madison & Merker, 2002; McAuley, 2010; Motz, Erickson, & Hetrick, 2013). Perception of rhythm is based on the entrainment of cortical oscillations to syllabic frequency, and this helps segment a continuous signal into quasi-regular chunks (Ding, Melloni, Zhang, et al., 2016; Batterink & Paller, 2017). As the period of neural oscillations in the theta frequency band (roughly corresponding to syllabic duration) may vary within physiologically determined limits, there is still room for variability in syllabic durations. The degree of this syllabic durational variability may differ across languages, accounting for cross-linguistic differences in syllabic rhythm. In addition, perception of rhythmicity is based on the regularity of pitch movements, for example, the repetition of sequences of rising or falling pitch contours (e.g., Thomassen, 1982; Handel, 1993; Niebuhr, 2009; Cumming, 2011a; b; Barry, Andreeva, & Koreman, 2009). These pitch contours represent phrasal prominence and define phrasal and sentential structure (Bourguignon, De Tiege, de Beeck, et al., 2013; Ding, Patel, Chen, et al., 2017; Nespor & Vogel, 2008). Tonal alternations also allow brain-to-sound coupling in

frequency bands that correspond to the rate of pitch accents, and these slower neural oscillations further support speech processing (Bourguignon et al., 2013; Ding et al., 2017; Molinaro & Lizarazu, 2018).

Cross-linguistic differences (Ramus & Mehler, 1999; White & Mattys, 2007), the characteristics of pathological speech (Liss, White, Mattys, et al., 2009; Henrich, Lowit, Schalling, & Mennen, 2006) and developmental aspects (Ordin & Polyanskaya 2014; 2015; van Maastricht, L., Krahmer, E., Swerts et al., in press) related to syllabic rhythm have received a great deal of attention. However, while recurrent patterns of tonal events do contribute to the perception of rhythmicity and may lead speakers of different languages to use different strategies for speech processing, cross-linguistic differences in tonal rhythm remain at best underexplored. Our work aims to determine what specific differences in tonal rhythm lead to perceived rhythmic differences between languages. We have based our study on the notion of macro-rhythm, or tonal rhythm, introduced by Jun (2014) as a component of prosodic typology. In the following, we explain our theoretical point of departure – Jun's (2014) prosodic typology – in more detail.

Jun (2014) proposed a prosodic typology based on the interplay of macro- vs. micro-rhythms, co-existing within the same language at different timescales. Micro-rhythm exists on the syllabic and sub-syllabic (consonantal and vocalic intervals) timescales and is related to phonotactic constraints and certain aspects of lexical prosody (e.g., how lexical stress is phonetically realized in a language, i.e., the relative contribution of prosodic parameters to the manifestation of lexical stress). Phonotactic constraints and lexical prosody either enhance or inhibit durational ratios between stressed and unstressed syllables and vowels. Emergent differences in speech timing pertain to so-called stress-, syllable-, and mora-timing (Ramus et al.,

1999): a higher degree of stress-timing manifests in larger durational ratios and a lower proportion of vocalic material in the speech signal.

Macro-rhythm is operationalized as a tonal rhythm created by regular alternations of pitch tones in a phrase. In the proposed typology, macro-rhythm is one of the three phonological parameters suggested for classifying languages into typological groups, the other two being *prominence type* (head vs. head/edge vs. edge) and *word prosody* (whether lexical prominence is manifested by tone, by stress, both by stress and tone, or neither by stress nor by tone, Jun, 2014: 535). This typology is grounded on the phonological analysis of languages within the autosegmental-metrical (AM) framework (Pierrehumbert, 1980; Beckman & Pierrehumbert, 1986; Goldsmith, 1990; Ladd, 2008). In this investigation, we decided to examine only head-prominence languages with lexical stress (to keep *prominence type* and *word prosody* parameters in both languages under study the same, and avoiding their potential interaction with macro-rhythm). This allowed us to focus exclusively on macro-rhythm, and to use the term *pitch accent* to refer strictly to post-lexical prosody.

Macro-rhythm is created by F0 fluctuations related to post-lexical prosody, in particular, by the distribution and phonetic realization of pitch accents[1]. Macro-rhythm classifies languages based on their prosodic structure and on the type, inventory and distribution of pitch accents, analyzed within the AM framework (Jun, 2014). Thus, macro-rhythm depends on 1) the size of the tonal inventory, 2) the most frequent pitch accent type(s), and 3) the number of pitch accents per phonological word. Languages with weaker macro-rhythm have a larger inventory of possible

[1] Following the tradition of the autosegmental-metrical approach (Ladd, 2008), by pitch accent we mean a local intonational feature which is associated with a particular prominent syllable. Pitch accents consist of a high (H) or low (L) F0 target (or a combination of H and L targets), indicating the relative highs (H) and troughs (L) of the F0 contour.

phonological tones, especially pitch accents that can be used in phrase-medial positions, than languages with stronger macro-rhythm. In languages with the strongest tonal rhythm, each phonological word bears a pitch accent (has a phonological tone associated with it), thus the ratio of the number of phonological tones to the number of phonological words is close to 1. Finally, languages with stronger macro rhythm have both high and low tones among the most frequently realized tones in spoken speech, which leads to alternating high and low tones, while languages with weaker macro-rhythm tend to prefer either high or low tones exclusively, and this preference leads to sequences of consecutive high or low tones in spoken speech. As is evident from this description, the factors that determine the strength of macro rhythm in this framework rely on phonological categories and phonological analysis, not on quantitative data.

It is important to test whether macro-rhythm constitutes a typological parameter using phonetic quantification. Phonological theory informs empirical research regarding what should be tested, while phonetic data, in turn, feeds back into phonological theory, either confirming or updating it. Our goal was to test macro-rhythm, a typological parameter that relies on the *phonological* analysis of over 50 languages (Jun, 2014), against *phonetic* data. This would be an important step in confirming the typology of tonal rhythm and better understanding not only cross-linguistic differences in language structure, but also how speech processing strategies (e.g., segmentation) may be tuned to different macro-rhythms in different languages. The schematic pitch contours (**Figure 1**), based on Jun (2014: 525), illustrate differences pertaining to the phonological properties of macro-rhythm strength.

INSERT FIGURES 1a AND 1b SOMEWHERE HERE

Jun (2014: 528) split languages into three macro-rhythm groups as follows: languages with strong (Italian, Spanish and Catalan, Swedish, Greek, Egyptian Arabic, etc.), medium

(English, German, Lebanese Arabic, Dutch, etc.) and weak (Cantonese and Mandarin) macro-rhythm[2]. Jun (2014: 534) clearly states that boundaries between typological classes are gradual, and that languages are spread over a continuum, with some languages exhibiting relatively stronger or weaker macro-rhythm than others.

The main objective of this study was to test phonological theory against phonetic data. Thus, we first translated phonological theory into concrete phonological hypotheses, such that each hypothesis generated a specific phonetic prediction. Then, we mapped these phonetic predictions onto the values of the proposed measures for macro-rhythm (see Table 1 for an overview).

One of the main criteria for strong macro-rhythm in Jun (2014) is the alternation of H and L *phonological* tones. Alternating H and L phonological tones cause larger fluctuations of F0 and result in wider frequency ranges between tones. Within the AM framework, the F0 contour on Figure 1b (bottom) would be annotated as a sequence of phonological H tones. A sequence of consecutive H tones leads to smaller F0 fluctuations (schematic contour 2 in Jun, 2014: 525). However, in our study, we use H and L labels to label *phonetic* events, namely, F0 peaks and valleys. Therefore, the L and H labels in Figures 1a and 1b refer to F0 high and low turning points, not to phonological tones. Alternating high and low tones are more likely to result in larger F0 excursions between F0 maxima and minima turning points compared to sequences of high or low tones.

