Attention, Perception, & Psychophysics (2021) 83:1861–1877
https://doi.org/10.3758/s13414-021-02251-y

What you hear first, is what you get: Initial metrical cue presentation modulates syllable detection in sentence processing

Anna Fiveash[1,2] · Simone Falk[3,4,5] · Barbara Tillmann[1,2]

Accepted: 14 January 2021 / Published online: 11 March 2021

Abstract

Auditory rhythms create powerful expectations for the listener. Rhythmic cues with the same temporal structure as subsequent sentences enhance processing compared with irregular or mismatched cues. In the present study, we focus on syllable detection following *matched* rhythmic cues. Cues were aligned with subsequent sentences at the syllable (low-level cue) or the accented syllable (high-level cue) level. A different group of participants performed the task without cues to provide a baseline. We hypothesized that unaccented syllable detection would be faster after low-level cues, and accented syllable detection would be faster after high-level cues. There was no difference in syllable detection depending on whether the sentence was preceded by a high-level or low-level cue. However, the results revealed a priming effect of the cue that participants heard *first*. Participants who heard a high-level cue first were faster to detect accented than unaccented syllables, and faster to detect accented syllables than participants who heard a low-level cue first. The low-level-first participants showed no difference between detection of accented and unaccented syllables. The baseline experiment confirmed that hearing a low-level cue first removed the benefit of the high-level grouping structure for accented syllables. These results suggest that the initially perceived rhythmic structure influenced subsequent cue perception and its influence on syllable detection. Results are discussed in terms of dynamic attending, temporal context effects, and implications for context effects in neural entrainment.

Keywords Rhythm · Speech · Language · Syllables · Attending · Entrainment

Introduction

Temporal regularities present within music and speech allow listeners to create expectations and to predict upcoming events. Musical rhythms provide clear expectations about when an upcoming event should occur (McAuley, 2010), allowing for easy synchronization and the ability to dance or move along with music. Speech rhythm is temporally less regular, but appears to contain temporal regularities in patterns of stressed and unstressed syllables (Arvaniti, 2009, 2012; Pitt & Samuel, 1990; Varnet, Ortiz-Barajas, Erra, Gervain, & Lorenzi, 2017). Stressed[1] syllables in speech are more important to the comprehension of a sentence than unstressed syllables, as they contain more relevant information to the understanding of the sentence (Aylett & Turk, 2004; Calhoun, 2010). Therefore, when listening to an incoming speech stream, being able to predict when the next stressed syllable will occur can facilitate sentence processing and understanding (e.g., Brown, Salverda, Dilley, & Tanenhaus, 2015; Dilley & McAuley, 2008). It has been suggested that attention is not stable over time, but rather fluctuates rhythmically to facilitate temporal prediction, expectation, and attending to information-relevant elements of the signal (Jones, 1976, 2016; Large, 2008).

✉ Anna Fiveash
anna.fiveash@inserm.fr

1 Lyon Neuroscience Research Center, CRNL, CNRS, UMR5292, INSERM, U1028, F-69000 Lyon, France

2 University of Lyon 1, F-69000 Lyon, France

3 Department of Linguistics and Translation, University of Montreal, Montreal, Canada

4 BRAMS, International Laboratory for Brain, Music and Sound Research, University of Montreal, Montreal, Canada

5 Laboratoire Phonétique et Phonologie, CNRS UMR 7018, Université Sorbonne Nouvelle Paris-3, Paris, France

[1] Note that we use the words *stress(ed)* and *accent(ed)* interchangeably in the current manuscript.

An influential theoretical framework that proposes how attention is allocated to predictable points in time is the dynamic attending theory (DAT; Jones, 1976, 2016; Large & Jones, 1999). The DAT suggests that endogenous neural oscillations entrain in phase to an external rhythmic stimulus. It also suggests that endogenous neural oscillations persist after the external stimulus has stopped, providing a neural basis as to how the rhythm of a prime or preceding cue can influence subsequent perception both within modalities (Barnes & Jones, 2000; Jones, Johnston, & Puente, 2006; Jones, Moynihan, MacKenzie, & Puente, 2002; Kösem et al., 2018; McAuley & Kidd, 1998) and between modalities (Bolger, Trost, & Schön, 2013; Brochard, Tassin, & Zagar, 2013; Fotidzis, Moon, Steele, & Magne, 2018; ten Oever, Schroeder, Poeppel, van Atteveldt, & Zion-Golumbic, 2014). In hierarchical stimuli such as music and speech, the DAT suggests that neural oscillations entrain at multiple, nested levels of the metric hierarchy, providing a neural basis for the tracking of hierarchical structure and the benefit of metric binding (Jones, 2016, 2019). Attending is suggested to be future-oriented or analytic (Drake, Jones, & Baruch, 2000; Jones & Boltz, 1989). Future-oriented attending refers to attention directed to levels higher than the referent level (i.e., the dominant level or tactus) of the stimulus. Analytic attending refers to attention directed at or below the referent level of a stimulus. Importantly, future-oriented attending allows for temporal prediction, as well as the generation of expectancies based on higher-order structures, whereas analytic attending results in a focus on local processing.

Rhythm and the metric hierarchy

Both music and speech contain patterns of strong and weak elements that are structured according to a metric hierarchy, allowing for temporal prediction and processing on multiple levels (Beier & Ferreira, 2018; Cummins & Port, 1998; Lerdahl & Jackendoff, 1983). In the case of music, the rhythm (the temporal patterning of individual elements) allows for the abstraction of beat and meter (Grahn, 2012; Kotz, Ravignani, & Fitch, 2018) that can either coincide with an acoustic event or be perceived without an event (Large, Herrera, & Velasco, 2015; Tal et al., 2017). Beats are perceptually organized within the metric hierarchy. For example, a waltz rhythm contains groups of three beats, and the first beat holds a stronger weight within the rhythm than the following two beats (Fujioka, Ross, & Trainor, 2015; McAuley, 2010). In music theory, beats that align with multiple levels of the metric hierarchy (i.e., the first beat within a waltz) are perceived as more salient within the rhythm compared with beats that do not align with multiple levels of the hierarchy (Fitch, 2013; London, 2012). For a simple rhythm with only one level (e.g., quarter notes), these notes can also be perceptually grouped into higher levels, even if there is no physical acoustic difference between the notes (e.g., Nozaradan, Peretz, Missal, & Mouraux, 2011; Phillips-Silver & Trainor, 2007). Thus, beat perception can be influenced in a top-down manner.

Although temporally less regular, speech also contains a metric hierarchy that is created by the combination of smaller speech elements (i.e., phonemes, syllables) into larger speech elements (i.e., words, sentences), which are grouped in different ways based on patterns of prominence and stress (Arvaniti, 2009; Giraud & Poeppel, 2012; Greenberg, Carvey, Hitchcock, & Chang, 2003). The description of speech rhythm in terms of prominence, grouping, and stress rather than the historical (empirically unsupported) categorization into stress-timed and syllable-timed languages allows for scientific comparison across different languages, and allows for a comparison with rhythm in music (Arvaniti, 2009; Ding et al., 2017). Speech rhythm has been shown to be predictable (Beier & Ferreira, 2018), with faster processing of stressed compared with unstressed syllables (Cutler & Foss, 1977; Gow & Gordon, 1993), and facilitated processing of predicted stressed syllables compared with syllables that are not predicted to be stressed (Cutler, 1976; Pitt & Samuel, 1990).

The processing of metric hierarchies of music and speech rhythm can also be observed in the brain. Multiple, nested neural oscillations have been observed that track different divisions of the beat in music (Fujioka et al., 2015; Large et al., 2015; Nozaradan, 2014; Stupacher, Wood, & Witte, 2017; Tierney & Kraus, 2013c), and different linguistic units (i.e., phonological, syllable, stressed syllable, and phrasal levels) in speech (Ding, Melloni, Zhang, Tian, & Poeppel, 2016; Giraud & Poeppel, 2012). For both types of material, there is evidence that these neural oscillations do not just passively track the acoustic elements in the signal, but are involved in higher-level, top-down processes. For music, neural oscillations have been shown in response to a beat level (Fiveash et al., 2020b; Tal et al., 2017), and to an imagined metric level (Nozaradan et al., 2011), even when these are not present (or are weakly present; Fiveash et al., 2020b) in the stimulus. For speech, neural oscillations have been shown to track cognitively relevant information (such as phrasal groupings) that is not represented acoustically, but is reliant on comprehension (Ding et al., 2016). These results show that multiple levels of the metric hierarchy are reflected in the brain response to music and speech stimuli and are influenced by top-down processes.

One prediction of the DAT is that neural oscillations should persist once the external stimulus has stopped. Evidence for the persistence of neural oscillations after an entrained prime or cue suggests that the brain continues to predict upcoming events, and that a previously entrained stimulus can affect subsequent perception (Canette et al., 2020; Fiveash, Bedoin, Lalitte, & Tillmann, 2020a; Gross et al., 2013; Hickok, Farahbod, & Saberi, 2015; Trapp, Havlicek, Schirmer, & Keller, 2018). Of relevance to the current study is that rhythmic cueing studies have shown an influence of

rhythmic stimuli on subsequent speech perception (e.g., Cason, Astésano, & Schön, 2015; Cason & Schön, 2012; Falk, Lanzilotti, & Schön, 2017a; Gould, McKibben, Ekstrand, Lorentz, & Borowsky, 2015).

Effects of cue regularity on subsequent speech processing

Research has shown that a rhythmic cue that matches the rhythm of a subsequent sentence facilitates processing within that sentence compared with an irregular or mismatching cue. At the phoneme level, participants were faster to detect phonemes presented in a nonsense word (Cason & Schön, 2012) and at the end of a sentence (Cason et al., 2015) when the preceding cue was aligned (on the beat) with the phoneme presentation, or matched the stress pattern of the sentence, respectively. Cason and Schön (2012) also showed that the P300 and N100 event-related potential components measured with electroencephalography (EEG) were enhanced for phonemes presented off-the-beat compared with on-the-beat, suggesting a larger expectancy violation for off-beat phonemes. Preceding stress cues (e.g., alternating strong and weak tones) have also been shown to affect reading speed of subsequent words depending on whether the stressed tone was aligned with the stress cue in the word or not (Gould et al., 2015; Gould, Mickleborough, Ekstrand, Lorentz, & Borowsky, 2017). Further, the alignment of finger taps with regularly recurring accented syllables (Falk & Dalla Bella, 2016) and the continuation of isochronous tapping that is congruent with the incoming speech signal (Falk, Volpi-Moncorger, & Dalla Bella, 2017b) enhance detection of word changes compared with misaligned or incongruent tapping. This set of results suggests that temporal regularity in a preceding cue can enhance the processing of subsequent speech compared with irregular or misaligned cues.

