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

Michael Wagner

McGill University

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

These supplemental materials provide additional detail on the experiments and their analysis, including full model formulas for the regression models, and tables detailing the statistical results. In addition, stimuli, code, and data have been posted on OSF (Wagner, 2021)

Experiment 1: Production

Additional Information

All experiments were run under the McGill ethics protocol REB#: 401-0409 ('La prosodie: production, perception, et differences interlinguistique').

Participants were seated in a sound-attenuated booth and recorded with a Logitech H390 headset. The target word sequences were visually presented, with word boundaries marked by spaces and stress by accents (e.g., "bága bága bága"), and participants were told that 'the accent indicates where the stress should be. So for example, the word "apple" is pronounced: ápple. In other words, AH-pull, not ah-PULL. And the word "attempt" is pronounced attémpt: ah-TEMPT, not AH-tempt.' Participants were also told to 'Please speak as naturally as possible, as if you were talking to a friend (despite of the fact that these are not real words!).' Each utterance consisted of 6 identical syllables. There were 4 different syllables, *ba*, *bo*, *na*, and *no*. Words with intensity and duration measures further than 2.5 standard deviations from the mean were excluded, which amounted to 4.1% of the data. These were mostly words with unrealistic measures (e.g., intensity $>$ 90db), probably due to the imperfect intensity extraction of the speech software (Praat), or imperfect alignments in the case of duration.

The production experiment included other stimuli, not analyzed here: The experiment also looked at words with three syllables with initial and final stress. A second part of the experiment looked at the production of syllable sequences in which word boundaries were not visually shown with spaces. There were excluded here since they are not directly relevant.

The aligner created a segment-by-segment and word-by-word annotation of the data. Note that each syllable was treated as if it had been a separate word, in order to avoid a bias due to expectations about the word-initial/word-final realization encoded in the acoustic models of the aligner. An example alignment can be found in Figure 1.

Models

Linear mixed effect models were fit for the z-scores of log duration and intensity using the R package lme4 (Bates, Mächler, Bolker, & Walker, 2015), as described in the methods section. The formulas of the

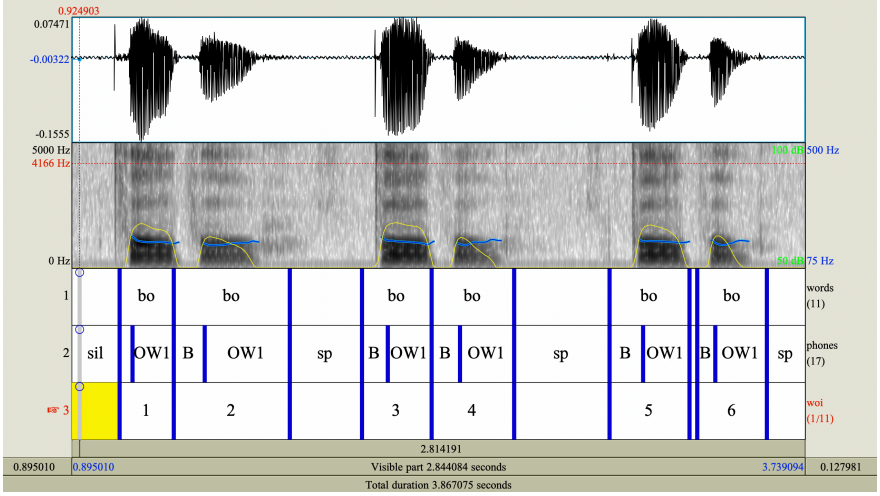

Figure 1. Example alignment for an utterance of *bóbo bóbo bóbo*

intensity model (z-score of intensity measured in dB) and the duration model (the z-scores of the log of the duration of each syllable) were the following:

```
modelIntensity = lmer(maxIntensity.std ~
                        Stressed.vs.Unstressed*Initial.vs.FinalSyllable  +
                        PositionInUtterance
                        + (1|item)
                        + (Stressed.vs.Unstressed+Initial.vs.FinalSyllable||participant),
                        data = proData)


modelDuration = lmer(duration.std ~
                       Stressed.vs.Unstressed*Initial.vs.FinalSyllable  +
                       PositionInUtterance
                       + (1|item)
                       + (Stressed.vs.Unstressed+Initial.vs.FinalSyllable||participant),
                       data = proData)
```

The results of the intensity and duration models are summarized in Table 1. An earlier version of the manuscript used THE POSITION OF THE WORD WITHIN THE UTTERANCE instead of the POSITION OF THE SYLLABE WITHIN THE UTTERANCE. However, this does not address the potential compound between *position of the syllable within the word* (the variable of main interest), the utterance-level declination of acoustic values like intensity and pitch.

	Intensity	Duration
(Intercept)	0.04 (0.12)	0.01 (0.09)
Stressed.vs.Unstressed	$-0.40\ (0.05)^{***}$	$-0.42\ (0.07)^{***}$
Initial.vs.FinalSyllable	$-0.05\ (0.02)^{*}$	$0.25\ (0.07)^{***}$
PositionInUtterance	$-0.15\ (0.02)^{***}$	$-0.03\ (0.02)$
Stressed.vs.Unstressed:Initial.vs.FinalSyllable	$-0.02\ (0.04)$	$0.08\ (0.05)$

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

Table 1
Models for the measures of intensity and duration from the production stimuli.

