NeuroImage 200 (2019) 490–500

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Hierarchical organization of melodic sequences is encoded by cortical entrainment

Lucas S. Baltzell [a,*], Ramesh Srinivasan [a,b], Virginia Richards [a]

[a] Department of Cognitive Sciences, University of California, Irvine, 3151 Social Sciences Plaza, Irvine, CA, 92687, USA
[b] Department of Biomedical Engineering, University of California, Irvine, 3151 Social Sciences Plaza, Irvine, CA, 92687, USA

ARTICLE INFO

Keywords:
Entrainment
EEG
Music
Language
Hierarchical
Cortical

ABSTRACT

Natural speech is organized according to a hierarchical structure, with individual speech sounds combining to form abstract linguistic units, and abstract linguistic units combining to form higher-order linguistic units. Since the boundaries between these units are not always indicated by acoustic cues, they must often be computed internally. Signatures of this internal computation were reported by Ding et al. (2016), who presented isochronous sequences of mono-syllabic words that combined to form phrases that combined to form sentences, and showed that cortical responses simultaneously encode boundaries at multiple levels of the linguistic hierarchy. In the present study, we designed melodic sequences that were hierarchically organized according to Western music conventions. Specifically, isochronous sequences of "sung" nonsense syllables were constructed such that syllables combined to form triads outlining individual chords, which combined to form harmonic progressions. EEG recordings were made while participants listened to these sequences with the instruction to detect when violations in the sequence structure occurred. We show that cortical responses simultaneously encode boundaries at multiple levels of a melodic hierarchy, suggesting that the encoding of hierarchical structure is not unique to speech. No effect of musical training on cortical encoding was observed.

1. Introduction

When fluent in a language, listeners can effortlessly segment continuous acoustic input into discrete linguistic units. To understand the auditory mechanisms that support this segmentation, neural correlates of the segmentation process have been investigated using a number of different paradigms in the last two decades, and one of the central questions has been the degree to which the neural mechanisms recruited for the segmentation of speech are also recruited for the segmentation of non-linguistic stimuli. In this introduction, we will begin by reviewing a body of ERP (event-related potentials) literature suggesting that similar neural mechanisms are recruited for the segmentation of speech and non-speech stimuli. We will then review the literature suggesting that cortical oscillations play a role in the speech segmentation process, and consider the claim that cortical oscillations support the segmentation of speech at multiple, simultaneous timescales, thereby encoding its hierarchical structure. Finally, we will review the literature suggesting that cortical oscillations also play a role in the segmentation of music, and hypothesize that the neural encoding of hierarchical structure might also be observed for hierarchically organized melodic sequences.

1.1. Background

In a landmark study, Saffran et al. (1996) constructed a set of discrete three-syllable sequences from a set of nonsense syllables. They concatenated these discrete sequences to form a continuous stream of syllables, and demonstrated that human infants can segment the discrete sequences from the continuous stream. In the absence of any acoustic cues indicating the boundaries between discrete sequences, this segmentation was based entirely on the statistical relationships between the syllables. This finding had profound implications for models of language acquisition, having suggested that the segmentation of sound sequences into discrete words might be a product of statistical learning. It was soon shown however, that this effect was not limited to linguistic stimuli, as both infants and adults were able to segment discrete sequences from continuous streams of tones as accurately as they could segment discrete sequences from continuous streams of syllables by identifying which sounds tended to occur in succession (Saffran et al., 1999).