Fewer pitch accent types and regular alternations between high and low phonological tones, typical of languages with stronger macro-rhythm, is reflected in more regular distribution

[2] All these languages are from the head-prominence group because we have restricted our work to this typological class (Jun, 2014: 535).

of F0 targets in the temporal domain. Thus, stronger macro-rhythm should be represented by less variation in the duration of F0 peak-to-peak and valley-to-valley intervals. Further, if the number of phonological tones in the inventory of a particular language is high, then the shape of pitch accents in phrase-medial positions may be very different, resulting in higher variation in the slope of F0 rises and falls. Jun (2014: 538) proposed a *variation index* (*MacR_Var*) to quantify the strength of macro-rhythm, which is the sum of the standard deviations (SD) in mean durations of the F0 peak-to-peak intervals, mean durations of valley-to-valley intervals, slopes of F0 rises and slopes of F0 falls. F0 slopes depend on how fast F0 rises and falls. Here we have substituted SD in the range of frequencies covered by rising F0 contours and range of frequencies covered by falling F0 contours for Jun's SD in slope. This change is motivated by the fact that larger excursion size corresponds to steeper F0 rise, if the distance between adjacent F0 peaks and valleys is constant. Higher scores of *MacR_Var* should indicate a weaker degree of macro-rhythm. As the shape of the pitch contour can be represented by the slope of F0 rise and F0 fall and by variability in the distances between peaks and valleys, the sum of SDs potentially captures an otherwise intangible parameter - the similarity of pitch contours within intonational units. This is a great advantage of the proposed macro-rhythm metric. On the other hand, *MacR_Var* merges the variability from the temporal and frequency domains. A comparison of two hypothetical languages (*A* and *B*) could potentially reveal that *A* exhibits a larger temporal variability between F0 turning points, thereby demonstrating it has weaker macro-rhythm in the temporal domain than *B*. At the same time, *A* might exhibit alternating H and L tones, leading to a larger range of frequencies covered by rising and falling F0 contours and to a similar degree of variability in F0 excursions, thereby demonstrating that it has stronger macro-rhythm than *B*. It is also possible that *MacR_Var* might not reveal any differences in macro-rhythm between languages because

variability differences in the temporal and frequency domains cancel each other out. Such hypothetical but plausible situations are depicted by the schematic F0 pitch contours in Figure 2. Therefore, in addition to the proposed metric, we need to find a method to distinguish between macro-rhythm in the temporal and frequency domains. We propose to apply the variability measures normally used for segmental and syllabic variability (to quantify differences in syllabic rhythm) to inter-tone differences both in frequency and in time. For our study, we selected the coefficient of variation (Varco, the SD divided by the mean) and nPVI (1) measures (Grabe & Low, 2002) to capture global and pairwise variability in the distribution of pitch accents in time and in magnitude of F0 excursions.

(1)

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

where $m$ is the number of measurements and $d$ is the score of the $k^{th}$ measurement in the array, and dividing the difference between two consecutive measurements by the mean of these measurements is the normalization factor for speech rate (in the temporal domain) and for speaker- and gender-specific differences in the F0 range.

INSERT FIGURE 2 SOMEWHERE HERE

The normalized measures were chosen because they are robust to idiosyncratic differences in speech rate (affecting the temporal distribution of pitch accents) and in mean F0 (affecting the magnitude of pitch accents). For the same reason, we calculated the *MacR_Var* index by substituting the SD with the Varco measure. Until now, very few attempts have been made to apply rhythm measures to tonal events (with some notable exceptions, e.g., Cumming, 2011a;b). This study bridges this gap.

Jun (2014) also proposed a *MacR_Freq* index to quantify the frequency of H/L alternations within phonological words. A strong macro-rhythm exhibits scores closer to 1, indicating one pitch accent per phonological word within a phonological phrase. Languages with stronger macro-rhythm tend to have a pitch accent on every phonological word (Jun, 2014: 526), leading to a higher number of tonal units. Jun (2014) considered the data collected on speech read in a declarative style; we worked on spontaneous speech, which sometimes leads to false starts, repetitions of the same words and other peculiarities which make it difficult to clearly decide whether or not a started but interrupted fragment should be considered to be a completed phonological word. Identifying boundaries between phonological words that do not have clear phonetic correlates is an additional problem. Segmenting speech into phonological words relies on human judgement and knowledge of the prosodic phonology of a particular language. Consequently, using the *MacR_Freq* index without adapting it to the objectives of the study would entail comparing phonological systems between languages rather than phonetic measures that differ across languages due to structural differences. Therefore, we decided to analyze the number of F0 turning points, defined as local pitch targets, which are phonetic primitives of pitch contour (e.g., Xu, 2005), rather than the frequency of H/L alternations. This allowed us to use purely phonetic data to test the phonological hypothesis related to the alternation of phonological H and L tones. As our main objective was to provide a phonetic underpinning for Jun's typology, we preferred to avoid testing phonological theory against phonological data. Calculating the number of *phonetic* tones per unit, on the other hand, is a more robust method for natural speech, and the measures are purely phonetic.

F0 turning points were determined by finding points on the pitch contour where F0 changed direction (from falling to rising or from rising to falling, automatically detected) and the

difference between two consecutive F0 turning points was equal to or greater than 2 semitones (ST). Please refer to the methods section for more details. The nature of the material accounts for our decision to move from the frequency to the quantitative measure, which is phonetically more stable and objective. This allowed us to consider interrupted fragments as phonological words independently of researchers' subjective decisions, or of any prior knowledge of the prosodic phonology of a particular language. At the same time, it provides information comparable to the frequency of tones per phonological phrase in the phonological analysis (i.e., a larger number of H/L alternations is reflected in a larger number of local pitch targets), and allows for the development of algorithms to automatically quantify macro-rhythm.

As far as we are aware, only one previous study has provided evidence that differences in the phonetic realization of prominence may be accounted for by differences in macro-rhythm (Burdin, Phillips-Bourass, Turnbull, et al., 2015). Burdin et al. (2015) analyzed focus marking in English, Guaraní, Moroccan Arabic, and K'iche, and concluded that acoustic differences in focus marking reflect cross-language differences in macro-rhythm. In our study, we wanted to provide quantitative data on how differences in macro-rhythm could be measured, and to provide a phonetic confirmation for Jun's assertion that macro-rhythm is a prosodic typological parameter. We analyzed Italian and English. According to Jun's typology (2014), Italian is predicted to have strong macro-rhythm, while English is predicted to exhibit medium macro-rhythm (Jun, 2014). We expected the following cross-language phonetic differences to reflect differences in phonological parameters pertaining to macro-rhythm. As compared to English, Italian

1) exhibits more evenly dispersed F0 turning points in the temporal domain;

2) exhibits wider F0 excursions (differences in F0 between consecutive F0 target points);

3) has a larger number of F0 target (turning) points per intonational unit;

4) exhibits more similar sub-tonal units, which is reflected in lower scores on the variation index for macro-rhythm (*MacR_Var*).

## *II. METHOD*

### A. PARTICIPANTS

Ten female English and ten female Italian speakers were recruited to take part in the experiment. Male speakers were not included in the sample because males and females may differ in the use of pitch range even when using the same linguistic structures (Ordin & Mennen, 2017). All participants were monolingual speakers from monolingual families. English native speakers were recruited in Bangor, UK; Italian speakers were recruited in Padova, Italy. All participants were matched in age (from 20 to 28 years old) and social background: all were university students. Italians came from the Veneto region and were speakers of standard Italian, though their accent was representative of the regional Veneto pronunciation (as confirmed by the second author, a phonetician and a native speaker of Italian, from the Veneto region). The native English speakers all spoke "Southern Standard British English" (SSBE). They had grown up outside Wales and had come to Bangor as students less than one year before the recording. To select only SSBE speakers we followed the formal procedure that Mennen, Schaefler and Docherty (2012) used to select their SSBE subjects from the local population of university students in Edinburgh: from the recordings we selected words that are indicative of SSBE (Wells, 1982) and asked two native speakers of British English to evaluate the words based on vowel quality and the presence or absence of rhoticity. All participants received a small monetary compensation for their time.

## B. MATERIAL AND PROCEDURE

An animated cartoon of the fable 'The Fox and the Crow' was used as a speech elicitation prompt. The cartoon was retrieved online (https://www.youtube.com/watch?v=vt3HP4VWuH0), shortened (to 88 sec.), and muted. The video was preceded by a short summary of the story shown on a computer screen. The text can be found in the appendix. The participants were informed that they would first read a short story displayed on a computer screen, and then watch the cartoon showing this story. The participants were told that they would then have to retell the story based on the cartoon they watched, so should pay attention to the details of the story. The language of the story was either English or Italian, depending on the participants' native language. The summary was used to elicit similar words from each participant and thus obtain comparable speech material. The cartoon was accompanied only by music, and contained no verbal information presented either auditorily or visually. The cartoon was shown two times. After watching the cartoon, the participants were asked to retell the story in detail to an audience of three listeners. The listeners did not interact with the speakers, and their presence served only a pragmatic purpose: we told the speakers that the listeners had not watched the cartoon and later would have to answer a few questions about the content of the story. The speakers' performance was recorded using a Diversity Wireless Microphone System ACT-311/ACT 312 for the audio, and a professional JVC 3CCD camera for the video recording. The audio recordings were taken at a 44kHz sampling rate, at 8 bit, in mono, in PCI WAV format. The video recordings were taken in MOV format, 50 frames per second.

