INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Parsing speech for grouping and prominence, and the typology of rhythm

Michael Wagner, Alvaro Iturralde Zurita, and Sijia Zhang

McGill University, Canada

chael@mcgill.ca, alvaro.iturraldezurita@mail.mcgill.ca, sijia.zhang2@mail.mcgill.ca

Abstract

Humans appear to be wired to perceive acoustic events rhythmically. English speakers, for example, tend to perceive alternating short and long sounds as a series of binary groups with a final beat (iambs), and alternating soft and loud sounds as a series of trochees. This generalization, often called the 'Iambic-trochaic Law' (ITL), although viewed as an auditory universal by some, has been argued to be shaped by language experience. Earlier work on the ITL had a crucial limitation, in that it did not tease apart the percepts of grouping and prominence, which the notions of iamb and trochee inherently confound. We explore how intensity and duration relate to percepts of prominence and grouping in six languages (English, French, German, Japanese, Mandarin, and Spanish). The results show that the ITL is not universal, and that cue interpretation is shaped by language experience. However, there are also invariances: Duration appears relatively robust across languages as a cue to prominence (longer syllables are perceived as stressed), and intensity for grouping (louder syllables are perceived as initial). The results show the beginnings of a rhythmic typology based on how the dimensions of grouping and prominence are cued.

Index Terms: speech segmentation, prominence, intonation, prosody, stress, phrasing

1. Introduction

Humans often impose a rhythmic interpretation on sound sequences when are on a time scale comparable to the acoustic events in human language. English speakers tend to perceive alternating short and long tones as series of iambs, but alternating soft and loud tones as a series of trochees [1, 2]. This generalization, often called the 'Iambic-trochaic Law' (ITL), following [3], also applies in speech [4, 5].

While some take the ITL to be a universal of auditory processing [4], others found that at least the duration-side of the ITL works differently in languages like Japanese [6, 7] , French [5], Spanish, and Zapotec [8], suggesting that rhythm perception is shaped by language experience. This cross-linguistic variation has been attributed to crosslinguistic differences in word order [7] or stress-systems [5], among other factors.

The prior work on the ITL has a crucial limitation. If one parses a sequence of sounds (e.g. iterations of the syllables 'ba' and 'ga') into binary groups, there are (at least) 4 potential percepts (e.g., BAga, baGA, GAba, gaBA). Prior research, however, used binary forced choice tasks to investigate the phenomenon. Most studies asked participants, in some way or other, about which foot they heard (e.g., *Did you hear [X x] or [x X]?* [2, 9, 4, 7], a task which confounds prominence and grouping. Other studies, including the acquisition literature on the ITL, used speech segmentation tasks (e.g., 'Did you hear *baga* or *gaba*?') [10, 11, 12, 13, 8, 14]. This task leaves open where prominence was perceived. Here, we ask participants two forced choices instead, one about grouping, and one about prominence, thereby teasing apart the two dimensions.

In English, the ITL emerges from the rational use of the cue distribution, assuming that listeners parse the signal along the dimensions of prominence and grouping [15]. Production experiments have shown that in English, intensity and duration correlate when they cue prominence, but anti-correlate when encoding grouping, both in phrases [16, 17] and in words [15]. The duration effect of grouping is due to well-known effects of word- and phrase-final lengthening [18]. Intensity cues to grouping anti-correlate with duration because intensity decreases throughout phrases and words, and resets at the beginning of a new phrase [19, 16, 17] or word [15]. These effects do not reduce to utterance-level downdrift. They are related to initial strengthening [20], but go far beyond their effect on the first segment. The ITL can hence be accounted for as follows [15]: Listeners perceive an exceptionally long sound as final and prominent, leading to the percept of iambs; and an exceptionally loud sound as as initial and prominent, leading to the percept of trochees.

Here, we report on a crosslinguistic comparison between six different languages of how such speech sequences are perceived: European French, German, Japanese, Mandarin, Mexican Spanish, and North American English. The goal is to replicate the English results in [15], and to extend this work to establish the beginnings of a rhythmic parsing typology, by quantifying how listeners of different languages parse the signal for prominence and grouping.