The underlying neural mechanism behind the rhythmic cueing effect is suggested to be the entrainment of endogenous neural oscillations to the rhythmic cue, and the persistence of these neural oscillations after the stimulus has stopped. To investigate this hypothesis, Falk, Lanzilotti, and Schön (2017a) presented participants with either a regular cue that matched the syllable and accent structure of a subsequent sentence, or an irregular cue that did not match the structure. Phase-locking of neural oscillations was enhanced at frequencies present in the sentence stimuli after a regular cue compared with an irregular cue, suggesting the influence of sustained neural oscillations. The continuation of neural oscillations has also been shown by Gordon, Magne, and Large (2011). Participants were presented with a metrical rhythm followed by a sung sentence that was congruently or incongruently accented in relation to the rhythm. Following the sung sentence, participants performed a lexical decision task on a visually presented word. Beta band power for sung syllables was enhanced when the metrical rhythm aligned with the accented syllables, and target words were detected faster (and with increased alpha and beta power) after a congruent rhythm and sentence compared with an incongruent pairing. These results support behavioral findings of rhythmic cueing and suggest that neural oscillations contribute to the observed temporal cueing effects.

The results presented above are all based on the comparison of a regular or matching cue to an irregular or mismatching cue, and with the matching cue investigating only one level of the metric hierarchy. Such designs do not allow for an investigation of more subtle hierarchical metrical structure and its effects across a trial or experimental session. The investigation of metric structure is relevant, as meter perception is also influenced by rhythmic context and the initially perceived metric level. For example, previous research has shown that the initial perception of a certain stimulus type can influence subsequent perceptual groupings throughout an experimental session. These *temporal context effects* suggest that stimuli presented first can have an *attractive* effect on subsequent perception, whereby a following stimulus is perceived as being similar to the initial stimulus (Snyder, Schwiedrzik, Vitela, & Melloni, 2015). Attractive effects are suggested to recruit higher-level cognitive processes, whereby previous experience is integrated with the current stimulus, resulting in an altered perception of that stimulus. A recent EEG study has also shown a persistence of metrical structure depending on the order rhythms were presented. Rhythms presented in the order of *most regular* to *least regular* (ambiguous) resulted in a longer persistence of meter-related neural responses compared with rhythms presented in the order of the *least regular* to *most regular*, suggesting a persistence of the initially perceived meter (Lenc, Keller, Varlet, & Nozaradan, 2019).

The current experiment presented only matching cues. This manipulation allows us to investigate whether the difference between matched/mismatched cues in previous work is only created by a *cost* of the mismatched condition. As we are focusing on the matched condition with two possible types of matching, we make the hypothesis that there is actually a *benefit* due to the matching (in previous studies) and we here further investigate the nature of this effect. Relatedly, presenting only matching cues avoids the potential concern that the correspondence between a matching/regular cue is more acoustically pleasing than a mismatched/irregular cue, resulting in an effect based on arousal rather than sustained entrainment. The current experiment therefore aimed for a more subtle test of the predictions of the DAT. In addition, we directly investigated whether initial metrical biases induced by the first cue heard persist across the experimental session, or whether perceived metrical level influences perception only in subsequently presented sentences.

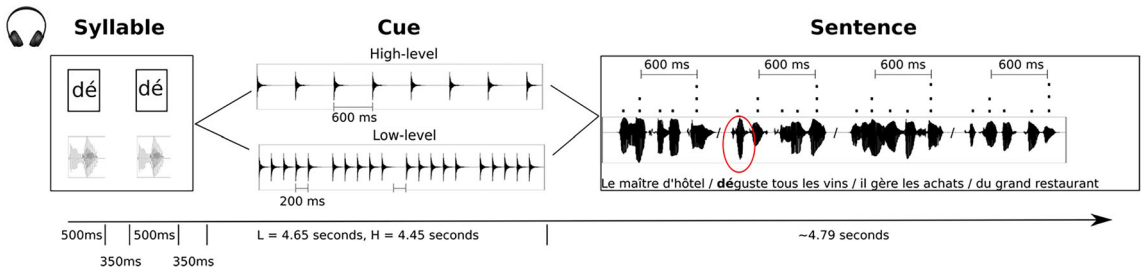

Fig. 1 Example of stimulus timing and presentation within a trial. All stimuli were presented auditorily. The amplitude envelopes of an example syllable, the two cues, and an example sentence are displayed. The target syllable (here unaccented) was presented twice in a row, then participants heard a low-level (L) or a high-level (H) cue, followed directly by a sentence. There were 600 milliseconds (ms) between each accented syllable, aligning to each tone in the high-level cue (i.e., H tones were presented every 600 ms), and every second and fifth tone of the smaller groups in the low-level cue (i.e., L tones were presented every 200 ms). The slashes (/) in the sentences indicate a phrase boundary. The small dots above each syllable in the sentence indicate the stress received by each syllable: one dot refers to unaccented syllables, two dots refer to secondary stressed syllables, and three dots refer to primary stressed syllables. The target syllable is indicated in boldface in the written sentence and circled in red in the corresponding waveform. The text of the sentence is for illustrative purposes and is not to scale, as the spoken and written shapes do not directly align (see Supplementary Material for sound file examples). (Color figure online)

The current study

In the current syllable detection study, participants were presented with a target syllable, followed by a cue matching at a metrically low level (L) or a metrically higher level (H), followed by a sentence in which the target syllable had to be detected (see Fig. 1). Rhythmic cues consisted of isochronous percussive tones that were designed to align with the timing of each syllable (L) or with the timing of accented syllables only (H). Our aim was to investigate whether cueing different levels of the metric hierarchy (low or high) with regular, matched cues could differently influence subsequent syllable detection within a sentence. A syllable detection task was chosen (i.e., instead of a phoneme detection task) so that the same unit could be manipulated to occur at a low-level or high-level of the metric hierarchy (i.e., in an unaccented or accented position). To enhance the effect of the cue before each sentence, and in line with previous cueing experiments (e.g., Cason et al., 2015; Falk & Dalla Bella, 2016; Falk, Lanzilotti, & Schön, 2017a), we presented rhythmic cues in blocks of trials (here, 10 trials per block).

Rhythmic cues were designed to correspond to and direct attention to events at either the low-level *syllable* rate (i.e., containing one beat corresponding to every syllable) or the high-level *accented syllable* rate (i.e., containing one beat corresponding to every accented syllable) of subsequently presented sentences. We predicted that *unaccented* syllables would be detected faster after a L cue compared with an H cue. We also predicted that *accented* syllables would be detected faster after an H cue compared with a L cue, as the H cue should impose a higher-level grouping structure, resulting in enhanced prediction and emphasis on the accented syllables. However, there is the alternative possibility that metrical interpretations established early based on experimental context tend to persist over time when no conflicting sensory information intervenes (e.g., Lenc et al., 2019). Hence, as all the cues and sentences were compatible with the same temporal and metrical hierarchical framework, one could predict that the first block of cues presented sets the stage for following metrical interpretations. For example, if participants heard an H cue first, they might be likely to perceive subsequent sentences as well as L cues with a higher-level grouping structure. If participants heard a L cue first, they might be likely to perceive subsequent sentences and L and H cues with a focus on the lower-level grouping structure. As switching attending between hierarchical levels is suggested to require more cognitive resources than remaining at the same attending level (Drake et al., 2000; Jones & Boltz, 1989), subsequent cues may be processed at the initially perceived level, reducing potential local cueing effects.

The current study consists of two experiments: the main experiment with H and L rhythmic cues (Experiment 1) and a baseline experiment without rhythmic cues (Experiment 2).

Experiment 1

Method

Participants

Forty native French-speaking participants were recruited through the University of Lyon and social media (M_{age} = 22.65, *SD* = 2.08 years; range: 18–27 years; 32 women). Participants had a range of musical training (*M* = 3.2, *SD* = 5.36 years of lessons[2]; range: 0–17 years). Eighteen of these participants reported that they currently played (*n* = 12) or have played an instrument in the past (*M* = 7.12, *SD* = 6.43 years of lessons). One participant reported being dyslexic[3] and

[2] Years of music playing (i.e., including but not limited to lessons) showed a similar outcome (*M* = 3.6 years, *SD* = 6.21; range: 0–19 years).

no participants reported hearing, cognitive, or neurological conditions or impairment.

Although it is still difficult to calculate effect sizes for linear mixed models (as the effect size is related to the number of observations, rather than the number of participants), we used G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) to calculate the number of participants necessary for a repeated-measures analysis of variance (ANOVA)-based analysis, with alpha = 0.05, power = 0.80, and a medium effect size (f = 0.25), as suggested by Cunningham and McCrum-Gardner (2007). This calculation suggested 24 participants were necessary to detect an effect. To align with previous related research (e.g., n = 32 in each group in Falk, Volpi-Moncorger, & Dalla Bella, 2017b), and to enhance power to 0.95, we tested 40 participants (n = 36 was suggested for power of 0.95). All participants were tested before data was analyzed, to avoid optional stopping (Rouder, 2014; Simmons, Nelson, & Simonsohn, 2011).

Design

The current experiment was a 2 (rhythmic cue: low-level, high-level) × 2 (accent: accented, unaccented syllable) within-subjects design. All the verbal material used in the experiment was in French. Each trial consisted of an auditory target syllable (i.e., the syllable to be detected), a cue (L or H), and a sentence. The task was to listen to these stimuli and detect the target syllable within each sentence as fast as possible. The target syllable was always contained within the sentence (i.e., there were no trials where the target syllable was not present). An equal proportion of L and H cues were paired with the sentences containing accented and unaccented target syllables. Trials with the same cue type were presented in blocks of 10 to heighten the perception of the cued level (e.g., 10L 10H 10L 10H 10L 10H), similarly to blocked designs from previous priming/cueing experiments (e.g., Cason et al., 2015; Cason & Schön, 2012; Falk & Dalla Bella, 2016; Falk, Lanzilotti, & Schön, 2017a). Cue order (whether the experiment started with a block of low-level trials or high-level trials first) was counterbalanced across participants. The pairing of each sentence with an L or H cue was also counterbalanced across participants and across cue order. Trial presentation was pseudorandomized while maintaining blocks of the same cue type.

Stimuli

Sentences Sixty sentences from Falk, Volpi-Moncorger, and Dalla Bella (2017b) were used in this study. These sentences were specifically designed to contain four accentual phrases (corresponding to the groups in the cues) consisting each of five syllables. Within each five-syllable accentual phase, the second and last syllables received more emphasis than the others (i.e., were accented). The second syllable in each group displayed a *secondary* accent (i.e., an initial rise in French), the last syllable in each group a *primary* accent (i.e., most of them final rises in French), and there was a 600 millisecond (ms) interonset interval (IOI) between accented syllables. Sentences were produced within this speech rhythm by a native French speaker, who was cued by a metronome at 600-ms IOI before production of each sentence started. To ensure regularity of interaccent intervals, the recorded sentences were slightly adjusted by manually lengthening or shortening silences or mid-vowel portions using Praat (Boersma, 2001) if necessary. More information on stimulus construction can be found in Falk, Volpi-Moncorger, et al. (2017), and example sentences in the Supplementary Material. Silence was added to the beginning of each sentence (as in Falk, Lanzilotti, et al., 2017) to ensure that the first accented syllable occurred 600 ms after the final note in the H cue, and 600 ms after the penultimate note in the L cue. Sentences were 4.79 seconds on average (SD = 0.08, range: 4.69–4.98 seconds). Sound files were exported at 44100 Hz, 16 bits per sample.

