Two-dimensional parsing of the acoustic stream explains the iambic-trochaic law

Michael Wagner

McGill University

Author Note

To appear in *Psychological Review*. © 2021, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: https://doi.org/10.1037/rev0000302. The preprint along with the stimuli, data, and code of the project are available on the OSF project page: http://doi.org/10.17605/OSF.IO/RWBYH (Wagner, 2021).

Abstract

In a sequence of otherwise equal sounds, listeners tend to hear a series of trochees (groups of two sounds with an initial beat) when every other sound is louder; they tend to hear a series of iambs (groups of two sounds with a final beat) when every other sound is longer. The paper presents evidence that this so-called 'Iambic-Trochaic Law' (ITL) is a consequence of the way listeners parse the signal along two orthogonal dimensions, grouping (*Which tone is first/last?*) and prominence (*Which tone is prominent?*). A production experiment shows that in speech, intensity and duration correlate when encoding prominence, but anticorrelate when encoding grouping. A model of the production data shows that the ITL emerges from the cue distribution based on a listener's predicted decisions about prominence and grouping respectively. This, and further predictions derived from the model, are then tested in speech and tone perception. The perception results provide evidence that intensity and duration are excellent cues for grouping and prominence, but poor cues for the distinction between iamb and trochee per se. Overall, the findings illustrate how the ITL derives from the way listeners recover two orthogonal perceptual dimensions, grouping and prominence, from a single acoustic stream.

Keywords: perception; rhythm; prominence; grouping; speech; speech segmentation

Two-dimensional parsing of the acoustic stream explains the iambic-trochaic law

Prominence, grouping, and the Iambic-Trochaic Law

Sequences of tones and syllables are often perceived as rhythmically grouped. Even if all tones or syllables in a sequence are acoustically identical and equally spaced (henceforth 'equisound sequences'), listeners tend to perceive rhythmic grouping (Bååth, 2015; Bolton, 1894; Vos, 1973; Woodrow, 1909). Acoustic differences in the alternate tones affect the perceived rhythm in systematic ways. Bolton (1894, p. 232) discovered that in tone sequences, 'If the recurrent difference is one of intensity, the strongest impression comes first in the group and the weaker ones after. If the recurrent difference is one of duration, then longest impression comes last.' Bolton furthermore reports that listeners hear the louder/longer sound in such alternating sequences as more prominent, or 'accentuated,' within their group. Bolton's overall generalization is then that alternately louder sounds are perceived as group-initial and prominent, leading to the percept of a sequence of trochees (binary groups with initial prominence); alternately longer sounds are perceived as group-final and prominent, leading to the percept of a sequence of iambs (binary groups with final prominence). [1]

These effects have been replicated in many studies on tone sequences for English (Hay & Diehl, 2007; Iversen, Patel, & Ohgushi, 2008; Rice, 1992; Woodrow, 1909), and have also been observed in studies on speech (Bhatara, Boll-Avetisyan, Unger, Nazzi, & Höhle, 2013; Hay & Diehl, 2007). Today this generalization is usually referred to as the 'Iambic-Trochaic Law' (ITL) following Hayes (1995), even though the law is more general than its name suggests. When every third tone is long, listeners tend to perceive that tone as final and prominent and perceive a sequence of groups of three tones with final

[1] 'Prominence' is used here to refer to the intuition that within a group of sounds, one or more sounds often appear to stand out as perceptually foregrounded, whereas others appear to remain more in the background. In the case of music such perceived prominences are often referred to as 'beats', in the case of language they are referred to as 'stress.'

prominence, i.e. anapests (Vos, 1977; Woodrow, 1909); when every third tone is loud, listeners tend to perceive a sequence of groups of three tones with initial prominence, i.e. dactyls (Woodrow, 1909). The ITL has been found to apply in speech and in non-speech, and even in other sensory domains, for example in stimulation with electric shocks (Woodrow, 1909), and in the visual domain (Miner, 1903; Peña, Bion, & Nespor, 2011). It remains under debate whether it is universal (e.g. Hay & Diehl, 2007), or whether it is due to language experience (Bhatara et al., 2013; Iversen et al., 2008), and if so to what extent (cf. Boll-Avetisyan, Bhatara, Unger, Nazzi, & Höhle, 2020). ITL-like effects based on pitch have even been reported in other species, such as rats (De la Mora, Nespor, & Toro, 2013), and zebra finches (Spierings, Hubert, & Ten Cate, 2017).

Although well-established for over a hundred years, the source of the ITL remains unclear. One reason is arguably that prior studies have not sufficiently teased apart the dimension of grouping (*Which sound is first/last?*) from the dimension of prominence (*Which sound is more prominent?*). Most studies on the ITL have used a task which asked, in some way or other, whether listeners heard trochees or iambs (e.g. Bell, 1977; Hay & Diehl, 2007; Rice, 1992; Vos, 1977). For example, in the first study looking at the ITL in speech (Hay & Diehl, 2007), participants listened to sound sequences consisting of repetitions of the syllable *ga*, and were asked 'whether the rhythm consisted of a strong sound followed by a weak sound, or, alternatively, a weak sound followed by a strong sound.' Some studies used a visual representation of two stimuli, which represented whether the first or second was more prominent (Bhatara et al., 2013; Boll-Avetisyan, Bhatara, Unger, Nazzi, & Höhle, 2016; Boll-Avetisyan et al., 2020; Iversen et al., 2008; Kusumoto & Moreton, 1997; Molnar, Carreiras, & Gervain, 2016), or asked participants to tap the perceived rhythm with their hands (Rice, 1992). This type of 'foot-decision task' conflates the dimensions of prominence and grouping. This is illustrated in Table 1.

When a participant reports hearing a series of iambs, they may have heard the uneven sounds as prominent and final or the uneven sounds; when they report hearing a

	Even tone last	*Odd tone last*
Even tone prominent	(x X) (x X) (x X) ...	x) (X x) (X x) (X...
Odd tone prominent	(X x) (X x) (X x) ...	X) (x X) (x X) (x...

Table 1

Four ways to perceive a sequence of sounds as a sequence of groups of two sounds, depending on the dimension of grouping (Which sound is first/last?|) and prominence (Which sound is more prominent?). The top left and bottom right sequences are sequences of iambs, the bottom left and top right ones are sequences of trochees.

series of trochees, they may have heard the uneven sounds as prominent and initial or the even ones. Asking for the perceived foot (iamb or trochee) hence underdetermines what a listener actually perceived, since it only establishes on which diagonal in Table 1 their percept falls, but it does not establish the precise cell.

Other studies on the ITL, especially looking at ITL effects in infants, used a speech/tone segmentation task (Abboub, Boll-Avetisyan, Bhatara, Höhle, & Nazzi, 2016; Bhatara, Boll-Avetisyan, Agus, Höhle, & Nazzi, 2016; Bion, Benavides-Varela, & Nespor, 2011; Crowhurst, 2016; Crowhurst & Teodocio Olivares, 2014; Hay & Saffran, 2012; Molnar, Lallier, & Carreiras, 2014; Yoshida et al., 2010). For example (Bion et al., 2011) exposed children to sequences of syllables, and then measured how long children kept looking in the direction of a bisyllabic test stimulus. The response to the stimuli revealed whether listeners had parsed the signal so that the presented two-syllable sequence represented a word or two syllables straddling a word boundary. As already noticed in Crowhurst and Teodocio Olivares (2014, 54), a segmentation task has the advantage that in contrast to the foot-choice task, it does not confound the dimensions of grouping and prominence. Crowhurst and Teodocio Olivares (2014) and Crowhurst (2016) adopted a task from Höhle, Bijeljac-Babic, Herold, Weissenborn, and Nazzi (2009), who alternated the syllables *ba* and *ga* in their speech sequences, and asked listeners whether they heard

the word *baga* or *gaba*. The results provided evidence that 'being able to perceive natural groupings does not necessarily depend on being able to locate stressed syllables' (Crowhurst & Teodocio Olivares, 2014, p. 88). However, the speech segmentation task, just like the foot choice task, ultimately also only narrows the percept down to two out of four outcomes: it establishes the column in Table 1.

How problematic is this indeterminacy of the tasks used in prior studies? According to Woodrow (1909, p. 60), the grouping effect and the prominence effect correlate so closely that 'the conclusion seems fairly safe that they are the result of mental operations which have about the same basis. In other words, the statement by the subject, that certain sounds form an iambic group, is equivalent to the statement that he has perceived a shorter interval before the louder or the longer sound than after it, and the statement that the sounds form a trochaic group means that the subject has perceived a shorter interval after the louder or longer sound than before it.' Similarly, studies using the foot task have often assumed that if every other sound is louder/longer, the variation in perception will lie in whether that sound is heard as final or initial, but that the louder/longer sound will always be perceived as the more prominent sound.

The results of the perception study, however, will show that this is not the case. Grouping and prominence do not necessarily go hand in hand. Even with the more extreme manipulations of duration and intensity, there is variability not just in which sound is perceived as initial, but also which sound (the longer/louder or the shorter/softer one) is perceived as prominent. [2] The correlation between prominence and grouping were

[2] The particular task in Kusumoto and Moreton (1997), Iversen et al. (2008), and Molnar et al. (2016) used a choice between iambs and trochees that visually representated the duration and loudness iconically in different ways (e.g., longer dash for duration, bigger blot for intensity). So in a way, listeners were not just asked whether they heard iambs or trochees, but were also 'told' that they were supposed to hear the longer/louder sound as prominent. It's not clear though that listeners can successfully detect the acoustic manipulation and distinguish it from their percept of prominence, and as we will see, the percept of prominence does not necessarily align with the location of the phonetic cue.

also put into question by Crowhurst and Teodocio Olivares (2014) and Crowhurst (2016), at least with respect to duration. However, given that their task only established grouping, it could not actually establish how the two decisions relate to each other.

That prominence and grouping can be dissociated should not be surprising, given that in speech, grouping and prominence are in principle orthogonal to each other. Imagine a sequence consisting of iterations of the syllables *up* and *set*. A listener may perceive it as one of four words, *UPset*, *upSET* and *SETup*, or *set UP*, each of which have a different meaning. These four options (two of which have trochaic stress, two iambic stress) correspond to the four cells in Table 1. How listeners interpret the cues can only be fully understood if we consider all four possibilities. In order to do so, the perception experiments adopt the speech segmentation task, and combine it with a second task, in order to get at the perceived prominence. This way, one can establish both the *row* and the *column* of Table 1 corresponding to a listener's percept.

The hypothesis proposed here is that the ITL is a consequence of the fact that listeners parse the signal along two in principle orthogonal dimensions, grouping and prominence. In the literature on speech prosody, it is standard to assume that grouping and prominence are separate dimensions, and that they are at least to some extent orthogonal to each other (Ladd, 2008). However, the consequences of this for a listener have not been sufficiently explored. To see how this can lead to an explanation for the ITL, we need to look at the distribution of cues for the two dimensions. The first experiment reported here is therefore a production study. The crucial finding is that intensity and duration *correlate* when they encode syllable stress, but *anti-correlate* when they encode the position of a syllable within a word.

This means that listeners can interpret whether an increase in one cue encodes prominence or grouping by interpreting it relative to changes in the other cue. The relationship between the cues and its potential role in perception has not been previously noticed when looking at word segmentation, although a similar relation between the cues

was observed for the encoding of prosodic phrasing in a production study reported in Wagner and McAuliffe (2017, 2019). A rational listener can make use of the cue distribution observed in production when deciding about the grouping and prominence structure of an utterance, just as the cue distribution for a phonemic contrast is directly predictive of listener's response to a categorization task (Clayards, Tanenhaus, Aslin, & Jacobs, 2008). When a syllable is longer than expected given its segmental content, but intensity lower, listeners can interpret this as a cue to word-finality. Conversely, when intensity is higher than expected but duration is not, then this provides a cue for word-initiality. If both are higher than expected, this can be interpreted as a cue for prominence.

The ITL can now be explained as follows: If every other syllable in a sequence is sufficiently long, this provides a cue to group-finality and prominence; if every other syllable is sufficiently loud, this provides evidence for group-initiality and prominence.

This paper uses both production and perception experiments to tests the predictions of the hypothesis that listeners parse the signal along two dimensions by exploiting the cue distribution, and also the more specific hypothesis about how to explain the ITL. The production study serves to test the assumptions about the cue distribution. The production results are then used to illustrate that perceptual ITL effects can be predicted directly from the cue distribution in the production data. The predictions derived from the production results are then tested in perception experiments on speech and tone sequence.

For the perception stimuli, both duration and intensity were manipulated, sometimes within the same stimuli, following Bolton (1894); Crowhurst and Teodocio Olivares (2014), and Crowhurst (2016). The hypothesis proposed here predicts that if both cues are manipulated on the same syllable, listeners should be consistent in their prominence choice (since both cues are consistent), but closer to chance in their grouping choice (since the cues are inconsistent for grouping); if both cues are manipulated on different syllables, then listeners should be consistent in their grouping decision, but closer to chance in their

prominence decision. The perception experiments also show that intensity and duration are actually quite poor as cues for the difference between iambs and trochees.

That intensity cues initiality and duration finality is already part of the ITL, but the statement of the ITL ties these effects inherently to the percept of prominence by attributing an inference to a certain foot type. The explanation for the ITL proposed here, by contrast, is based on the idea that listeners try to explain the acoustic signal by attributing aspects of it to two separate sources, prominence and grouping, that are in principle orthogonal. This is parallel to other situations in which an auditory scene is analyzed as being composed of different sources (Bregman, 1990). Take the phenomenon of auditory streaming (Bregman & Campbell, 1971), where listeners analyze a series of tones alternating in frequency as being composed of two different sequences, each one consisting of a sequence of tones of equal frequency. The present hypothesis, by contrast, posits two 'sources' that correspond to two dimensions of organization of a single sound sequence. The same cues are used to encode both dimensions, so in order to explain the acoustic stimulus, listeners attribute some aspects to prominence and some to grouping, in an attempt to find a coherent 'auditory description' for the signal (Bregman, 1981).

Experiment 1: Speech production

Intensity and duration are well known to play a role in encoding whether a syllable is stressed in English. Stressed syllables are both longer and louder compared to non-stressed syllables (Beckman, 1986; Chrabaszcz, Winn, Lin, & Idsardi, 2014; Fry, 1958; Lehiste, 1970). This is true whether the stressed syllable is initial or final within a word (Beckman, 1986; Chrabaszcz et al., 2014).

Duration also plays a role in encoding how syllables are grouped into words. Word-final syllables are lengthened, to a degree that depends on the strength of the prosodic boundary following the word (Beckman & Edwards, 1990; Klatt, 1975; Oller, 1973; Wightman, Shattuck-Hufnagel, Ostendorf, & Price, 1992), though see (Turk &

Shattuck-Hufnagel, 2000) for conflicting findings. Expectations about durational cues to word segmentation are used already by 6-month old infants (Shukla, White, & Aslin, 2011) to segment speech. The role of intensity in encoding grouping is less established. In two perception studies, Streeter (1978) found that intensity was a less important cue for grouping compared to other acoustic cues. Production evidence in Wagner and McAuliffe (2017, 2019), however, found that words at the beginning of a phrase have significantly higher intensity than words later in a phrase, suggesting that intensity could in principle be an important cue for phrasing. While intensity generally drops throughout an utterance (see Pierrehumbert 1979 for the observation, and perceptual evidence that listeners compensate for an expected downdrift in intensity), this utterance-level downdrift alone cannot explain this effect: The results in Wagner and McAuliffe (2017, 2019) showed that loudness partially resets when a new prosodic phrase begins (see also Poschmann & Wagner, 2016).