The formulas for pitch model (semitone of maximum pitch computed relative to 100Hz), the models for F1 (z-score of first formant) and F2 (z-score of second formant), and for silence (z-score of log duration) were identical, except that the model for silence had a simplified random effect structure, in order to avoid a singular fit:

```
modelPitch = lmer(maxPitch.std ~
                      Stressed.vs.Unstressed*Initial.vs.FinalSyllable  +
                      PositionInUtterance
                      + (1|item)
                      + (Stressed.vs.Unstressed+Initial.vs.FinalSyllable||participant),
                      data = proData)
modelF1 = lmer(F1.std ~
                      Stressed.vs.Unstressed*Initial.vs.FinalSyllable  +
                      PositionInUtterance
                    + (1|item)
                    + (Stressed.vs.Unstressed+Initial.vs.FinalSyllable||participant),
                    data = proData)


modelF2 = lmer(F2.std ~
                  Stressed.vs.Unstressed*Initial.vs.FinalSyllable  +
                  PositionInUtterance
                + (1|item)
                + (Stressed.vs.Unstressed+Initial.vs.FinalSyllable||participant),
                data = proData)


modelSilence = lmer(silence.std ~
```

```
                  Stressed.vs.Unstressed*Initial.vs.FinalSyllable  +
                  PositionInUtterance
                + (1|item)
                + (Initial.vs.FinalSyllable||participant),
                data = proData)
```

The results of the models of these other predictors are summarized in Table 2.

	Pitch	F1	F2	Silence
(Intercept)	−0.02 (0.09)	−0.00 (0.20)	−0.01 (0.21)	0.02 (0.12)
Stressed.vs.Unstressed	$-0.15\ (0.05)^{***}$	$-0.26\ (0.04)^{***}$	0.05 (0.03)	0.00 (0.03)
Initial.vs.FinalSyllable	0.09 (0.08)	$0.08\ (0.03)^{**}$	$-0.08\ (0.03)^{**}$	$0.38\ (0.08)^{***}$
PositionInUtterance	$-0.14\ (0.03)^{***}$	−0.00 (0.02)	$-0.06\ (0.02)^{*}$	$-0.31\ (0.03)^{***}$
Stressed.vs.Unstressed:Initial.vs.FinalSyllable	$-0.13\ (0.06)^{*}$	0.09 (0.05)	−0.02 (0.05)	0.10 (0.06)

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

Table 2

Models for the measures of pitch, F1, F2, and silence from the production stimuli.

Independent Components Analysis

Branislav Gerazimov and Bill Idsardi (p.c.) independently suggested to conduct an Independent Components Analysis, in addition to the principal components analysis that was already part of the paper. The plots for the components derived by the ICA are given in Fig. 2.

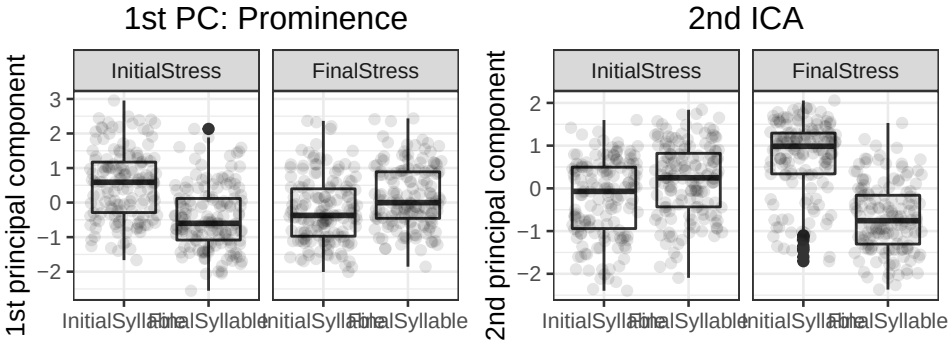

Figure 2. Indpendent components based on an ICA of the intensity and duration measures

It seems that the first component of the ICA shows large stress-related differences for words with initial stress, and only a relatively small difference for words with final stress. The second component shows large stress-dependent differences for words with final stress, and only a relatively small differences for words with initial stress. In a way, this makes the distinction between iamb and trochee seem to matter

more, but not in that there are cues that directly distinguish the words based on foot type, but only in that stress is marked with one cue for one type of word, and with a different cue in the other type of word. It's hard to make sense of this as a way of parsing the signal, even if one only wanted to parse it for foot type. By comparison, the PCA reported in the main paper leads to components that (mostly) map to two separate dimensions, grouping and prominence.

Experiment 2: Predictions for perception

Additional Information

In order to be able to compute predictions from the production experiment, a data frame was created in which half of the productions were coded as if speakers had initiated the production in the middle of the word. To do so, the data was simply doubled, and one half recoded. The whole dat frame was then reorganized, such that for each utterance, acoustic measures for even and uneven syllables were on the same line. The models then tried to predict from the acoustics whether (i) uneven or even syllables were word-initial within that sequence, and (ii) uneven or even syllables were stressed. To do so, Two logistic regression models for grouping and for prominence were fit.