Intuitions about rhythmic differences between languages have long been reported, but attempts to quantify these have encountered difficulties. For example, the intuition that there are syllable-timed and stress-timed languages has been argued to be due to a confluence of several orthogonal dimensions of variation, such as variation in syllable complexity and degree of vowel reduction [21]. Existing quantitative measures of rhythm (e.g. [22]) have been criticized as tapping phonotactics or phonemic differences rather than what we intuitively call 'rhythm (see [23], i.a.). The typology based on the parsing of the dimensions of grouping and prominence proposed here may provide a better quantitative map of cross-linguistic differences in what we might intuitively want to call 'rhythm.'

2. Methods

We conducted six perception experiments, one for each language, using the ProsodylabExperimenter [24], a javascript tool which makes use of JsPsych [25] to facilitate running online experiments. Participants listened to sequences of syllables, and then answered questions. They were recruited on the crowdsourcing website Prolific (English, Spanish, French, German) and Crowdworks (Japanese). Mandarin speakers were recruited by the third author through social media in China and among McGill students. In English, we ran 54 participants (23 female/mean age 33/mean years of musical training 1.9); German: 59 (17/31/2); French 48 (13/29/1.9); Spanish 50 (12/26/1.2); Mandarin 18 (13/23/2.0); Japanese 57 (28/42/1.4). All instruc-

http://dx.doi.org/10.21437/Interspeech.2021-1684

tions were in the language under study.

The sound sequences were modeled after [10, 8, 14, 15] and alternated the syllables *ba* and *ga*. To create the sequences, we synthesized the two syllables *ba* and *ga* using Amazon Polly. We used a Praat [26] script to create syllable sequences out of these. We first scaled each syllable to an average intensity of 70dB, a length of 240ms, and a constant pitch at 100Hz. The syllables were then concatenated to from the base-words *baga* and *gaba*. To create a particular speech sequence, we then manipulated the syllables in these words for intensity (increasing average intensity by 0, 3, 6, or 9dB) and duration (increasing duration by 0, 40, 80, or 120 ms). Each cue was manipulated only on one of the two syllables of the word for a given word. So a given syllable could be 3 steps louder than the other, or three steps softer, creating a 7-point scale, and similarly for duration, for a total of 49 different manipulations of each *baga* and *gaba*. Sequences were created out of these words by repeating the manipulated word 12 times. We added a fade-in ramp at the beginning of the sequence (based on a half-cosine function), and superimposed white noise, which faded out while the sound sequence faded in, in order to decrease order effects. Each participant encountered both the *gaba* and *baga* baseline, and one version of each of the other 48 manipulation steps, 25% from the *baga* set, and 25% from the *gaba* set, for a total of 50 trials.

On each trial, listeners had to answer two binary forced choice questions. Half the participants were first asked *Which word did you hear?* with choices *baga* and *gaba*. And then *Which syllable in the word did you hear as prominent?*, with choices *BAga* or *baGA*, if they responded *baga* to the first question. The other half answered the questions in the opposite order. We varied the order of the answer buttons for each question randomly between participants, but used the same order within participant to avoid confusion.

For English, Spanish, and German, 'prominence' was explained as lexical word stress. For the other languages, we explained prominence based on an an example involving contrast, which was illustrated using minimal pairs that differed in their first or second syllable respectively.

Listeners also filled out a language questionnaire before the experiment, as well as a music questionnaire and post-experiment questionnaire asking about what they thought the experiment was about. They also participated in a modified version of head-phone screener task in [27] to ensure they were actually wearing headphones, as we requested. We do not have the space to report on the questionnaire results here. All experiments, except for the one on German and Mandarin, were preregistered on OSF, as part of the project https://osf.io/v25kd/.

3. Results

The results, illustrated in Figure 1, replicate the perception findings [15] for English, but also extend them to 5 more languages. We see that in English, listeners make consistent prominence choices when intensity and duration correlate, since in that case there are consistent cues for prominence (bottom left and top right corner of heatmap), and are closer to chance when they anti-correlate, leading to a 'diagonal of uncertainty' in the heatmap (response proportions at 50% are colored green).