During the recordings, participants were positioned in front of the camera, at a two-meter distance; the recordings were made from a frontal view. The recordings of the English speakers were made in an anechoic chamber; the recordings of the Italian speakers were made in a sound-

attenuated room. The video recordings were made for a different study and were not analyzed here (however, the participants were aware that they were being recorded, and all participants signed a consent form explicitly giving permission to make video recordings of their performance). To interact with the participants in the UK, we recruited a native English speaker (a student from the Film Studies Department, who also served as a professional technician and made the recordings during the experiment). A native Italian speaker from the Veneto region, a student of linguistics, interacted with the participants in Padova and made the recordings.

**C. ANNOTATION**

The retellings lasted approximately one minute per participant (English: mean 50.92 sec, SD 20.66; Italian: 50.68 sec, SD 15.99, and the difference in mean duration between the groups was not significant, $t(18)=.03$, $p=.977$). The entire audio extracts were annotated. The audio signals were analyzed acoustically to extract data on the speakers' F0 patterns.

Prior to the annotation of the phonetic F0 turning points, pitch contour stylization was performed. We used the modified stylization method proposed by Mennen et al. (2012). For the reader's convenience, we describe the main steps of this procedure (the reader can refer to Mennen et al. 2012 for the full description and step-by-step justification of the original procedure). Initially we created an F0 manipulation object in Praat (v.5.4.01) for each audio file (Boersma, 2001). This object represents the F0 contour as a succession of F0 points. The autocorrelation algorithm embedded in Praat was applied for F0 estimation. We set up the software to obtain the pitch point every 10ms, with 500Hz and 50Hz as the ceiling and bottom values of the F0 contour. Afterwards, we applied a 2-semitone (ST) stylization using the standard Praat algorithm. As a result, this simplified pitch contour was marked by pitch points which fell a minimum of 2 SD above or below the line connecting 2 consecutive points. After this step, we

visually inspected the remaining points and corrected the artefacts caused by the pitch estimation algorithm. The rationale behind this pre-labelling step was to correct pitch estimation algorithm errors, and create a simplified representation of the F0 contour (without spurious F0 fluctuations caused by pitch estimation errors, e.g., octave shifts, spurious F0 fluctuations following word-initial plosive consonants, etc.). The artifacts were corrected manually, based on visual inspection and auditory comparisons between the corrected contours and the originals. For the auditory comparison, the stylized pitch tier (as a Praat object) was used to compute the hummed sound object, with humming corresponding to the F0 contour. The same humming sound object was also computed from the unstylized contour. Both sounds were auditorily compared to make sure that the new stylized contour corresponded auditorily to the original. Further, the new contour was used to re-synthesize the utterance applying the PSOLA algorithm embedded in Praat. Again, an auditory comparison between the original and the resynthesized utterances was performed to make sure that the stylization led to perceptually equivalent results after artifact removal. Figure 3 displays a sample pitch trace before and after manual correction, and two corrected artifacts as examples.

INSERT FIGURE 3 SOMEWHERE HERE

We next applied the MOMEL (Modelling Melody) algorithm for intonational analysis (Hirst & Espesser, 1993; Hirst, DiCristo & Espesser, 2000; Hirst, 2007) to the stylized F0 contours. MOMEL is based on Fujisaki's (1997) prosodic model: the actual acoustic F0 contour is an output of the interplay between a slow-varying phrase component with a declination function and faster-varying accent components that model local intonational events (pitch accents). However, the resulting contour is further modulated by short-term perturbations caused by segmental – consonantal and vocalic – characteristics: close vowels are produced with higher

F0 than open vowels; vocal folds vibration is completely inhibited on voiceless consonants leading to the absence of pitch tracks. These unavoidable phonetic perturbations are not intended when the F0 targets are planned during speech production. Yet, in consequence, certain unintended, unplanned F0 peaks can appear due to interaction with segmental realizations. Such short-term perturbations may conceal the planned F0 target points, that is, the point where the F0 contour reaches the planned – not actual acoustic – targets and turns in the opposite direction (high (H) and low (L) turning points). This makes it difficult, for example, to compare the physical acoustic F0 contours of the utterances "*I said that for mama*" and "*I said that for papa*". As the F0 contour cannot be estimated on voiceless plosives, the actual F0 maxima and minima will be different on these two contours. In the "for mama" case, the F0 targets are clear, in the "for papa" case, the contour is interrupted at /p/. Yet, the intended F0 targets are similar in both cases. The MOMEL algorithm was developed to account for the fact that intonational contours are planned during speech production independently of planning for segmental strings, and that the planned F0 targets are not always acoustically realized when targets are aligned with voiceless segments. The MOMEL algorithm aims to interpolate the F0 contour so as to detect the F0 targets even when these targets cannot be estimated. We applied the linear spline function to interpolate the target F0 points. The quadratic spline function was not necessary because the contours had already been stylized in Praat, using a more flexible and robust procedure, as described above. The detected F0 turning points were labeled as a sequence of alternating high (H) and low (L) turning points on the F0 contour. Figure 4 shows two sample sentences, approximately comparable in duration, spoken by an Italian speaker (4a) and by an English speaker (4b), with annotations.

INSERT FIGURE 4a and 4b SOMEWHERE HERE

Next, speech was split into prosodic phrases based on auditory impressionistic analysis, visual inspection of pitch tracks and spectrograms in Praat, as well as semantic and discourse factors. Phrase boundaries were primarily determined by syntactic and semantic factors such as focus and given vs. new information (Nespor & Vogel, 2008: 187-205), taking into consideration that the boundaries of prosodic phrases often match the syntactic structure of the utterance (Nespor & Vogel, 2008; Selkirk, 1986; Pierrehumbert & Hirschberg, 1990). Further, we made sure that the phrases were characterized by temporal, rhythmic and intonational unity. Strong changes in intonational structure, tempo, and longer pauses were additional markers of phrase transitions. The mean number of prosodic phrases did not differ between language groups (English: 10.5, SD 5.3; Italian: 10.4, SD 4.3, p=.962).

**D. MEASUREMENTS**

The F0 turning points, i.e., F0 targets, were automatically detected by the MOMEL algorithm, ensuring the reproducibility of the results and protecting against any unintentional bias that could arise during manual annotation by researchers informed about the experimental hypothesis. The detected targets were labelled H (high targets, interpolated F0 peaks) and L (low targets, interpolated F0 valleys). All scores were calculated separately for each phrase containing at least three turning points. First, we measured peak-to-peak, valley-to-valley, and target-to-target interval durations to calculate the variability in the temporal distribution of F0 turning points using the Varco and nPVI metrics. In the frequency domain, we calculated the magnitude of F0 changes between successive F0 targets and estimated the variability in the magnitude of F0 excursions. The variability in the magnitude of F0 rises and falls, together with the variability in the temporal distribution of H and L F0 target points, was used to calculate the *MacR_Var* index as a proxy for the similarity of the sub-tonal units. Finally, we compared the number of F0

turning points per intonational unit in English and Italian. Please refer to Table 1 for more details on how phonological properties are mapped onto the phonetic predictions, and how these predictions are captured by macro-rhythm measures.

## *III. RESULTS*

### A. MACRO-RHYTHM IN ENGLISH AND ITALIAN IN THE TEMPORAL DOMAIN

Given the relatively low number of speakers and considering that inter-speaker variation is a major contributor to prosodic variation (Cangemi, El Zarka, Wehrle, Baumann, & Grice, 2016), it was possible that any cross-linguistic differences we found might have been driven by individual differences between speakers. In order to establish whether any differences found between languages would hold true independently of interspeaker variation, we constructed mixed effects models (in SPSS version 18) with *Language* (English vs. Italian) as a fixed factor and *Speaker* as a random factor. The random factor is a qualitative variable whose levels are randomly drawn from a population of levels. By introducing the speaker as a random factor, we assumed that the speakers of two groups were randomly sampled from a population of Italian and a population of English speakers. This means that the scores of the rhythm metrics would vary around the mean of the population, if a different set of speakers were to be sampled from the same populations. This approach also has the advantage of being robust to the assumption of normality (cf. Norman, 2010); thus, the normality tests are not reported. We constructed the model with random intercepts for each speaker. By allowing the intercepts to be random, we assumed that the intercepts for each speaker could vary around the model mean, i.e., $\text{MetricScore}_{Xj} = (b_0+u_{oj})+b_1*\text{Language}_j + \varepsilon$, where the $\text{MetricScore}_X$ observation value, i.e., the value of a metric

of a participant X with native language *j*, and $u_{oj}$ is the component that measures variability of the intercepts around the mean of a language *j* (Italian or English), and ε is the error (deviation from the model of a certain observation). Consequently, $(b_0+u_{oj})$ estimates the intercept of the overall model AND the variability of speakers' intercepts around the model.