By doubling the data, the p-values in these models may be inflated, since they essentially assume that twice the number of productions were collected. So this will artificially increase power. This seems acceptable here, since the p-values of these particular models were not of interest for the purposes of this paper. These models just serve to illustrate the predictions from the cue distribution.

In order to check whether the precise procedure of how the data set was created mattered, predictions were also computed in a different way, without doubling the data. Instead, a random sample of 50% of half the data was recoded. This was done twice. This lead to qualitatively very similar predictions, although there was slight variation depending which sample was recoded. Since the method of creating the training data set did not appear to matter, the method of doubling the data seemed best, since it avoids the risk of introducing any accidental biases.

Models: Computing predictions for perception from production

Logistic regression models using general linear models (with the R command GLM) with a binomial distribution were used to predict grouping and prominence from the acoustic data. For the predictions used in the main text, the models included intensity, duration, and four additional cues (duration, intensity, pitch, F1, F2, silence). The formula for the prominence model was the following:

```
predProminence = glm(Prominence ~
                           maxIntensity.Even + maxIntensity.Odd +
                           logDuration.Even + logDuration.Odd  +
                           logSilence.Even + logSilence.Odd  +
                           F1.Even + F1.Odd +
                           F2.Even + F2.Odd +
                           maxPitch.Even + maxPitch.Odd +
                           WordPosition,
                         family = "binomial", data=predictData
    )
```

The formula for the grouping model was the following:

```
predGroup = glm(Grouping ~
                  maxIntensity.Even + maxIntensity.Odd +
                  logDuration.Even + logDuration.Odd  +
                  logSilence.Even + logSilence.Odd  +
                  F1.Even + F1.Odd +
                  F2.Even + F2.Odd +
                  logMaxPitch.Even + logMaxPitch.Odd +
                  WordPosition,
                family = "binomial", data=predictData
       )
```

A summary of the models is given in Table 3.

The PREDICT-function was then used to calculate log odds for the responses for the stimuli in the perception experiment based on these models. These were calculated based on the actual acoustic measures from the perception stimuli.

The correlation between the predicted prominence choice in the all-cues model and the actually perceived prominence was 0.34. The correlation between the predicted grouping choice in the all-cues model and the actually perceived grouping was 0.33.

By contrast, the correlation between the prediction foot choice derived from these models and the actual foot choice was only 0.09.

	Prominence	Grouping
(Intercept)	−0.00 (3.64)	−0.00 (1.70)
maxIntensity.Even	−0.44 (0.06)***	−0.15 (0.02)***
maxIntensity.Odd	0.44 (0.06)***	0.15 (0.02)***
logDuration.Even	−4.89 (0.89)***	1.95 (0.39)***
logDuration.Odd	4.89 (0.89)***	−1.95 (0.39)***
logSilence.Even	0.30 (0.12)*	0.33 (0.06)***
logSilence.Odd	−0.30 (0.12)*	−0.33 (0.06)***
F1.Even	−0.01 (0.00)*	0.01 (0.00)**
F1.Odd	0.01 (0.00)*	−0.01 (0.00)**
F2.Even	0.00 (0.00)	−0.00 (0.00)**
F2.Odd	−0.00 (0.00)	0.00 (0.00)**
maxPitch.Even	−0.01 (0.00)	
maxPitch.Odd	0.01 (0.00)	
WordPosition.L	0.00 (0.44)	−0.00 (0.22)
WordPosition.Q	0.00 (0.43)	−0.00 (0.20)
logMaxPitch.Even		−0.01 (0.04)
logMaxPitch.Odd		0.01 (0.04)

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

Table 3
Models for computing predictions from production data for perception experiments using all cues

Simpler models based only on duration and intensity

In addition to the predictions reported in the main paper, predictions calculated based on simpler models that only used intensity and duration measures as predictors are reported below. Since the ITL is about the effects of intensity and duration, one may wonder whether the predictions based on the all-cues model are in fact mostly driven by other cues. The formulas of the simpler models were the following:

```
predProminenceSimple = glm(Prominence ~ maxIntensity.Even + maxIntensity.Odd +
                             duration.Even + duration.Odd + WordPosition,
                           family = "binomial", data=predictData
#
predGroupSimple = glm(Grouping ~ maxIntensity.Even+ maxIntensity.Odd
                + duration.Even + duration.Odd  +  WordPosition,
                   family = "binomial", data=predictData
)
```

A summary of the simpler models is given in Table 4.

Fig. 3 shows that qualitatively, the predictions of the simple models were similar to those of the all-cues model, which are plotted in the main text:

The correlation between the predicted prominence choice in the simple model and the actually perceived prominence in the speech experiment is 0.31, so a bit lower than in the case of the all-cues model.