When the cues anti-correlate, intensity and duration give consistent cues for grouping (top left and bottom right corner of heatmaps), while responses are closer to chance when the two cues correlate, leading to a diagonal of uncertainty perpendicular to the one observed for prominence.

Similar patterns are observed crosslinguistically, but the importance of each cue for a particular decision varies to some degree. For example, duration plays less of a role in cueing grouping in Japanese and Mandarin.

We can reconstruct the foot choice (*Iamb or trochee?*) from the grouping and prominence decision. The original ITL pattern is observable in the plots for the foot decisions in English and German, where extreme manipulations of only duration (the end points of the middle rows of the heatmaps) lead to iambic responses, and extreme manipulations of only intensity (the end points of the middle columns) to a trochaic pattern. The ITL pattern is not attested, however, in the other languages, and overall mostly trochees were perceived. The most extreme case was Japanese in this regard. The choice between iamb and trochee shows little systematicity compared to the highly systematic prominence and grouping decisions. Instead, the data is better understood as the result of the grouping and prominence decisions.

We fit logistic ME regression models for the prominence and grouping decision for each language using the R-package *lme4* [28], summarized in Tables 1 and 2. Predictors were scaled in order to make effect sizes comparable across different cues, and dependent variables were scaled to make effect sizes comparable across models.

We see that intensity and duration significantly affect both decisions in all languages, except that duration was not significant as a predictor of the grouping decision in Japanese and Mandarin. The direction of all coefficients for intensity and duration for each decision was the same across languages, with one exception: There was a very small, non-significant coefficient in the opposite duration for intensity in the grouping decision in Japanese. The size of the effects, however, varies substantially by language.

In all languages, the grouping decision (baga.vs.gaba in the model) is a significant predictor of the prominence decision, and the prominence decision (ba.vs.ga) is a significant predictor of grouping. This is an interesting difference to the results reported in [15], where the decisions also informed each other, but only by way of an interaction. Our fade-in manipulation did not completely remove order effects, since there was a significant effect of underlying syllable order (ba.vs.gaStart) on grouping in all languages.

We visualized the differences between languages by plotting the coefficients from the models for both decisions in each language in Fig. 2. While cue interpretation across languages is not universal, the plots illustrate that the effect of intensity on the grouping decision, and the effect of duration on the prominence decision, are comparatively consistent across languages.

To assess whether the apparent language differences were significant, we fit an Omnibus model that included interactions of various predictors with Language, which we do not report in full due to lack of space. This model included interactions between Language and intensity and Language and duration, as well as the underlying order (*baga* vs. *gaba*). The effect of duration on prominence perception was significantly different from English only in German, where it was bigger; the effect of intensity on prominence was significantly different from English only in Japanese, where it was bigger.

With respect to grouping, the effect of duration on grouping was significantly different from English in Spanish (replicating findings in [14]), Mandarin, and Japanese (in all three, the effect was in the same direction but smaller in size). The effect of intensity on grouping was not significantly different from English in any language, but approached significance in German ($p < 0.055$), where the trend was that the effect was bigger.

4. Conclusions

Our results replicate the perception results for English reported in [15]. The ITL pattern is observed in English and German, but not in the other languages. Foto choice in general did not relative in straightforward to the cues under investigation. This is compatible with the conclusion in [15], that the notions of iamb and trochee are epiphenomenal to the explanation of the ITL. A much clearer pattern emerges when looking at how the cues convey the dimensions of prominence and grouping. There was cross-linguistic variation in how intensity and duration are interpreted, but also some consistency. All languages had a sizeable effect in the same direction for how duration affects prominence, and how intensity affects grouping.

In all languages the two decisions mutually influence each other, as expected if grouping and prominence provide competing explanations for the cues. This is similar to other perceptual domains, for example judgments about the size and distance of an object, which mutually inform each other, since they 'explain' overlapping aspects of the incoming cues, even if they are in principle orthogonal dimensions.