Separate models for the temporal distribution of H F0turning points, L F0 turning points and all consecutive turning points were constructed. Each phonological tone is phonetically realized as an F0 target, thus variability in the temporal distribution of all consecutive F0 targets supposedly captures any cross-linguistic differences with respect to the regularity of alternating H and L tones. However, as specified in the typology we were testing (Jun, 2014), languages differ in the type of their most frequent tones. In English, for example, the most frequent type of phonological tone is H* (corresponding to a F0 maximum point), while in Italian both high and low phonological tones, including L+H*, L*, and H* are frequent (see D'Imperio, 2002; Jun, 2014 for an overview). Thus, the distribution of F0 valleys in Italian is likely to be more regular than in English because F0 minima are specifically planned as F0 targets associated with prominent syllables by Italian speakers. In English, on the other hand, F0 valleys between two adjacent H tones are not phonological tones. They are phonetic by-products of small F0 dips between consecutive high phonological tones and are not under the speaker's active control. To account for this potential difference, we analyzed the temporal distribution of L and H F0 turning points separately.

The estimates and model statistics are given in Table 2. The results showed that Italians exhibited a more regular distribution of F0 targets, as reflected by lower scores in the variability measures for the temporal distribution of F0 turning points (Figure 5). However, only the cross-linguistic difference in nPVI between L F0 targets and between all F0 turning points reached

significance. This pattern of results confirms the prediction that macro-rhythm in Italian is stronger than in English. It is worth noting that cross-linguistic differences in macro-rhythm are detected in pairwise variability only, and global variability in the distribution of tonal targets, as captured by the Varco measure, is not significant.

INSERT FIGURE 5 SOMEWHERE HERE

## B. THE MAGNITUDE OF F0 CHANGES BETWEEN SUCCESSIVE F0 TARGET POINTS

The same model was constructed for the magnitude of F0 fluctuations between F0 targets. The analysis (Table 2) revealed that the range of frequencies covered between successive F0 targets is larger in Italian than in English (Figure 6a). Larger magnitude F0 changes between consecutive H and L targets also indicates that phonological H and L tones alternate (we are not talking about variability in F0 magnitude, but about F0 minima turning points (L) following F0 maxima turning points (H). This result and interpretation are in line with the *phonological* assumption that Italian is more rhythmic at the macro-rhythm level than English.

INSERT FIGURE 6(a-b-c) SOMEWHERE HERE

## C. THE NUMBER OF F0 TARGET TURNING POINTS IN ITALIAN AND ENGLISH

The same model was constructed for the number of F0 target points (Figure 6b). The analyses revealed that, even controlling statistically for utterance length, the number of F0 turning points is significantly larger in Italian than in English (Table 2). No correlations between the number of target turning points and the values of rhythm metrics for macro-rhythm (across speakers) were detected. This result, again, conforms to the prediction.

### D. VARIATION INDEX IN ITALIAN AND ENGLISH AS A PROXY ESTIMATE FOR CROSS-LANGUAGE F0 CONTOURS SIMILARITY

Finally, we calculated the *MacR_Var* index for each IP, and, as in the previous analyses, constructed the model with *Speaker* as a random factor, to find out whether the difference in *MacR_Var* between the two languages was significant. This difference, however, did not turn out to be significant. We will later discuss possible causes why the *MacR_Var* measure does not differ between Italian and English, although Italian is characterized as more macro-rhythmic than English.

## *IV. DISCUSSION*

Our results revealed that macro-rhythm can indeed be quantified using acoustic parameters in the temporal and frequency domains. In general, the results confirmed cross-linguistic differences in tonal rhythm and supported the prosodic typology proposed by Jun (2014). Jun (2014) categorized Italian as a language with strong macro-rhythm, and English as part of a group of languages with medium macro-rhythm. This classification should supposedly be reflected in (1) higher regularity in the temporal distribution of H and L F0 turning points within utterances, (2) wider F0 excursions, (3) more frequent F0 peaks and valleys, and (4) more similar sub-tonal units in Italian compared to English. We quantified temporal regularity in the distribution of H and L tonal targets with the help of the variability measures applied to peak-to-peak and valley-to-valley intervals; excursion width was quantified as the range of frequencies covered between consecutive F0 valleys and peaks; and the frequency of tonal events was accounted for by comparing the number of tonal events per IP, factoring in the length of the IP during the analysis. The data confirmed that Italian indeed exhibits stronger macro-rhythm than English. Some of

these cross-language differences in the acoustic correlates of macro-rhythm are exemplified in Figures 4a and 4b. The difference in the shape of sub-tonal units, as captured by the proposed metric *MacR_Var*, was not significant. This might indicate that the metric was not actually capturing the tonal similarity in the shape of the F0 contours properly, and a different approach to quantify the similarity of consecutive F0 rising tones and the similarity of consecutive F0 falling tones needs to be developed. It is more likely, however, that the metric was not capturing cross-linguistic differences for the particular language pair we tested. *MacR_Var* supposedly depends on the similarity of the F0 rising and falling slopes, which in turn hinges on the variability of different phonological tones in the language inventory (please note that here we are not talking about the variability of tones as used in speech, but rather about the variability of tones in a language's phonological inventory). The sizes of the tonal inventories in English and Italian are roughly similar (see Jun, 2014 for an overview of the tonal inventories of different languages). As the *MacR_Var* measure was proposed as a proxy to capture cross-linguistic differences stemming from differences in the size of language-specific inventories of phonological tones, it is not surprising that *MacR_Var* does not differ for English and Italian. However, for a different pair of languages, e.g., Egyptian Arabic or Swedish (languages with a strong macro-rhythm and a very limited inventory of phonological tones) on the one hand, and English or German (languages with weaker macro-rhythm and a large inventory of phonological tones) on the other hand, the *MacR_Var* measure would potentially be more informative. However, this awaits further empirical verification.

Importantly, the *MacR_Var* metric combines the measures of tonal variability both in the frequency (as the variability in F0 excursion size between successive F0 turning points) and in the temporal (as the temporal distribution of F0 turning points) domains. Although macro-rhythm

indeed develops simultaneously in both domains, the planning involved in temporal variation and frequency variation are independent. Rhythmic abilities can be selectively disrupted only in the temporal or the tonal domain, leaving processing in the unaffected domain undamaged (Wilson et al., 2002; Fries & Swihart, 1990). Modelling work on the prosody of natural languages has also shown that local pitch targets and pitch range are separate types (and thus represented by separate parameters in models, e.g., the PENTA model) of phonetic primitives of speech melody, stemming from articulatorily controlled parameters (Xu, 2005). The distribution of tonal events across time and the magnitude of F0 changes are independent cues for perceived rhythmicity (Cumming, 2011a). The contribution of variation in the temporal domain to tonal rhythm is independent and can even oppose the contribution of variation in the frequency domain (Cumming, 2011a), e.g., with stronger macro-rhythm in one domain and weaker macro-rhythm in the other domain. Hence the *MacR_Var* metric, which combines both dimensions into one measure, may not be informative for certain pairs of languages. We tentatively suggest that exploring tonal rhythm in these two domains separately might result in more insights than combining all measures into one metric, especially when comparing languages with similar sized inventories of phonological tones.

Interestingly, the distribution of low F0 turning points was more informative regarding cross-language differences than was the distribution of high F0 targets. This result, as we suggested earlier, can be explained by the types of the most frequent phonological tones. In Italian, L* and L+H* tones are much more frequent than in English, hence Italian speakers need to purposefully plan F0 valleys, which should be associated with prominent syllables. In English, on the contrary, many of the F0 valleys are small pitch dips between consecutive F0 peaks representing H* tones. Such F0 minima are by-products of articulation that result from

purposeful planning of a sequence of H tones. As these minima are not purposefully planned and controlled, and do not have to align with prominent syllables, they are likely to be more variable temporal distribution.

This overall successful enterprise in quantifying cross-language differences in macro-rhythm relies on a semi-automatic approach to annotate and detect F0 target points, including F0 targets which are not acoustically realized and F0 targets which cannot be estimated. This approach ensures that the results are not affected by subjective decisions or any unintentional bias caused by awareness of the test hypothesis, an annotator's experience, or the theoretical framework within which the annotator was trained. This makes the results stable and reproducible. Using the MOMEL algorithm for future work quantifying macro-rhythm is advisable. The algorithm can be used on stylized or on raw F0 contours estimated in Praat, and this choice is subjective, depending on the available resources and the amount of speech material that is to be analyzed. We believe that prior stylization based on Mennen et al.'s (2012) approach is a useful step; however, it is not required by MOMEL. Empirical evidence is needed to estimate whether the algorithm will give different outputs for stylized and raw contours.