One aim of the production study was to establish whether this cue distribution also applies at the word-level (as opposed to phrase-level), and hence can be used to segment the acoustic stream. According to the hypothesis, it is the relation between these two cues that explains the ITL pattern. However, there are other acoustic features that also encode prominence and grouping. Pitch is an important cue to word and phrase-level stress (Beckman, 1986; Chrabaszcz et al., 2014; Fry, 1958; Price, Ostendorf, Shattuck-Hufnagel, & Fong, 1991), as well as for speech segmentation (Jusczyk, Cutler, & Redanz, 1993) and phrase-level grouping (Ladd, 2008; Price et al., 1991). Prominence has also been reported to be cued by the precise articulation of vowels based on their formants, which reflect tongue position and degree of jaw opening (Beckman, Edwards, & Fletcher, 1992; De Jong, 1995; Mo, Cole, & Hasegawa-Johnson, 2009). While in this paper, cues other than duration and intensity are not of direct interest, taking them into account will still be useful in understanding the effects of intensity and duration for the perception study.

Methods

Participants were recorded on sequences of disyllabic nonce words varying in stress. The stimuli consisted of three repetitions of disyllabic nonsense words, each consisting of identical syllables. For example, on a given trial a participant may have been asked to say *bába bába bába*. The stimuli varied in whether the words had initial or final stress, and also in the phonemic content of the syllables. Participants were instructed to say the sequences as naturally as possible, as if they were real words. Trial order was pseudo-random, so that the same stress condition and the same phonemic content did not occur more than twice in a row. A total of 18 native speakers of North American English were recorded, but only data of 16 were included here, because the recording did not work for the other two. 14 were Canadian, and 2 from the US. 14 were female and two male. All participants were recorded in a sound-attenuated booth at McGill, in Montréal, using a headset. Participants filled out a language questionnaire and a music questionnaire. Most participants spoke French in addition to English to some degree, ranging from beginner-level to fluent. Participants had a median of 7 years of music lessons.[3]

The data were hand-checked by two RAs for speech errors, and were automatically aligned using the Montréal-Forced aligner (McAuliffe, Socolof, Mihuc, Wagner, & Sonderegger, 2017), with acoustic models trained on the Libri-speech corpus. The automatic alignment created a word-by-word and segment-by-segment annotation of the speech signal. In order to avoid any word-boundary-related biases in the alignment due to the acoustic models, every utterance was transcribed as if it contained only monosyllabic words. So rather than feeding the aligner the transcription *baba baba baba*, the transcription *ba ba ba ba ba ba* was used. The aligner had the ability to annotate a silence if there was a gap between words that was bigger than expected giving the surrounding

[3] The influence of language and music background on the results in this and the other two experiments is not explored here, but the complete questionnaire information is released as part of the published data. See the supplemental materials for more information, and the reasoning behind not exploring this further here.

segments. Given that each syllable was annotated as if it were a word, this could result in the annotation of a short silence within a word if, say, the closure duration of [b] was longer than usual. An annotated silence could also reflect an intended pause between words. If an utterance began with a stop (e.g., *ba...*) the aligner would still annotate a certain amount of silence as the closure duration of [b], based on the expected value for closure duration in the acoustic model. An example alignment is provided in the supplemental materials. Acoustic measures were extracted for each syllable using the speech software Praat (Boersma & Weenink, 1996) with a script, using different settings for formant extraction based on the gender of the speaker. In addition to duration and intensity of the syllables,t maximum pitch within a syllable (in semitones relative to a reference frequency of 100Hz), the first and second formant of the vowels (i.e.,the resonant frequencies of the vowel tract, which depend on tongue and jaw positioning), and the silence following a syllable were also measured.

For each measure, linear mixed effects models were fit using the R package lme4 (Bates, Mächler, Bolker, & Walker, 2015). P-values were computed based on the Satterthwaite approximation using the R package lmerTest (Kuznetsova, Brockhoff, & Christensen, 2018). Each of the models included stress and position of the syllable within the word (initial or final), and their interaction. They also included the position of the syllable within the utterance (as a numeric factor), as well as random effects for item and participant. All predictors were converted into z-scores, in order to avoid spurious collinearity. The dependent variables were turned into z-scroes, so that one can compare effect size across cues. The random effect for participants also included slopes for the acoustic measures as well as word position. More details, including full models, for this and the later experiments can be found in the supplemental materials.[4]

---

[4] See supplemental materials for model formulas. The stimuli, code, and data for all three experiments are posted on OSF http://doi.org/10.17605/OSF.IO/RWBYH (Wagner, 2021).

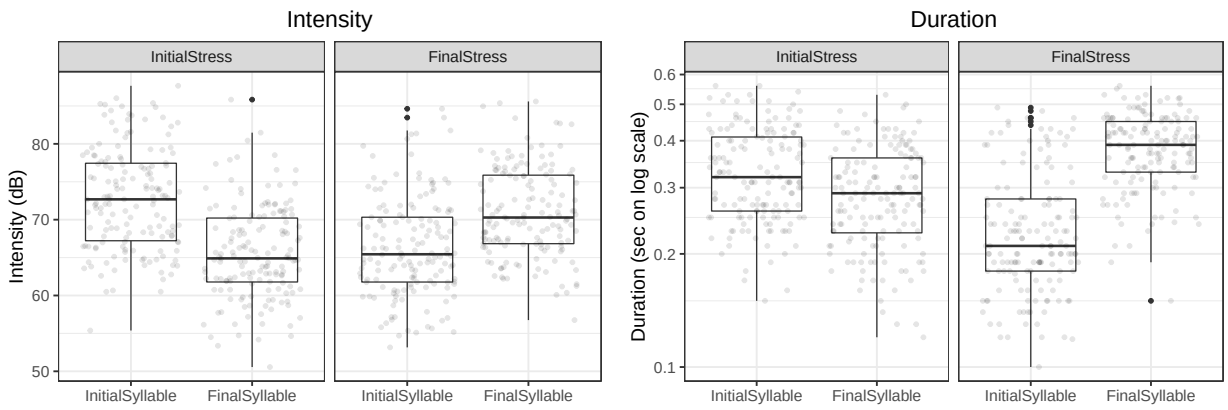

Figure 1. Duration (sec on log scale) and intensity (dB) of each syllable within the disyllabic words, depending on whether they have initial stress (trochees) or final stress (iambs).

Results

Figure 1 shows measures of maximum intensity and duration of the two syllables depending on which syllable was stressed. The intensity plot shows that that stressed syllables, both in trochaic and iambic words, are louder than unstressed syllables. This effect was significant (β = -0.4; s.e. = 0.05; $p < 0.001$). The plot suggests that this intensity difference is greater in the case of trochaic words. However, this is due to a separate effect of the position within a word. Initial syllables show higher intensity and final syllables lower intensity. This effect was smaller than the effect of stress, but also significant (β = -0.05; s.e. = 0.02; $p < 0.05$). These two main effects compound in the case of trochees, leading to a greater intensity difference between stressed and unstressed syllable, while they counteract each other in the case of iambs, resulting in a smaller difference. There was no significant interaction, however, (β = -0.02; s.e. = 0.04; $p < 0.58$), suggesting that these two effects are purely additive.

In addition to the decrease of intensity within words, there was also a significant overall decrease in intensity throughout the utterance, such that syllables occurring later in the utterance were less loud (β = -0.15; s.e. = 0.02; $p < 0.001$), confirming that there

tends to be an utterance-level downdrift effect (cf. Pierrehumbert, 1979).

When looking at duration, we see a similar additive pattern of the effects of grouping and prominence. There is a significant effect of stress on duration, of about equal size as the effect observed for intensity (β = -0.42; s.e. = 0.07; $p < 0.001$). There is also an effect of position of the syllable within the word ($\beta = 0.25$; s.e. = 0.07; $p < 0.001$), which is bigger than the effect observed for intensity. Crucially, this effect has the opposite direction compared to the intensity effect, as is expected, if it is a result of word-final lengthening. Since final syllables are lengthened, there appears to be a greater difference in duration between stressed and unstressed syllables in iambs, but this is mostly just because in iambs, increase in duration due to stress applies in the same syllable as the increase due to word-finality.

There was also a trend toward a small interaction between the effect of stress and position within the word, such that final stressed syllables were lengthened more than initial stressed syllables, but it did not reach significance ($\beta = 0.08$; s.e. = 0.05; $p < 0.07$).

There was no significant effect of the position of the syllable within the utterance. This may seem surprising, since one might have expected phrase-final lengthening at the utterance end.[5] The result is in line, however, with some prior studies that found that phrase-final lengthening is comparatively small at the utterance end, and is much greater before a phrasal juncture *within* an utterance (Wagner, 2005; Wagner & McAuliffe, 2019). This suggests the phrasing-related duration increases (as opposed to word-related duration increases) are better characterized as pre-boundary lengthening, rather than as utterance final lengthening.

While duration and intensity are the two cues at the center of the ITL and the focus of this study, they are not the only acoustic cues relevant in the encoding of prominence

[5] A separate model was used whether position of the word within the utterance had an effect, which also showed no significant effect. This second model excluded the position of the syllable within the utterance, in order to avoid collinearity.

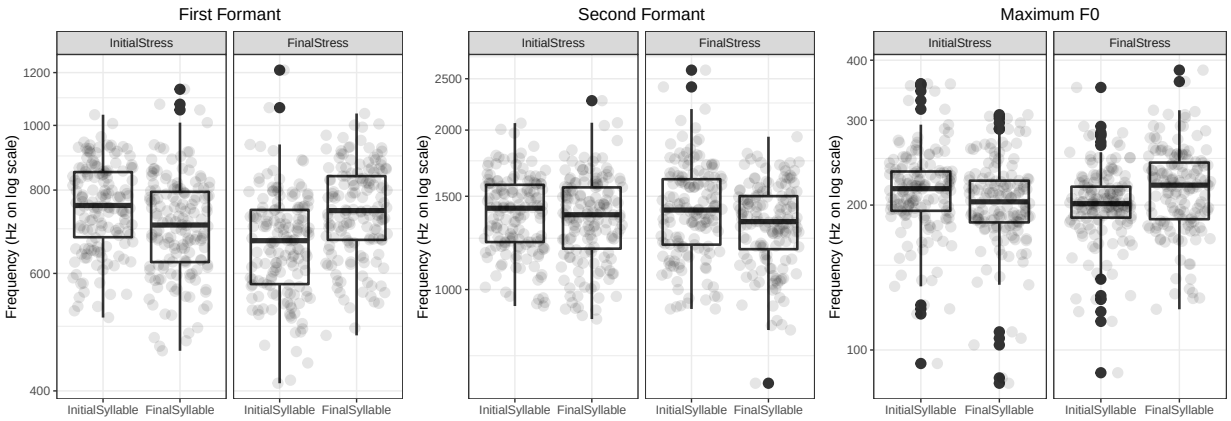

Figure 2. First and second formant (Hz) and maximim F_0 (Hz),in disyllabic words depending on whether they have initial stress (trochees) or final stress (iambs).

and grouping. , This is illustrated in Figure 2. The first plot shows that maximum F_0 is higher on syllables that are stressed. This is expected since in English, stressed syllables tend to carry a pitch accent with a high tonal target. The effect of prominence on F_0 was significant (β = -0.15; s.e. = 0.05; $p < 0.001$). F_0 can also play a role in cueing grouping. Ladd (1988), for example, found that pitch accents are scaled according to how syllables and words are grouped within an utterance, and there are also tonal events at prosodic junctures. In the present data, the pitch difference between stressed and unstressed syllables of the measure of maximal F_0 appears bigger in iambic words compared to trochaic words. This interaction between prominence and grouping was significant (β = -0.13; s.e. = 0.06; $p < 0.02$), showing that grouping did affect F_0 in this data, even if the main effect of position within the word was not significant ($\beta = 0.09$; s.e. = 0.08; $p < 0.27$). In addition, there was also an effect of position of the word within the utterance such that F_0 was lower in later syllables (β = -0.14; s.e. = 0.03; $p < 0.001$), reflecting the well-known phenomenon of F_0 declination throughout an utterance (Liberman & Pierrehumbert, 1984).

The remaining two plots in Figure 2 illustrate the role of formants in cueing prominence and phrasing. It is well known that the first formant tends to be higher when a syllable is stressed, an effect sometimes attributed to a greater jaw opening (Beckman et

al., 1992; Mo et al., 2009), and sometimes to hyperarticulation of the tongue height of the vowel (De Jong, 1995), resulting in higher F1 values for non-high vowels. [6] Stress indeed led to a significant increase of F1 (β = -0.26; s.e. = 0.04; $p < 0.001$).

There was furthermore a significant effect of grouping on F2 (β = -0.08; s.e. = 0.03; $p < 0.02$), such that F2 was lower in word-final position. Lower F2 often indicates a tongue position further to the back. This study involved back vowels ([ɑ] and [o]), so this result could be due to hyperarticulation. When a vowel is longer, as in word-final environments, the tongue has more time to move to the back target. This interpretation is compatible with findings in Cho (2005), who show that phrase finally, [a] has a lower F2 (tongue position further back) and the front vowel [i] has a higher F2 (tongue position further front).

Furthermore, there was an effect of position of the syllable in the utterance such that later syllables had a lower F2 (β = -0.06; s.e. = 0.02; $p < 0.02$). This is reminiscent of spectral declination patterns over the course of an utterance observed in Italian by (cf. Vayra & Fowler, 1992).

Finally, a model was for the duration of silence following a syllable (not plotted here; see supplemental materials for the full model). Pauses and their duration are also a cue for grouping, such that syllables straddling word boundaries are separated by a larger amount of silence (β = 0.38; s.e. = 0.08; $p < 0.001$). There was also an effect of position of the syllable in the utterance (β = -0.31; s.e. = 0.03; $p < 0.001$). Since pauses after the last word were omitted from analysis, this effect suggests that pauses were more likely between the first two words than between the last two words.

---

[6] Since the experiment stimuli only involved words with non-high vowels, this data does not allow us to distinguish these two interpretations.

Discussion

The results of Experiment 1 provide evidence for the premise of the hypothesis of how to explain the ITL: Duration and intensity provide important cues both for stress and for grouping, and, crucially, the relation between the two cues depends on the dimension they encode. Both cues increase when they encode the prominence of stressed syllables, but they anticorrelate when encoding grouping. These word-level effects observed here are parallel to the phrase-level effects observed in Wagner and McAuliffe (2017, 2019).