If these results hold up crosslinguistically, they have important implications. Prosody has been argued to help solve the bootstrapping problem that language learners face when trying to parse the signal into words and phrases (e.g. [29]). Given the results here, learners could use the more invariant cue (duration for prominence; intensity for grouping) to parse the signal. This could also help explain why adults can segment speech into words even in languages they don't know [30]. Syllabification from the signal is already possible (e.g. [31]), and taking cue relation into account might make it possible to parse the syllables along the dimensions of grouping and prominence in addition.

The results shed new light on existing findings about the acquisition of the ITL. ITL-effects for intensity have been found to be cross-linguistically more robust and acquired earlier compared to duration effects [11, 12, 13, 32, 33, 34]. However, these results are largely based on speech segmentation tasks, which tap grouping, where we found intensity to be more robust. Maybe early on, children interpret duration mostly as a cue to prominence, and intensity as a cue for grouping.

The mutual influence of grouping and prominence decisions suggests that segmenting speech into words, whether by a language learner or an AI algorithm, may actually be easier if one tries to parse for prominence at the same time. This is consistent with the finding in [35] that making assumptions about stress helps with speech segmentation. However, the particular proposal [35], that each word in a language carries a single stress, is not realistic even for English, since word often have multiple stress and sometimes carry more than one pitch accent. Rather than 'building in' a particular assumptions about the number of stresses per word, it might be sufficient to build in the prior assumption that grouping and prominence (possibly along with other dimensions) are what we listen for when listening to speech. The cue distribution for each dimension could be established based on single-word or single-phrase utterances, and then extrapolated to more complex utterances.

The experiments had various limitations. First, we have only looked at a few languages, and a broader typological perspective has to be taken to come to firmer conclusions. Second, we used the same stimuli across languages, which were originally designed based on English. They will probably not sound native-like in other languages. Third, we looked at intensity and duration, but there are also pitch and spectral cues for both prominence and grouping, which would deserve equal attention, even if they are not part of what is referred to as the ITL, and we also did not manipulate pause duration between sounds. It might also be worthwhile to consider alternative measures of vocal effort than intensity, since intensity is subject to large and irrelevant variation in more realistic listening conditions. There are also methodological limitations, given differences in recruitment and sample size, especially with respect to Mandarin.

Finally, our results do not provide any insight yet as to why the languages differ in the way they do. [7] argued in a head-final language like Japanese, there are more constituents with a long-before-short pattern compared to English, hence length will be interpreted a cue to initiality rather than finality. This is compatible with the new finding here that the difference between English and Japanese indeed resides in the grouping dimension. Earlier studies used the foot choice task, which confounds grouping and prominence. This account assumes that based on contingent properties of the given language, a given cue may convey opposite information about a particular source, in this case grouping.

Another possibility, however, is that differences arise because listeners attribute aspects of the signal to different sources. How cues are interpreted has been found to depend on linguistic background, even in non-speech stimuli. For example, pitch modulates the percept of duration differently crosslinguistically [36]. Such differences may reveal that cues can be attributed to different explanations. For example, when judging the duration of a stimulus, speakers might attribute some proportion of the length to the effects of the intonational tune, which they might regard as task irrelevant. Whether a given cue (for example pitch), will be disregarded in a duration estimation task may depend on the effect of intonation on duration in the language, or conversely on whether pitch is used as a cue to phonemic length. In our experiments, duration may be disregarded as a cue to grouping or prominence because it is interpreted as conveying phonemic information. Japanese distinguishes short and long vowels, or rather vowels that are monomoraic versus bi-moraic. Surplus length could be attributed to bi-moraic vowels instead of grouping. If these are perceived as more prominent, this would leave duration effects on prominence intact. In Mandarin, length might be attributed to the effect of lexical tone [37], rather than as a cue to grouping. So depending on the linguistic structures present in their language, listeners might posit different 'auditory descriptions' [38] that explain the signal. Some of these confounds can be controlled more easily in production studies, and we are currently conducting nonce-word production studies, similar to the English production study reported in [15], to see whether the cue distribution in production matches the effects in perception in a given language.