Within the broader picture, our results cast light on previous findings showing that a listener's native language determines the relative significance of F0 and durational ratios in the perception of speech rhythm. Cumming (2011a) showed that (Swiss) German listeners pay much more attention to durational ratios than to tonal cues, while French listeners pay equal attention to durational and tonal cues. It follows that local F0 fluctuations are much more important for the perception of rhythm in French (a language that is rhythmically similar to Italian) than in German (a language that is rhythmically similar to English). Also, the use of pitch and durational cues for rhythmic grouping and segmentation is affected by linguistic experience (Bhatara, Boll, Unger,

Nazzi & Hoehle, 2013; Molnar, Carreiras & Gervain, 2016; Ordin, Polyanskaya, Laka & Nespor, 2017; Tyler & Cutler, 2009), most probably because durational and pitch cues are employed differently as stress and as final boundary markers across languages.

Moreover, speakers of different languages also employ different means to generate rhythmic patterns (Mori, Hori, & Erickson, 2014; Cumming, 2011a; Niebuhr, 2009), and depending on these strategies and the phonetic means used to implement rhythmic patterns in their native language, speakers might weigh acoustic cues differently in their perception of speech rhythm. Mori, Hori and Erickson (2014) investigated rhythm in English produced by American speakers (L1 speech) and Japanese speakers (L2 speech). They found that American speakers constructed rhythmic patterns by alternating long and short vocalic intervals, while Japanese speakers implemented rhythmic patterns through recursive high-low F0 fluctuations. The authors explained that this was due to the preferred strategies for generating rhythmic patterns in the native languages of the English L1/L2 speakers.

These data related to the perception and production of rhythmic patterns can also be explained within the proposed prosodic typology with recourse to the idea of multiple coupled rhythms. As reviewed above, speech exhibits several rhythms at different timescales. Syllabic rhythm operates at the syllabic and sub-syllabic level timescales and is characterized by the durational variability of vocalic and consonantal intervals, syllables, morae (called micro-rhythm in Jun, 2005; 2014). Tonal rhythm manifests by means of tonal alternations and develops at a larger timescale (macro-rhythm). We suggest that all languages exploit the same set of acoustic cues to manifest rhythmic patterns at a certain timescale (e.g., durations of speech intervals for micro-rhythm and F0 excursions for macro-rhythm). Cross-linguistic differences are not due to the set of exploited acoustic cues, but rather to how languages assign relative importance and

perceptual salience to rhythms at different timescales. If a language exhibits strong macro-rhythm, then it is more likely that recursive high-low F0 alternations are more salient than the durational alternations of syllables and sub-syllabic units. If a language exhibits weak macro-rhythm, then durational ratios at the syllabic timescale are perceptually more salient than tonal alternations at a larger timescale. French and Japanese are categorized in Jun (2014) as having stronger macro-rhythm than English or German, and therefore tonal events are expected to outweigh cues pertaining to the durational ratios of sequential speech constituents. This explanation is in line with Jun (2014: 537), who suggested that "the strength of a language's macro-rhythm seems to be inversely correlated with the strength of phonetic realization of stress".

Speech rhythm is widely explored in many fields and has applications beyond the scope of phonetics and phonology. Firstly, speech rhythm plays an important role in speech processing and in language acquisition. For example, rhythm determines segmentation strategies (Cutler, Mehler, Norris, & Segui, 1986; Nazzi, Iakimova, Bertoncini, et al., 2006). A listener selectively attends to utterances with a particular rhythmic pattern distinct from other patterns; this helps the listener focus on the utterances of a particular speaker in the presence of competing, temporally overlapping utterances. Thus, people find it more difficult to focus on a single speech stimulus (i.e., conversation) when a competing stimulus exhibits the same or similar rhythm (Reel & Hicks, 2012). Abnormal rhythms disrupt inter-speaker entrainment in speech (Borrie & Liss, 2014), intelligibility (Tajima et al., 1997) and increase the perceived foreign accent of L2 utterances (Polyanskaya et al., 2017). Rhythmic patterns can be diagnostic of pathological speech produced by patients with ataxic, flaccid-spastic, hypokintetic and hyperkinetic dysarthria (Liss, White, Mattys, et al., 2009). Interestingly, such studies have mainly focused on syllabic rhythm

in the temporal domain (i.e., micro-rhythm), while macro-rhythm (i.e., tonal rhythm) remains understudied. Our work suggests that future studies on the role of rhythm should not be restricted to syllabic rhythms. We suggest that more insights can be obtained in future studies by factoring in the characteristics of tonal rhythm.

## *V. CONCLUSIONS, LIMITATIONS AND FURTHER INVESTIGATION*

Research on cross-linguistic rhythmic differences and on the role of rhythm in speech processing and production should also investigate tonal, or macro-rhythm, both in the temporal and the frequency domains. Our work provides empirical support, based on acoustic analysis, for the model of prosodic typology proposed by Jun (2014).

We understand that typological statements need to be contextualized by a broader range of languages. Our support for the proposed prosodic typology, which takes macro-rhythm into account as one of its major parameters, is preliminary because it only includes one pair of languages that are considered rhythmically different with regard to macro-rhythm. Another limitation relates to bridging between phonological hypotheses and phonetic predictions, and mapping these translations between phonetics and phonology onto macro-rhythm measures: each step has many-to-many dependencies. For example, the relationship between the strength of macro-rhythm and the magnitude of F0 excursions is not linear. Thus, the difference in the excursion size between Italian and English is indeed substantial and reflects phonological differences due to the presence or absence of H/L alternations. However, this phonetic difference could also emerge if two languages exhibited H/L alternating phonological tones, but one was characterized by a large and the other only by a medium difference between adjacent F0 peaks and valleys. Importantly, regular alternations between larger F0 excursions produce the

impression of stronger rhythmicity (Cumming, 2011b). These perceptually informed differences in tonal rhythm are not mapped onto phonological properties related to macro-rhythm in Jun's (2014) typology, yet they are indeed captured by one of the metrics we are proposing. Further work is needed to add another level of dependencies, related to perceived rhythmicity, to phonology/acoustic mapping.

The novelty and significance of our work, however, is that our quantitative *acoustic* data confirm a prosodic typology based on a *phonological* analysis within the AM framework (Jun, 2014). At the same time, the calibration of tonal rhythms in other languages is necessary before this proposed typology can be considered to have been verified by empirical data, and, if necessary, refined accordingly. Our work is the first step in this direction. We think that our results are a promising signal that this enterprise is worthwhile, and that the proposed typological distinction is testable, and is based on concrete acoustic correlates related to the linguistic structure of certain languages.

## *VI. APPENDIX – THE TEXTS READ BY THE SUBJECTS*

ENGLISH:

"The crow and the fox"

A Crow was hungry. He was flying around looking for food. He found a piece of cheese on a window sill. He took it and flew to a tree to eat it. A fox came by, saw the cheese and decided to steal it from the crow. The fox thought: "If the crow opens its beak, the cheese will fall". So the fox began to flatter the crow: "I bet you have a beautiful voice". The crow was flattered and started to sing. The cheese fell down. And the fox jumped, caught it and walked away.

ITALIAN:

“Il corvo e la volpe”

C’era una volta un corvo affamato che volava in cerca di cibo. Vide un pezzo di formaggio sul davanzale di una finestra. Lo prese e volò su un albero per mangiarlo. Ma arrivò una volpe che vide il formaggio e decise di rubarlo al corvo. La volpe pensò: “Se il corvo apre il becco, il formaggio cadrà.” Quindi la volpe iniziò a fargli dei complimenti: “Scommetto che hai una bellissima voce”. Il corvo, lusingato, iniziò a cantare. Ma così il formaggio cadde a terra. Allora la volpe saltò, lo prese e se ne andò via.


## *ACKNOWLEDGEMENT*

The research was supported by the IKERBASQUE – Basque Foundation for Science. The authors also acknowledge financial support from the Spanish Ministry of Economy and Competitiveness (MINECO), through the “Severo Ochoa” Programme for Centres/Units of Excellence in R&D, and through Juan de la Cierva formacion fellowship for junior researchers to the first author.


## *VII. REFERENCES*


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

Batterink, L., & Paller, K. (2017). Online neural monitoring of statistical learning. *Cortex, 90*, 31-45.

Beckman, M., & Pierrehumbert, J. (1986). Intonational structure in English and Japanese. *Phonology 3*, 255-309.

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(5)*, 3828-3843.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5:9/10, 341-345.

Borrie, S., & Liss, J. (2014). Rhythm as a coordinating device: Entrainment with disordered speech. *Journal of Speech, Language and Hearing Research 57*, 815-824.