	Prominence	Grouping
(Intercept)	−0.00 (2.39)	−0.00 (1.27)
maxIntensity.Even	−0.44 (0.04)***	−0.15 (0.02)***
maxIntensity.Odd	0.44 (0.04)***	0.15 (0.02)***
logDuration.Even	−3.52 (0.59)***	2.80 (0.33)***
logDuration.Odd	3.52 (0.59)***	−2.80 (0.33)***
WordPosition.L	0.00 (0.32)	0.00 (0.18)
WordPosition.Q	0.00 (0.31)	−0.00 (0.17)

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

Table 4
Simpler models for computing predictions from production data for perception experiments, just based on intensity, duration, and word position

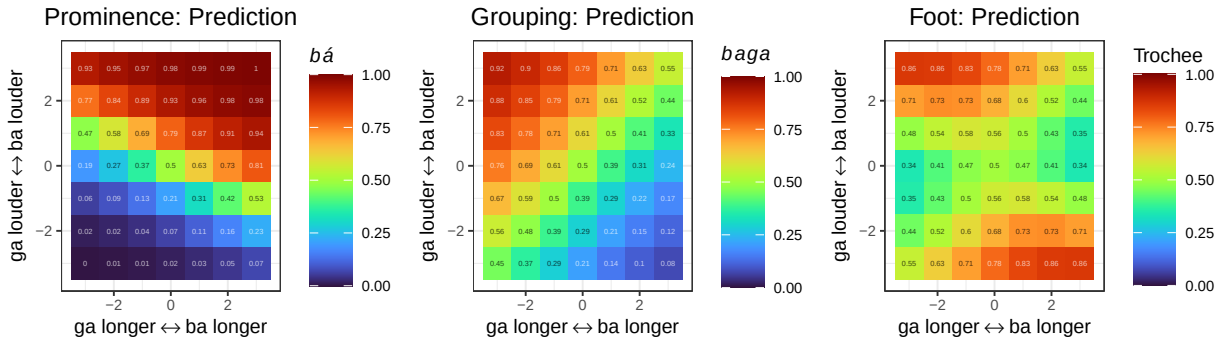

Figure 3. Heatmaps of the predictions for the responses in the grouping decision (left), the prominence decision (middle), as well as for the foot decision (right), based on intensity, duration, and word position, omitting the additional acoustic cues.

The correlation between the predicted grouping choice in the simple model and the actually perceived grouping is 0.35, so slightly higher than in the case of the all-cues model. Overall, the all-cues model fared better in predictions of combinations of both decisions, but not by much.

Direct predictions for foot choice

The predictions for the foot choice discussed so far have been computed from the predicted prominence and grouping respectively. These predictions do not show a close match to the perception data in the speech experiment (which did not directly ask about the foot percept), and the tone experiment (which did directly ask about the foot percept). Here's a potential concern, however: Would the production data give a better prediction for the foot choice if we had computed these predictions directly from the acoustics? This was tested (in response to a related question by a reviewer about the speech perception experiment) by fitting an additional logistic model for the foot choice to the production data. The model formulas used to compute these direct predictions are the following:

```
predFootSimple = glm(Foot ~ maxIntensity.Even + maxIntensity.Odd +
                             duration.Even + duration.Odd + WordPosition,
                         family = "binomial", data=predictData
)
#
predFoot =  glm(Foot ~
                   maxIntensity.Even + maxIntensity.Odd +
                   logDuration.Even + logDuration.Odd  +
                   logSilence.Even + logSilence.Odd  +
                   F1.Even + F1.Odd +
                   F2.Even + F2.Odd +
                   logMaxPitch.Even + logMaxPitch.Odd +
                   WordPosition,
                 family = "binomial", data=predictData
)
```

The model tables are given below. None of the predictors for the foot choice come out significant. This is compatible with the result from the perception studies, were foot choice was also not successfully predicted from intensity and duration.

	FootSimple	FootAllCues
(Intercept)	0.21 (1.09)	−0.86 (1.39)
maxIntensity.Even	−0.01 (0.01)	−0.00 (0.02)
maxIntensity.Odd	−0.01 (0.01)	−0.00 (0.02)
logDuration.Even	−0.33 (0.25)	−0.45 (0.29)
logDuration.Odd	−0.33 (0.25)	−0.45 (0.29)
WordPosition.L	0.04 (0.15)	0.09 (0.18)
WordPosition.Q	−0.03 (0.14)	−0.08 (0.15)
logSilence.Even		−0.06 (0.05)
logSilence.Odd		−0.06 (0.05)
F1.Even		−0.00 (0.00)
F1.Odd		−0.00 (0.00)
F2.Even		0.00 (0.00)
F2.Odd		0.00 (0.00)
maxPitch.Even		0.00 (0.00)
maxPitch.Odd		0.00 (0.00)

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

Table 5
Models for computing predictions from production data for perception experiments for the foot decision, based on a simple model (intensity, duration, word position), and a more complex model with additional acoustic cues as fixed effects

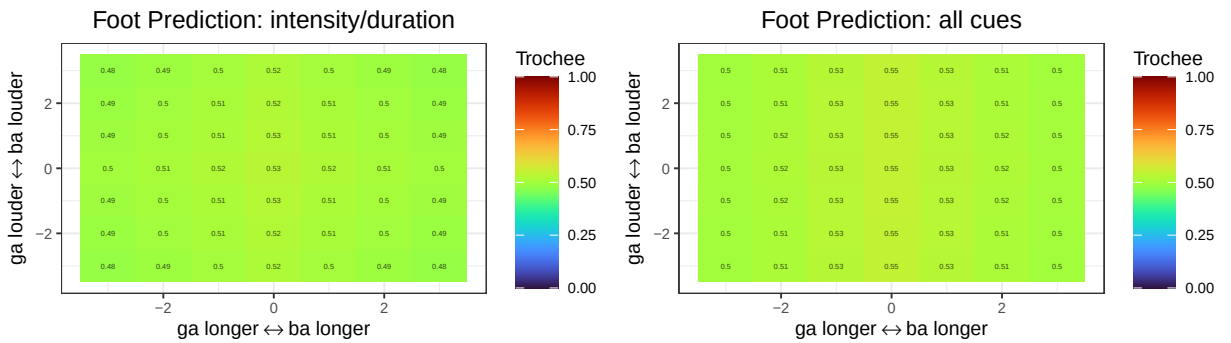

Figure 4. Predictions for the foot decision, directly computed from the acoustic features, based on the simple model (left) and the all-cues model (right).