Syllables Within these sentences, we selected 60 content words (containing two to four syllables) for the syllable detection task. Syllables were selected based on (1) syllable structure (92% of syllables had a CV or CVC structure, the remaining syllables were CCV); (2) syllable position in the word (for both accented and unaccented syllables, half of the target syllables were on the first syllable of the word, and half of the target syllables were on the second/final syllable of the word); (3) syllable position in the sentence (target syllables were distributed over the four accentual phrases); (4) syllable accent (for accented syllables, half carried a secondary accent, and half carried a primary accent); and (5) uniqueness (each syllable was uniquely identifiable and only occurred once in the sentence). (See Table 1 for the distribution of syllables.) An independent-samples t-test (equal variances not assumed) confirmed that there was no systematic difference between the temporal onset (measured from the start of the sentence) of accented (M = 1.72 s, SD = 1.65 s) and unaccented (M = 2.46 s, SD = 1.33 s) syllables, $t(31.74) = 1.72$, $p = .10$, across the sentences. Twenty accented and 40 unaccented syllables were identified as the to-be-detected target syllables. There were more unaccented than accented target syllables because (1) there were more available options for unaccented syllables within each sentence, and (2) this choice allowed the location of the target syllables to be less predictable.

To create the auditory syllable prompt before the start of each trial, each syllable was generated by using automatic

[0] Note that the same pattern of results was observed if the dyslexic participant was excluded, so they were kept in the analyses.

Table 1 Parameters for the syllables included in the syllable detection task

Type	Structure		Position in word		Position in sentence				Accent level	
	CV/CVC	CCV	1st syll.	2nd syll.	1	2	3	4	Secondary	Primary
Unaccented	36	4	20	20	9	10	11	10	NA	NA
Accented	19	1	10	10	10	4	3	3	10	10
Total	55	5	30	30	19	14	14	13	10	10

Note. There were 40 unaccented and 20 accented syllables in total presented to each participant. CV = consonant-vowel structure; syll. = syllable

text-to-speech synthesis. The syllable was phonetically written into Natural Reader (www.naturalreaders.com) and pronounced by the French text-to-speech voice *Alice*. To ensure the syllables were pronounced as expected, three native French speakers listened to the syllables and transcribed them. Any discrepancies between the transcribed syllable and the expected syllable were discussed and the prompt was altered phonetically until the syllable sounded correct to the native French speakers. These syllable prompts were then exported into 500-ms wav files, and the maximum amplitude of each syllable was normalized in loudness (and DC offset removed) with Audacity.

Cues The low-level (one beat every syllable) and high-level (one beat every accented syllable) cues were designed based on the regular cues used in Falk, Lanzilloti, et al. (2017), which matched the rhythmic structure of the sentences. Both cues were created with the software GarageBand (Apple, Version 10.2.0) using the percussion instrument *coffee-shop*, which consisted of a sharp onset and a quick decay. The spectral envelope of the cues can be seen in Fig. 1 and heard in the Supplementary Material. Because the notes were played with a percussion timbre, each note had the quality of a high-pitched drum sound (~390 Hz) and lasted for 200 ms. We will refer to these sounds as tones. Tones used for L and H cues had the same intensity and timbre across the entire sequence (see Supplementary Material and Fig. 1 for examples of the stimuli).

Low-level cues contained four groups of five tones (see Fig. 1), with a 200-ms IOI between tones and a 200-ms silence after every group of tones, corresponding to the syllables in the sentences. The total duration was 4.65 seconds.

High-level cues contained four groups of two tones (see Fig. 1), with a 600-ms interval between tones, corresponding to the accented syllables in the sentences. The cues were designed so that there was a 600-ms pause between the onset of the last tone of the cue and the perceptual onset (estimated by the algorithm of Cummins & Port, 1998) of the first accented syllable in the sentences. The total duration was 4.45 seconds (200-ms shorter than the L cue because the cue started directly on the first accented note).

Procedure

After signing the information and consent form, participants were told that they would hear a syllable, a rhythmic sequence, and then a sentence, and to press a button on the keyboard as soon as they detected the syllable in the sentence. They were not informed that there were differences between accented and unaccented syllables. There were two practice trials—one with a L cue and one with an H cue. Practice trial presentation order corresponded to the cue order condition: Participants in the L-first condition heard the L cue first in the practice, followed by the H cue. The experimenter ensured that the participant understood the instructions, and then the trials began. For each trial, a fixation cross appeared on the screen for 1 second, and then the target syllable sounded twice, with a 350-ms silence between the syllables. Following the second presentation of the syllable, there was a 350-ms silence before the cue started to play (L or H). The sentence played directly after the cue had finished (see Fig. 1). If participants detected the syllable, a screen appeared asking participants to press space bar for the next trial. If participants did not indicate that they had heard the syllable, they were asked to press a key to confirm that they had not heard the syllable. They then continued to the following trial. This procedure continued for 60 trials, with one break in the middle. The experiment was conducted on a MacBook Pro laptop, running MATLAB 2018a, using PsychToolbox (Version 3.0.14) and lasted approximately 20 minutes. Participants wore headphones for the duration of the task. At the end of the experiment, participants filled out a musical background questionnaire that collected background information about musical experience to investigate this potential influence on cue effects.

Data analysis

Syllables were marked as undetected if the participant confirmed that they did not hear the syllable at the end of a trial.

Detection times Syllable onset times were marked in Praat (Boersma, 2001) and were used to determine response time (RT) in milliseconds from the beginning of each target

syllable. Any negative RT values (where the participant indicated that they had heard the syllable before syllable onset) were removed.[4] These removed values ranged from 0% to 15% per participant (M = 5.17%, SD = 3.17%). The average RT across all conditions for each participant was also calculated, and one participant who was more than three standard deviations above the mean group RT (M = 561.47 ms, SD = 131.20, outlier > 955.06 ms) was excluded from all analyses (for both RT and undetected syllables). After negative RTs, undetected syllable responses, and trials with no responses were removed, there was a total data loss of between 5% and 25% of trials for each participant (M = 13.88%, SD = 5.23%). Note that an ANOVA on the proportion of trials removed in each condition showed that there were no differences depending on cue, $F(1, 38) = 0.82$, $p = .37$, or cue order, $F(1, 38) = 0.02$, $p = .88$, and no interactions between cue, accent, or cue order (all p values > .55). There was a significant effect of accent condition, $F(1, 38) = 7.22$, $p = .001$, as more trials were removed in the unaccented (M = 15%, SD = 8.5%) compared with the accented (M = 11%, SD = 10.6%) condition.

Generalized linear mixed models The proportion of trials in which participants did not detect the target syllable and the RTs for detected syllables were analyzed in R (R Core Team, 2018) using the *lme4* package for linear mixed models (Bates, Mächler, Bolker, & Walker, 2015). Linear mixed models were used as they allowed for analysis at the trial level, while controlling for random effects of participant and sentence (Baayen, Davidson, & Bates, 2008). Linear mixed-model approaches are considered to be more sensitive and to have more power than traditional ANOVA-based approaches, and simulations comparing mixed effects models to traditional F tests have shown that mixed models contain higher power while minimizing the Type I error rate (Baayen et al., 2008). This analysis is also particularly powerful in situations with missing data or unequal group sizes (i.e., with our different number of trials for accented and unaccented conditions), as the trial-by-trial approach does not result in participant averages made up of different numbers of trials and can model the variance within each distribution (Baayen et al., 2008).

Because both the undetected syllables and the RT data were not normally distributed, generalized linear mixed models (GLMM) were employed (Lo & Andrews, 2015). For undetected syllables, a GLMM with a binomial distribution was used because the response was categorical. For RT data, Lo and Andrews (2015) suggest to use a gamma or inverse Gaussian distribution with an identity link to model the raw data within a skewed distribution without having to transform the data to satisfy mathematical assumptions. To determine which distribution was most appropriate for the current data, we used the Cullen and Frey graph in the *fitdistrplus* package (Delignette-Muller & Dutang, 2015) with a nonparametric bootstrap procedure to model skew and kurtosis values under a number of distributions. This graph, as well as model comparisons with the two distributions, showed that the gamma distribution was the most appropriate for the current data set, so this distribution was used for all RT models. The GLMMs were fitted with the maximum likelihood method based on a Laplace approximation, as implemented in *lme4*. The *car* package (Fox & Weisberg, 2011) was used for significance testing of individual effects within the models (based on Type III Wald chi-squared tests), and the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2017) was used for comparing between models. Akaike information criterion (AIC) was compared between models to assess model fit.

Models were built from the most basic effects up to a more elaborated model. For both undetected syllables and RT models, the base model included the fixed effects of cue (low-level, high-level), accent condition (accented, unaccented), and their interaction. Participant (n = 39, 19 in the low-level-first condition, 20 in the high-level-first condition) and sentence (n = 60) were included as random effects in all models, as these variables were expected to vary randomly (Baayen et al., 2008). Because there was a very low proportion of undetected syllables, we did not elaborate further on the undetected syllable model to avoid overfitting the data, except to investigate the fixed effect of years of private music lessons.

For the RT model, we added the fixed effect of cue order (including interactions) to the base model. The maximal model with all interactions did not converge, so interaction terms that were not contributing significantly to the model were removed in a step-wise manner, starting with the three-way interaction (to avoid overfitting; Brysbaert & Stevens, 2018). Fixed effects of years of music lessons and syllable position within the sentence (first half, second half) were then separately added to the final model. The effect of syllable position within a sentence was included in the model to investigate whether the typical linguistic context effect of faster syllable detection toward the end of the sentence would occur (as in Montgomery, 2000; Montgomery, Scudder, & Moore, 1990; Planchou et al., 2015; Simpson, Peterson, Casteel, & Burgess, 1989), indicating that participants were performing the task as expected. Post hoc comparisons of significant effects were investigated using the *emmeans* package (Lenth, Singmann, Love, Buerkner, & Herve, 2019; version 1.4.3.01), which uses

[4] Note that early RTs (<150 ms) were not removed because they were very rare (20 RTs in total, <1 % of the data) and may have reflected predictive processing within the sentence (19/20 RTs occurred on *primary [strongly] accented* syllables that were on the *second syllable of the word*). An analysis based on the data set without these early values confirmed the main effects and interactions observed in the total set, with additional main effects of cue (p = .045) in the base model and cue order (p = .03) in the extended model that were modulated by the higher-order interactions and thus did not change result interpretation.