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.

Cangemi, F., El Zarka, D., Wehrle, S., Baumann, S., & Grice, M. (2016). Speaker-specific intonational marking of narrow focus in Egyptian Arabic. *Proceedings of Speech Prosody* 2016, Boston.

Cumming, R. (2011a). The language-specific interdependence of tonal and durational cues in perceived rhythmicality. *Phonetica 68(4)*, 1–25.

Cumming, R. (2011b). Perceptually informed quantification of speech rhythm in pairwise variability indices. *Phonetica 68*, 256-277.

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

D'Imperio, M. (2002). Italian intonation: An overview and some questions. *Probus 14(1)*, 37–69.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience 19*, 158–164.

Fries, W., & Swihart, A. (1990). Disturbance of rhythm sense following right hemisphere damage. *Neuropsychologia 28*, 1317-1323.

Fujisaki, H. (1997). Prosody, models, and spontaneous speech. In Y. Sagisaka, N. Campbell & N. Higuchi (Eds.). *Computing Prosody: Computational Models for Processing Spontaneous Speech* (pp. 27-42). New York: Springer Verlag.

Goldsmith, J. (1990). *Autosegmental and Metrical Phonology*. Oxford: Blackwell.

Grabe, E., & Low, L. (2002). Acoustic correlates of rhythm class. In C. Gussenhoven & N. Warner (Eds.). *Laboratory Phonology 7* (515-546). New York: Mouton de Gruyter.

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

Henrich, J., Lowit, A., Schalling, E., & Mennen, I. (2006). Rhythmic disturbance in ataxic dysarthria: a comparison of different measures and speech tasks. *Journal of Medical Speech Language Pathology*, *14*, 291-296.

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* (pp.1233-1236). Saarbrücken, Germany.

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: M. Horne (Ed.). *Prosody: Theory and Experiment* (pp. 51-87). Dordrecht: Kluwer Academic Publishers.

Jun, S.-A. (2005). Prosodic Typology. In Jun, S.-A. (ed.) *Prosodic Typology: The Phonology of Intonation and Phrasing*. Oxford: Oxford University Press. 430-458.

Jun, S.-A. (2014). Prosodic typology: By prominence type, word prosody, and macro-rhythm. In: Sun-Ah Jun (Ed.). *Prosodic Typology II: The Phonology of Intonation and Phrasing* (pp. 520-540). Oxford: Oxford University Press.

Liss, J., White, L., Mattys, S., Lansford, K., Lotto, A., Spitzer, S., Caviness, J. (2009). Quantifying speech rhythm abnormalities in the dysarthrias. *Journal of Speech, Language and Hearing Research 52*, 1334-1352.

McAuley, J. D. (2010). Tempo and rhythm. In M. R. Jones, R. R. Fay, & A. N. Popper (Eds.). *Springer Handbook of Auditory Research 36: Music Perception*. New York: Springer Verlag.

Mennen, I., Schaeffler, F. & Docherty, G. (2012). Cross-language difference in f0 range: a comparative study of English and German. *Journal of the Acoustical Society of America 131(3)*, 2249-2260.

Molinaro, N., & Lizarazu, M. *2018). Delta (but not theta)-band cortical entrainment involves speech-specific processing. European Journal of Neuroscience 48(7), 2642-2650.

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.

Motz, B. A., Erickson, M. A., & Hetrick, W. P. (2013). To the beat of your own drum: cortical regularization of non-integer ratio rhythms toward metrical patterns. *Brain and Cognition 81*, 329–336.

Nazzi,T.,Iakimova,G.,Bertoncini, J.,Frédonie,S.,and 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.

Nespor, M., & Vogel, I. (2008). *Prosodic phonology*. Berlin: Mouton de Gruyter.

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

Nolan, F., & Jeon, H.-S. (2014). Speech rhythm: a metaphor? *Transactions of the Royal Society B* 369, 20130396. doi:10.1098/rstb.2013.0396

Ordin, M., & Mennen, I. (2017). Cross-linguistic differences in bilinguals' fundamental frequency ranges. *Journal of Speech, Language and Hearing Research 60*, 1493-1506.

Ordin, M., & Polyanskaya, L., (2014) Development of timing patterns in first and second languages. *System 42*, 244-257.

Ordin, M., & Polyanskaya, L., (2015) Acquisition of speech rhythm in a second language by learners with rhythmically different native languages. *Journal of Acoustical Society of America 138(2)*, 533-545.

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(5)*, 863-876.

Pierrehumbert, J. (1980). The phonology and phonetics of English intonation. PhD thesis, MIT. (available online at https://dspace.mit.edu/bitstream/handle/1721.1/16065/07492108-MIT.pdf?sequence=2, last accessed on 02/02/2019).

Pierrehumbert, J., & Hirschberg, J. (1990). The meaning of intonational contours in the interpretation of discourse. In P. R. Cohen, J. Morgan, & M. E. Pollack (Eds.), *Intentions in communication* (pp. 271–311). Cambridge, MA: MIT Press.

Polyanskaya, L., Ordin, M., & Busà, M. G. (2017) Relative salience of speech rhythm and speech rate on perceived foreign accent in a second language. *Language and Speech 60(3)*, 333-355.

Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America 105 (1)*, 512-521.

Reel, L., & Hicks, C. (2012). Selective auditory attention in adults: effects of rhythmic structure of the competing language. *Journal of Speech, Language and Hearing Research 55*, 89-104.

Selkirk, E. (1986). On derived domains in sentence phonology. *Phonology Yearbook 3*, 371-405.

Strogatz, S. (2004). *Sync.* London: Penguin Books.

Tajima, K., Port, R., & Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. *Journal of Phonetics 25(1),* 1-24.

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

Tyler, M., &Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *Journal of the Acoustical Society of America 126(1)*, 367–376.

van Maastricht, L., Krahmer, E., Swerts, M., & Prieto, P. (in press). Learning direction matters: A study on L2 rhythm acquisition by Dutch learners of Spanish and Spanish learners of Dutch. *Studies in Second Language Acquisition*. doi: 10.1017/S0272263118000062

Wells, J. (1982). *Accents of English*. Vol. 1 and 2. Cambridge: Cambridge University Press.

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

Wilson, S., Pressing, J., & Wales, R. (2002). Modelling rhythm function in a musician post-stroke. *Neuropsychologia 40*, 1494-1505.

Xu, Y. (2005). Speech melody as articulatorily implemented communicative functions. *Speech Communication 46*, 220-251.

FIGURE CAPTIONS:

**FIGURE 1a.** Schematic pitch contours that differ in the temporal domain of macro-rhythm (Jun, 2014: 525). The first contour (1a top) exhibits H and L turning points that are more evenly distributed in time and thus has a stronger macro-rhythm than the second contour (1a bottom).

**FIGURE 1b.** Schematic pitch contours that differ in the frequency domain of macro-rhythm (Jun, 2014: 525). The first contour (1b top) exhibits larger differences in frequency between successive H and L turning points and thus creates a stronger macro-rhythm than the second contour (1b bottom).

**FIGURE 2.** Schematic pitch contours representing a hypothetical situation when two languages (language A – above and language B – below) have differences in variability related to tonal rhythm, thus do not result in differences using the *MacR_Var* index. Language A exhibits high variability in the frequency domain (vertical arrows, corresponding to the magnitude of F0 changes, vary in length) and low variability in the temporal domain (horizontal arrows, corresponding to regularity in the distribution of F0 peaks, are quasi-equal), while language B exhibits low variability in the frequency domain and high variability in the temporal domain.

**FIGURE 3.** Sample pitch contour before and after stylization and correction for errors due to the pitch estimation algorithm.

**FIGURE 4 (a and b).** Typical F0 contours from two random sentences after labelling, Italian (4a) and English (4b). All turning points (T), high turning points (H) and low turning points (L) are labelled. Italian text reads: "…a beautiful voice and – uhhhm – so the crow begins to sing…"

**FIGURE 5.** Cross-language differences in the variability of intervals between consecutive F0 turning points (all – peak-to-valley and valley-to-peak intervals; H – peak-to-peak intervals; L – valley-to-valley intervals).

**FIGURE 6.a** Cross-language differences in the magnitude of F0 excursions, **b** cross-language differences in the number of F0 targets per intonational unit; **c** cross-language differences in the scores of the variation index calculated as the sum of the variation coefficient (*MacR_Var*).