The plots in Fig. 4 visualize the predictions for the foot decision based on these models. Both the foot model based on all cues and the model just based on intensity and duration results in predictions that are essentially at chance, and neither predicts iambic responsese for extreme manipulations of durations (the outermost cells in the middle row). By contrast, for the foot predictions derived from the grouping and prominence models, both for the simple model (plotted above) and the all-cues model (plotted in the main text), the ITL is correctly predicted for extreme manipulations of only duration and only intensity.

Across all stimuli, however, the direct predictions fare slightly better numerically than the indirect predictions (the simple model shows a correlation with the actual foot decision of 0.12), the all-cues model a correlation of 0.15), but since these predictions are essentially at chance, this means that none of the foot decision models fare very well. These correlations are very low compared to the correlations observed for the grouping and prominence choices.

Experiment 2: Speech Perceptionn

Additional Information

Participants were seated in a sound-attenuated booth and listened to the stimuli via a Logitech H390 headset. In the instructions, participants were told that 'your task is first to answer which words you heard (e.g., baga vs. gaba). You are then asked which syllable in the word was stressed. For example, in the word 'apple,' the first syllable is stressed (stressed marked with capital letters): (AHpple not appEL). But in the word 'attempt', the second syllable is stressed (attEMPT, not ATTempt).'

Biases

In the responses to the perception stimuli (plotted in Fig. 5 in the main text, not here), the responses to the baseline stimuli showed a slight bias toward generally hearing *BA* as initial (the center cell

in the grouping response is blue) and to hear *ga* as stressed (the center cell in the prominence is red). This could reflect prior knowledge about the distribution in the lexicon.

Lexical knowledge has been shown to be play an important role in speech segmentation (Mattys & Bortfeld, 2016), and the observed biases could be due to the distribution of words in the English lexicon. In the word list of the CMU dictionary (Weide, 1998), 7.1% of the words begin with [b], but only 4.3% with [g]. As for prominence, in words in which [ba] occurs it carries main word stress 73.8% of the time, but [ga] 86.1% of the time. This is compatible with the bias observed in the baseline condition.

However, in the overall data set, there was actually a slight bias toward hearing *ga* as initial, and no bias with respect to which syllable was stressed (*BAga* was heard 287 times; *baGA* 288 times; *GAba* 335 times; *gaBA* 340 times). Since the experiment was not designed to test for lexical top-down effects, these questions are not pursued further here.

It could be that the small trochee bias observed for the baseline stimulus was due to the biases in the grouping and prominence responses. If, say, the syllable *ba* (or the sylalble *ga*) was more likely to be word-initial and stressed than *ga*, then this could lead to more trochee responses. However, in that case more of the trochee responses in the baseline condition should have been one of *BAga* or *GAba*. Essentially, which trochee was chose was at chance, though: *BAga* was heard in 13 out of 27 cases. Since the baserate for *Gaba* was higher, this is compatible with a slight bias toward *Baga*. So it looks like the numerical bias toward trochees cannot be explained in this way, and maybe a genuine trochee bias. The number of observations was too low, however, to come to any firm conclusions.

Responses for baseline cases, and for the four corners of the heatmaps

Looking at the baseline condition where syllables are equally long and equally loud, (and baLonger and baLouder equal 0), we see that participants were roughly equally likely to choose any of the four outcomes, with a small bias for trochees:

	baLouder	baLonger	BAga	baGA	GAba	gaBA
1	-3.00	-3.00	3	13	9	0
5	-3.00	3.00	2	0	9	14
9	0.00	0.00	13	14	14	9
13	3.00	-3.00	7	13	0	5
17	3.00	3.00	10	2	1	12

Looking at the stimuli in which both duration and intensity were manipulated to their extreme values (either on the same or on different syllables), we see that the two outcomes that were chosen with the highest frequency were always the two outcomes that could explain all of the cues. For example, if *ga* was louder and longer (i.e., baLouder = -3; baLonger = -3), participants chose *baGA* 13 times (stress

explain loudness of *ga*; stress and finality explain length of *ga*); and they chose *GAba* 9 times (stress and initiality explain loudness of *ga*; stress explains length of length of *ga*). The next most frequent percept, *BAga*, was only chosen 3 times for that manipulation. That interpretation fails to explain why *ga* is louder than *ba*: It is neither initial nor stressed. The two most frequent outcomes in each row are always the ones that, based on the hypothesis, can explain all of the cues.

In other words, as predicted, there were prominence-stable stimuli in which the gruoping percept varied, and grouping-stable stimuli in which the prominence percept varied. These stimuli are similar to Necker-cubes, which also have two valid interpretations.