It has sometimes been argued that the greater durational asymmetry in iambic words (visible in Figure 1) provides evidence for a representational distinction between foot types, where iambs differ inherently from trochees in their phonetic realization (Hayes, 1995). But the present data shows that this apparent greater asymmetry of iambs is largely a consequence of the fact that the durational effects of grouping and prominence compound in the case of iambs, but counteract each other in the case of trochees. This is as expected if the tendency for a greater durational asymmetry in iambs observed across languages comes about due to final lengthening, as argued already in Revithiadou (2004) (see Hyde 2011 for a review of cross-linguistic correlations between foot type and syllable weight). There was non-significant a trend toward an interaction of the effect of prominence on duration with the position of the syllable within the word, suggesting that stress in iambs may be realized slightly differently compared to trocheesd, but by and large the effects are additive.

The relation between the cues may be the key to retrieving the two separate dimensions from the acoustic signal when perceiving speech. For example, one way to tease apart initial lengthening (Fougeron & Keating, 1997) from final lengthening (Wightman et al., 1992) would be to see whether there is a concomitant increase in intensity, and similarly, one can tease apart prominence-related lengthening from grouping-related lengthening in this way. Streeter (1978) had reported that intensity was only informative as a cue for grouping (in that case at the phrase level) when taking its relation to duration into account, and left this finding as a puzzle. We can now understand why this is: When

an increase in intensity is accompanied with an increase in duration, it provides different information than if it is not. When intensity is higher than expected and duration as well, this provides a cue for prominence; when intensity is higher and duration is lower, this provides a cue for word-initiality. Conversely, when intensity is lower than expected but duration is higher, then this provides a cue for word-finality.

According to the present hypothesis, listeners attribute deviations of the intensity and duration of a syllable from what would be expected given their phonemic content to grouping and prominence. It could in principle be that listeners first make a grouping decision (*Is the syllable initia/medial/final?*) and then make their prominence decision (*Is the syllable stressed?*) based on the unexplained residual; or they could first make the prominence decision, and then decide on grouping. However, it is possible that the cue relation makes the signal more directly interpretable. To illustrate how the cue distribution could be used to tease apart the dimensions, the two principal components of intensity and log duration are plotted in Fig. 3. Principal components analysis provides a different representation of the information of the two cues, and teases them apart into two orthogonal components.[7]

The left panel in Fig. 3 plots the first component, the shared information between duration and intensity. We see that this component distinguishes stressed from unstressed syllables, while it is mostly unaffected by grouping. The second, orthogonal component, by contrast, reflects grouping, and distinguishes initial from final syllables, while it is largely unaffected by the location of stress. Principal components analysis is 'unsupervised' in that the information of stress and position within the word are not used in the process. That two dimensions can be teased apart in this way suggests that listeners could in principle

[7] The components were computed using the R-package prcomp, based on a correlation matrix of duration (in sec) and intensity (in dB). The loadings for component 1 were 0.71 for both duration and intensity; the loadings for component 2 were -.71 for duration and 0.71 for intensity.

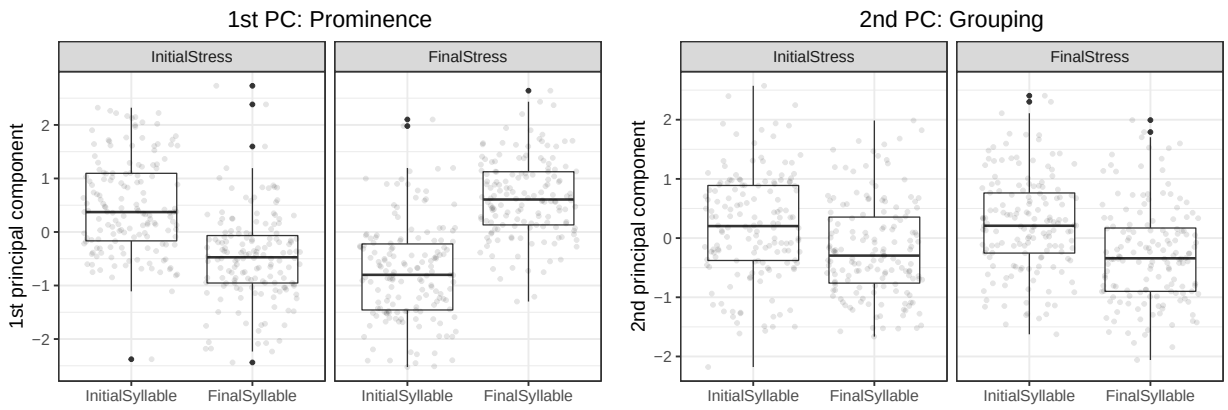

Figure 3. The first principal component of intensity and log duration distinguishes stressed from unstressed syllables; the second one initial from final syllables.

use this strategy to derive separate cues for each dimension.[8]

In sum, intensity and duration are important cues for both grouping and prominence, and their relation is crucial in disentangling the two dimensions. We also saw that these two cues are not the only relevant cues, since we also observed cues from vowel formants and pitch for both dimensions, which show effects of grouping and prominence of comparable size as those on intensity and duration. Especially the grouping-effects on formants would be worth closer scrutiny. There was an utterance-level effect, suggesting that there is a 'declination' at the spectral level over time, similar to pitch declination and downdrift (cf. Vayra & Fowler, 1992). And there was also a word-level effect, suggesting that formants are actively used to encode grouping, just as they encode prominence. It would also be interesting to look at the dynamic development of these cues over the course of the syllable into account. For example, pitch should be interpreted differently depending on where on the syllable a change occurs, since pitch modulation later in the syllable should be more likely to be attributed to so-called boundary tones, rather than to tonal

[8] An independent components analysis (ICA) was less successful in teasing apart the two dimensions of grouping and prominence. It returns two component that (idealizing a bit) track stress more in trochaic and iambic words respectively. See the supplemental materials for more information.

reflexes of stress. Similarly, the dynamic realization of duration and intensity could disentangle different dimensions to some extent (Beckman & Edwards, 1990; Edwards, Beckman, & Fletcher, 1991).

Model predictions for perception

The production data can be used to compute predictions for the perception of syllable sequences. In experiments used to establish the ITL, both in the prior literature and here, listeners hear sequences of tones or syllables. Listeners are then often asked to report whether they hear a sequence of iambs or trochees, or are asked a question about grouping. As discussed, in the experiments reported here listeners were instead asked two separate questions, in order to establish both their grouping percept and their prominence percept. From these two responses, one can reconstruct whether they heard iambs or trochees.

Logistic models on the production data were employed to compute predictions for both the grouping and prominence decisions. To do so, the production data was recoded as if the experiment had, in addition to varying the location of word stress, also varied whether speakers started speaking mid-word or at the beginning of a word. This was done by using the data twice, and coding one half as if speakers had initiated speaking in the middle of a word.[9] This data was then fit against three separate logistic mixed effect regression models, one for each of the three choices. These models can be used to predict the log odds for the three decisions about a given stimulus, given its acoustic properties, based on the overall distribution of the cues in the production data.

While the focus here is mostly the duration and intensity effects, which the original formulation of the ITL was based on, each model contained the relevant acoustic cues as fixed effects: intensity, duration, pitch, F1, and F2. The model also included the duration

[9] Coding randomly a random half of the data led to similar results, but using the data twice for this purpose avoides arbitrary differences due to the sample that was recoded. See supplemental materials for more details.

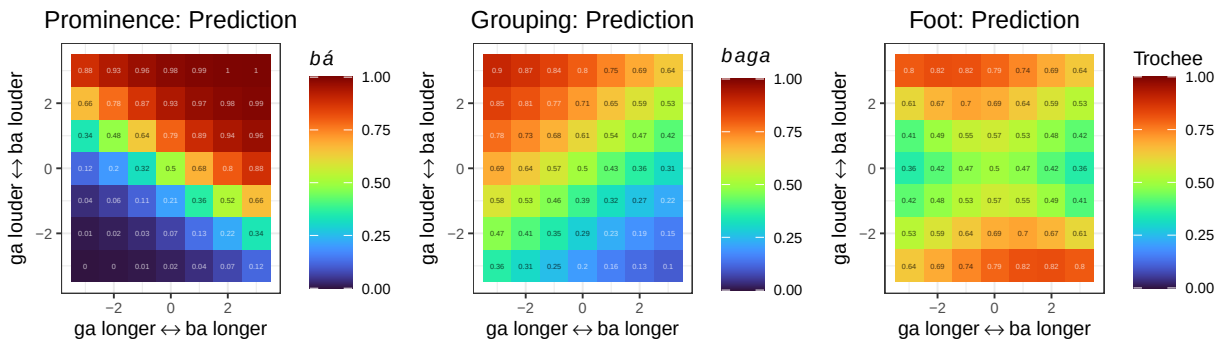

Figure 4. heat maps of the predictions for the responses in the grouping decision (left), the prominence decision (center), as well as for the foot decision (right).

of silence following a syllable in the model, as well as the position of the word within the utterance. [10]

In order to compute predictions from these models for the perception stimuli, a new data set was created, with acoustic specifications matching the perception stimuli. Every line in this data set corresponds to one stimulus from the perception study, and includes values for all the acoustic parameters in the model. Only the values for duration and intensity are varied between the stimuli according to the manipulation in the perception study, all other acoustic cues remain constant. All stimuli were coded as being utterancea-medial for calculating the predictions, since this seems most appropriate given the nature of the sequences listeners were going to listen to.

Figure 4 illustrates that the predictions based on the production data are in line with the hypothesis.[11]

---

[10] While the addition of cues other than intensity and duration improved the predictions, the predictions look qualitatively very similar if these cues are omitted, and this choice is non-essential to the argument of this paper. The predictions of the simpler model including only duration and intensity are plotted in the supplemental materials.

[11] The heat maps in the online version use a color scale designed to be compatible with color blindness that ranges from dark red (100%) via light green (50%) to dark blue (0%). The greyscale in the print version darkens with distance from chance (50%). The actual proportions are listed in each cell of the heat map.

The signature pattern is that there is symmetry along one diagonal for the prominence choice, and along the orthogonal diagonal for the grouping choice. This is just as expected based on the hypothesis: When the same syllable is both louder and longer (as in the bottom left and top right corners of the heat map), listeners should be likely to perceive that syllable as prominent, but they should be closer to chance with respect to which syllable they they think of as initial and final, since there are conflicting cues for grouping. When one syllable is louder and the other is longer (as in the top left and bottom right corner of the heat maps), listeners should be likely to perceive the louder syllable as initial and the longer one as final, but they should be uncertain about prominence in this case, since they receive conflicting cues for that dimension. The 'diagonals of uncertainty' do not exactly cut through the corners of the heat maps. When intensity is highest, the model predicts it to 'take over,' and mostly determine the outcome.

Given the model predictions for the grouping and the prominence decision, one can also calculate predictions for the choice between iamb vs. trochee.[12] If the same syllable was predicted to be perceived as prominent and initial, it was predicted to be heard as a trochee, otherwise as an iamb. This simulates the task that most studies on the ITL, in one way or another, have employed. The proportions predicted from the production model are plotted in the rightmost plot in Fig. 4.

The most important prediction relevant here is that of the ITL itself. When only duration is manipulated, as plotted in the middle row of the heat map, the prediction at the outer edges is that the sequences should tend to be perceived as iambs. When only intensity is manipulated, plotted in the middle column of each heat map, the prediction at the edges is that that the sequences should be perceived as trochees. In other words, given the distribution of cues in the production data, and assuming that listeners make rational choices for the grouping and prominence decisions given the cue distribution, the model

[12] I also computed predictions based on a model of the foot decision directly, which led to very poor predictions.

predicts the iambic-trochaic law. We now turn to experiments that test how these model predictions compare to actual perceptual responses.

Experiment 2: Speech perception

Our perception study differs from prior work in that participants were asked two questions, one to establish the perceived grouping, and one to establish the perceived prominence. Earlier experiments used either a foot-choice task or a segmentation task, which underdetermines the actual percept. The experiment involved sequences of the syllables *ba* and *ga*, and combined the speech segmentation task (Did you hear *baga* or *gaba*?) with with a second question which established in addition which syllable listeners heard as prominent in order to establish the full picture.

Methods

A total of 25 participants were run in the speech perception study. 14 were Canadian, and 11 from the US. 16 were female and the others male. The experiment was conducted in the same sound-attenuated booth at McGill, stimuli were played via a headset. Participants filled out a language questionnaire and a music questionnaire. As in the production study, most participants spoke French in addition to English to some degree, some of them fluently. Participants had a median of 3 years of music lessons.

For greater comparability, the stimuli were modeled after those in Crowhurst and Teodocio Olivares (2014) and Crowhurst (2016). Each stimulus consisted of sequences alternating the syllables *ba* and *ga*. Another advantage of using two different syllables is that it helps avoid an alternative interpretation of the sequence. When playing an example speech stimulus from Hay and Diehl (2007) (provided by Jessica Hay, p.c.), who used repetitions of the same syllable *ga*, to a class, several listeners including myself thought that the softer syllable sounded a bit like an echo of the louder one. This illusion could be the cause of a trochee response, given that echos are lower in intensity than the original acoustic event that caused them, and necessarily follow rather than precede it. The

manipulation is based on a smaller range of intensity than those in Crowhurst and Teodocio Olivares (2014); Hay and Diehl (2007) and Crowhurst (2016) in order to avoid this potential confound.

The syllables *ba* and *ga* were generated using Amazon's speech synthesizer Polly. [13] Using the resynthesis manipulation in Praat, the two syllables were normalized to a duration of 240ms, an average intensity of 70dB, and constant pitch of 120Hz. They were then concatenated to form bisyllabic words, *baga* and *gaba*.

The intensity and duration were manipulated on each of the two words using resynthesis (again using a script in Praat) ranging from a baseline level of 0 in intensity/duration change, and three additional steps, adding 3dB, 6dB, and 9dB for the intensity steps, and 40, 80, and 120ms for the duration steps. *ba* could be of equal intensity/duration as *ga*, or up to 3 steps louder/longer or softer/shorter, for a total of 7 different steps in intensity and duration. The manipulations were crossed, but a given cue was only manipulated on one of the two syllables. For example, if *ba* was manipulated for duration, then the *ga* was always at the baseline level for duration, but either *ba* or *ga* might have also been manipulated for intensity in addition. This means that there were a total of 49 different manipulations for a given word.

The 49 manipulated words were then each concatenated 12 times, keeping the spacing between syllables constant, resulting in 49 different sequences, with 24 syllables each. For reasons of design balance, we included the stimulus in which duration and intensity were at baseline in both syllables twice, leading to a total of 50 sequences.

This was done both for *baga* and *gaba*, so there were 2*50 stimuli. The underlying order has been found to matter (Hay & Diehl, 2007), so it was counterbalanced it within each participant. Underlying order should matter particularly for the grouping task, since listeners will be likely to take the first sound emerging from the noise to be the beginning of a word. A given participant listened to 50 different stimuli that included all

[13] Polly is an online text-to-speech synthesizer, accessed in September 2019.

manipulations, but 25 were drawn from the *gabagaba...*-set, and 25 from the *bagabaga...*-set. In order to reduce the effect of underlying order, the beginning of the sequence was faded in an on-rammp in the shape of a quarter sine function that lasted 800ms, and concomitantly white noise was superimposed that faded out over the same duration. This way, the sequence emerged from white noise, obscuring whether it had started with *ga* or *ba*. There was no off-ramp or white noise at the end, in order to keep the experiment more similar to Experiment 3 on tone, where as we will see, it was crucial that the sequence was cut off cleanly at the end. The presentation order of the stimuli was random, except that the randomization was constrained such that we at most played two stimuli in a row with the same underlying order.