A large body of literature has since emerged examining the neural correlates of segmentation within this statistical learning framework. Across a range of linguistic and non-linguistic stimuli, ERP signatures of

* Corresponding author. University of California, Department of Cognitive Sciences, 3151 Social Science Plaza, Irvine, CA, 92697, USA.
E-mail address: lucassbaltzell@gmail.com (L.S. Baltzell).

https://doi.org/10.1016/j.neuroimage.2019.06.054
Received 8 March 2019; Received in revised form 20 June 2019; Accepted 23 June 2019
Available online 27 June 2019

segmentation have been investigated by comparing ERP responses to statistically learned sequences vs. unlearned sequences, yielding two key findings. First, the N100 component indicates segmentation of syllable sequences (Sanders et al., 2002), sung syllable sequences (Francois and Schön, 2010), tone sequences (Abla et al., 2008), and sequences of incidental sounds (Sanders et al., 2009). The N100 is a largely obligatory ERP component generated in response to the onset of a sound, and the fact that it is larger for learned than for unlearned sequences suggests that the recognition of a sequence modulates its initial stages of processing.

Second, the N400 component indicates segmentation for some sequences but not for others. The N400 is typically associated with semantic processing (Friederici, 2002; De Diego Balaguer et al., 2007), and consistent with this role, the N400 response to sequences of syllables was significantly larger when the sequences were learned vs. unlearned (Sanders et al., 2002). Using the same paradigm though, Sanders et al. (2009) did not find an N400 response to learned incidental sound sequences compared to unlearned sequences. Using tone sequences, Abla et al. (2008) found that the N400 response indicated on-line sequence learning, with the highest performing sequence learners exhibiting an N400 response to learned sequences in the first block, intermediate sequence learners in the second/third blocks, and the poorest sequence learners not at all. Interestingly, the N400 response disappeared for highest performing sequence learners in the second/third blocks, suggesting that the N400 may be more indicative of the learning of tonal sequences than of their recognition (for a similar effect using speech stimuli, see Cunillera et al., 2009). While the specific role of the N400 component in segmentation is not entirely clear, it appears to be more dependent than the N100 on the type of sequence being segmented. The same can also be said for the later components (500–900 ms) reported by Francois and Schön (2010) using sequences of sung syllables (for review, see Schon & Francois, 2011).

Together, these key findings suggest that continuous streams of both speech and non-speech sounds can be segmented into discrete sequences based on the conditional probabilities between individual sounds, that ERP signatures of this segmentation process can be observed, and that the underlying neural mechanisms may be similar but not identical.

Research in the last two decades has also suggested that cortical oscillations entrain (phase-lock) to fluctuations in speech energy to support the parsing of the signal into discrete linguistic units (e.g. Ghitza, 2011; Giraud and Poeppel, 2012; Wilsch et al., 2018). This suggestion is based on the observation of spontaneous oscillations in local field potentials in the auditory cortex, and the fact that the phase of these oscillations reset in response to acoustic input (e.g. Lakatos et al., 2005; Giraud et al., 2007; Canolty et al., 2006). Furthermore, these oscillations can organize hierarchically such that the phase of low-frequency oscillations govern the amplitude of high-frequency oscillations, potentially supporting the hierarchical segmentation of speech (Buzsaki and Draguhn, 2004; Schroeder and Lakatos, 2009). Finally, temporal modulations in speech and music can be quite similar (Ding et al., 2017), and cortical oscillations are similarly thought to support meter perception (Large and Palmer, 2002). Models based on hierarchically-coupled oscillations have been used to predict tapping behavior to musical stimuli (e.g. Large and Palmer, 2002; Repp, 2005), and ongoing cortical oscillations have been shown to synchronize with isochronous rhythms (Fujioka et al., 2012).

Neural correlates of segmentation within this statistical learning framework have also been demonstrated in the steady state response (SSR), which can be more sensitive to ongoing cortical oscillations than the ERP response. Unlike the ERP, which is analyzed in the time domain, the SSR is measured as a peak in the EEG spectrum corresponding to an isochronous pattern in the stimulus, under the assumption that a stimulus presented at a particular frequency will drive a response at that frequency, and this response can be summarized as a peak at that frequency. Presenting isochronous streams of nonsense syllables composed of discrete sequences, both Buiatti et al. (2009) and Batterink and Paller (2017) found an SSR corresponding to the sequence presentation rate, separate from the presentation rate of individual syllables, suggesting that cortical oscillations may entrain to segmentation boundaries in the absence of acoustic cues.