Models

Mixed effects logistic regression models (using the GLMER command of the R package lme4) were used to analyze how the various factors influenced the prominence and grouping decisions in speech perception. Here's the model formula for the prominence decision:

```
modelProminenceSpeech = glmer(Prominence ~
                          baLouder*baLonger*baga.vs.gaba +
                          ba.vs.gaStart +
                          (baLouder+baLonger|participant),
                          family ='binomial',
                          control = glmerControl(optimizer = "bobyqa",
                                                 optCtrl=list(maxfun=120000)),
                          data = perData)
```

And here's the model formula used to analyze the grouping decision for the speech data:

```
# Grouping decision taking prominence into account
modelGroupingSpeech = glmer(Grouping ~
                        baLouder*baLonger*ba.vs.ga +
                        ba.vs.gaStart +
                        (baLouder+baLonger|participant),
                        family ='binomial',
                        control = glmerControl(optimizer = "bobyqa",
                                               optCtrl=list(maxfun=120000)),
                        data = perData)
```

And the model of the (reconstructed) foot decision:

```
# Foot decision model
modelFootSpeech =glmer(Foot ~
                         baLouder*baLonger +
                         ba.vs.gaStart +
                         (baLouder||participant),
                       family ='binomial',
                       control = glmerControl(optimizer = "bobyqa",
                                              optCtrl=list(maxfun=120000)),
                       data = perData)
```

The results of the models for the speech perception data are summarized in Table 6.

	Prominence	Grouping	Foot
(Intercept)	−0.05 (0.10)	0.23 (0.14)	−0.01 (0.11)
baLouder	−0.30 (0.06)***	−0.34 (0.06)***	−0.07 (0.04)
baLonger	−0.51 (0.08)***	0.50 (0.13)***	−0.05 (0.03)
baga.vs.gaba	0.23 (0.16)		
ba.vs.gaStart	−0.14 (0.13)	0.56 (0.14)***	−0.15 (0.12)
baLouder:baLonger	0.05 (0.02)*	−0.01 (0.02)	0.01 (0.01)
baLouder:baga.vs.gaba	−0.22 (0.08)**		
baLonger:baga.vs.gaba	−0.05 (0.08)		
baLouder:baLonger:baga.vs.gaba	−0.01 (0.04)		
ba.vs.ga		0.23 (0.16)	
baLouder:ba.vs.ga		−0.20 (0.08)*	
baLonger:ba.vs.ga		−0.05 (0.09)	
baLouder:baLonger:ba.vs.ga		−0.03 (0.04)	

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 6
Models for grouping and prominence for the perception experiment on speech

Bivariate Baysian analysis of Speech Perception Data

An arguably even better way to analyze the mutual dependence of the two decisions it to model them with a bivariate model, with a shared random effect for participant. This model does justice to the fact that the two decisions were not independent—they collected during the same trial, and our analysis shows that one decision influenced the other.

The brm package Bürkner (2017), which uses the Stan modeling language (Carpenter et al., 2017), gives us the means to to fit two decisions in a single model. The model used here is a bivariate Baysian mixed effects logistic regression model, which provides model estimates for both the grouping and the prominence decision.

This type of model also seems conceptually more adequate, apart from the adressing the issue of dependence in a better way: When listening to a stimulus, listeners make decisions about prominence, about grouping, about the identity of the phonemes, about the identity of the speaker and their mood, and so on. Our models often treat one decision as the dependent variable (e.g., grouping), and the other decisions as fixed effects (e.g., prominence) or random effects (e.g., speaker), in order to address certain research questions. But ultimately, it would be more adequate to find ways of modeling this kind of data which does justice to the fact that these are just multiple mutually constraining decisions. In the phenomenon under investigation, prominence and grouping are both relevant dependent variables, which mutually constrain each other. Modeling them in a single model does justice to their dependence.

The joint model for both decision was set up using the following formulae for the individual decisions:

```
prominence_fm <- bf(prominence ~
                      (baLouder+baLonger)*baga.vs.gaba +
                      (baLouder+baLonger|p|participant) +
                      ba.vs.gaStart) +
                      bernoulli()

grouping_fm <- bf(grouping ~
                      (baLouder + baLonger)*ba.vs.ga  +
                      (baLouder+baLonger|p|participant) +
                      ba.vs.gaStart) +
                      bernoulli()
```

I set mildly skeptical priors against the hypotheses, and in favor of all coefficients being equal to zero:

```
priorInteract <- c(
set_prior("normal(0, 10)", class="b", resp="prominence"),
set_prior("normal(0, 10)", class="b", ,resp="grouping")
)
```

The code for the bivariate model was as follows:

```
jointModel <- brm( prominence_fm +
                   grouping_fm +
                   set_rescor(FALSE),
```

```
                    data = perData,
                    chains = 4,
                    prior = priorInteract,
                    cores = 4)
```

The model appears to capture the variability in the data well, given the posterior-predictive checks illustrated in Fig. 5.