Does the emerging parsing typology based on grouping and prominence cues capture the notion of rhythmic differences between languages? While it may be that linguistic rhythm is really just a metaphoric extension of 'true' musical rhythm, as argued in [39], we think that the combination of prominence and grouping at least captures a core ingredient of what we experience as rhythm when we listen to speech.

5. Acknowledgements

Thanks also to the audiences at the Cuny sentence processing conference and at a colloquium at UPenn in March 2021. The research was conducted under McGill ethics protocol REB#: 401-0409, and funded by NSERC Discovery Grant RGPIN-2018-06153: *Three dimensions of sentence prosody*

Table 1: *Logistic ME models for prominence decision. Fixed effects included intensity and duration step, and their interaction with the grouping decision. A random effect for participant included slopes for intenisty and duration.*

	English	German	French	Spanish	Mandarin	Japanese
(Intercept)	0.10 (0.07)	−0.03 (0.09)	−0.22 (0.07)**	−0.12 (0.08)	0.24 (0.14)	−0.15 (0.09)
baLouder	−0.17 (0.04)***	−0.06 (0.03)*	−0.21 (0.04)***	−0.15 (0.04)***	−0.13 (0.05)**	−0.32 (0.05)***
baLonger	−0.41 (0.05)***	−0.64 (0.06)***	−0.40 (0.06)***	−0.35 (0.07)***	−0.60 (0.13)***	−0.39 (0.06)***
baga.vs.gaba	0.93 (0.11)***	1.17 (0.11)***	0.78 (0.11)***	0.82 (0.10)***	0.75 (0.19)***	2.74 (0.13)***
ba.vs.gaStart	0.12 (0.09)	0.19 (0.09)*	−0.05 (0.09)	−0.10 (0.09)	0.17 (0.17)	−0.27 (0.10)**
baLouder:baga.vs.gaba	−0.01 (0.05)	−0.08 (0.05)	−0.04 (0.05)	0.04 (0.05)	0.07 (0.09)	0.04 (0.06)
baLonger:baga.vs.gaba	−0.06 (0.05)	0.00 (0.06)	0.02 (0.06)	−0.04 (0.05)	0.17 (0.10)	−0.02 (0.06)

$^{***}p < 0.001; ^{**}p < 0.01; ^{*}p < 0.05$

Table 2: *Logistic ME models for grouping decision, which additionally includes the underlying syllable order (ba.vs.gaStart).*

	English	German	French	Spanish	Mandarin	Japanese
(Intercept)	−0.07 (0.09)	0.11 (0.15)	−0.29 (0.09)**	−0.23 (0.12)	−0.33 (0.14)*	−0.91 (0.19)***
baLouder	−0.27 (0.03)***	−0.45 (0.05)***	−0.22 (0.04)***	−0.26 (0.03)***	−0.31 (0.04)***	−0.26 (0.05)***
baLonger	0.56 (0.09)***	0.63 (0.09)***	0.41 (0.07)***	0.24 (0.06)***	0.10 (0.15)	−0.03 (0.10)
ba.vs.ga	0.89 (0.11)***	1.16 (0.12)***	0.80 (0.11)***	0.82 (0.11)***	0.68 (0.19)***	2.82 (0.14)***
ba.vs.gaStart	0.47 (0.09)***	1.16 (0.10)***	0.65 (0.10)***	0.41 (0.09)***	1.25 (0.17)***	0.18 (0.11)
baLouder:ba.vs.ga	−0.05 (0.05)	−0.18 (0.05)***	−0.06 (0.05)	−0.05 (0.05)	0.11 (0.09)	−0.02 (0.06)
baLonger:ba.vs.ga	−0.02 (0.06)	0.13 (0.06)*	0.04 (0.06)	−0.04 (0.05)	0.26 (0.10)**	−0.04 (0.07)