Table 2 Average undetected syllables across participants

	Low-level cue				High-level cue				Total			
	M	*SD*	%	Range	*M*	*SD*	%	Range	*M*	*SD*	%	Range
Accented (n = 20)	0.49	0.72	4.9	0–2	0.59	0.72	5.90	0–2	1.08	1.01	5.4	0–3
Unaccented (n = 40)	1.77	1.63	8.85	0–6	1.74	1.27	8.70	0–4	3.51	2.46	8.78	0–9
Total	2.26	1.90	7.5	0–7	2.33	1.61	7.77	0–6	4.59	2.94	7.65	0–12

Note. Percentage values were calculated by dividing the mean undetected syllables by the total amount of syllables in each category.

the estimates and standard errors within the GLMM to calculate whether there are significant differences between conditions. Reported p values were adjusted using the Tukey method for a family of estimates, as implemented in *emmeans*.

Results

Undetected syllables

On average, participants did not detect 4.59 out of 60 syllables (7.65%, SD = 2.94, range: 0–12), indicating that task performance was high. Average undetected syllables depending on cue and accent can be seen in Table 2. Probably linked to this ceiling performance, the GLMM showed no main effect of cue, $\chi^2(1, N = 39) = 0.54$, $p = .46$, no main effect of accent, $\chi^2(1, N = 39) = 2.56$, $p = .11$, and no interaction between cue and accent, $\chi^2(1, N = 39) = 0.44$, $p = .51$. If years of music lessons was added as a fixed effect to the model (including interactions), the same pattern of results was observed, and there was no main effect of years of music lessons, $\chi^2(1, N = 39) = 0.07$, $p = .80$, or other main effects or interactions (all p values > .11).

Syllable detection times

Base model: Cue and syllable type The base model (AIC = 27867) showed a main effect of accent, $\chi^2(1, N = 39) = 14.32$, $p < .001$: accented syllables were detected faster than unaccented syllables. The main effect of cue and its interaction with accent were not significant, $\chi^2(1, N = 39) = 1.58$, $p = .21$, and $\chi^2(1, N = 39) = 0.04$, $p = .84$, respectively (see Fig. 2a). These results suggest that there was no effect of cue type on subsequent syllable detection.

The role of cue order To investigate whether the initial cue, to which participants were presented first, affected the results, cue order was added as a fixed effect to the base model. The final model resulted in the fixed effects of cue, accent, cue order, and the Cue Order × Accent and Cue Order × Cue interactions. The addition of these effects significantly enhanced the model fit, $\chi^2(1, N = 39) = 6.96$, $p = .03$, and reduced the AIC value (AIC = 27,864), so they were kept in the model. The pattern of results suggests that the type of cue participants were initially presented with influenced RTs to accented versus unaccented syllables differently throughout the experiment (see Fig. 2b–c).[5] A Cue Order × Accent interaction, $\chi^2(1, N = 39) = 14.88$, $p < .001$ (see Table 3 for all contrasts) showed that participants who heard a high-level cue first detected accented syllables significantly faster than they detected unaccented syllables ($p = .003$, $SE = 17.4$). However, for participants who heard a low-level cue first, there was no difference between the detection of accented and unaccented syllables ($p = .48$, $SE = 13.9$). Between subjects, the detection of *accented* syllables was significantly faster for participants who heard a high-level cue first ($M = 559$ ms, $SD = 288$ ms) than for participants who heard a low-level cue first ($M = 656$ ms, $SD = 483$ ms), $p = .01$, $SE = 11.4$. However, the detection of *unaccented* syllables did not significantly differ between participants who heard a high-level cue first ($M = 659$ ms, $SD = 404$ ms) and participants who heard a low-level cue first ($M = 661$ ms, $SD = 353$ ms), $p = .98$, $SE = 14.5$.

A significant Cue Order × Cue interaction, $\chi^2(1, N = 39) = 19.85$, $p < .001$ revealed that RT was generally faster after a low-level cue than a high-level cue for participants who heard a low-level cue first, ($p = .002$, $SE = 8.33$). Participants who heard a high-level cue first did not show differences in RT depending on the cue that preceded the sentence. However, participants who had heard high-level cues at the beginning of the experiment were marginally faster at detecting syllables (both accented and unaccented) after a high-level cue than those who had heard a low-level cue first ($p = .067$, $SE = 13.73$). (See Table 4 and Fig. 2 for all contrasts.)

Additionally, there was a main effect of cue, $\chi^2(1, N = 39) = 12.95$, $p < .001$ which appeared to be modulated by the higher-order interactions presented above, as there was no significant difference between low-level and high-level cues when investigated without the interactions (estimate = −11.2, $SE = 7.98$, z ratio = −1.4, $p = .16$). The main effects of cue order, $\chi^2(1, N = 39) = 2.29$, $p = .13$, and accent, $\chi^2(1, N = 39)$

[5] It would be interesting to model the effect of these variables over time, and for primary versus secondary accents; however, we did not have enough data to run these analyses. A future experiment with more trials could further investigate how the current effects evolve over time and look more closely at potential differences between accent types.

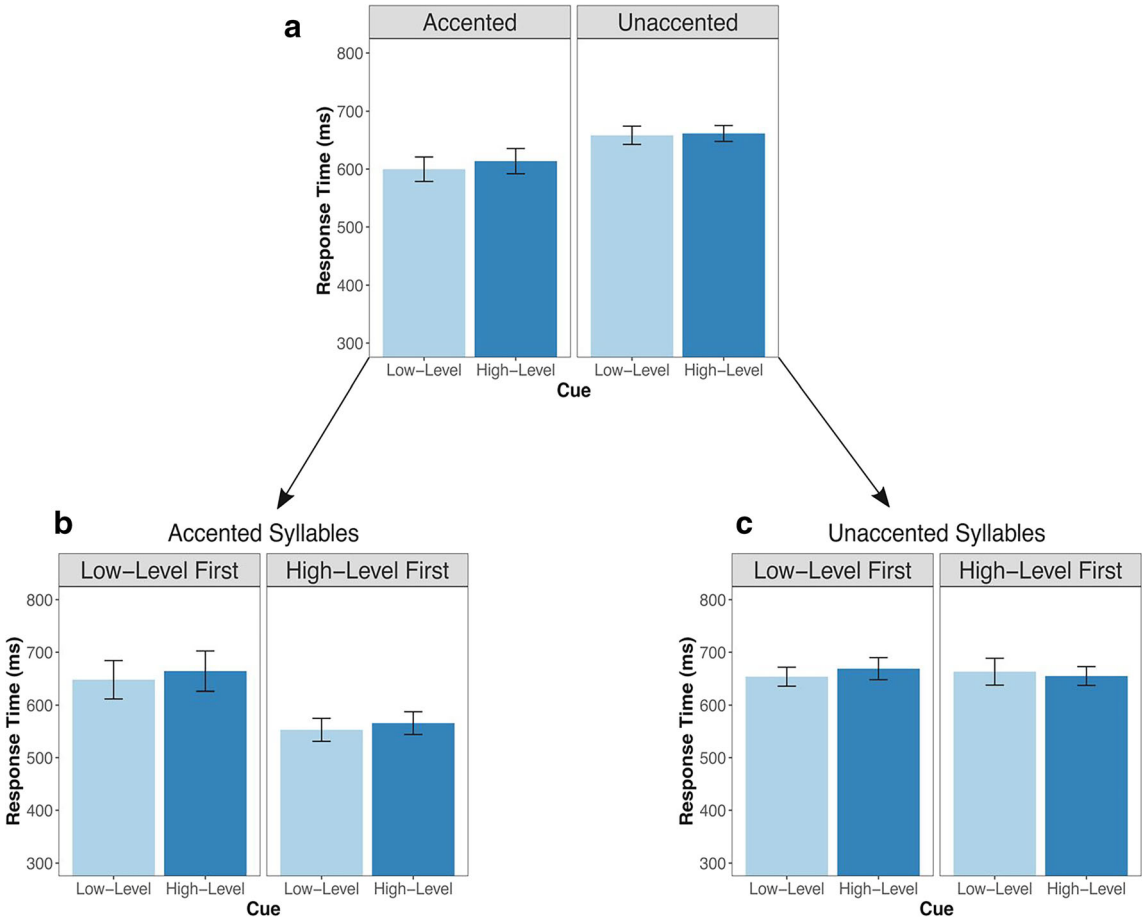

Fig. 2 Average response times (in milliseconds) for accented and unaccented syllables. **a** Representation of the base model with the factors of cue and accent. **b** and **c** Respectively, accented and unaccented syllable detection depending on whether the participant heard a low-level or a high-level cue first. Error bars represent one standard error around the mean

= 2.04, $p = .15$, were not significant. If years of music lessons was included as a fixed effect,[6] the same pattern of results was observed, and there was no main effect of years of music lessons, $\chi^2(1, N = 39) = 0.35$, $p = .56$.

Cue and sentence half It is a possibility that the syllable position within the sentence (first or second half of the sentence) could have affected RT performance. To investigate these potential effects combined with the cue, we added sentence half as a fixed effect into the final model presented above (including cue order). Based on the baseline results (see Experiment 2), we also added the interaction between sentence half and accent. A comparison of the model with sentence half compared with the model without sentence half showed that the addition of sentence half significantly improved the model, $\chi^2(2, N = 39) = 9.89$, $p = .007$, reducing the AIC value from 27,864 to 27,858. There was a significant main effect of sentence half, $\chi^2(1, N = 39) = 242.27$, $p < .001$, revealing that syllables in the second half of the sentence were detected faster than those in the first half of the sentence (see Fig. 3a). There was also a significant interaction between sentence half and accent, as in the baseline experiment, $\chi^2(1, N = 39) = 163.01$, $p < .001$. The interaction reflects a larger difference between RTs to accented and unaccented syllables in the second half of the sentence (estimate = −147.5, $SE = 14.6$, z ratio = −10.08, $p < .001$), compared with the first half of the sentence (estimate = −23.6, $SE = 10.2$, z ratio = −2.31, $p = .02$), though accented syllables were detected faster than unaccented syllables in both halves of the sentence (see Fig. 3b). The main effect of cue, $\chi^2(1, N = 39) = 12.49$, $p < .001$, interactions between cue order and accent, $\chi^2(1, N = 39) = 17.47$, $p < .001$, and cue order and cue, $\chi^2(1, N = 39) = 12.15$, $p < .001$ remained significant. The main effects of accent, $\chi^2(1, N = 39) = 0.16$, $p = .69$ and cue order, $\chi^2(1, N = 39) = 2.84$, $p = .09$, were nonsignificant.

[6] Note that interactions with years of music lessons could not be added to the model because there was not enough data for the model to converge.