After listening to a given sequence, participants first had to answer 'Which word did you hear?,' choosing between *baga* and *gaba*. They were then asked 'Which syllable did you hear as stressed?', and again had to choose between two options (e.g., if responded *baga* to the first, question, their choices were *BAga* vs. *baGA* for the second).

The two choices were analyzed using separate logistic mixed effects regression models. The models included the intensity and duration as well as their interaction, and the underlying order of the syllables *baga vs. gaba*. Each model also included the other decision, the prominence decision was used as a predictor in the grouping model, and the grouping decision as a predictor in the prominence model. The models also included a random effect for participant, with slopes for the acoustic predictors. [14]

---

[14] One weakness of fitting two separate model is that they treat the error of the two decisions as independent from each other. This raises some statistical and conceptual questions. To address this concern, a bivariate Bayesian model was fit in addition, using the brm-package (Bürkner, 2017; Carpenter et al., 2017). This model simultaneously fits both decisions in a single model, with a shared error term. Modeling the data in this way arguably provides a closer fit with what the listener actually does when making two mutually informative decisions. The Bayesian analysis provides evidence for the same conclusions reached here based on the independent models. See the Supplemental Materials for more information in these Bayesian models, and the results.

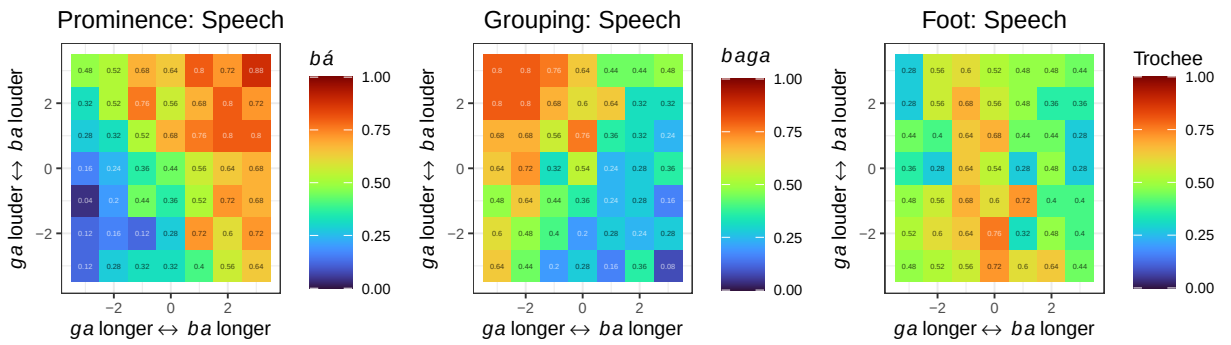

Figure 5. Heat map of the proportion of responses to the speech stimuli for the prominence decision, the grouping decision, and the foot decision. The foot decision was inferred from the responses to the grouping and prominence task.

Results: Speech perception

Figure 5 summarizes the responses to the grouping and prominence decisions. Let's first consider the prominence responses. When *ba* is both louder and longer than *ga*, listeners hear it as the more prominent of the two syllables most of the time; similarly, *ga* is perceived as more prominent when it is louder and longer. When *ba* is longer and *ga* is louder (the bottom right corner) or vice versa (the top left corner), the results are closer to chance, trending toward the longer stimulus being perceived as more prominent.

In the model for the prominence decision, main effects of duration (β = -0.51; s.e. = 0.08; $p < 0.001$) and intensity (β = -0.3; s.e. = 0.06; $p < 0.001$) were significant. There was also a significant interaction in how the intensity effect varied depending on the grouping decision (β = -0.22; s.e. = 0.08; $p < 0.001$), showing that the two decisions were not made independently. There was furthermore a significant interaction between intensity and duration, although the effect size was rather small ($\beta = 0.05$; s.e. = 0.02; $p < 0.01$). There was no effect on the prominence decision depending whether the underlying sequence started with *ba* or *ga*.

The grouping pattern also matches the predictions. When only duration was manipulated (the outer cells in the middle row of the heat map), we see that, as expected,

whichever syllable was longer tended to be perceived as final. More generally, lengthening a syllable makes it more likely to be perceived as final. The effect of duration on the grouping choice was statistically significant (β = 0.5; s.e. = 0.13; p < 0.001). When only intensity was manipulated (the outer cells in the middle column of each heat map), listeners tended to hear the louder syllable as initial. More generally, increased intensity resulted in syllables being more likely to be perceived as initial (β = -0.34; s.e. = 0.06; p < 0.001). These two effects replicate results reported for the grouping task in Crowhurst and Teodocio Olivares (2014) and Crowhurst (2016).

As expected, listeners mostly heard *baga* when *ba* was loud and *ga* long, and *gaba* when the opposite was true, while responses were closer to chance when both cues were increased on the same syllable. There was a significant interaction between the prominence choice and the effect of intensity, further confirming that the effect of intensity on the grouping decision depended on whether a syllable was perceived as stressed or unstressed (β = -0.2; s.e. = 0.08; p < 0.01). There was no significant interaction between intensity and duration in the grouping decision (β = -0.01; s.e. = 0.02; p < 0.57).

In addition, there was an effect of underlying syllable order on the grouping choice, such that listeners were more likely to report hearing *baga* when *ba* was the first audible syllable in the sequence, and *gaba* if it was *ga* (β = 0.56; s.e. = 0.14; p < 0.001). The initial ramp, overlayed with white noise, probably reduced this order effect, but did not completely remove it. This is to be expected if listeners are likely to just take the first audible sequence of two syllables to form a word.[15]

What would the results have looked like if the participants had been asked whether they heard iambs or trochees, as most prior studies on the iambic trochaic law have done?

---

[15] The responses to the baseline stimuli showed a slight bias toward hearing *ba* as initial (the center cell in the grouping response is slightly about 50%) and to hear *ga* as stressed (the center cell in the prominence is slightly below 50%). This could reflect top-down lexical knowledge (Mattys & Bortfeld, 2016). However, the overall data did not show the same biases. See the supplemental material for discussion and information how these biases relate to the distribution the CMU dictionary of English (Weide, 1998).

The foot choice can be reconstructed from the grouping and prominence responses together. This is plotted in the rightmost heat map in Fig. 5. The ITL predicts a trochaic response when only intensity is manipulated (the middle column), and an iambic response when only duration is manipulated (the middle row). The heat maps show that these basic ITL effects, first observed by Bolton, are borne out in the data.

The results can also be used to evaluate whether intensity and duration are more generally reliable as cues for the distinction between trochees and iambs, not just in the extreme cases where only one cue is manipulated to a maximum. This was tested using a mixed effects logistic regression withfoot choice as the dependent variable. The model showed no significant effect, not for intensity (β = -0.069; s.e. = 0.036; $p < 0.055$), and not for duration either (β = -0.05; s.e. = 0.03; $p < 0.1$) (see the supplemental materials for full models), even if there are trends. This shows that intensity and duration overall show at best very small effects on the choice of foot type, much smaller compared to their effects on the grouping and prominence decision.

Looking at the responses as a foot-choice is interesting also from a methodological point of view. As discussed, many earlier studies only manipulated either duration or intensity on a given stimulus, and then asked whether participants heard an iamb or a trochee. A common assumption was that there would be variability with respect to whether listeners hear iambs or trochees, but that in either case they would hear the longer/louder sound as the more prominent one. This turns out to be false. When only looking at the extreme cases of the intensity manipulation (the outermost cells of the middle row), they were 62% trochee responses (as expected based on the ITL, a proportion greater than chance). When looking at the remaining iambic responses, the louder sound was heard as prominent only in 47% of those cases. In the other cases, the excess intensity was attributed purely to group-initiality, and the softer sound was heard as prominent. Similarly, when looking at the 32% trochee-responses for the extreme duration manipulation (the outermost cells in the middle column, which as expected show an iambic

bias), the longer sound was heard as prominent 69% of the time. In the remaining cases, the excess duration was attributed to group-finality, and the shorter sound was heard as prominent. So if listeners did not respond in the predicted way to these extreme manipulations of a single cue, it could be either due to a different grouping decision, or to a different prominence decision, contradicting Woodrow (1909, 60)'s assumption, discussed above, that grouping and prominence go hand in hand.

Figure 5 also shows that for the baseline stimulus (the center cell), there was a weak trend for listeners to perceive a trochee (54% trochee responses), replicating earlier evidence for a trochee bias (e.g. Bhatara et al., 2013; Hay & Diehl, 2007). Such a trochee bias for the baseline condition could be due to a general trochee-preponderance in the lexicon. Among the disyllabic words of English, 70% have initial stress (Cutler & Carter, 1987). English listeners tend to infer that a new word begins when a stressed syllable is heard (Cutler & Butterfield, 1992; Cutler & Norris, 1988; Mattys, Jusczyk, Luce, Morgan, et al., 1999), a bias already present in English children (Jusczyk et al., 1993). Compatible with this, Bhatara et al. (2013) found a trochaic bias in German but not in French, supporting the idea that distributions in the lexicon is to blame. However, it should be noted that the number of observations in the baseline condition (2 from each of the 25 participants) and the observed bias are small. We can instead check for a trochee bias by looking at the entire data set. Participants overall heard trochees in 49.8% of the trials, and the intercept of the foot-decision model was not significant (β = -0.01; s.e. = 0.11; p < 0.93), suggesting that if the observed trochee bias in the baseline stimulus is systematic, this is not due to general trochee bias, but rather due to how equisound specifcially are interpreted.

Discussion

Our results replicate the basic ITL effects, but put them into a broader context. Yes, alternating long and short sounds tend to be perceived as iambs, and alternating loud and soft sounds tend to be perceived as trochees. But Bolton's ITL only captures a small

subset of the full pattern, and crucially fails to capture the pattern in the corners of the heat maps, where listeners are most consistent either in their prominence perception or in their grouping perception. Only by disentangling the task into two separate choices about prominence and grouping can we see the full pattern, and observe how listeners take advantage of the relation between the cues to retrieve the two dimensions from the signal.

It should be noted that the trochee rate with the most extreme intensity manipulation is not much higher than in the baseline condition. This is in line with prior studies where the intensity manipulation often lead to trochee rates not that different from the baseline condition. One potential explanation for this is offered in Woodrow (1909, 39): " [...] there are two separable objective factors tending to produce subjective accent, at least in some subjects. There is a tendency to accent the longer sound and also a tendency to accent the sound which seems to begin the group." We saw that this characterization is not sufficient, since intensity is also a cue for prominence and not just for initiality. Further experimentation is needed to understand the source of the bias for initial prominence in equisound sequences better. It is noteworthy that this bias in the perception of equisound sequences does not appear to be due to an overall trochee bias.

When the cues anti-correlate, they provide a consistent signal for grouping, while the responses are closer to chance when they correlate. This pattern is already visible in the plots reported in Crowhurst (2016, Fig. 4/5 on page 22), which shows grouping responses closer to chance when the cues correlate. This is just what we expect if duration and intensity encode prominence and grouping in an additive way. [16] The pattern for the

---

[16] Crowhurst (2016) concludes that there was an interaction between intensity and duration in the grouping decision, even though it was not significant in the reported model, just as in the present data. Crowhurst (2016) bases the conclusion that there was an interaction nevertheless on planned comparisons between certain manipulations. The interaction is attributed it to an auditory effect, namely that longer sounds can appear louder under certain circumstances, although it is not clear why this effect would only arise when the cues correlate. The additional comparisons reported in the paper, however, seem to tap the flattening of the s-shaped categorization functions. This s-shaped curve is already expected without an

prominence decision, which has not been tested in the prior literature on the ITL, shows the reverse pattern: When the cues correlate, it is relatively consistent, when they anti-correlate, it is closer to chance.

Together, this means that stimuli in which two cues are manipulated often had two valid interpretations, they were bistable. For example, if *ba* was both louder and longer, then this is in principle compatible with the percept *BAga*, since loudness can be attributed stress and/or initiality and duration to stress, and with the percept *gaBA*, since loudness can be attributed to stress and duration to stress and/or finality. Choosing *GAba*, by contrast, fails to explain the additional intensity of *ba* (it is neither initial not stressed); and choosing *baGA* fails to explain the additional length of *ba* (it is neither stressed nor final). This means that the prominence decision should be consistent (the compatible percepts *BAga* and *gaBA* have stress on *ba*), and the grouping response closer to chance. [17]

Such bistable stimuli arise when the same cues are used to make two in principle orthogonal decisions, and aspects of the signal can be attributed to either source. Visual analogues would be when we judge the size and distance of an object, or when we judge the color of an object and the color of the background light. This type of perceptual ambivalence has gained notoriety a few years ago with the viral image of a dress that some

---

interaction given that this is a binary response, so the planned comparisons arguably do not really provide evidence for an interaction effect. Crowhurst and Teodocio Olivares (2014) reports an interaction between intensity and duration in the main model, which is an interesting discrepancy to the results found in Crowhurst (2016), as well as to the results here.

[17] Teasing apart the two dimensions also offers a potential explanation for a curious finding in Woodrow (1909) for small manipulations of duration. According to Woodrow, small duration changes lead to a trochaic pattern instead of an iambic pattern. Woodrow argues that small increases in duration and perceived as increases in intensity. However, a small increase in duration could be attributed either to stress or to finality or to both. Given the four percepts, there are 3 stimuli that could explain a small increase in the length of *ba*: *BAga* (stress), *GAba* (finality), and *gaBA* (both). Two of these are trochaic. If with small increases participants choose at random between these, we expect 2/3 trochee responses. With longer manipulations, listeners will interpret the syllable to be both stressed and final and hence hear iambs.

people saw as white and some as blue, depending on whether they assumed a blueish light source, for example daylight, or assumed the scene was lit by artificial incandescent light (Lafer-Sousa, Hermann, & Conway, 2015). Another visual analogue is the Necker cube (Necker, 1832), which has two interpretations depending on perceptual decisions about the viewpoint and the relative depth of two surfaces. Warren and Gregory (1958) in fact observed that sequences of repeated words can have two stable percepts, for example repetitions of the word 'say' can be perceived as 'ace', and Warren (2008, 205) reports that repetitions of the word 'wellfare' can be perceived as 'farewell.' In this latter example, it is grouping that varies, and stress remains the same. In the present experiment, there are also bistable sequences varying in grouping, but also ones varying in prominence. In the stimuli with the most extreme manipulations for both cues (the four corners of the heat maps), listeners chose one of the two stimuli explaining both cues 87% of the time (expected by chance: 50%).[18]

Our model predictions based on the production data were based on the assumption that the prominence decision and the grouping decision are made independently, but the results show they are not.[19] The intensity effect is modulated by the other decision in each model. This makes sense if the two perceptual dimensions mutually compete to explain overlapping cues. It would be unexpected of one decision was not taken into account when making the other, just as it would be unexpected if our visual system decided that the ambient light is blue, but then not take that information into account when deciding on the color of the dress, or not to use our knowledge about the color of the dress when deciding on the ambient light. While we know that the perceptual system generally takes

---

[18] See the supplemental materials for the distribution of responses for these stimuli as well as the baseline condition, where the distribution was close to chance across the four outcomes.