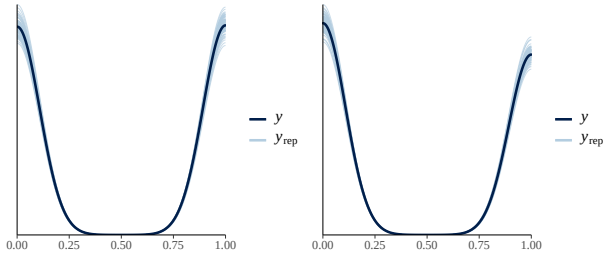

Figure 5. Posterior predictive check for the fit for prominence (left) and grouping (right) for the speech data.

Based on posterior samples from the model, the strength of the evidence for the relevant hypotheses can be evaluated. There is good evidence that greater intensity and greater duration make it more likely that a syllable is heard as prominent (for intensity: $p(\beta > 0) = 1$; for duration: $p(\beta > 0) = 1$.

There is also strong evidence that greater intensity and lower duration makes it more likely that a syllable is heard as initial (for intensity: $p(\beta > 0) = 1$; for duration: $p(\beta < 0) = 0.99975$.

Finally, the model suggests that there is strong evidence that when making the prominence decision, the effect of intensity depends on the grouping decision ($p(\beta > 0) = 0.98625$, and similarly, when making the grouping decision, the effect of intensity depends on the prominence decision ($p(\beta > 0) = 0.992$.

There was also strong evidence that for the grouping decision, the underlying order of the syllables in the sequence (*ba* first vs. *ga* first) mattered ($p(\beta > 0) = 0$.

Note that since in this model, each decision is dependent on the other, it is not possible to use the model to compute predictions for either decision, unless one specifies the choice for the respective other decision. In that sense, it is not a generative model.

It is possible that one could model listener-behavior by assuming that they make a 'best guess' decision by computing the log odds for each of the four possibilities, and choose the pair of decisions that maximizes the log odds. It would be interesting to explore a Baysian decision strategy of this kind in more detail, but this goes beyond the scope of this paper.

Experiment 3: Tone Perception

A mixed effects logistic regression models was used to analyze how the various factors influenced the prominence and grouping decisions in tone perception. Here's the model for the prominence decision for tone sequences:

```
modelProminenceTone = glmer(
  Prominence ~
  UnevenLouder * UnevenLonger * Uneven.vs.EvenFirst +
  (UnevenLouder + UnevenLonger ||
  participant),
  family = 'binomial',
  control = glmerControl(optimizer = "bobyqa",
                         optCtrl =
  list(maxfun = 120000)),
  data = perDataTone
)
```

And here's the model of the grouping decision for the tone data:

```
modelGroupingTone =  glmer(Grouping ~
  UnevenLouder * UnevenLonger * Uneven.vs.Even +
  (UnevenLouder + UnevenLonger || participant),
  family = 'binomial', data = perDataTone,
  control = glmerControl(optimizer = "bobyqa",
  optCtrl = list(maxfun = 120000))
)
```

And for the foot choice model:

```
modelFootTone =  glmer(
  FootChoice ~
    (UnevenLouder + UnevenLonger)  +
    (UnevenLouder + UnevenLonger |participant),
  family = 'binomial',
  control = glmerControl(optimizer = "bobyqa", optCtrl =
```

```
                          list(maxfun = 120000)),
  data = perDataTone
)
```

The results of the models for the tone perception data are summarized in Table 7.

	Prominence	Grouping	Foot
(Intercept)	0.11 (0.10)	0.34 (0.16)*	0.01 (0.11)
UnevenLouder	0.16 (0.05)**	0.20 (0.04)***	0.04 (0.03)
UnevenLonger	0.43 (0.07)***	−0.37 (0.12)**	0.11 (0.05)*
Uneven.vs.EvenFirst	0.29 (0.13)*		
UnevenLouder:UnevenLonger	−0.03 (0.02)	−0.01 (0.02)	
UnevenLouder:Uneven.vs.EvenFirst	0.06 (0.06)		
UnevenLonger:Uneven.vs.EvenFirst	0.23 (0.07)**		
UnevenLouder:UnevenLonger:Uneven.vs.EvenFirst	−0.10 (0.03)**		
Uneven.vs.Even		0.24 (0.13)	
UnevenLouder:Uneven.vs.Even		0.07 (0.07)	
UnevenLonger:Uneven.vs.Even		0.24 (0.07)***	
UnevenLouder:UnevenLonger:Uneven.vs.Even		−0.10 (0.04)**	

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 7
Models for grouping and prominence for the perception experiment on tone sequences

In addition to the models for the tone data, a model was fit for the speech and tone data combined. The figures in the main paper suggest that the tone choices may have been slightly different at least in that there was more uncertainty in the choices (the colours in the heatmaps for tones appear a bit paler). This could be due to the fact that the intensity manipulation was less extreme in the tone cases, but it could also be because of a more general difference, for example that listeners have much more experience processing speech than music.

The models indeed show some significant differences in how intensity and duration affected the prominence, grouping, and foot choices, respectively. Since there were differences for the effect of duration as well, this is not entirely due to the smaller steps that were used for the intensity manipulation in tones. These overall models confirm that intensity and duration are poor predictors of foot choice, although there were differences in how they relate to foot choice depending on whether the stimulus was speech or tones (since the interactions came out significant).