Table 3 Cue Order × Accent interaction

Contrast	Estimate	SE	ratio	*p* value
L first, Accented – L first, Unaccented	−19.8	13.9	−1.43	.48
L first, Accented – H first, Accented	35.2	11.4	3.10	.01*
L first, Accented – H first, Unaccented	−25.1	21.1	−1.19	.63
L first, Unaccented – H first, Accented	55.0	16.7	3.30	.005*
L first, Unaccented – H first, Unaccented	−5.3	14.5	−0.37	.98
H first, Accented – H first, Unaccented	−60.3	17.4	−3.47	.003*

Note. L = low-level, H = high-level; comparisons are collapsed across cue type. * indicates significant contrasts at the p < .05 level.

Experiment 2

A baseline experiment was run on a second group of participants to investigate response time to accented and unaccented syllables without cues. All materials, computers, software, procedures, and analysis were the same, the only difference was that the cues were removed (and instructions and analysis adapted accordingly). The baseline experiment took approximately 10 minutes.

Method

Participants

Twenty native French speaking participants were recruited through the Lyon Neuroscience Research Centre and social media (M_{age} = 30.4, *SD* = 13.22 years; range: 19–61 years; 13 women). Participants had a range of musical training (*M* = 4.3, *SD* = 5.23 years of lessons; range: 0–14 years).[7] Eleven of these participants reported that they currently played (*n* = 8) or have played an instrument in the past (*M* = 8.6, *SD* = 4.84 years of lessons). Two participants reported having had speech therapy for reading, but no participants were dyslexic. No participants reported cognitive or neurological conditions or impairments.

RT cleaning

Negative removed RTs ranged from 1.67% to 11.67% per participant (*M* = 5.67%, *SD* = 2.57%). After negative RTs, undetected syllable responses, and trials with no responses were removed, there was a total data loss of between 3.3% and 43% of trials for each participant (*M* = 12.75%, *SD* = 8.58%). A paired-samples *t* test on the proportion of trials removed in each condition showed that there were more unaccented syllables removed (*M* = 13.25%, *SD* = 10.55%) compared with accented syllables (*M* = 5.5%, *SD* = 8.26%), $t(19) = 3.64$, $p = .002$, $d = 0.81$.

Results

Undetected syllables

On average, participants did not detect 3.2 out of 60 syllables (5.3%, *SD* = 2.63, range: 0–10), corresponding to an average percentage of 2.8% of accented syllables and 6.6% of unaccented syllables that were not detected.

Syllable detection times

Base model The base model (AIC = 14373) showed a main effect of accent, $\chi^2(1, N = 20) = 4.98$, $p = .03$: accented syllables were detected faster than unaccented syllables (see Fig. 4a). If years of music lessons was added as a fixed effect (including interactions), there was still a main effect of accent, $\chi^2(1, N = 20) = 5.16$, $p = .02$, no main effect of music lessons, $\chi^2(1, N = 20) = 1.59$, $p = .21$, and no interaction, $\chi^2(1, N = 20) = 1.86$, $p = .17$.

Sentence half Sentence half and interactions were included into the base model. This analysis revealed a significant main effect of sentence half, $\chi^2(1, N = 20) = 117.04$, $p < .001$, and an interaction between sentence half and accent, $\chi^2(1, N = 20) = 31.15$, $p < .001$. The main effect of accent was no longer significant, $\chi^2(1, N = 20) = 2.57$, $p = .11$. The interaction reveals that detection of accented syllables was significantly faster than the detection of unaccented syllables when syllables were in the second half of the sentence (estimate = −154.1, *SE* = 31.8, *z* ratio = −4.84, $p < .001$), but not in the first half of the sentence (estimate = −39.4, *SE* = 24.6, *z* ratio = −1.60, $p = .11$, suggesting that prediction for accented syllables might build up over time (see Fig. 4b).

Comparison with Experiment 1

Data analyses Data from Experiment 1 and Experiment 2 were combined to investigate effects of cue order (baseline, H first,

[7] An independent-samples *t* test on years of private music lessons showed no difference between musical training of participants in the baseline experiment compared to the main experiment, $t(38.32) = 0.72$, $p = .47$.

Table 4 Cue Order × Cue interaction

Contrast	Estimate	*SE*	*z* ratio	*p* value
L first, L cue – L first, H cue	−29.97	8.33	−3.60	.002*
L first, L cue – H first, L cue	−3.84	11.44	−0.34	.99
L first, L cue – H first, H cue	3.79	14.28	0.27	.99
L first, H cue – H first, L cue	26.13	14.40	1.82	.27
L first, H cue – H first, H cue	33.77	13.73	2.46	.067
H first, L cue – H first, H cue	7.63	9.67	0.79	.86

Note. Comparisons are collapsed across accent condition; L = low-level, H = high-level. * indicates significant contrasts at the $p < .05$ level.

L first), and accent condition (accented, unaccented) on syllable RT. The cue order model contained the between-subjects effect of cue order and the within-subjects effect of accent condition as fixed effects (including their interaction). The three-way interaction between cue, cue order, and accent was not possible because the baseline condition did not have any cue information. We did not further model differences between baseline and H or L cues as there were no differences depending on cue in the first experiment. Random effects structures and distributions were identical to the previous analyses. Comparisons and multiple comparisons correction were run with *emmeans* as described in the analysis section of Experiment 1.

Cue order and accent There was a main effect of cue order, $\chi^2(1, N = 20) = 8.95$, $p = .01$, a main effect of accent, $\chi^2(1, N = 59) = 72.89$, $p < .001$, and an interaction between cue order and accent, $\chi^2(1, N = 59) = 46.95$, $p < .001$. The interaction revealed the following pattern of results: for participants who heard L cues first, there was no difference between detection of accented and unaccented syllables ($p = .30$, $SE = 10.92$), whereas for the baseline participants (estimate = −61.2, $SE = 7.17$, z ratio = −8.54, $p < .001$) and those who heard an H cue first (estimate = −49.4, $SE = 10.07$, z ratio = −4.91, $p < .001$), accented syllables were detected faster than unaccented syllables. Hearing a L cue first appeared to have specifically disrupted detection of accented syllables, as participants who heard a L cue first were slower than both the baseline participants (estimate = −23.87, $SE = 8.36$, z ratio = −2.86, $p = .01$) and the H-first participants (estimate = 33.95, $SE = 11.99$, z ratio = 2.83, $p = .01$) at detecting accented syllables. There was no difference in accented syllable detection between the baseline and H-first groups (estimate = 10.08, $SE = 7.19$, z ratio = 1.40, $p = .34$). For the unaccented syllables there were no significant differences between conditions, though in comparison to the baseline condition, the L-first (estimate = 26.05, $SE = 11.89$, z ratio = 2.19, $p = .07$) and the H-first (estimate = 21.93, $SE = 10.53$, z ratio = 2.08, $p = .09$) conditions were marginally faster (see Fig. 5). This pattern of results shows that hearing L cues first was detrimental to the detection of accented syllables throughout the experiment.

General discussion

The current study was designed to determine whether directing attention to a low or high hierarchical level of a regular rhythmic cue could selectively influence syllable detection in a subsequently presented sentence at the syllable level or the accented syllable level. We predicted that *unaccented* syllable detection would be faster after low-level cues compared with high-level cues, and that *accented* syllable detection would be faster after high-level cues compared with low-level cues. We also considered the possibility that the cue participants were presented with first might impact subsequent metrical interpretation, and thereby syllable detection. Results from Experiment 1 provided evidence for this second potential outcome. There was no difference in accented or unaccented syllable detection depending on whether an L or H cue preceded the sentence. However, it was revealed that the cue participants were presented with first (and heard throughout the first 10 trials of the experimental session) influenced syllable detection throughout the experiment, suggesting a longer-lasting effect of rhythmic cue and primed attending style throughout the experiment.

Previous studies have found an effect of aligned versus nonaligned (Cason & Schön, 2012; Falk & Dalla Bella, 2016; Gould et al., 2015; Gould et al., 2017), matched versus nonmatched (Cason et al., 2015; Gordon et al., 2011), or regular (and matched) versus irregular (Falk, Lanzilotti, & Schön, 2017a) cues on subsequent speech processing. Our current experimental manipulation was more subtle, as both cue types were aligned and matched with the subsequent sentence. The cues differed in which structural level of the metric hierarchy they cued—notably, the *low-level* syllable or *high-level* accented syllable level. Therefore, both cues were congruent with the sentence structure, resulting in the potential for perception at multiple hierarchical levels.

Participants who started Experiment 1 with L cues were selectively slower at detecting accented syllables compared with participants who started the experiment with H cues and compared with baseline results in Experiment 2. Experiment 1 participants who heard the H cues first and Experiment 2 baseline participants showed the classic finding of faster RT for accented compared with unaccented syllables (Cutler, 1976; Cutler & Foss, 1977; Gow & Gordon, 1993; Pitt & Samuel, 1990; Shields, McHugh, & Martin, 1974). This pattern of results suggests that the initial cue that was perceived in Experiment 1 influenced the perception and grouping of the subsequent stimuli. Initial cue perception then influenced syllable detection, such that participants who were

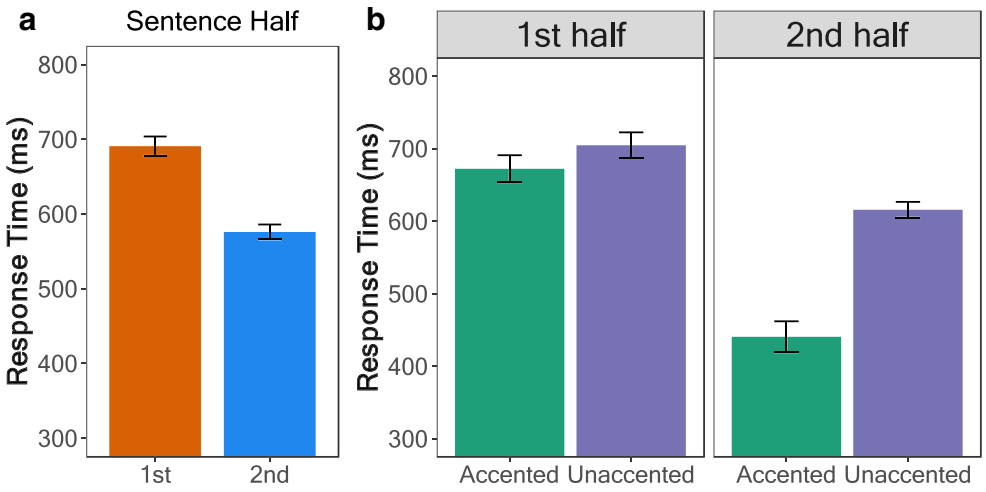

Fig. 3 **a** Participants were significantly faster to detect syllables in the second half of the sentence compared with the first half. **b** Faster detection of accented compared with unaccented syllables occurred most strongly in the second half of the sentence. Error bars represent one standard error around the mean

first exposed to the L cue were not able to benefit from this higher-level structure.