[19] An exploration of the predictions based on models that do take the other decision into account in each model, or based on models that mutually fit both decisions, goes beyond the scope of this paper, but would clearly be of interest.

context into account when making decisions about linguistic percepts (McMurray & Jongman, 2011), what is particularly interesting here is how the two decisions mutually constrain each other.

The perception results differ to some extent from the predictions, for example we see that the 'diagonals of uncertainty' in the perception heat maps cut right through the very corners of the heat maps, whereas according to the predictions derived from the production data, greatest uncertainty is expected on a diagonal to pass below the most extreme intensity manipulations, with intensity taking over for the most extreme intensity manipulations. It may be that listeners use intensity less than predicted since in real life, an intensity difference between two syllables can come about for many reasons. For example, a speaker or listener could move their head slightly during the utterance (loudness drops with the square of the distance of the sound source; it is lowered by obstructing objects (e.g. your head when you're not facing the speaker); and it will be lower if the speaker's mouth does not face you). Other cues like pitch, duration, and vowel formants are not susceptible to environmental conditions in this way, they are entirely under speaker control. [20]

Intensity and cues did not prove to be reliable cues for the foot decision. They only lead to a consistent foot choice in the cases in which a single cue is manipulated to a large degree, where a high degree of intensity is interpreted to cue both stress and initiality, and a high degree of duration as cueing both stress and finality. The fit between the predicted responses for the foot decision and the actual responses looks less good than for the other two decisions. This is unsurprising if intensity and duration are really cues for prominence and grouping, and these are the basic perceptual decisions that a listener makes, while the

[20] Crowhurst and Teodocio Olivares (2014) and Crowhurst (2016) found that intensity has a stronger effect on grouping perception than duration (prompting the 'intensity wins hypothesis'). Which cue is stronger may just depend on the degree of manipulation, however. As acknowledged in Crowhurst and Teodocio Olivares (2014), the difference of 12dB used in the paper is huge compared to the variability observed in natural speech (and the 9dB used here may also be very high compared to typical stress and grouping induced variability).

choice in foot type is derivative. A deviation in one decision (grouping or prominence) from the expectation will result in a perceiving a different foot type, but a deviation in both will result in hearing the same foot type again. There is therefore no monotonous relationship between the error in their decision and the actual response. Intensity and duration are simply not reliable as cues for the distinction between iambs and trochees per se, and did not come out as significant predictors for the foot choice here, and neither did their interaction.

Some of the prior research took the iamb/trochee distinction to be at the heart of the ITL (e.g. Hayes, 1995). Hay and Diehl (2007), e.g., discussed the possibility that listeners parse the signal first for iambic/trochaic patterns as a general strategy of speech segmentation. Our results, however, suggest that when it comes to explaining the ITL, the notions of *iamb* vs. *trochee* are in fact epiphenomenal.

Of course, it could also be that the foot-decision led to less reliable results because participants were not directly asked about the foot choice in the experiment. After all, in this experiment, the foot choice was reconstructed from the prominence and grouping choices. Experiment 3 will address this concern.

Experiment 3: Tone perception

The ITL was originally observed for tone sequences. While we know independently from linguistic research that the two dimensions of grouping and prominence each play a separate role in speech, it is not obvious that they would in tone sequences. An experiment was conducted to see whether they do, and whether the proposed explanation for the ITL carries over to non-speech stimuli.

It could be that both dimensions play a role in non-speech stimuli because our perceptual system somehow parses all incoming information along both dimensions. Or maybe it does so at least for those types of sounds that have a plausible interpretation as having been produced by the motor movements of an organism, such as music and speech.

Tone sequences are sometimes related to music, and Lerdahl and Jackendoff (1983) propose that music is also cognitively organized into a grouping structure, which parallels linguistic constituent structure, and a metrical structure, which represents the location of beats and is similar to stress and accentuation (see Katz, 2018; Patel, 2007, for detailed discussions). Performed music shows some similarities to speech in how these dimensions are phonetically cued. Grouping or 'phrasing' is encoded in music performance by an increase in duration at phrase endings, with greater degrees of lengthening observed at greater junctures (Repp, 1992; Todd, 1985). And Gabrielsson (1987, p. 98) observed that in piano performances the 'termination of each phrase is thus associated with diminishing amplitude.' The presence of a beat is often encoded by an increase in intensity, similar to the effect of stress in speech, and affects music performance independently of the grouping structure (Drake & Palmer, 1993). These findings suggest that the hypothesis under considration here might extend to music, maybe also tone sequences, even if clearly neither is perceived as speech. In order to establish whether two-dimensional parsing can indeed account for the ITL such non-speech stimuli, a perception experiment on tone sequences was conducted.

Methods

A total of 30 participants were run in the speech perception study. 20 were Canadian, and 10 from the US. 25 were female, and the others male. The experiment was conducted under the same circumstances as the experiment on speech. Participants again filled out a language questionnaire and a music questionnaire. As in the production study, most participants spoke French in addition to English to some degree, some of them fluently. Participants had a median of 5 years of music lessons.

One difficulty with testing the hypothesis on tones is that in a sequence of otherwise identical tones, there are no direct parallels to the grouping and the prominence tasks that used in the case of speech. Tones of different frequency cannot easily be used to mimic the

difference between *ba* and *ga*, since frequency itself has been shown to have ITL-like effects (De la Mora et al., 2013; Rice, 1992), although they are less systematic (Rice, 1992; Woodrow, 1911). To get at the grouping of the tones, participants were instead asked whether the final group of two tones they heard was interrupted by the end of the sound file, or whether it was flush with the end. For this task, it was crucial therefore that there was no off-ramp with overlayed noise at the end, since that would have obscured whether the last group was interrupted or not. The second question participants were asked was whether they heard the first or the second tone within a group as prominent, essentially the foot-choice task.[21] From the combined answers of these questions, we were able to code which tone, even or odd, was heard as prominent.

The tone stimuli were similar to the speech stimuli in Experiment 2, in that intensity and duration were each manipulated in steps on each tone. We generated the sequences using a Praat script. Each individual tone was generated from a formula based on a simple sine function. Each tone had a frequency of 500Hz, and was faded in and faded out over the course of 10ms with a quarter sine/cosine function, in order to prevent perceptual artifacts at their edges.

The baseline duration of a tone was 100ms with manipulation steps at 140, 180, and 220ms, matching the durations of the speech manipulation. The baseline intensity was 77dB, with manipulation steps at 79, 81, and 83dB. The intensity steps were less extreme than in the speech stimuli, since matching the intensity steps seemed to create an overly strong intensity effect based on intuitions by the members of a class that the stimuli were played to. The sequences were created by first generating a sequence of two tones. As in the case of the speech stimuli, a given cue was only ever manipulated on one of the two tones, but both duration and intensity could be manipulated within a stimulus (either on

[21] An alternative task could have been to ask whether the final tone they heard was a prominent one or a non-prominent one, but this task was considered harder by several students I played the stimuli to, especially in combination with the grouping task we use.

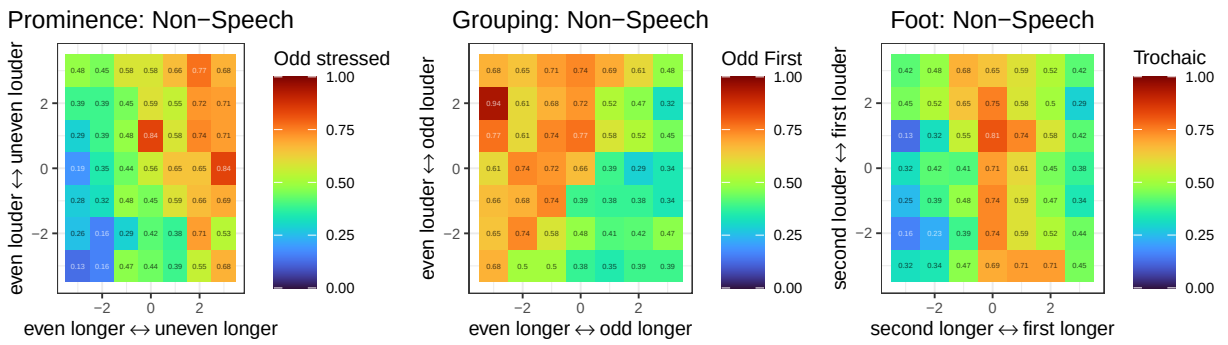

Figure 6. Heat map of the proportion of responses to the tone stimuli for the prominence decision, the grouping decision, and the foot decision. The prominence decision was inferred from the responses to the grouping and the foot decision task.

the same tone or on the other tone), again making for 49 combinations. The manipulated two-tone sequence was then repeated 8 times, creating a sequence of 16 tones, each separated by a pause of 125ms. For the tone sequences, less repetitions were used than for speech, since with longer sequences, it seemed harder to evaluate whether the last group was flush with the end. Again, the baseline sequence was run twice, for a total of 50 stimuli. Presentation order was random. The stimuli had an 800ms on-ramp with overlayed white noise at the beginning of the stimulus, but no off-ramp at the end.

The statistical analysis was parallel to that in Experiment 2 (full models are provided in the supplemental materials).

Results

The plots in Figure 6 summarize the results. Overall, the observed pattern is qualitatively very similar to the pattern observed for speech in Experiment 2, even if the responses are overall a bit closer to chance (as reflected by the greener colour palette).

The grouping responses are most consistent when duration and intensity were manipulated on different syllables, providing consistent cues to grouping (toward the top left and bottom right), and closer to chance when the two cues were manipulated on the same syllable, providing inconsistent cues to grouping (toward the bottom left and top

right). Both the intensity effect ($\beta = 0.19$; s.e. $= 0.04$; $p < 0.001$) and the duration effect ($\beta =$ -0.36; s.e. $= 0.12$; $p < 0.001$) on the grouping decision were significant. There was also a significant interaction between the perceived prominence and the effect of duration, such that the duration of uneven syllables was less of a cue for finality when they were heard as stressed ($\beta = 0.24$; s.e. $= 0.07$; $p < 0.001$), showing that the two decisions were not entirely independent.

There was a small bias in the baseline case (the center cell) toward hearing the uneven tones as initial. This is probably simply an order effect. In Experiment 2, which counterbalanced whether the sequence began with *ba* or *ga*, we observed a significant order effect in the statistical model. In the case of tones, order effects cannot be tested for in this way, since all tones had identical frequency. However, the coefficient for the intercept of the grouping model was significantly different from 0, and this arguably means that tone order had a significant effect also here ($\beta = 0.33$; s.e. $= 0.16$; $p < 0.04$).

The prominence percept as reconstructed from the grouping response and the foot choice. For example, if a participant reported that the last group of tones was flush with the end and they heard iambic prominence, then the even tones were prominent. The qualitative pattern is again very similar to the pattern found in speech in Experiment 2. Responses were most consistent when duration and intensity were manipulated on the same syllable, providing consistent cues for prominence, compared to when they were manipulated on different syllables, when they provide conflicting cues and responses were closer to chance. Intensity ($\beta = 0.16$; s.e. $= 0.05$; $p < 0.001$) and duration ($\beta = 0.43$; s.e. $= 0.07$; $p < 0.001$) each significantly affected the prominence decision. There was also a main effect of grouping, such that listeners were more likely to hear a tone as prominent if it was initial ($\beta = 0.31$; s.e. $= 0.13$; $p < 0.02$). Finally, there was also an interaction between the grouping decision and the effect of duration ($\beta = 0.23$; s.e. $= 0.07$; $p < 0.001$).

The rightmost heat map in Figure 6 illustrates the responses for the iambic-trochaic question. We can see that when only intensity was manipulated, listeners tended to hear

trochees, and when only duration was manipulated, they tended to hear iambs, replicating the basic ITL effect.

In the absence of any acoustic differences between the tones (the baseline condition shown in the center cell of the heat map), listeners showed a slight bias toward hearing trochees(71% trochee responses for baseline), just as in the case of speech. It should be noted, however, that there was again no overall trochee bias when considering the entire set (50% trochee responses overall).

A model for the foot decision was fit as well, in order to see whether intensity and duration generally provide reliable cues for the distinction between iambic and trochaic sequences. The model showed a significant effect of duration, but with an effect size that was only about 1/3 of the effect of duration on the grouping and prominence decision respectively ($\beta = 0.11$; s.e. $= 0.05$; $p < 0.02$), and there was no significant effect of intensity, just as in the case of speech.

Discussion

The results of Experiment 3 for tone sequences are qualitatively very similar to those of Experiment 2 for speech sequences. This provides evidence that tone sequences, too, are parsed along the dimensions of prominence and grouping. We also saw evidence that the two decisions are not independent from each other, although the precise way they affect each other was different from the case of speech in that the effect of duration rather than intensity was modulated by the other decision. Another parallel to the speech case is that there was again a bias toward hearing trochees in the baseline equisound sequence, although, as in the case of speech, there was no such bias in the overall data.

There was also a difference to Experiment 2: The results overall seem a bit closer to chance, which is reflected in the figure in the slightly paler colors in the heat maps when compared to the heat maps for Experiment 2. In the case of intensity, this could be because a less extreme intensity manipulation was used, but there was also a difference in

the effect of duration (see the supplemental materials for an overall regression model pooling the two experiments, which shows significant interactions between the acoustic predictors and the speech/tone distinction). Possible reasons for these discrepancy between the perception of tones and speech could be the differences in the speech and tone tasks, or the fact that the tone sequences were slightly shorter. However, Bhatara et al. (2016) also found weaker effects in their tone experiment than Bhatara et al. (2013) in their speech task (judging by the figures), and they had identical tasks for both types of stimuli.[22] Maybe the effects are smaller because participants have much more experience listening to speech compared to tones. [23]

Finally, the foot choice again was showed again a less consistent relation to the acoustic cues compared to grouping and prominence, just as in Experiment 2. This was the case despite the fact that here, participants were directly asked about their foot choice and their grouping choice, while the prominence choice was reconstructed from the grouping and foot task. Nevertheless, the acoustic cues showed a closer correspondence to their prominence choice than to their foot choice. This supports the hypothesis proposed here, that prominence and grouping are the 'true' perceptual dimensions along which participants parse, while the foot type is not what participants listen for specifically.

The superficial similarity between the responses for speech and for tones does not necessarily mean that the same mechanisms are at play. A reviewer suggested that maybe in speech, the dimensions at play are prominence and grouping, while with tones, it is meter and grouping. Musical meter is a pattern of regularly occurring beats, and listeners take meter as part what a performer intends to convey. While speech shows a tendency

---

[22] Thanks to a reviewer for pointing out the parallel to the two prior studies.