In an elegant extension of this work, Ding et al. (2016) demonstrated independent SSRs to multiple levels of a linguistic hierarchy. They presented a stream of monosyllabic words that combined to form hierarchically-nested two-word phrases and four-word sentences. They observed an SSR to both phrase and sentence presentation rates while simultaneously observing an SSR to the presentation rate of individual words. Scalp-recorded cortical responses have been repeatedly shown to follow the acoustic envelope (e.g. Ding and Simon, 2012; Doelling et al., 2014), and it is likely that the SSR following word boundaries reflects an acoustic envelope-following response, at least in large part. However, in the absence of any acoustic cues indicating phrasal and sentential boundaries, the SSRs following these linguistic boundaries must reflect an internal segmentation process. Critically, the fact that SSRs can be simultaneously recorded following multiple levels of the linguistic hierarchy suggests that these cortical responses may reflect the hierarchical segmentation of linguistic information.

While this kind of hierarchical segmentation may be specific to speech, hierarchical segmentation might also emerge from more generic grouping properties of the auditory system. Indeed, the ERP literature discussed above suggests that in the absence of semantic or syntactic value, speech sounds and non-speech sounds show similar neural signatures of segmentation. Nozaradan et al. (2011) found that when participants were asked to mentally segment a continuous stream of pure tones into discrete metrical units composed of two or three tones, an SSR was observed at the segmentation rate in addition to the rate at which individual tones were presented (see also, Nozaradan et al., 2012; Brochard et al., 2003). As with Ding et al. (2016), while the SSR following individual tones likely reflects an acoustic envelope-following response, the SSR following the metrical subdivision imposed by the listener must reflect an internal segmentation process.

An important difference between the result of Ding et al. (2016) and Nozaradan et al. (2011) is that Ding et al. reported an SSR to multiple levels of the linguistic hierarchy in addition an acoustic envelope-following response, while Nozaradan et al. reported an SSR to a single level of metrical organization in addition to an acoustic envelope-following response. The acoustic envelope-following response seems to be externally driven, and the extent to which it depends on linguistic features of the stimulus is unclear at best (Zoefel and VanRullen, 2016; Baltzell et al., 2017). Indeed, Ding et al. observed an SSR to the word presentation rate (but not the phrase or sentence presentation rates) even when the stimuli were presented in a foreign language. Alternatively, the metrical segmentation reported by Nozaradan et al. as well as the phrasal and sentential segmentation reported by Ding et al. were internally driven, depending entirely on non-acoustic features of the stimulus. Nozaradan et al. observed only a single “internally-driven” SSR then, while Ding et al. observed a pair of internally-driven SSRs at multiple timescales. Because the stimuli used by Ding et al. were isochronous, if only a single SSR was observed, it would be possible to interpret this SSR as reflecting a metrical rather than linguistic segmentation process.

1.2. Objectives

The primary goal of the present study was to construct a hierarchically-organized acoustic stimulus that was not linguistically meaningful, and to record simultaneous SSRs to multiple levels of the hierarchy. To do this, we focused on the hierarchical structure inherent in Western music, and constructed a set of sung nonsense syllables with a specific melodic/harmonic structure. This construction was designed to maximize the perception of structure while minimizing the acoustic differences between segments in the structure. While the mapping between linguistic structure and speech acoustics can be effectively arbitrary, the same is not true for music, where structure cannot be divorced from acoustic context (Lerdahl and Jackendoff, 1985; Jackendoff and

Lerdahl, 2006; Jackendoff, 2009).

To the extent that listeners segment these melodic stimuli according to their hierarchical structure, we expect to observe SSRs at segment boundaries. Furthermore, to the extent that musically trained participants have more experience with explicit harmonic listening (or harmonic analysis), we might expect stronger signatures of segmentation. However, expectancies based on passive exposure to Western music were expected to exist in all participants, regardless of musical training. To this end, neural signatures of harmonic expectancy violations have been observed for non-musicians (Koelsch et al., 2000), as have neural signatures of the segmentation of tonal sequences (discussed above), and we expected SSRs to segment boundaries to be present in participants with varying degrees of musical training.