$^{***}p < 0.001; ^{**}p < 0.01; ^{*}p < 0.05$

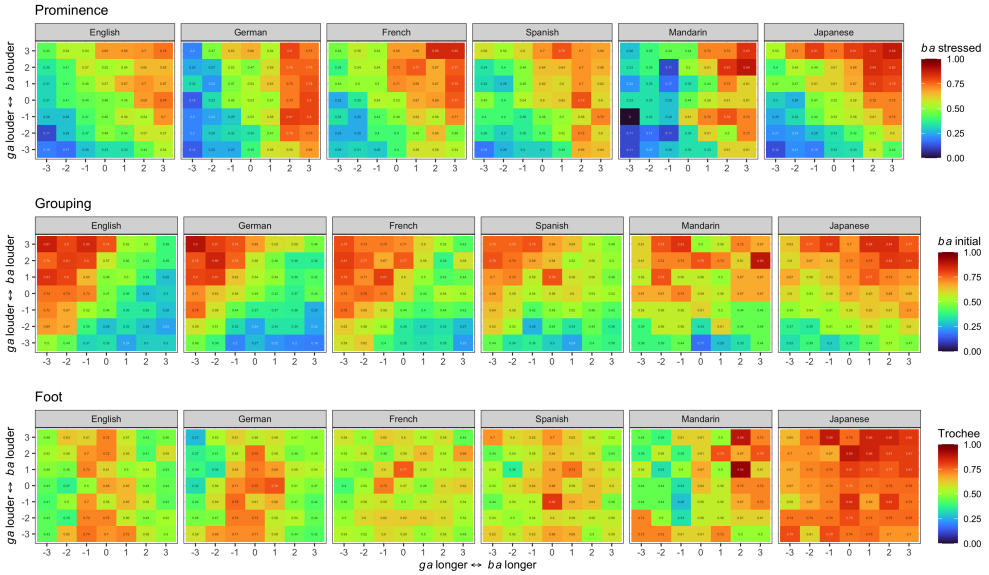

Figure 1: *Responses for each language for the prominence and grouping decisions, and the foot decision reconstructed from those two responses. The heatmap uses color scheme from dark red (100%) via green to dark blue (0%). The degrees of shading are color-blind compatible, but the polarity (above or below 50%) will be lost. Zoom in for actual proportions*

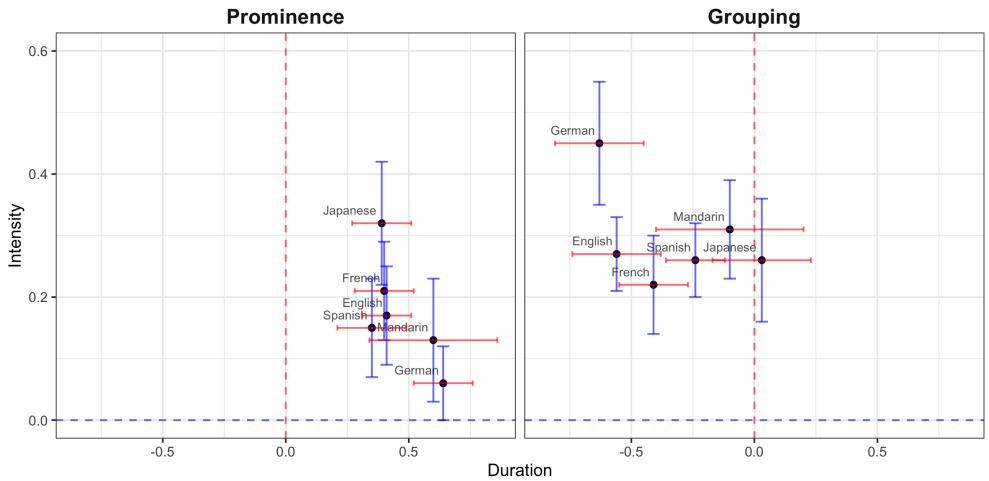

Figure 2: *The coefficients from logistic MER models for the individual languages. Coefficients correspond to the predicted change in log odds given a unit change in intensity/duration. Duration (x-axis) and intensity (y-axis) coefficients are shown for the prominence decision (left) and the grouping decision (right). Error bars show 2*se estimated by the logistic models.*

6. References

[1] T. L. Bolton, "Rhythm," *The American Journal of Psychology*, vol. 6, no. 2, pp. 145–238, 1894.