[23] Related to this, I checked whether the amount of musical training affected the size of the effects in the tone experiment(cf. Boll-Avetisyan, Bhatara, & Höhle, 2017; Boll-Avetisyan et al., 2016) , but did not find such a difference. Of course, maybe the experiment did not have enough participants to look at such individual differences systematically, and maybe even musicians have much more experience with speech stimuli than with music. See the supplemental materials for more information.

toward metrical regularity, meter is not an organizational principle of speech in the same way it is in music (Patel, 2007). The difference between the use of meter in music and the use of prominence in speech can be illustrated by looking at a hypothetical sequence that contains an 'odd-ball' with an unexpected prominence relation:

(1) . Tones: (X x) (X x) (X x) (x X) (X x) . Syllables: (BA ga) (BA ga) (BA ga) (ba GA) (BA ga)

The fourth group in each example shows a deviant prominence pattern. The interpretation of the deviation in the sequence in (a) could be that the performer made an error, or maybe that they introduced this variation intentionally as a fun violation of the expectation. The musical beat might still be perceived on the first tone in this case. By contrast, if we understand the syllables in (b) as someone's attempt at listing words, we would interpret the deviation in the fourth group as nothing more than intending to convey the word baGA instead of the word BAga. In a list of words, there is normally no top-down expectation that the pattern of prominence be recurrent.

So it is possible that cues are used differently when listeners make inferences about an abstract meter, and they do so in music, but not in speech. However, rather than thinking of meter (in music) and prominence (in speech) as different dimensions, it may be more productive to think of meter as top-down expectations about upcoming prominences, to the point where a perceived beat does not necessarily have to align with actual prominent acoustic event. Such top-down expectations are generally present in (at least certain kinds of) music, but absent or at least weaker in speech. In Experiment 2, however, prominences in the speech sequences actually were predictable: Participants were told that each sequences consisted of repetitions of a words, so regularity was expected. There is evidence from the psycho-acoustic literature that an acoustic event that is expected is processed differently from one that cannot be anticipated (Kuroda, Tomimatsu, Grondin, & Miyazaki, 2016), as originally hypothesized already in Woodrow (1909). This means that

to further tease this apart, more experiments will be necessary, e.g. looking at sequences of speech and tones that are less regular. It could also be interesting to see how responses change if one presents the speech stimuli as performances of a dadaist poem, in which case the metrical regularity might be taken as part of the intent rather than an accidental emergent property of particular words chosen. Or conversely, one could present the tone sequence as a sequence of words encoded in Morse code, in which case the metrical regularity as such might not be taken as part of the intent, but rather as emergent from the repetition of the intended words.

General Discussion

A listener's perception of an acoustic signal reflects an attempt of the perceptual system to find a plausible 'auditory description' for the signal (Bregman, 1981). Similar to the phenomenon of auditory streaming, and other situations in which an auditory scene is composed from multiple sources (Bregman, 1977, 1990; Bregman & Campbell, 1971), the results suggest that listeners attribute the properties of the incoming acoustic signal to two separate causes. The decisions about prominence and grouping are similar to certain decisions in the visual domain, for example the decisions about the size and distance of an object, or the decisions about the color of an object and the hue of the background light. In each of these cases, there are two mutually constraining perceptual decisions at play explaining the same or at least overlapping cues.

Our perception results replicate the basic ITL effects, but provide a new explanation for why they arise. Alternating long and short sounds tend to be perceived as iambs, because increased duration is a cue for both prominence and finality, and sufficiently increasing duration on every other sound will lead to the percept of a series of iambs. Alternating loud and soft sounds tend to be perceived as trochees because increased loudness is a cue for both prominence and initiality, and sufficiently increasing the intensity of every other sound will lead to the percept of a series of trochees.

However, the cases in which one cue is manipulated in an extreme way are only special cases of a more general pattern, and they are not representative of the full generalization. We saw that similar to a Necker cube, there are often two stable interpretations of the signal, differing either only in grouping or in prominence, depending on the relation between the cues. But in these cases, the choice between iambs and trochees was close to chance. Intensity and duration are generally poor cues for the distinction between iambs and trochees, but excellent cues for grouping and prominence. To see the full pattern, two separate tasks were needed, whereas previous studies only used a single choice which left the exact percept underdetermined. Listeners disentangle the two dimensions by interpreting one cue relative to the other. When they correlate, they are interpreted as cues for prominence, where they anticorrelate, as cues for grouping. This offers a recipe for how listeners differentiate between multiple effects on a given cue. For example, this may also explain how listeners can tease apart the effects of word-initial lengthening observed in Fougeron and Keating (1997) from those of final lengthening Wightman et al. (1992).

The experiments had various limitations. For the perception studies, only intensity and duration an duration were manipulated. It would be interesting to explore other cues. As we saw in the production study, F_0 or and formants are also involved in cueing both dimensions. Also, following earlier studies, this study only looked at nonce word perception, but the same patterns are predicted for actual words (e.g., *UPset*, *upSET* and *SETup*, or *set UP*).

Furthermore, the cue distribution at the heart of the ITL is not restricted to word segmentation, and could be tested at other levels. The cue distribution for sentence-level prominence and phrasing reported was shown to be very similar at the phrase level when producing sentences (Wagner & McAuliffe, 2017, 2019). Fitzroy and Breen (2020) found a distribution of intensity cues in performances of texts with a poetic meter that are compatible with these results. (Herman, 2000) found that intensity decreases at the ends of

discourse moves, which is expected given the phenomenon of intensity downdrift (cf. Pierrehumbert, 1979), suggesting that a similar cue distribution helps disentangle discourse structure. This suggests that the parsing pattern we observed for prominence and grouping at the word level may in fact be part of a very general parsing strategy for speech, across different levels of organization.

Another limitation of this study is that it restricted participants to two binary choices, but it would be interesting to offer a richer set of options. In addition of the four options offered here, listeners might have perceived groups of two sounds of equal prominence, or sounds of alternating prominence that are not grouped, or an unstructured sequence of sounds without any prominence or grouping differentiation. Looking at these options more closely might be revealing. Some listeners might also have attributed the extra length of a syllable to encode phonemic information. In a language with a phonemic vowel length distinction, excessive duration might be taken to reflect vowel length. In a language in which length is not used to distinguish vowels, e.g. Spanish, an unduly long vowel could still be taken as a sequence of two vowels, and a *...bagabaga..* sequence could be interpreted as a *abaga* or *agaba*.[24]

Perhaps the main question that we are left with is the following: While the results show how the ITL emerges from the cue distribution observed for the two separate dimensions of grouping and prominence in speech, the results do not tell us about the source of this cue distribution. The results show a correlation between perception and production, but which way the does the causal relation go?

One possibility is that low-level psycho-acoustic effects explain the perceptual effects of the cues, and that speakers design their utterances to exploit these effects in order to achieve the intended effects. Woodrow (1909), for example, found in tone sequences, a relatively longer sound will make the previous interval between sounds appear shorter, and

[24] Kusumoto and Moreton (1997) also considered how phonotactic cross-linguistic differences might influence results of ITL judgments in discussing their results for Japaense and English.

a relatively louder sound will make the following interval appear shorter. This alone, if true, could explain the grouping effect: Based on the Gestalt principle of proximity (Wertheimer, 1923), sounds separated by perceptually longer intervals should be perceived as groups. And the observed prominence effect could then simply be driven by the energy of the manipulated sound. Both duration and intensity increase energy and hence contribute to perceiving those sounds as figure rather than ground.

This line of reasoning runs into problems, however. The effects that Woodrow (1909) speculates to be responsible for the perceived grouping appear not to be observed when grouping is not an issue, as for example in sequences of two tones. If the second of two tones is longer, the interval between the two tones is rated as longer than when the tones are of equal size (Hasuo, Nakajima, Osawa, & Fujishima, 2012). Similarly, in sequences of three sounds an interval appears longer if a longer sound follows (Hasuo, Nakajima, & Hirose, 2011; Kuroda et al., 2016). Listeners more generally perceive the timing of an event not based on its physical onset, but rather depending on its center of gravity, or its 'perceptual center,' which for longer sounds falls later relative to their onset (Howell, 1988; Marcus, 1981; Morton, Marcus, & Frankish, 1976). Perceptual center effects have been observed both in speech (Morton et al., 1976) and in tones (Terhardt & Schütte, 1976). P-center effects have also been observed for intensity, such that increasing the intensity of a sound brings the p-center closer to the onset (Howell, 1988; Terhardt & Schütte, 1976). Crucially, the direction of the perceptual center effects for duration and intensity is the *opposite* of what would be needed in order to explain the grouping effect observed in the ITL.

Woodrow (1909) actually already noticed these patterns, and and concluded that the effects of both intensity and duration change depending on whether a listener *expects* a louder/longer sound or not. "A more intense stimulus, if unexpected, causes a relative underestimation of the interval preceding it; if expected (or regularly recurrent) a relative underestimation of the interval following it," (Woodrow, 1909, p.11). And he notes similar

interactions for duration. Woodrow's observation anticipates p-center theory, but also the finding that perceptual effects in sequences of three sounds are very different depending on whether these three sounds are repeated to form a longer sequence, and the pattern hence predictable (Kuroda et al., 2016). The ability to predict an upcoming sound, however, arguably relies on having analyzed a sequence into groups. It seems then that the perceived increase in interval length before loud tones and after long tones is not the *cause* of the perceived grouping, but rather its *effect*.

And indeed, many previous studies have found that intervals between groups are perceived as longer than intervals within groups (Bolton, 1894; Fraisse, 1956; Geiser & Gabrieli, 2013; Miner, 1903; Thorpe & Trehub, 1989; Woodrow, 1909), even holding duration and intensity constant (Geiser & Gabrieli, 2013). This effect has been called a 'duration illusion', and has been found in children of about 9 months already (Thorpe & Trehub, 1989). Listeners are also less accurate in estimating the lengths of intervals between groups than within groups (Fitzgibbons, Pollatsek, & Thomas, 1974; Fraisse, 1956; Geiser & Gabrieli, 2013; Thorpe & Trehub, 1989).

If perceptual laws such as p-center effects cannot explain the grouping effect of intensity and duration, maybe we need to look for an explanation in production instead. One possibility is that the cue distribution for grouping is a result of the dynamics of motor planning in speech production. Final lengthening in speech, for example, may in part be a result of general constraints on motor planning. It has also been related to specific constraints of speech production planning. Production planning proceeds hierarchically, such that higher units like words are planned before the fine phonetic details of syllables are fully fleshed out (Sternberg, Monsell, Knoll, & Wright, 1978). Word-final lengthening has been characterized as a way of 'buying time' for word-level planning for an upcoming word while the fine phonetic motor commands of the current word is being planned and executed (Oller, 1973). And the observed intensity decrease throughout an utterance is sometimes attributed to the drop in sub-glottal pressure while exhaling (Vaissière, 1983).

Maybe such constraints on the production system are the source of the ITL. This might also explain why there are similar results in speech and in non-speech, if both speech and tone sequences were treated by listeners as having been produced by the motor movements of an organism. It has long been noted that music performers, in order to not sound mechanical, have to deviate from the score to create a natural-sounding rhythm (Bengtsson & Gabrielsson, 1983). These deviations in part reflect constraints on motor planning, but they are also expressive, and have been argued to to be used evoke associations with physical motion (Gabrielsson, 1987; Kronman & Sundberg, 1984; Repp, 1992; Todd, 1992). Honing (2003) related phrase-final ritardando in music to the 'exertion of a continuous breaking force.' Other forms of physical interpretations of musical events are reviewed in Schlenker (2019). An account of the cue distribution in terms of motor-planning-based or other physical interpretations of the cause of the sound does not explain, however, why the same pattern should be observed in the tactile domain when perceiving electro shocks at regular intervals (Woodrow, 1909), or in the visual domain (Miner, 1903; Peña et al., 2011).

An explanation of the cue distribution in terms of attention also seems conceivable. Starting with Bolton (1894) andWoodrow (1909), many researchers have linked rhythm perception and ITL effects to attention, and recent models often relate rhythm to perceptual oscillators and their role in modulating attention (Brochard, Abecasis, Potter, Ragot, & Drake, 2003; Jones & Boltz, 1989; Large, 2008; van Noorden & Moelants, 1999). Metrical prominence in music (Fitzroy & Sanders, 2015; Jones & Boltz, 1989) and speech (Pitt & Samuel, 1990) has been shown to correlate with greater allocation of attention. With respect to subjective rhythm in equisound sequences, this literature has focused on the perception of prominence (Bååth, 2015). It is conceivable, though, that grouping could be accounted for in such models as well. A louder sound, for example, may attract attention, and this may lead to the inference that a new word is beginning. Also, overdue events that occur later than expected receive more attention (Jones & Boltz, 1989; Kim & McAuley, 2013), so lengthening or increasing interval length between sounds might be used

to draw attention to a following sound and lead to the inference that a new word begins. It is not clear how exactly such attention-based accounts of rhythm could fully rationalize the observed acoustic cue distribution, but this might be a promising avenue to pursue.

Finally, it is also possible that listeners have simply learned the pattern from their experience with their native language. While at least some phonetic aspects of how grouping and prominence are encoded may be universal (Vaissière, 1983), there is also substantial variation in both, for example intensity is used to mark accentual prominence in English but not in Japanese (Beckman, 1986). Such differences could explain cross-linguistic variation in ITL effects. That alternating louder and softer tones leads to the percept of trochees has been argued to be cross-linguistically robust, and has been shown for Japanese (Iversen et al., 2008; Kusumoto & Moreton, 1997), French (Bhatara et al., 2013; Hay & Diehl, 2007), German (Bhatara et al., 2013), and Spanish (Crowhurst, 2016; Crowhurst & Teodocio Olivares, 2014). The effect of duration, on the other hand, has been found to vary between these languages, suggesting that it is modulated by language experience (see Crowhurst, 2019, for a review).

Our understanding of the cross-linguistic variation of ITL effects is limited, however, since the prior literature has not sufficiently teased apart grouping from prominence. Most prior studies used the foot-decision tasks, which as we saw conflates these two dimensions. We cannot tell from these responses whether the observed variation is mostly seen in the grouping or in the prominence dimensions. For example, Japanese listeners show variability in whether they prefer long-short or short-long grouping (Iversen et al., 2008; Kusumoto & Moreton, 1997; Yoshida et al., 2010), a fact that Iversen et al. (2008) attributed to differences in word order (function words precede lexical words in English but follow them in Japanese). This would be tantamount to a difference in the grouping function. Bhatara et al. (2013) argued that French differed from German because of differences in the prominence function. Both studies, however, used a foot-choice task, which can actually not establish whether the observed differences where due to grouping or prominence.

The same issue arises when looking at evidence from developmental studies (Abboub et al., 2016; Bion et al., 2011; Trainor & Adams, 2000; Yoshida et al., 2010), or work on bilinguals (Molnar et al., 2014). The finding in this literature is that the intensity-side of the ITL is attested early on in development, and across languages, while the duration-side is acquired later and less robust cross-linguistically. Since this literature mostly uses a segmentation task, which only gets at grouping, these findings that duration is less robust across languages and developmental stages may only pertain to the dimension of grouping. It is not clear whether when it comes to prominence, the duration cue is also more malleable by language experience.