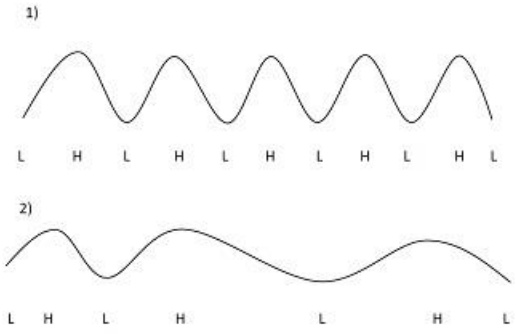
**Table 1.** Phonological hypotheses and phonetic predictions related to cross-language differences in tonal rhythm.

| **Phonological theory: languages differ in tonal rhythm (also referred to as macro-rhythm)**<br><br>**Tonal rhythm depends on** | **Phonological hypothesis related to Italian and English** | **Phonetic consequences** | **Predictions for the scores of the macro-rhythm metrics** |
|---|---|---|---|
| Type of the most frequent tone:<br><br>alternating high and low phonological tones are present in languages with stronger macro-rhythm; a sequence of high or a sequence of low phonological tones are more frequent in languages with weaker macro-rhythm. | In Italian, L*, H* and L+H* are frequent pitch accents, hence high and low phonological tones are more likely to be alternating than in English, where H* is most frequent and thus a sequence of high phonological tones is less likely to alternate with low phonological tones. | Larger F0 span between consecutive F0 turning points in Italian than in English.<br><br>Larger variability in temporal distribution of F0 maxima and F0 minima turning points in English than in Italian | Larger F0 magnitude between consecutive F0 maxima and F0 minima in Italian than in English.<br><br>Lower scores of nPVI and Varco metrics for F0 peak-to-peak and valley-to-valley intervals in Italian than in English |
| Size of the tonal inventory:<br><br>Larger inventory of mid-phrase-medial pitch accents are more characteristic of languages with weaker macro-rhythm | English has more different types of phrase-medial pitch accents in the phonological inventory than Italian. | Similar shape of sub-tonal units (i.e., similar slope of F0 rising and F0 falling contours).<br><br>Larger variability in temporal distribution of F0 maxima and F0 minima turning points in English than in Italian | *MacR_Var* scores are higher in English than in Italian.<br><br>Lower scores of nPVI and Varco metrics for F0 peak-to-peak and valley-to-valley intervals in Italian than in English. |
| Number of tones per phonological word:<br><br>Languages with strongest macro-rhythm tend to have one pitch accent per phonological word, languages with weaker macro rhythm have less pitch accents than phonological words. | Larger number of phonological tones per phonological word in Italian than in English | Larger number of F0 turning points per intonational unit in Italian than in English. | |

**Table 2.** Parameter estimates of the linear mixed models

| | Temporal domain | | | | | | Frequency domain | Number of F0 targets per IP | Shape of tonal contours |
|---|---|---|---|---|---|---|---|---|---|
| | **nPVI (H)** | **nPVI (L)** | **nPVI (all)** | **Varco (H)** | **Varco (L)** | **Varco. (all)** | **F0 excursion width** | | **MacR_Var** |
| ***b*** | -.053 | -.133 | -.1 | -.013 | -.09 | -.029 | 8.96 | 2.74 | .173 |
| **SE** | .065 | .052 | .035 | .047 | .04 | .034 | 3.82 | .69 | .127 |
| ***t*-stat.** | -.814 | -2.55 | -2.886 | -.288 | -1.93 | -.859 | 2.35 | 3.97 | 1.37 |
| ***p*** | .417 | .023 | .004 | .771 | .072 | .391 | .03 | .001 | .204 |
| **95% CI** | [-.18:.08] | [-.25:-.02] | [-.17:-.03] | [-.11:.79] | [-.19:.01] | [-.1:.04] | [.97:16.96] | [1.29:4.21] | [-.113:.459] |

FIGURE 1a (left) and 1b(right)



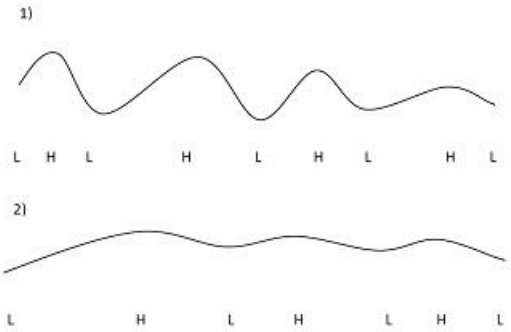



FIGURE 2.

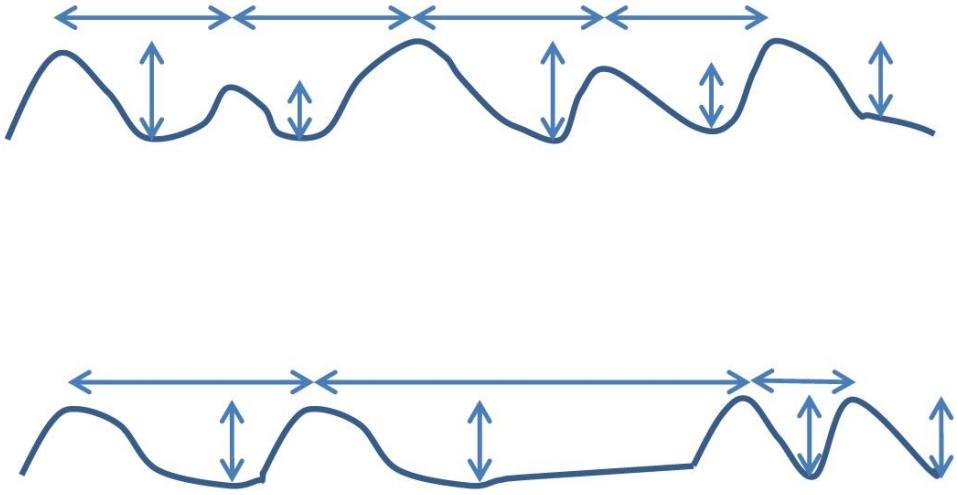
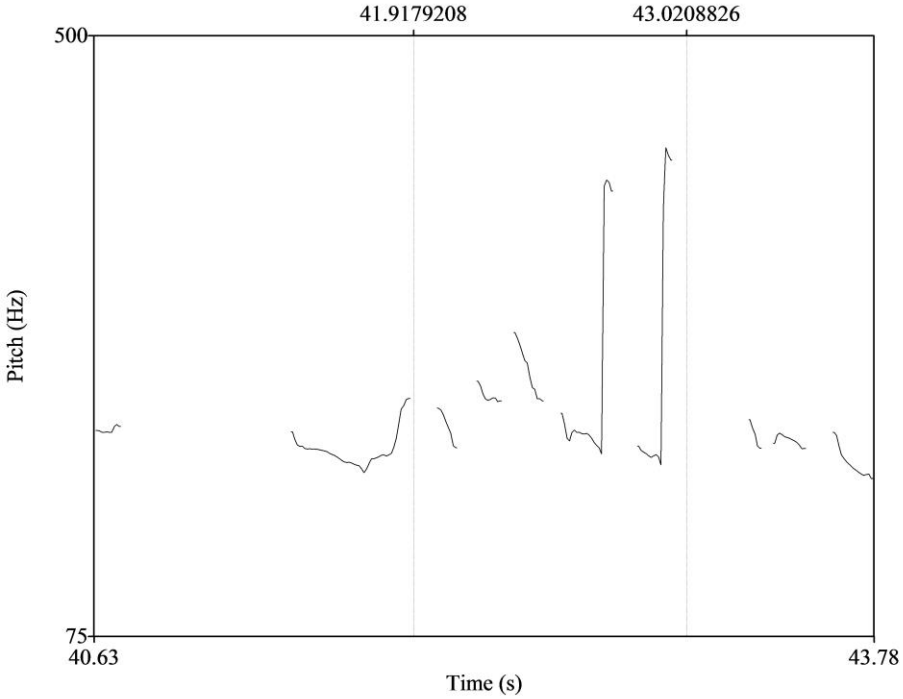


FIGURE 3.