[2] H. Woodrow, "A quantitative study of rhythm: The effect of variations in intensity, rate and duration," *Archives of Psychology*, no. 14, pp. 1–66, 1909.

[3] B. Hayes, *Metrical Stress Theory: Principles and Case Studies*. Chicago: University of Chicago Press, 1995.

[4] J. F. Hay and R. L. Diehl, "Perception of rhythmic grouping: Testing the iambic/trochaic law," *Perception and Psychophysics*, vol. 69, no. 1, pp. 113–122, 2007.

[5] A. Bhatara, N. Boll-Avetisyan, A. Unger, T. Nazzi, and B. Höhle, "Native language affects rhythmic grouping of speech," *The Journal of the Acoustical Society of America*, vol. 134, no. 5, pp. 3828–3843, 2013.

[6] K. Kusumoto and E. Moreton, "Native language determines the parsing of nonlinguistic rhythmic stimuli," *The Journal of the Acoustical Society of America*, vol. 102, no. 5, pp. 3204–3204, 1997.

[7] J. R. Iversen, A. D. Patel, and K. Ohgushi, "Perception of rhythmic grouping depends on auditory experience," *The Journal of the Acoustical Society of America*, vol. 124, no. 4, pp. 2263–2271, 2008.

[8] M. Crowhurst and A. Teodocio Olivares, "Beyond the iambic-trochaic law: the joint influence of duration and intensity on the perception of rhythmic speech," *Phonology*, vol. 31, no. 01, pp. 51–94, 2014.

[9] C. Rice, "Binarity and ternarity in metrical theory: Parametic extensions," Ph.D. dissertation, University of Texas, Austin, 1992.

[10] B. Höhle, R. Bijeljac-Babic, B. Herold, J. Weissenborn, and T. Nazzi, "Language specific prosodic preferences during the first half year of life: Evidence from German and French infants," *Infant Behavior and Development*, vol. 32, no. 3, pp. 262–274, 2009.

[11] R. A. Bion, S. Benavides-Varela, and M. Nespor, "Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences," *Language and speech*, vol. 54, no. 1, pp. 123–140, 2011.

[12] K. A. Yoshida, J. R. Iversen, A. D. Patel, R. Mazuka, H. Nito, J. Gervain, and J. F. Werker, "The development of perceptual grouping biases in infancy: A Japanese-English cross-linguistic study," *Cognition*, vol. 115, no. 2, pp. 356–361, 2010.

[13] J. F. Hay and J. R. Saffran, "Rhythmic grouping biases constrain infant statistical learning," *Infancy*, vol. 17, no. 6, pp. 610–641, 2012.

[14] M. Crowhurst, "Iambic-trochaic law effects among native speakers of Spanish and English," *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, vol. 7, no. 1, pp. 1–41, 2016.

[15] M. Wagner, "Two-dimensional parsing explains the iambic-trochaic law," *Psychological Review*, 2021.

[16] M. Wagner and M. McAuliffe, "Three dimensions of sentence prosody and their (Non-)Interactions," in *Proceedings of Interspeech 2017 in Stockholm*, 2017.

[17] ——, "The effect of focus prominence on phrasing," *Journal of Phonetics*, vol. 77, pp. 1–26, 2019.

[18] D. K. Oller, "The effect of position in utterance on speech segment duration in English," *The Journal of general psychologyurnal of the Acoustical Society of America*, vol. 54, no. 5, pp. 1235–1247, 1973.

[19] C. Poschmann and M. Wagner, "Relative clause extraposition and prosody in German," *Natural Language & Linguistic Theory*, vol. 34, no. 3, pp. 1021–1066, 2016.

[20] P. Keating, T. Cho, C. Fougeron, and C.-S. Hsu, "Domain-initial strengthening in four languages," in *Phonetic Interpretation: Papers in Laboratory Phonology VI*, J. Local, R. Ogden, and R. Temple, Eds. Cambridge, UK: Cambridge University Press, 2003, pp. 143–161.