Cross-linguistic differences in encoding the two dimensions may also inform the literature 'rhythm classes' in novel ways (Dauer, 1983; Ramus, Nespor, & Mehler, 1999). Existing ways of trying to quantify cross-linguistic differences in rhythm have been met with criticism, since these measures seem to relate more to differences at the level of phonotactics and phoneme inventory (Arvaniti, 2012; Dauer, 1983). Grouping and prominence, however, arguably capture core aspects of rhythm. If languages systematically vary in the way the cues that encode the dimensions of grouping and prominence are distributed, this would have fundamental consequences for how these languages are parsed, and may better capture typological differences in what we intuitively call rhythm.

We also need a better understanding of the cognitive functions of the two dimensions in speech and tone/music processing. Morgan, Edwards, and Wheeldon (2014), for example, showed that if stress in words is marked with intensity (as opposed to pitch), there is better recall for trochaic words compared to iambic words. They concluded that "intensity is a critical acoustic factor for trochaic grouping." This conclusion seems incompatible with the results reported here. However, this finding is also expected if participants took intensity to cue word initiality as well as stress, which would lead to the wrong parse in the case of iambic words. If the unstressed syllable in iambic words included a level intensity compatible with it being the initial syllable of a word, the trochee

bias should disappear.

Looking at the cognitive functions of grouping and prominence could also shed some light on whether either of the decisions, grouping or prominence, is in some sense 'prior.' Further studying why exactly equisound sequences are perceeived as rhythmic might be particularly revealing. Attributing the source of this to grouping seems plausible: Suppose that our perceptual system first imposes an arbitrary grouping on an equisound sequence, maybe as a chunking strategy used to build an efficient representation of the sequence in memory (Miller, 1956). Rhythmic grouping has been shown to interact with memory, such that musical notes that form a group are more easily remembered than notes that span group boundaries (Dowling, 1973). Similar effects exist in speech. Reeves, Schmauder, and Morris (2000) show that grouping a list into stress groups (anapests or dactyls) improves recall, so imposing grouping even in the absence of stress could simply be a mnemonic. A chunking effect could also be a consequence of the auditory representation of the signal in the phonological loop (Baddeley, 1986).

Now, once an arbitrary grouping is imposed on an equisound sequence, the listener might use prior expectations about grouping and prominence to make a prominence decision. Since group-final sounds tend to be less loud and longer (at least in speech), the grouping could lead to the inference that the initial sound within each group is too long and too quiet compared to the final tone. Depending on the size of the expected effects of grouping on duration and loudness, this could tip the balance to perceiving the initial or final sound as more prominent than the other.

A prominence-first model also seems plausible, however. The oscillator account of the rhythmic perception of equisound sequences (Bååth, 2015; Large, 2008) could explain why we hear some sounds as more prominent than others. This by itself does not but it fails to explain the percept of grouping. But based on the perceived prominence, our perceptual system might interpret the acoustic cues taking prior expectations about the effects of prominence into account. A listener could infer that the non-prominent sound is louder and

longer than expected, and this could lead to an inference about it being initial or final within a group. We know that speech can have a salient prominence structure that crosscuts grouping: Many poetic forms require a certain meter, for example Shakespeare's iambic pentameter, and we could perceive this regularity even if we didn't know the exact word boundaries. It is not clear that a parallel exists for grouping. There is no poetic form in English that requires, for example, that every other syllable has to coincide with a word boundary. This suggests that we can impose and recognize a prominence pattern without having parsed for grouping.

Either scenario, prominence-first or grouping-first, could shed light on the intuition that it easy to hear an equisound sequence (or at least ones within the time-range used in this study) as either unstructured or grouped, but it seems hard to hear them as grouped without also hearing a prominence asymmetry, or to hear a prominence asymmetry without hearing the sequence as grouped. [25] Of course, it could also be that neither of the decisions is consistently 'prior,' and their relative influence on the each other depends on the task. To understand their precise relation, further experimentation and modeling are needed.[26] Crucially, however, future studies of subjective rhythm will need to tease apart the dimensions of grouping and prominence.

Acknowledgments

This research has been presented at two talks at the University of Pennsylvania in March 2021, one of which was part of the CUNY sentence processing conference. A

[25] This is another potential parallel to Necker cubes, which (depending on the angle at which one looks at the cube), can be seen as flat, or as three dimensional, but if the latter, the both decisions, the one about depth and about viewpoint, have to be made, since otherwise one would end up with in incoherent percept.

[26] It could also be that both directions, grouping-first and prominence-first, are 'tried out', and the analysis (in this case, the grouping and prominence decisions) explaining the sensory data best 'wins.' The bivariate Bayesian models based on the brm package reported in the supplemental materials could be used to explore this hypothesis.

four-page article reporting on an online replication based of the speech perception experiment with a cross-linguistic extension to five more languages has been accepted to the Interspeech conference 2021 in Brno, Czechia. A preprint of the manuscript along with stimuli, data, and code are available as part of an OSF project: http://doi.org/10.17605/OSF.IO/RWBYH (Wagner, 2021).

I would like to thank Albert Bregman for crucial advice and encouragement at the outset of this project, and detailed comments on an earlier draft. Thanks for comments on different aspects of this work by Mara Breen, Meghan Clayards, Ahren Fitzroy, Edward Flemming, Branislav Gerazov, Bill Idsardi, David Kleinschmidt, Jonah Katz, Mark Liberman, Kris Onishi, Jon Sakata, Philippe Schlenker, and Morgan Sonderegger, as well as to the feedback of the participants of a McGill class on the topic in the fall of 2019, and in particular to Ian McNeice and Louise Steben for their work on the production experiment. Thanks also to Jessica Hay and to Megan Crowhurst for sharing the stimuli from their experiments. Finally, thanks to Megan Jezewski for conducting the first pilot experiments in 2017, and various other members of prosodylab for help with running the experiments. All errors are my own. The experiments reported here were conducted under McGill ethics protocol REB#: 401-0409. This research was funded by NSERC Discovery Grant RGPIN-2018-06153: *Three dimensions of sentence prosody*

References

Abboub, N., Boll-Avetisyan, N., Bhatara, A., Höhle, B., & Nazzi, T. (2016). An exploration of rhythmic grouping of speech sequences by French- and German-learning infants. *Frontiers in Human Neuroscience*, *10*, 292. doi: 10.3389/fnhum.2016.00292

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics*, *40*(3), 351–373. doi: 10.1016/j.wocn.2012.02.003

Bååth, R. (2015). Subjective rhythmization: A replication and an extension. *Music Perception*, *33*(2), 244–254. doi: 10.1525/mp.2015.33.2.244

Baddeley, A. D. (1986). *Working memory*. Oxford University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. doi: 10.18637/jss.v067.i01

Beckman, M. E. (1986). *Stress and non-stress accent*. Dordrecht: Foris. doi: 10.1515/9783110874020

Beckman, M. E., & Edwards, J. (1990). Lengthenings and shortenings and the nature of prosodic constituency. In M. E. Beckman & J. Kingston (Eds.), *Papers in laboratory phonology I—between the grammar and physics of speech* (p. 152). Cambridge, UK: Cambridge University Press. doi: 10.1017/cbo9780511627736.009

Beckman, M. E., Edwards, J., & Fletcher, J. (1992). Prosodic structure and tempo in a sonority model of articulatory dynamics. In G. J. Docherty & D. R. Ladd (Eds.), *Papers in laboratory phonology II: gesture, segment, prosody* (pp. 68–119). Cambridge: Cambridge University Press. doi: 10.1017/cbo9780511519918.004

Bell, A. (1977). Accent placement and perception of prominence in rhythmic structures. In L. Hyman (Ed.), *Studies in stress and accent* (p. 1-13). Department of Linguistics, USC.

Bengtsson, I., & Gabrielsson, A. (1983). Analysis and synthesis of musical rhythm. In J. Sundberg (Ed.), *Studies of music performance* (Vol. 39, pp. 27–60). Royal Swedish Academy of Music.

Bhatara, A., Boll-Avetisyan, N., Agus, T., Höhle, B., & Nazzi, T. (2016). Language experience affects grouping of musical instrument sounds. *Cognitive Science*, *40*(7), 1816–1830. doi: 10.1111/cogs.12300

Bhatara, A., Boll-Avetisyan, N., Unger, A., Nazzi, T., & Höhle, B. (2013). Native language affects rhythmic grouping of speech. *The Journal of the Acoustical Society of America*, *134*(5), 3828–3843. doi: 10.1121/1.4823848

Bion, R. A., Benavides-Varela, S., & Nespor, M. (2011). Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences. *Language and speech*, *54*(1), 123–140. doi: 10.1177/0023830910388018

Boersma, P., & Weenink, D. (1996). *PRAAT, a system for doing phonetics by computer. Report 132.* (Institute of Phonetic Sciences of the University of Amsterdam)

Boll-Avetisyan, N., Bhatara, A., & Höhle, B. (2017). Effects of musicality on the perception of rhythmic structure in speech. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *8*(1). doi: 10.5334/labphon.91

Boll-Avetisyan, N., Bhatara, A., Unger, A., Nazzi, T., & Höhle, B. (2016). Effects of experience with L2 and music on rhythmic grouping by French listeners. *Bilingualism: Language and Cognition*, *19*(5), 971–986. doi: 10.1017/S1366728915000425

Boll-Avetisyan, N., Bhatara, A., Unger, A., Nazzi, T., & Höhle, B. (2020). Rhythmic grouping biases in simultaneous bilinguals. *Bilingualism: Language and Cognition*, *23*(5), 1070–1081. doi: 10.1017/s1366728920000140

Bolton, T. L. (1894). Rhythm. *The American Journal of Psychology*, *6*(2), 145–238. doi: 10.2307/1410948

Bregman, A. S. (1977). Perception and behavior as compositions of ideals. *Cognitive Psychology*, *9*(2), 250–292. doi: 10.1016/0010-0285(77)90009-3

Bregman, A. S. (1981). Asking the "what for" question in auditory perception. In M. Kubovy & J. R. Pomerantz (Eds.), *Perceptual organization* (pp. 99–118). Routledge.

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. MIT press. doi: 10.7551/mitpress/1486.001.0001

Bregman, A. S., & Campbell, J. (1971). Primary auditory stream segregation and perception of order in rapid sequences of tones. *Journal of Experimental Psychology*, *89*(2), 244–249. doi: 10.1037/h0031163

Brochard, R., Abecasis, D., Potter, D., Ragot, R., & Drake, C. (2003). The "ticktock" of our internal clock: Direct brain evidence of subjective accents in isochronous sequences. *Psychological Science*, *14*(4), 362–366. doi: 10.1111/1467-9280.24441

Bürkner, P.-C. (2017). brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, *80*(1), 1–28. doi: 10.18637/jss.v080.i01

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., ... Riddell, A. (2017). Stan: a probabilistic programming language. *Grantee Submission*, *76*(1), 1–32.

Cho, T. (2005). Prosodic strengthening and featural enhancement: Evidence from acoustic and articulatory realizations of /ɑ, i/ in english. *The Journal of the Acoustical Society of America*, *117*(6), 3867–3878. doi: 10.1121/1.1861893

Chrabaszcz, A., Winn, M., Lin, C. Y., & Idsardi, W. J. (2014). Acoustic cues to perception of word stress by English, Mandarin, and Russian speakers. *Journal of Speech, Language, and Hearing Research*, *57*(4), 1468–1479. doi: 10.1044/2014_jslhr-l-13-0279

Clayards, M., Tanenhaus, M., Aslin, R., & Jacobs, R. (2008). Perception of speech reflects optimal use of probabilistic speech cues. *Cognition*, *108*(3), 804–809. doi: 10.1016/j.cognition.2008.04.004

Crowhurst, M. (2016). Iambic-trochaic law effects among native speakers of Spanish and English. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *7*(1), 1-41. doi: 10.5334/labphon.42

Crowhurst, M. (2019). The iambic/trochaic law: Nature or nurture? *Language and*

Linguistics Compass, *14*(1), 1-16. doi: 10.1111/lnc3.12360

Crowhurst, M., & Teodocio Olivares, A. (2014). Beyond the iambic-trochaic law: the joint influence of duration and intensity on the perception of rhythmic speech. *Phonology*, *31*(01), 51–94. doi: 10.1017/s0952675714000037

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of memory and language*, *31*(2), 218–236. doi: 10.1016/0749-596x(92)90012-m

Cutler, A., & Carter, D. (1987). The predominance of strong initial syllables in the english vocabulary. *Computer Speech and Language*, *2*, 133–142. doi: 10.1016/0885-2308(87)90004-0

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human perception and performance*, *14*(1), 113-121. doi: 10.1037/0096-1523.14.1.113

Dauer, R. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, *11*(1), 51–62. doi: 10.1016/s0095-4470(19)30776-4

De Jong, K. J. (1995). The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. *The Journal of the Acoustical Society of America*, *97*(1), 491–504. doi: 10.1121/1.412275

De la Mora, D. M., Nespor, M., & Toro, J. M. (2013). Do humans and nonhuman animals share the grouping principles of the iambic–trochaic law? *Attention, Perception, & Psychophysics*, *75*(1), 92–100. doi: 10.3758/s13414-012-0371-3

Dowling, W. J. (1973). Rhythmic groups and subjective chunks in memory for melodies. *Perception & Psychophysics*, *14*(1), 37–40. doi: 10.3758/bf03198614

Drake, C., & Palmer, C. (1993). Accent structures in music performance. *Music Perception*, *10*(3), 343–378. doi: 10.2307/40285574

Edwards, J., Beckman, M. E., & Fletcher, J. (1991). The articulatory kinematics of final lengthening. *The Journal of the Acoustical Society of America*, *89*, 369-382. doi:

10.1121/1.400674

Fitzgibbons, P. J., Pollatsek, A., & Thomas, I. B. (1974). Detection of temporal gaps within and between perceptual tonal groups. *Perception & Psychophysics*, *16*(3), 522–528. doi: 10.3758/bf03198581

Fitzroy, A. B., & Breen, M. (2020). Metric structure and rhyme predictability modulate speech intensity during child-directed and read-alone productions of children's literature. *Language and Speech*, *63*(2), 292–305. doi: 10.1177/0023830919843158

Fitzroy, A. B., & Sanders, L. D. (2015). Musical meter modulates the allocation of attention across time. *Journal of Cognitive Neuroscience*, *27*(12), 2339–2351. doi: 10.1162/jocn_a_00862

Fougeron, C., & Keating, P. (1997). Articulatory strengthening at edges of prosodic domains. *Journal of the Acoustical Society of America*, *101*(6), 3728–3740. doi: 10.1121/1.418332

Fraisse, P. (1956). *Les structures rythmiques*. Louvain: Studia Psychologica, Publications Universitaires de Louvain.

Fry, D. (1958). Experiments in the perception of stress. *Language and Speech*, *1*(2), 126–152. doi: 10.1177/002383095800100207

Gabrielsson, A. (1987). Once again: The theme from Mozart's piano sonata in A major. In A. Gabrielsson (Ed.), *Action and perception in rhythm and music* (pp. 81–103). Royal Swedish Academy of Music.