It might be argued that an alternative explanation as to why hearing L cues first may have disrupted accented syllable detection is that participants perceptually grouped L cues into a structure that did not match the sentences (e.g., a three-beat rhythm with a pause at the end of each group of two, rather than groupings of five tones and a pause), compared with a potential binary perception of the H cues. This explanation appears unlikely for three reasons. First, the L cues had a strong grouping structure, with five tones presented quickly in succession, and a silence between each group of tones. According to Gestalt grouping principles, such a pattern should enforce grouping boundaries based on similarity (the tones were all identical) and proximity (groups of tones were close together, separated by a pause). Second, prior evidence has shown a general preference for binary compared with ternary perception (Fujioka, Fidali, & Ross, 2014; Povel, 1981), suggesting that if subdivisions had been perceived within the groups of tones, they were more likely to be binary, in line with perceptual groupings that may have occurred for the H cue. Third, the cues directly matched the sentence structures that had clear accentual phrase groupings. Within the experimental context and repeated trial structure, it is likely that participants perceived the cues as they were intended rather than using greater cognitive energy to impose alternative grouping structures, such as ternary meter.

A more plausible explanation is that participants who heard low-level cues first were primed for analytic attending, whereas participants who heard high-level cues first were primed for future-oriented attending (Drake et al., 2000; Jones & Boltz, 1989). When a low-level grouping structure (i.e., the low-level cues) was presented at the beginning of Experiment 1, participants' attention and subsequent attending may have been

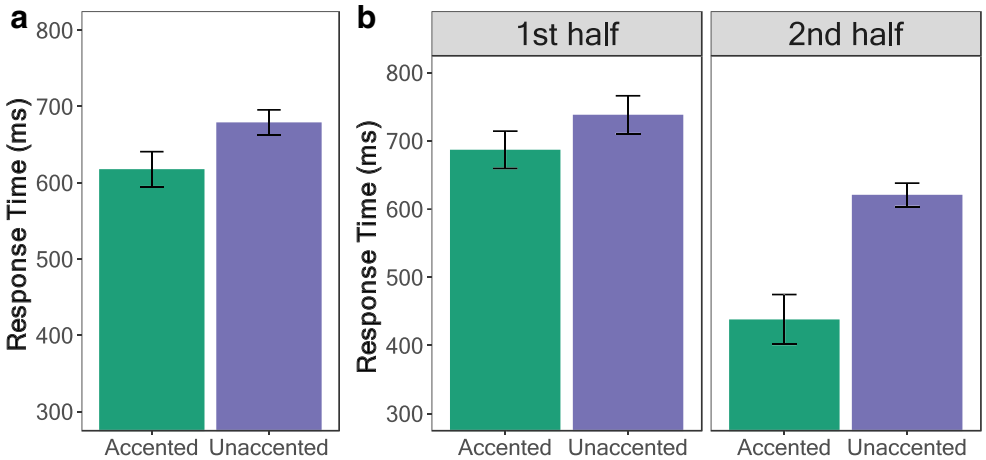

Fig. 4 **a** Baseline response times for accented and unaccented syllables. **b** Response times for accented and unaccented syllables depending on whether the syllable was in the first half or the second half of the sentence. Error bars reflect one standard error either side of the mean

directed to the lower, analytic level, where attention was equally distributed across all syllables. If this was the case, then participants who heard L cues first would not have been able to benefit from the higher-level structure of the sentences (i.e., the accents) to predict upcoming syllables. This interpretation is supported by the baseline experiment (Experiment 2), which showed that hearing a L cue first selectively disrupted the RT to accented syllables. It also appears that the first cue heard influenced how the subsequent cues within the experiment were perceived. This interpretation is consistent with the finding that participants who heard the low-level cues first in Experiment 1 were generally faster at detecting syllables (both unaccented and accented) when a L cue preceded the sentence compared with an H cue, as the low-level grouping structure and analytic attending style may have been reinforced. In addition, compared with baseline (Experiment 2), detection of *unaccented* syllables was marginally improved for participants who heard H or L cues first, suggesting a potential broader benefit of matched rhythmic cues regardless of the hierarchical level cued. These findings therefore support the concept of nested hierarchical oscillations and shows how future-attending and analytic attending could be primed within the DAT framework.

The order effects observed in Experiment 1 may also have resulted from temporal context effects (Snyder et al., 2015), whereby the initially perceived cues elicited an attractive effect, resulting in a grouping of the subsequent cues according to the perception of the initial cues. Similar effects of presentation order across the experimental session have been shown for pitch structure in music (Bigand, Poulin, Tillmann, Madurell, & D'Adamo, 2003) and metric perception in rhythm (Lenc et al., 2019), and previously heard differences between two sound streams have been shown to influence perception of a subsequent sound stream into one or two streams (e.g., Bregman, 1990; Snyder, Carter, Lee, Hannon, & Alain, 2008; Snyder, Holder, Weintraub, Carter, & Alain, 2009). This phenomenon can also be observed in other modalities—notably, vision (see Snyder et al., 2015). Future research could thus complement our research line with a perceptual experiment based on our rhythmic cues, with the goal to investigate the initially perceived grouping structure, and to study whether the strength of this grouping structure changes depending on the first cue block participants were exposed to. Further, based on the here observed influence of cue order on subsequent perception, future research could manipulate this block design to investigate systematically how it might impact task performance. Future research could investigate whether (1) the block design is critical to obtaining or maintaining effects of rhythmic cueing and cue order, and (2) similar results would occur without blocking the stimuli, or would be enhanced, reduced, or even eliminated with more or less trials in each block.

Based on these results, we suggest that the reinforcement of the grouping structure of the low-level or high-level cue at the beginning of Experiment 1 resulted in a selective, relative perceptual enhancement of the low-level or high-level structure within the cue throughout the experiment, with consequences for subsequent cue perception. This interpretation fits with research suggesting that sensory evidence actively accumulates to resolve uncertainty about upcoming events (Koelsch, Vuust, & Friston, 2018). Considering the strong regularity of the cues, once the initial cue was repeatedly heard and could be easily predicted, subsequent cues may have been influenced by the perception of the first cues. It would be particularly interesting to use EEG to investigate whether embedded neural oscillations are elicited differently by the two types of cues, and notably, whether the strength of these oscillatory levels differs depending on the initial cue participants

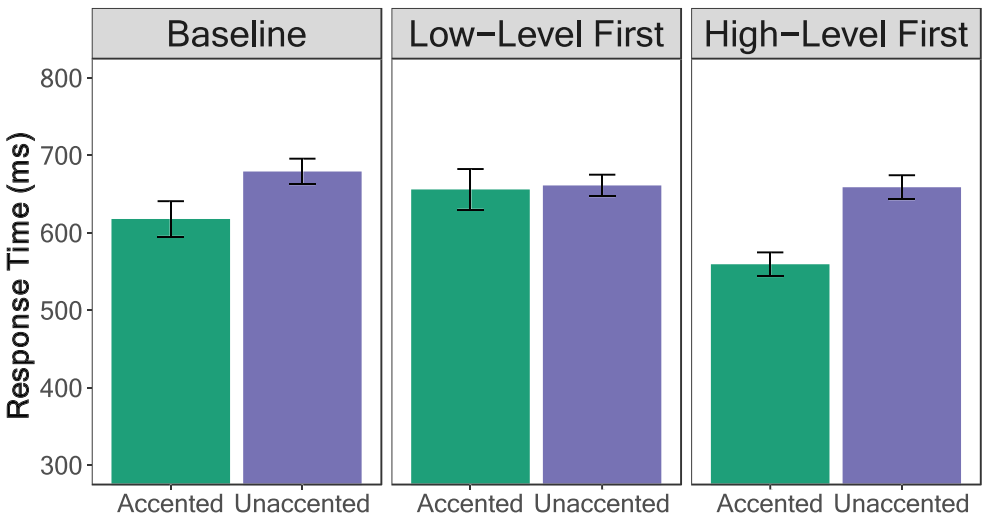

Fig. 5 Mean response times for accented and unaccented syllables depending on whether participants performed the baseline experiment or were in the low-level first or high-level first group of the main experiment. Error bars represent one standard error either side of the mean

were presented with. Research has shown that imagining either a binary or ternary meter on top of an isochronous rhythmic sequence results in a brain response elicited at this meter frequency (Nozaradan et al., 2011). It is therefore possible that perception of a dominant frequency can be selectively enhanced depending on top-down processes. Future research could investigate the current paradigm with naturally spoken sentences; however, it may be necessary to adapt the cues to the naturally spoken sentence structure. This manipulation would be particularly interesting considering that temporal prediction might occur in both rhythmic and nonrhythmic stimuli (Rimmele, Morillon, Poeppel, & Arnal, 2018).

The current findings also show that syllables were detected faster when they were in the second compared with the first half of the sentence, and that the accented syllable advantage occurred primarily in the second half of the sentence. Faster syllable detection in the second half of the sentence could be explained by linguistic context effects and/or foreperiod effects. Linguistic context effects would suggest that the increase in linguistic contextual information over time should enhance predictions about upcoming words (and syllables), thereby reducing RT when more information has been accumulated in the sentence (Montgomery, 2000; Montgomery et al., 1990; Simpson et al., 1989). It might be argued that foreperiod effects could have occurred because participants were aware that all sentences contained the to-be-detected syllable, so attention likely increased toward the end of the sentence as it became more probable over time that the syllable would occur (as suggested in Planchou et al., 2015). Previous work has suggested that foreperiod and dynamic attending can work in parallel, but reflect separate cognitive processes (Jones, Hsu, Granjon, & Waszak, 2017). The finding that accented syllables were detected faster than unaccented syllables in the second half of the sentence in particular suggests a buildup over time of dynamic attending and expectation.

The traditional foreperiod effects appear to be reflected by generally faster syllable detection in the second half of the sentence. However, it should be noted that the foreperiod or linguistic context effects cannot explain our current result pattern related to accented/unaccented syllables because (1) there was no difference between temporal occurrence of accented and unaccented syllables across sentences (i.e., faster detection of accented syllables cannot be explained by accented syllables occurring more often toward the end of the sentence); (2) a different result pattern occurred depending on the cue the participants heard first, suggesting that the results could not be based only on the distribution of the syllables themselves; and (3) there was no consistent pairing of accented compared with unaccented syllables with a high-level or low-level cue, suggesting that the results are not based on a confound in the experimental material. Our results therefore fit nicely into the literature, reflecting effects of both dynamic attending and foreperiod effects, with dynamic attending also explaining the larger benefit to accented syllable detection in the second half of the sentence. They also validate our syllable detection task, as participants' expectations appeared to grow across sentences, as would be expected. Future research could aim to tease apart foreperiod and dynamic attending effects by varying sentence length (and therefore predictability of when the syllable will occur) and adding catch trials where there is no syllable to detect so that expectation does not necessarily increase throughout the sentence.