Bivariate Baysian analysis of Tone Perception Data

A joint model for both decisions was again fit using BRM. The formulas for the two decisions were as follows:

	Prominence	Grouping	Foot
(Intercept)	0.05 (0.07)	0.07 (0.10)	−0.02 (0.08)
baLouder	0.24 (0.04)***	0.23 (0.03)***	−0.01 (0.02)
baLonger	0.42 (0.05)***	−0.42 (0.09)***	0.04 (0.02)
Speech.vs.Tone	0.10 (0.13)	0.73 (0.16)***	0.03 (0.14)
baLouder:Speech.vs.Tone	−0.16 (0.07)*	−0.15 (0.06)*	0.11 (0.05)*
baLonger:Speech.vs.Tone	−0.02 (0.09)	−0.23 (0.11)*	0.18 (0.05)***

***$p < 0.001$; **$p < 0.01$; *$p < 0.05$

Table 8
Models for grouping and prominence for the perception experiment for speech and tone combined.

```
prominenceT_fm <- bf(prominence ~
                       (UnevenLouder + UnevenLonger) * Uneven.vs.EvenFirst +
                       (UnevenLouder + UnevenLonger|p|participant)) +
                      bernoulli()


groupingT_fm <- bf(grouping ~
                       (UnevenLouder + UnevenLonger) * Uneven.vs.Even +
                       (UnevenLouder + UnevenLonger|p|participant)) +
                      bernoulli()
```

Mildly skeptical priors were set against the hypotheses, and in favor of all coefficients being equal to zero:

```
priorInteract <- c(
set_prior("normal(0, 10)", class="b", resp="prominence"),
set_prior("normal(0, 10)", class="b", ,resp="grouping")
)
```

Here is code for the bivariate model:

```
jointModelT <- brm( prominenceT_fm +
                    groupingT_fm +
                    set_rescor(FALSE),
                    data = perDataTone,
                    chains = 4,
                    prior = priorInteract,
                    cores = 4)
```

The model appears to capture the variability in the data well, given the posterior-predictive checks illustrated in Fig. 6.

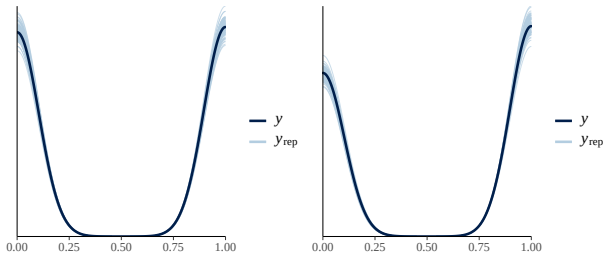

Figure 6. Posterior predictive check for the fit for prominence (left) and grouping (right) for the tone data.

Based on posterior samples from the model, the strength of the evidence for the relevant hypotheses was evaluated. There is good evidence that greater intensity and greater duration make it more likely that a tone is heard as prominent (for intensity: $p(\beta > 0) = 0.99675$; for duration: $p(\beta > 0) = 1$.

There is strong evidence that greater intensity and lower duration makes it more likely that a tone is heard as initial (for intensity: $p(\beta > 0) = 0.99975$; for duration: $p(\beta < 0) = 0.99625$.

Finally, the model suggests that there is strong evidence that when making the prominence decision, the effect of duration depends on the grouping decision ($p(\beta > 0) = 0.99975$, and similarly, when making the grouping decision, the effect of duration depends on the prominence decision ($p(\beta > 0) = 0.99975$.

Further participant information

Each participant filled out a language questionnaire and a music questionnaire. For one participant in Experiment 1, the language questionnaire information was somehow not recorded.

This questionnaire information was collected because it is clear language and music background might be relevant to the task. However, there was no particular hypothesis about how these might affect the outcome. Prior literature on the ITL often reports musical experience and ability (e.g. Bhatara, Boll-Avetisyan, Unger, Nazzi, & Höhle, 2013; Geiser & Gabrieli, 2013). (Boll-Avetisyan, Bhatara, Unger, Nazzi, & Höhle, 2016) found musical experience only to be relevant for bilingual participants, and boll17 found that musical aptitude, but not musical experience, to be relevant in monolingual speakers.

I decided not to use these questionnaires to explore individual differences here for three reasons: (i) Given the small sample sizes in these studies ($n < 35$ participants in each experiment), I felt that results about individual differences might not be representative; (ii) I didn't have any specific hypothesis about how these factors might distort the results exactly, so any reported effects would be post hoc; (iii) the number of languages spoken and especially the amount of musical ability might correlate with other factors that the questionnaires didn't measure, e.g. socio-economic status, education level, and attention and

work-ethic-related individual differences. For example,someone with ADHD is presumably less likely to tough out 8 years of piano lessons, and attention variability might very well influence this type of perception task, given that many prior authors have related the ITL to question of attention allocation.

The main goal with collecting the questionnaire data was to create a data set that can be used to look for patterns, that then can be tested in future experiments. I did check whether participants with musical abilities are better at the tone task, and show less random results, but this does not appear to have been the case. I report here that, oddly, participants with high musical ability were more likely to hear the even tones as loud and initial than participants with lower musical ability, possibly suggesting a smaller order effect. If this effect will replicate in the future, it would be worthwhile to try to find an explanation.

So for the present purposes, this study simply releases the questionnaire information as part of the public data set, so that one can explore potential effects and thus conjure up hypotheses for future experiments.

References

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

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

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

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.

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

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.

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

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