Geiser, E., & Gabrieli, J. D. (2013). Influence of rhythmic grouping on duration perception: a novel auditory illusion. *PLoS One*, *8*(1). doi: 10.1371/journal.pone.0054273

Hasuo, E., Nakajima, Y., & Hirose, Y. (2011). Effects of sound-marker durations on rhythm perception. *Perception*, *40*(2), 220–242. doi: 10.1068/p6846

Hasuo, E., Nakajima, Y., Osawa, S., & Fujishima, H. (2012). Effects of temporal shapes of sound markers on the perception of interonset time intervals. *Attention, Perception, & Psychophysics*, *74*(2), 430–445. doi: 10.3758/s13414-011-0236-1

Hay, J. F., & Diehl, R. L. (2007). Perception of rhythmic grouping: Testing the iambic/trochaic law. *Perception and Psychophysics*, *69*(1), 113–122. doi: 10.3758/bf03194458

Hay, J. F., & Saffran, J. R. (2012). Rhythmic grouping biases constrain infant statistical learning. *Infancy*, *17*(6), 610–641. doi: 10.1111/j.1532-7078.2011.00110.x

Hayes, B. (1995). *Metrical stress theory: Principles and case studies*. Chicago: University of Chicago Press.

Herman, R. (2000). Phonetic markers of global discourse structures in English. *Journal of Phonetics*, *28*(4), 466–493. doi: 10.1006/jpho.2000.0127

Höhle, B., Bijeljac-Babic, R., Herold, B., Weissenborn, J., & Nazzi, T. (2009). Language specific prosodic preferences during the first half year of life: Evidence from German and French infants. *Infant Behavior and Development*, *32*(3), 262–274. doi: 10.1016/j.infbeh.2009.03.004

Honing, H. (2003). The final ritard: On music, motion, and kinematic models. *Computer Music Journal*, *27*(3), 66–72. doi: 10.1162/014892603322482538

Howell, P. (1988). Prediction of p-center location from the distribution of energy in the amplitude envelope: I. *Perception & Psychophysics*, *43*(1), 90–93. doi: 10.3758/bf03208978

Hyde, B. (2011). The iambic-trochaic law. In C. Ewen, E. Hume, M. van Oostendorp, & K. Rice (Eds.), *The blackwell companion to phonology* (Vol. 2, pp. 1052–1077). Oxford: Blackwell-Wiley. doi: 10.1002/9781444335262.wbctp0044

Iversen, J. R., Patel, A. D., & Ohgushi, K. (2008). Perception of rhythmic grouping depends on auditory experience. *The Journal of the Acoustical Society of America*, *124*(4), 2263–2271. doi: 10.1121/1.2973189

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological review*, *96*(3), 459–491. doi: 10.1037/0033-295x.96.3.459

Jusczyk, P. W., Cutler, A., & Redanz, N. J. (1993). Infants' preference for the

predominant stress patterns of English words. *Child Development*, *64*(3), 675–687. doi: 10.2307/1131210

Katz, J. (2018). *Grouping in music and language.* Retrieved from `http://ling.auf.net/lingbuzz/003938` (Ms. West Virginia University)

Kim, E., & McAuley, J. D. (2013). Effects of pitch distance and likelihood on the perceived duration of deviant auditory events. *Attention, Perception, & Psychophysics*, *75*(7), 1547–1558. doi: 10.3758/s13414-013-0490-5

Klatt, D. H. (1975). Vowel lengthening is syntactically determined in a connected discourse. *Journal of Phonetics*, *3*, 129–140. doi: 10.1016/s0095-4470(19)31360-9

Kronman, U., & Sundberg, J. (1984). Is the musical ritard an allusion to physical motion? *KTH Quarterly Progress and Status Report*, *25*(2-3), 126–141.

Kuroda, T., Tomimatsu, E., Grondin, S., & Miyazaki, M. (2016). Perceived empty duration between sounds of different lengths: Possible relation with repetition and rhythmic grouping. *Attention, Perception, & Psychophysics*, *78*(8), 2678–2689. doi: 10.3758/s13414-016-1172-x

Kusumoto, K., & Moreton, E. (1997). Native language determines the parsing of nonlinguistic rhythmic stimuli. *The Journal of the Acoustical Society of America*, *102*(5), 3204–3204. doi: 10.1121/1.420936

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2018). lmerTest: Tests for random and fixed effects for linear mixed effect models. *Journal of statistical software*, *82*(13), 1-26. doi: 10.18637/jss.v082.i13

Ladd, D. R. (1988). Declination and 'reset' and the hierarchical organziation of utterances. *Journal of the Acoustical Society of America*, *84*(2), 530–544. doi: 10.1121/1.396830

Ladd, D. R. (2008). *Intonational phonology* (2nd ed.). Cambridge: Cambridge University Press. doi: 10.1017/cbo9780511808814

Lafer-Sousa, R., Hermann, K. L., & Conway, B. R. (2015). Striking individual differences in color perception uncovered by 'the dress' photograph. *Current Biology*, *25*(13),

R545–R546. doi: 10.1016/j.cub.2015.04.053

Large, E. W. (2008). Resonating to musical rhythm: Theory and experiment. In S. Grondin (Ed.), *The psychology of time* (pp. 189–232). West Yorkshire: Emerald.

Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA: MIT Press.

Lerdahl, F., & Jackendoff, R. S. (1983). *A generative theory of tonal music*. MIT Press.

Liberman, M., & Pierrehumbert, J. (1984). Intonational variance under changes in pitch range and length. In M. Aronoff & R. Oehrle (Eds.), *Language sound structure* (pp. 157–233). Cambridge, Ma.: MIT Press.

Marcus, S. (1981). Acoustic determinants of perceptual center (P-center) location. *Perception & Psychophysics*, *30*(3), 247–256. doi: 10.3758/bf03214280

Mattys, S. L., & Bortfeld, H. (2016). Speech segmentation. In G. Gaskell & J. Mirkovic (Eds.), *Speech perception and spoken word recognition* (pp. 55–75). New , NY: Routledge.

Mattys, S. L., Jusczyk, P. W., Luce, P. A., Morgan, J. L., et al. (1999). Phonotactic and prosodic effects on word segmentation in infants. *Cognitive psychology*, *38*(4), 465–494. doi: 10.1006/cogp.1999.0721

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal forced aligner: trainable text-speech alignment using Kaldi. In *Proceedings of Interspeech 2017 in Stockholm.* doi: 10.21437/Interspeech.2017-1386

McMurray, B., & Jongman, A. (2011). What information is necessary for speech categorization? harnessing variability in the speech signal by integrating cues computed relative to expectations. *Psychological Review*, *118*(2), 219. doi: 10.1037/a0022325

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81. doi: 10.1037/0033-295x.101.2.343

Miner, J. B. (1903). Motor, visual and applied rhythms. An experimental study and a

revised explanation. *The Psychological Review: Monograph Supplements*, *5*(4). doi: 10.1037/h0093001

Mo, Y., Cole, J., & Hasegawa-Johnson, M. (2009). Prosodic effects on vowel production: evidence from formant structure. In *Proceedings of the tenth annual conference of the international speech communication association (interspeech) in brighton, uk.*

Molnar, M., Carreiras, M., & Gervain, J. (2016). Language dominance shapes non-linguistic rhythmic grouping in bilinguals. *Cognition*, *152*, 150–159. doi: 10.1016/j.cognition.2016.03.023.

Molnar, M., Lallier, M., & Carreiras, M. (2014). The amount of language exposure determines nonlinguistic tone grouping biases in infants from a bilingual environment. *Language Learning*, *64*(s2), 45–64. doi: 10.1111/lang.12069

Morgan, J. L., Edwards, S., & Wheeldon, L. R. (2014). The relationship between language production and verbal short-term memory: The role of stress grouping. *The Quarterly Journal of Experimental Psychology*, *67*(2), 220–246. doi: 10.1080/17470218.2013.799216

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (p-centers). *Psychological Review*, *83*(5), 405. doi: 10.1037/0033-295x.83.5.405

Necker, L. A. (1832). LXI. observations on some remarkable optical phænomena seen in Switzerland; and on an optical phænomenon which occurs on viewing a figure of a crystal or geometrical solid. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *1*(5), 329–337. doi: 10.1080/14786443208647909

Oller, D. K. (1973). The effect of position in utterance on speech segment duration in English. *The Journal of general psychologyurnal of the Acoustical Society of America*, *54*(5), 1235–1247. doi: 10.1121/1.1914393

Patel, A. D. (2007). *Music, language, and the brain.* Oxford University Press. doi: 10.1093/acprof:oso/9780195123753.001.0001

Peña, M., Bion, R. A., & Nespor, M. (2011). How modality specific is the iambic–trochaic

law? Evidence from vision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(5), 1199–1208. doi: 10.1037/a0023944.supp

Pierrehumbert, J. (1979). The perception of fundamental frequency declination. *The Journal of the Acoustical Society of America*, *66*(2), 363–369. doi: 10.1121/1.383670

Pitt, M. A., & Samuel, A. G. (1990). The use of rhythm in attending to speech. *Journal of Experimental Psychology: Human perception and performance*, *16*(3), 564. doi: 10.1037/0096-1523.16.3.564

Poschmann, C., & Wagner, M. (2016). Relative clause extraposition and prosody in German. *Natural Language & Linguistic Theory*, *34*(3), 1021–1066. doi: 10.1007/s11049-015-9314-8

Price, P. J., Ostendorf, M., Shattuck-Hufnagel, S., & Fong, C. (1991, 12). The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, *90*(6), 2956–2970. doi: 10.3115/112405.112738

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, *73*(3), 265–292. doi: 10.1016/s0010-0277(00)00101-3

Reeves, C., Schmauder, A. R., & Morris, R. K. (2000). Stress grouping improves performance on an immediate serial list recall task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1638–1654. doi: 10.1037/0278-7393.26.6.1638

Repp, B. H. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei". *The Journal of the Acoustical Society of America*, *92*(5), 2546–2568. doi: 10.1121/1.404425

Revithiadou, A. (2004). The Iambic/Trochaic law revisited. *Leiden Papers in Linguistics*, *1*, 37–62.

Rice, C. (1992). *Binarity and ternarity in metrical theory: Parametic extensions* (Unpublished doctoral dissertation). University of Texas, Austin.

Schlenker, P. (2019). *Musical meaning within super semantics.* (Ms. Institut Jean-Nicod

and New York University)

Shukla, M., White, K. S., & Aslin, R. N. (2011). Prosody guides the rapid mapping of auditory word forms onto visual objects in 6-mo-old infants. *Proceedings of the National Academy of Sciences*, *108*(15), 6038–6043. doi: 10.1073/pnas.1017617108

Spierings, M., Hubert, J., & Ten Cate, C. (2017). Selective auditory grouping by zebra finches: testing the iambic–trochaic law. *Animal Cognition*, *20*(4), 665–675. doi: 10.1007/s10071-017-1089-3

Sternberg, S., Monsell, S., Knoll, R., & Wright, C. (1978). The latency and duration of rapid movement sequences: Comparisons of speech and typewriting. In G. E. Stelmach (Ed.), *Information processing in motor control and learning* (pp. 117–152). Academic Press. doi: 10.1016/B978-0-12-665960-3.50011-6

Streeter, L. A. (1978, 12). Acoustic determinants of phrase boundary perception. *The Journal of the Acoustical Society of America*, *64*(6), 1582-1592. doi: 10.1121/1.382142

Terhardt, E., & Schütte, H. (1976). Akustische Rhythmus-Wahrnehmung: Subjektive Gleichmäßigkeit. *Acustica*, *35*(2), 122–126.

Thorpe, L. A., & Trehub, S. E. (1989). Duration illusion and auditory grouping in infancy. *Developmental Psychology*, *25*(1), 122. doi: 10.1037/0012-1649.25.1.122

Todd, N. (1985). A model of expressive timing in tonal music. *Music Perception*, *3*(1), 33–58. doi: 10.2307/40285321

Todd, N. (1992). The dynamics of dynamics: A model of musical expression. *The Journal of the Acoustical Society of America*, *91*(6), 3540–3550. doi: 10.1121/1.402843

Trainor, L. J., & Adams, B. (2000). Infants' and adults' use of duration and intensity cues in the segmentation of tone patterns. *Perception & Psychophysics*, *62*(2), 333–340. doi: 10.3758/bf03205553

Turk, A. E., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics*, *28*(4), 397–440.

Vaissière, J. (1983). Language-independent prosodic features. In *Prosody: Models and measurements* (pp. 53–65). Springer. doi: 10.1007/978-3-642-69103-4_5

van Noorden, L., & Moelants, D. (1999). Resonance in the perception of musical pulse. *Journal of New Music Research*, *28*(1), 43–66. doi: 10.7551/mitpress/4812.003.0005

Vayra, M., & Fowler, C. A. (1992). Declination of supralaryngeal gestures in spoken italian. *Phonetica*, *49*(1), 48–60.

Vos, P. G. (1973). *Waarneming van metrische toonreeksen* (Unpublished doctoral dissertation). University of Nijmegen.

Vos, P. G. (1977). Temporal duration factors in the perception of auditory rhythmic patterns. *Scientific Aesthetics/Sciences de l'Art*, *1*, 183–199.

Wagner, M. (2005). *Prosody and recursion* (Doctoral dissertation, MIT). Retrieved from `http://dspace.mit.edu/handle/1721.1/7582`

Wagner, M. (2021). *Two-dimensional parsing explains the iambic-trochaic law—stinmuli, data, and code.* (OSF repository) doi: 10.17605/OSF.IO/RWBYH

Wagner, M., & McAuliffe, M. (2017). Three dimensions of sentence prosody and their (Non-)Interactions. In *Proceedings of Interspeech 2017 in Stockholm.* doi: 10.21437/Interspeech.2017-1500

Wagner, M., & McAuliffe, M. (2019). The effect of focus prominence on phrasing. *Journal of Phonetics*, *77*. doi: 10.1016/j.wocn.2019.100930

Warren, R. M. (2008). *Auditory perception: an analysis and synthesis*. Cambridge University Press. doi: 10.1017/cbo9780511754777

Warren, R. M., & Gregory, R. L. (1958). An auditory analogue of the visual reversible figure. *The American Journal of Psychology*, *71*(3), 612-613. doi: 10.2307/1420267

Weide, R. (1998). *The CMU pronunciation dictionary, release 0.6.* Carnegie Mellon University.

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt II. *Psychologische Forschung*, *4*(1), 301–350. doi: 10.1007/bf00410640

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M., & Price, P. J. (1992). Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of the Acoustical Society of America*, *92*, 1707–1717. doi: 10.1121/1.402450

Woodrow, H. (1909). A quantitative study of rhythm: The effect of variations in intensity, rate and duration. *Archives of Psychology*(14), 1–66.

Woodrow, H. (1911). The role of pitch in rhythm. *Psychological Review*, *18*(1), 54. doi: 10.1037/h0075201

Yoshida, K. A., Iversen, J. R., Patel, A. D., Mazuka, R., Nito, H., Gervain, J., & Werker, J. F. (2010). The development of perceptual grouping biases in infancy: A Japanese-English cross-linguistic study. *Cognition*, *115*(2), 356–361. doi: 10.1016/j.cognition.2010.01.005