The current results have implications for short-term and long-term metrical cueing and rhythmic training to influence subsequent phonological processing of accented or stressed syllables for speech-impaired populations. The finding that attending style can be influenced by an initially perceived cue suggests the potential for priming over a longer period of time than just directly before a sentence. The current experiments showed a sustained cost of hearing the L cue first in the detection of accented syllables. There appeared to be no direct benefit of hearing the H cue first, as participants were already able to benefit from the higher-level grouping structure of these sentences (i.e., enhanced accented syllable detection). Considering that our participants were typically developed adults, it is possible that their detection of accented syllables was already at ceiling level and their behavioral performance could not be improved. However, for participants with deficits tracking the speech envelope, training or priming a future-oriented attending style may enhance detection of accented syllables, which is valuable because stressed syllables contain more informational content within sentences compared with unstressed syllables (Altman & Carter, 1989; Calhoun, 2010). Rhythmic training could particularly help children with developmental dyslexia, who show impairments in neural tracking of the speech signal (Goswami, 2011), and deficits in stressed syllable processing (Barry, Harbodt, Cantiani, Sabisch, & Zobay, 2012; Jiménez-Fernández, Gutiérrez-Palma, & Defior, 2015). Further, more long-term rhythmic training in abstracting high-level rhythmic structures in music could also have potential applications to the processing of higher-level metric structures in speech (i.e., stress patterns, phrasal boundaries), considering the strong connections between music rhythm and speech processing (Beier & Ferreira, 2018; Tierney & Kraus, 2013a, 2013b). However, potential benefits of long-term training and whether it is possible to prime attending style for longer than an experimental session will need to be tested in future research.

Conclusion

The current results suggest that the initial perception of a rhythmic cue can influence subsequent sentence perception throughout an experimental session, regardless of the cue type

immediately preceding each individual sentence. The presentation of a low-level or high-level cue at the beginning of the experiment may have encouraged analytic-oriented or future-oriented attending styles (in line with DAT) that then persisted across the experiment and influenced subsequent cue and sentence perception. These results can also be interpreted in line with temporal context effects, whereby initial perception of the cue draws subsequent perception closer to the initially perceived grouping structure. The current study has implications for the use of metrical cueing and rhythmic training to direct attention to higher-level grouping structures within speech processing.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13414-021-02251-y.

Acknowledgements We thank Etienne Gaudrain, Oussama Abdoun, and Romain Bouet for advice, help, and discussions about the generalized linear mixed-model approach, Daniele Schön for advice on cue and trial construction, Simone Dalla Bella for collaboration on sentence stimuli construction, and Pauline Ortega for collecting the data for the baseline experiment. This research was supported by a grant from Agence Nationale de la Recherche (ANR-16-CE28-0012-02) to BT, and also in part by a grant to BT from the National Institutes of Health Common Fund under award DP2HD098859 (PI: R. L. Gordon), through the Office of Strategic Coordination/Office of the NIH Director. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The team *Auditory Cognition and Psychoacoustics* is part of the LabEx CeLyA (Centre Lyonnais d'Acoustique, ANR-10-LABX-60).

Declarations

Ethics statement This study was run in accordance with the Declaration of Helsinki, and was approved by the French Ethics committee *Comité de Protection des Personnes* (CPP) SUD-EST II. All participants provided written, informed consent.

References

Altman, G., & Carter, D. (1989). Lexical stress and lexical discriminability: Stressed syllables are more informative, but why? *Computer Speech & Language*, *3*(3), 265–275. https://doi.org/10.1016/0885-2308(89)90022-3

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica*, *66*(1/2), 46–63. https://doi.org/10.1159/000208930

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

Aylett, M., & Turk, A. (2004). The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, *47*(1), 31–56. https://doi.org/10.1177/00238309040470010201

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Barnes, R., & Jones, M. R. (2000). Expectancy, attention, and time. *Cognitive Psychology*, *41*(3), 254–311. https://doi.org/10.1006/cogp.2000.0738

Barry, J. G., Harbodt, S., Cantiani, C., Sabisch, B., & Zobay, O. (2012). Sensitivity to lexical stress in dyslexia: A case of cognitive not perceptual stress. *Dyslexia*, *18*(3), 139–165. https://doi.org/10.1002/dys.1440

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

Beier, E. J., & Ferreira, F. (2018). The temporal prediction of stress in speech and its relation to musical beat perception. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.00431

Bigand, E., Poulin, B., Tillmann, B., Madurell, F., & D'Adamo, D. A. (2003). Sensory versus cognitive components in harmonic priming. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(1), 159–171. https://doi.org/10.1037/0096-1523.29.1.159

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–347.

Bolger, D., Trost, W., & Schön, D. (2013). Rhythm implicitly affects temporal orienting of attention across modalities. *Acta Psychologica*, *142*(2), 238–244. https://doi.org/10.1016/j.actpsy.2012.11.012

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound.* MIT Press.

Brochard, R., Tassin, M., & Zagar, D. (2013). Got rhythm . . . for better and for worse. Cross-modal effects of auditory rhythm on visual word recognition. *Cognition*, *127*(2), 214–219. https://doi.org/10.1016/j.cognition.2013.01.007

Brown, M., Salverda, A. P., Dilley, L. C., & Tanenhaus, M. K. (2015). Metrical expectations from preceding prosody influence perception of lexical stress. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(2), 306–323. https://doi.org/10.1037/a0038689

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 9. https://doi.org/10.5334/joc.10

Calhoun, S. (2010). How does informativeness affect prosodic prominence? *Language and Cognitive Processes*, *25*(7/9), 1099–1140. https://doi.org/10.1080/01690965.2010.491682

Canette, L.-H., Fiveash, A., Krzonowski, J., Corneyllie, A., Lalitte, P., Thompson, D., Trainor, L., Bedoin, N., & Tillmann, B. (2020). Regular rhythmic primes boost P600 in grammatical error processing in dyslexic adults and matched controls. *Neuropsychologia*, *138*, 107324. https://doi.org/10.1016/j.neuropsychologia.2019.107324

Cason, N., Astésano, C., & Schön, D. (2015). Bridging music and speech rhythm: Rhythmic priming and audio–motor training affect speech perception. *Acta Psychologica*, *155*(0), 43–50. https://doi.org/10.1016/j.actpsy.2014.12.002

Cason, N., & Schön, D. (2012). Rhythmic priming enhances the phonological processing of speech. *Neuropsychologia*, *50*(11), 2652–2658. https://doi.org/10.1016/j.neuropsychologia.2012.07.018

Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics*, *26*(2), 145–171. https://doi.org/10.1006/jpho.1998.0070

Cunningham, J. B., & McCrum-Gardner, D. E. (2007). Power, effect and sample size using G*Power: Practical issues for researchers and members of research ethics committees. *Evidence Based Midwifery, 5*(4), 132–136.

Cutler, A. (1976). Phoneme-monitoring reaction time as a function of preceding intonation contour. *Perception & Psychophysics*, *20*(1), 55–60. https://doi.org/10.3758/BF03198706

Cutler, A., & Foss, D. J. (1977). On the role of sentence stress in sentence processing. *Language and Speech*, *20*(1), 1–10. https://doi.org/10.1177/002383097702000101

Delignette-Muller, M. L., & Dutang, C. (2015). fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software*, *64*(1), 1–34. https://doi.org/10.18637/jss.v064.i04

Dilley, L. C., & McAuley, J. D. (2008). Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language*, *59*(3), 294–311. https://doi.org/10.1016/j.jml.2008.06.006

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. https://doi.org/10.1038/nn.4186

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews*. https://doi.org/10.1016/j.neubiorev.2017.02.011

Drake, C., Jones, M. R., & Baruch, C. (2000). The development of rhythmic attending in auditory sequences: Attunement, referent period, focal attending. *Cognition*, *77*(3), 251–288. https://doi.org/10.1016/S0010-0277(00)00106-2

Falk, S., & Dalla Bella, S. (2016). It is better when expected: Aligning speech and motor rhythms enhances verbal processing. *Language, Cognition and Neuroscience*, *31*(5), 699–708. https://doi.org/10.1080/23273798.2016.1144892

Falk, S., Lanzilotti, C., & Schön, D. (2017a). Tuning neural phase entrainment to speech. *Journal of Cognitive Neuroscience*, *29*(8), 1378–1389. https://doi.org/10.1162/jocn_a_01136

Falk, S., Volpi-Moncorger, C., & Dalla Bella, S. (2017b). Auditory-motor rhythms and speech processing in French and German listeners. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00395

Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191.

Fitch, W. T. (2013). Rhythmic cognition in humans and animals: Distinguishing meter and pulse perception. *Frontiers in Systems Neuroscience*, *7*, 68. https://doi.org/10.3389/fnsys.2013.00068

Fiveash, A., Bedoin, N., Lalitte, P., & Tillmann, B. (2020a). Rhythmic priming of grammaticality judgments in children: Duration matters. *Journal of Experimental Child Psychology*, *197*, 104885. https://doi.org/10.1016/j.jecp.2020.104885

Fiveash, A., Schön, D., Canette, L.-H., Morillon, B., Bedoin, N., & Tillmann, B. (2020b). A stimulus-brain coupling analysis of regular and irregular rhythms in adults with dyslexia and controls. *Brain and Cognition*, *140*, 105531. https://doi.org/10.1016/j.bandc.2020.105531

Fotidzis, T. S., Moon, H., Steele, J. R., & Magne, C. L. (2018). Cross-modal priming effect of rhythm on visual word recognition and its relationships to music aptitude and reading achievement. *Brain Sciences*, *8*(12). https://doi.org/10.3390/brainsci8120210

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (2nd ed.). SAGE Publications.