[21] R. Dauer, "Stress-timing and syllable-timing reanalyzed," *Journal of Phonetics*, vol. 11, no. 1, pp. 51–62, 1983.

[22] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 73, no. 3, pp. 265–292, 1999.

[23] A. Arvaniti, "The usefulness of metrics in the quantification of speech rhythm," *Journal of Phonetics*, vol. 40, no. 3, pp. 351–373, 2012.

[24] M. Wagner, "ProsodylabExperimenter. a spreadsheet-in spreadsheet-out tool for running online experiments," 2021. [Online]. Available: https://github.com/prosodylab/prosodylabExperimenter

[25] J. R. De Leeuw, "jspsych: A javascript library for creating behavioral experiments in a web browser," *Behavior research methods*, vol. 47, no. 1, pp. 1–12, 2015.

[26] P. Boersma and D. Weenink, "PRAAT, a system for doing phonetics by computer. Report 132." 1996, institute of Phonetic Sciences of the University of Amsterdam.

[27] K. J. Woods, M. H. Siegel, J. Traer, and J. H. McDermott, "Headphone screening to facilitate web-based auditory experiments," *Attention, Perception, & Psychophysics*, vol. 79, no. 7, pp. 2064–2072, 2017.

[28] D. Bates, M. Mächler, B. Bolker, and S. Walker, "Fitting linear mixed-effects models using lme4," *Journal of Statistical Software*, vol. 67, no. 1, pp. 1–48, 2015.

[29] T. Christophe, Anne andn Guasti and M. Nespor, "Reflections on phonological bootstrapping: Its role for lexical and syntactic acquisition," *Language and cognitive processes*, vol. 12, no. 5-6, pp. 585–612, 1997.

[30] A. D. Endress and M. D. Hauser, "Word segmentation with universal prosodic cues," *Cognitive psychology*, vol. 61, no. 2, pp. 177–199, 2010.

[31] O. Räsänen, G. Doyle, and M. C. Frank, "Pre-linguistic segmentation of speech into syllable-like units," *Cognition*, vol. 171, pp. 130–150, 2018.

[32] M. Molnar, M. Lallier, and M. Carreiras, "The amount of language exposure determines nonlinguistic tone grouping biases in infants from a bilingual environment," *Language Learning*, vol. 64, no. s2, pp. 45–64, 2014.

[33] A. Bhatara, N. Boll-Avetisyan, T. Agus, B. Höhle, and T. Nazzi, "Language experience affects grouping of musical instrument sounds," *Cognitive Science*, vol. 40, no. 7, pp. 1816–1830, 2016.

[34] N. Abboub, N. Boll-Avetisyan, A. Bhatara, B. Höhle, and T. Nazzi, "An exploration of rhythmic grouping of speech sequences by French- and German-learning infants," *Frontiers in Human Neuroscience*, vol. 10, p. 292, 2016.

[35] C. D. Yang, "Universal grammar, statistics or both?" *Trends in cognitive sciences*, vol. 8, no. 10, pp. 451–456, 2004.

[36] J. Šimko, D. Aalto, P. Lippus, M. Włodarczak, and M. Vainio, "Pith, perceived duration and auditory biases: Comparison among languages," in *18th International Congress of Phonetic Sciences, Glasgow, Scotland, UK, August 10-14, 2015*, 2015.

[37] C. L. Yu, Alan, "Tonal effects on perceived vowel duration," in *Papers in Laboratory Phonology*, C. Fougeron, B. Kühnert, M. D'Imperio, and N. Vallée, Eds. Mouton De Gruyter, 2010, vol. 10, pp. 151–168.

[38] A. S. Bregman, "Asking the "what for" question in auditory perception," in *Perceptual organization*, M. Kubovy and J. R. Pomerantz, Eds. Routledge, 1981, pp. 99–118.

[39] F. Nolan and H.-S. Jeon, "Speech rhythm: a metaphor?" *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 369, no. 1658, 2014.