41.9179208
43.0208826
500
75
Pitch (Hz)
40.63
43.78
Time (s)

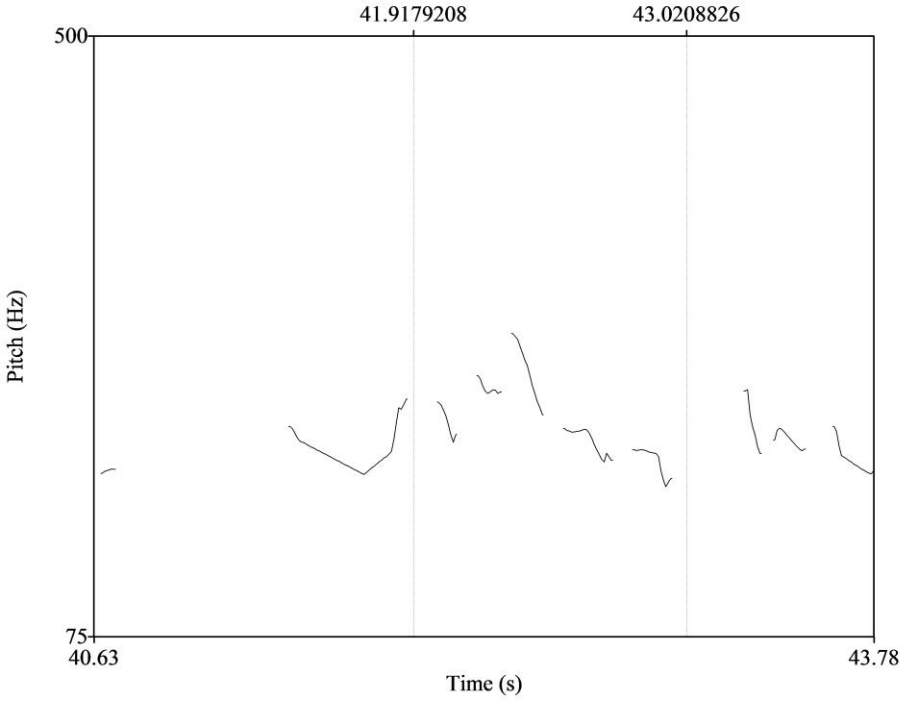

41.9179208
43.0208826
500
75
Pitch (Hz)
40.63
43.78
Time (s)

FIGURE 4a (above) and 4b (below).

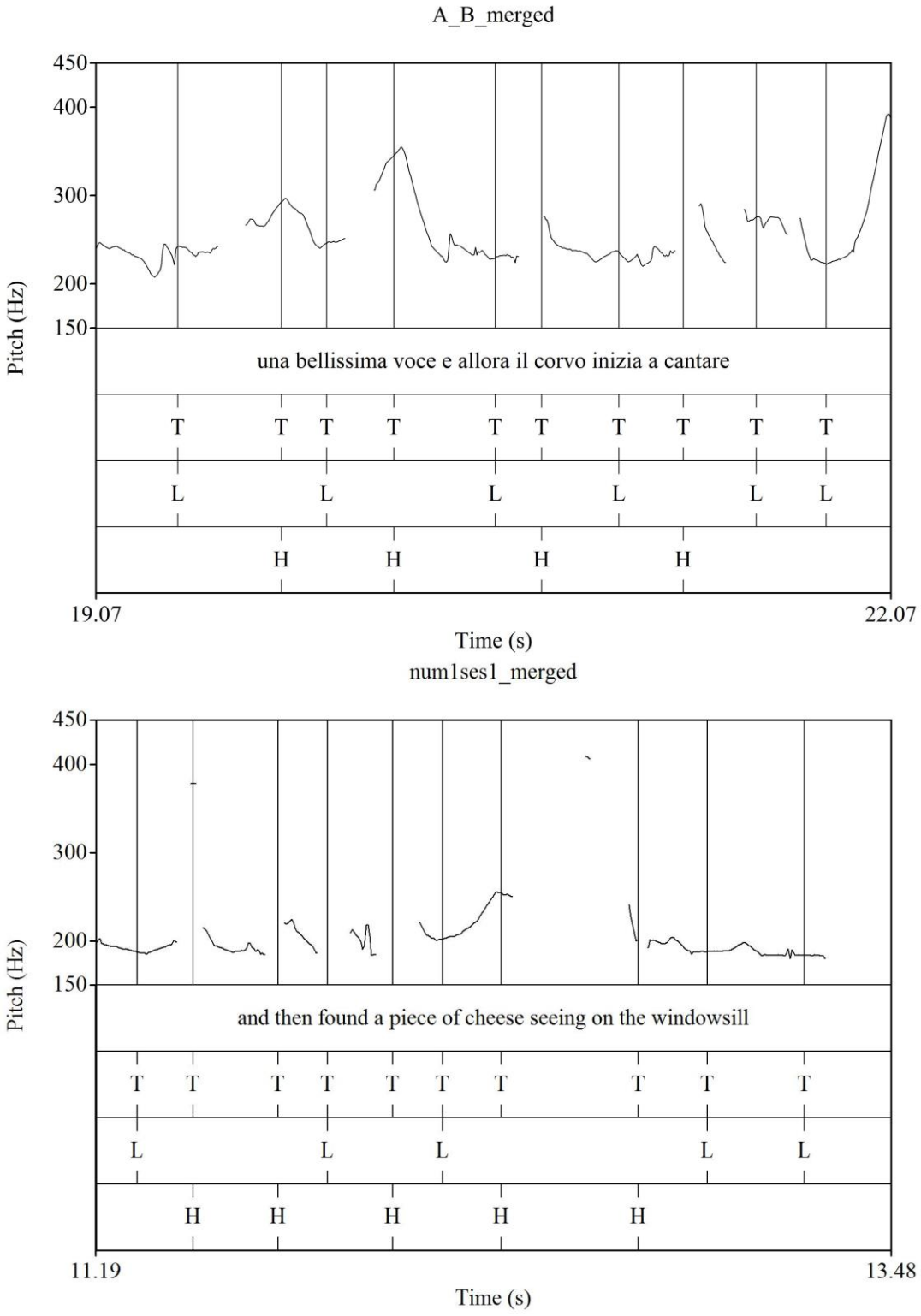


FIGURE 5.

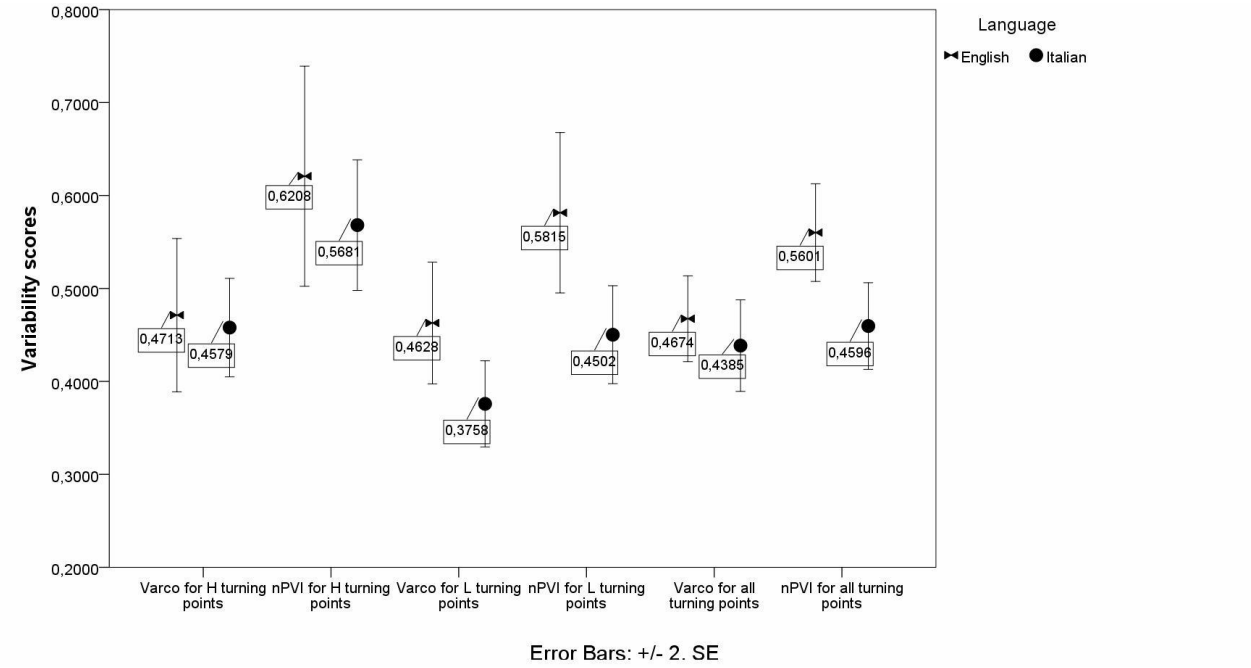

Language
English
Italian
Variability scores
0,8000
0,7000
0,6000
0,5000
0,4000
0,3000
0,2000
0,4713
0,4579
0,6208
0,5681
0,4628
0,3758
0,5815
0,4502
0,4674
0,4385
0,5601
0,4596
Varco for H turning points
nPVI for H turning points
Varco for L turning points
nPVI for L turning points
Varco for all turning points
nPVI for all turning points
Error Bars: +/- 2. SE

FIGURE 6a (above left), 6b(above right), 6c (below left), 6b(below right).