Fujioka, T., Fidali, B. C., & Ross, B. (2014). Neural correlates of intentional switching from ternary to binary meter in a musical hemiola pattern. *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01257

Fujioka, T., Ross, B., & Trainor, L. J. (2015). Beta-band oscillations represent auditory beat and its metrical hierarchy in perception and imagery. *Journal of Neuroscience*, *35*(45), 15187–15198. https://doi.org/10.1523/JNEUROSCI.2397-15.2015

Giraud, A., & Poeppel, D. (2012). Cortical oscillations and speech processing: Emerging computational principles and operations. *Nature Neuroscience*, *15*(4), 511–517. https://doi.org/10.1038/nn.3063

Gordon, R. L., Magne, C. L., & Large, E. W. (2011). EEG correlates of song prosody: A new look at the relationship between linguistic and musical rhythm. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00352

Goswami, U. (2011). A temporal sampling framework for developmental dyslexia. *Trends in Cognitive Sciences*, *15*(1), 3–10. https://doi.org/10.1016/j.tics.2010.10.001

Gould, L., McKibben, T., Ekstrand, C., Lorentz, E., & Borowsky, R. (2015). The beat goes on: The effect of rhythm on reading aloud. *Language, Cognition and Neuroscience*, 1–15. Advance online publication. https://doi.org/10.1080/23273798.2015.1089360

Gould, L., Mickleborough, M. J. S., Ekstrand, C., Lorentz, E., & Borowsky, R. (2017). Examining the neuroanatomical and the behavioural basis of the effect of basic rhythm on reading aloud. *Language, Cognition and Neuroscience*, *32*(6), 724–742. https://doi.org/10.1080/23273798.2016.1271135

Gow, D. W., & Gordon, P. C. (1993). Coming to terms with stress: Effects of stress location in sentence processing. *Journal of Psycholinguistic Research*, *22*(6), 545–578. https://doi.org/10.1007/BF01072936

Grahn, J. A. (2012). Neural mechanisms of rhythm perception: Current findings and future perspectives. *Topics in Cognitive Science*, *4*(4), 585–606. https://doi.org/10.1111/j.1756-8765.2012.01213.x

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech—A syllable-centric perspective. *Journal of Phonetics*, *31*(3), 465–485. https://doi.org/10.1016/j.wocn.2003.09.005

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLOS Biology*, *11*(12), e1001752. https://doi.org/10.1371/journal.pbio.1001752

Hickok, G., Farahbod, H., & Saberi, K. (2015). The rhythm of perception: Entrainment to acoustic rhythms induces subsequent perceptual oscillation. *Psychological Science*, *26*(7), 1006–1013. https://doi.org/10.1177/0956797615576533

Jiménez-Fernández, G., Gutiérrez-Palma, N., & Defior, S. (2015). Impaired stress awareness in Spanish children with developmental dyslexia. *Research in Developmental Disabilities*, *37*, 152–161. https://doi.org/10.1016/j.ridd.2014.11.002

Jones, A., Hsu, Y.-F., Granjon, L., & Waszak, F. (2017). Temporal expectancies driven by self- and externally generated rhythms. *NeuroImage*, *156*, 352–362. https://doi.org/10.1016/j.neuroimage.2017.05.042

Jones, M. R. (1976). Time, our lost dimension: Toward a new theory of perception, attention and memory *Psychological Review*, *83*(5), 323–355.

Jones, M. R. (2016). Musical time. In S. Hallam, I. Cross, & M. Thaut, *The Oxford Handbook of Music Psychology* (2nd ed.). Oxford University Press.

Jones, M. R. (2019). *Time will tell*. Oxford University Press.

Jones, M. R., & Boltz, M. (1989). Dynamic attending and responses to time. *Psychological Review*, *96*(3), 459–491.

Jones, M. R., Johnston, H. M., & Puente, J. (2006). Effects of auditory pattern structure on anticipatory and reactive attending. *Cognitive Psychology*, *53*(1), 59–96. https://doi.org/10.1016/j.cogpsych.2006.01.003

Jones, M. R., Moynihan, H., MacKenzie, N., & Puente, J. (2002). Temporal aspects of stimulus-driven attending in dynamic arrays. *Psychological Science*, *13*(4), 313–319. https://doi.org/10.1111/1467-9280.00458

Koelsch, S., Vuust, P., & Friston, K. (2018). Predictive processes and the peculiar case of music. *Trends in Cognitive Sciences*, *23*(1), 63–77. https://doi.org/10.1016/j.tics.2018.10.006

Kösem, A., Bosker, H. R., Takashima, A., Meyer, A., Jensen, O., & Hagoort, P. (2018). Neural entrainment determines the words we hear. *Current Biology*, *28*(18), 2867–2875.e3. https://doi.org/10.1101/175000

Kotz, S. A., Ravignani, A., & Fitch, W. T. (2018). The evolution of rhythm processing. *Trends in Cognitive Sciences*, *22*(10), 896–910. https://doi.org/10.1016/j.tics.2018.08.002

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(1), 1–26. https://doi.org/10.18637/jss.v082.i13

Large, E. W. (2008). Resonating to musical rhythm: Theory and experiment. In S. Grondin (Ed.), *Psychology of time*. Emerald Group.

Large, E. W., Herrera, J. A., & Velasco, M. J. (2015). Neural networks for beat perception in musical rhythm. *Frontiers in Systems Neuroscience*, *9*. https://doi.org/10.3389/fnsys.2015.00159

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review*, *106*(1), 119–159.

Lenc, T., Keller, P. E., Varlet, M., & Nozaradan, S. (2019). Hysteresis in the selective synchronization of brain activity to musical rhythm. *BioRxiv*, 696914. https://doi.org/10.1101/696914

Lenth, R, Singmann, H., Love, J., Buerkner, P., & Herve, M. (2019). emmeans: Estimated marginal means, aka least-squares means [Computer software]. https://cran.r-project.org/web/packages/emmeans/index.html

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

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01171

London, J. (2012). *Hearing in time: Psychological aspects of musical meter* (2nd ed.). Oxford University Press.

McAuley, J. D. (2010). Tempo and rhythm. In M. R. Jones (Ed.), *Music perception*. Springer.

McAuley, J. D., & Kidd, G. R. (1998). Effect of deviations from temporal expectations on tempo discrimination of isochronous tone sequences. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(6), 1786–1800.

Montgomery, J. W. (2000). Relation of working memory to off-line and real-time sentence processing in children with specific language impairment. *Applied Psycholinguistics*, *21*(1), 117–148.

Montgomery, J. W., Scudder, R. R., & Moore, C. A. (1990). Language-impaired children's real-time comprehension of spoken language. *Applied Psycholinguistics*, *11*(3), 273–290. https://doi.org/10.1017/S0142716400008894

Nozaradan, S. (2014). Exploring how musical rhythm entrains brain activity with electroencephalogram frequency-tagging. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*. https://doi.org/10.1098/rstb.2013.0393

Nozaradan, S., Peretz, I., Missal, M., & Mouraux, A. (2011). Tagging the neuronal entrainment to beat and meter. *The Journal of Neuroscience*, *31*(28), 10234–10240. https://doi.org/10.1523/jneurosci.0411-11.2011

Phillips-Silver, J., & Trainor, L. J. (2007). Hearing what the body feels: Auditory encoding of rhythmic movement. *Cognition*, *105*(3), 533–546. https://doi.org/10.1016/j.cognition.2006.11.006

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), 564–573. https://doi.org/10.1037//0096-1523.16.3.564

Planchou, C., Clément, S., Béland, R., Cason, N., Motte, J., & Samson, S. (2015). Word detection in sung and spoken sentences in children with typical language development or with specific language impairment. *Advances in Cognitive Psychology*, *11*(4), 118–135. https://doi.org/10.5709/acp-0177-8

Povel, D.-J. (1981). Internal representation of simple temporal patterns. *Journal of Experimental Psychology: Human Perception and Performance*, *7*(1), 3–18. https://doi.org/10.1037/0096-1523.7.1.3

R Core Team (2018) R: A language and environment for statistical computing [Computer software]. https://www.R-project.org/

Rimmele, J. M., Morillon, B., Poeppel, D., & Arnal, L. H. (2018). Proactive sensing of periodic and aperiodic auditory patterns. *Trends in Cognitive Sciences*, *22*(10), 870–882. https://doi.org/10.1016/j.tics.2018.08.003

Rouder, J. N. (2014). Optional stopping: No problem for Bayesians. *Psychonomic Bulletin & Review*, *21*(2), 301–308. https://doi.org/10.3758/s13423-014-0595-4

Shields, J. L., McHugh, A., & Martin, J. G. (1974). Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. *Journal of Experimental Psychology*, *102*(2), 250–255. https://doi.org/10.1037/h0035855

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simpson, G. B., Peterson, R. R., Casteel, M. A., & Burgess, C. (1989). Lexical and sentence context effects in word recognition. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *15*(1), 88–97. https://doi.org/10.1037/0278-7393.15.1.88

Snyder, J. S., Carter, O. L., Lee, S.-K., Hannon, E. E., & Alain, C. (2008). Effects of context on auditory stream segregation. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(4), 1007–1016. https://doi.org/10.1037/0096-1523.34.4.1007

Snyder, J. S., Holder, W. T., Weintraub, D. M., Carter, O. L., & Alain, C. (2009). Effects of prior stimulus and prior perception on neural correlates of auditory stream segregation. *Psychophysiology*, *46*(6), 1208–1215. https://doi.org/10.1111/j.1469-8986.2009.00870.x

Snyder, J. S., Schwiedrzik, C. M., Vitela, A. D., & Melloni, L. (2015). How previous experience shapes perception in different sensory modalities. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/fnhum.2015.00594

Stupacher, J., Wood, G., & Witte, M. (2017). Neural entrainment to polyrhythms: A comparison of musicians and non-musicians. *Frontiers in Neuroscience*, *11*. https://doi.org/10.3389/fnins.2017.00208

Tal, I., Large, E. W., Rabinovitch, E., Wei, Y., Schroeder, C. E., Poeppel, D., & Zion Golumbic, E. (2017). Neural entrainment to the beat: The "missing-pulse" phenomenon. *The Journal of Neuroscience*, *37*(26), 6331–6341. https://doi.org/10.1523/JNEUROSCI.2500-16.2017

ten Oever, S., Schroeder, C. E., Poeppel, D., van Atteveldt, N., & Zion-Golumbic, E. (2014). Rhythmicity and cross-modal temporal cues facilitate detection. *Neuropsychologia*, *63*(Supplement C), 43–50. https://doi.org/10.1016/j.neuropsychologia.2014.08.008

Tierney, A., & Kraus, N. (2013a). Music training for the development of reading skills. In M. M. Merzenich, M. Nahum, & T. M. Van Vleet (Eds.), *Progress in brain research* (Vol. 207, pp. 209–241). Elsevier. https://doi.org/10.1016/B978-0-444-63327-9.00008-4

Tierney, A., & Kraus, N. (2013b). The ability to tap to a beat relates to cognitive, linguistic, and perceptual skills. *Brain and Language*, *124*(3), 225–231. https://doi.org/10.1016/j.bandl.2012.12.014

Tierney, A., & Kraus, N. (2013c). Neural responses to sounds presented on and off the beat of ecologically valid music. *Frontiers in Systems Neuroscience*, *7*. https://doi.org/10.3389/fnsys.2013.00014

Trapp, S., Havlicek, O., Schirmer, A., & Keller, P. E. (2018). When the rhythm disappears and the mind keeps dancing: Sustained effects of attentional entrainment. *Psychological Research*, 1–7. Advance online publication. https://doi.org/10.1007/s00426-018-0983-x

Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. (2017). A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America*, *142*(4), 1976–1989. https://doi.org/10.1121/1.5006179