To directly compare SSRs to hierarchically-organized melodic stimuli with SSRs to hierarchically-organized linguistic stimuli in the same participants, we replicated the design of Ding et al. (2016) with a modification that allowed parallels with the melodic stimuli. We also constructed stimuli with a hierarchical semantic rather than syntactic organization. We did this by constructing two separate pairs of semantically-related tri-syllabic words and combining these pairs in an alternating order. These hierarchically-organized semantic stimuli were constructed for two reasons. First, the semantic hierarchy we constructed was fundamentally different from the syntactic hierarchy governing the sentence stimuli. While syntactic hierarchies contain dependencies within segments such that the meaning of one element can depend on another (syntactic hierarchies are "headed"), the semantic couplets we constructed do not contain dependencies between the two words. In this sense they are more similar to metrical hierarchies than to melodic/harmonic (prolongational) hierarchies in music, since melodic/harmonic hierarchies are headed (Jackendoff, 2009). Towards the goal of characterizing which structured stimuli can induce SSRs to multiple structural levels, we wanted to examine non-syntactic in addition to syntactic hierarchies for linguistic stimuli. To the extent that participants segment tri-syllabic words into distinct pairs, we expect an SSR corresponding to the pair boundary. Second, while Ding et al. (2016) recorded an SSR to the presentation rate of individual words, word boundaries were present in the acoustic envelope, making it difficult to determine the extent to which the SSR was linguistically driven. We removed this confound by using tri-syllabic words, allowing us to measure the SSR to internally computed word boundaries (see Buiatti et al., 2009; Batterink and Paller, 2017). To the extent that participants are segmenting individual words from their constituent syllables, we expect an SSR corresponding to the word boundary.

2. Materials and methods

2.1. Participants

All experimental procedures were approved by the Institutional Review Board of the University of California, Irvine. Twenty English-speaking listeners (11 female; age range: 18–61; mean age = 26; st dev = 9.9) with normal hearing (thresholds ≤ 20 dB HL from 250 to 8000 Hz) participated in the study. Listeners were recruited via posters placed around the UC Irvine campus. Twelve of these listeners reported having musical training, and the other eight reported having no musical training. Of these twelve listeners with musical training, eight reported having formal theory training. No listeners reported having absolute pitch. Participants were not screened for handedness.

2.2. Speech stimuli

Following Ding et al. (2016), sentences were constructed that consisted of a noun phrase (NP) followed by a verb phrase (VP). Each phrase contained three monosyllabic words, and each sentence contained two phrases (six monosyllabic words). Individual words were presented isochronously at 3 Hz, which means that individual linguistic phrases were presented at 1 Hz, and sentences were presented at 0.5 Hz (Fig. 1a). Individual words were generated with the Klatt text-to-speech synthesizer in Praat (Version 6.0.36) using an American male talker. The mean fundamental frequency (f_0) for the sentence stimuli was 91.5 Hz, with a range of 75–104 Hz. A total of 14 sentences were generated for this experiment, and are listed in Table 1.

Semantic "couplets" were constructed that consisted of a semantically related pair of tri-syllabic words. Individual syllables were presented at 3 Hz, which means that tri-syllabic words were presented at 1 Hz, and semantic couplets were presented at 0.5 Hz (Fig. 1b). Individual syllables were generated with the same text-to-speech synthesizer used to generate the individual words of the sentence stimuli. The mean f_0 for the semantic couplet stimuli was 91 Hz. Two semantic couplets were constructed (four words), and these four words were selected based on their high frequency of use according to the Corpus of Contemporary American English (COCA), and the fact that they are content (rather than function) words. Furthermore, while an objective semantic similarity analysis was difficult due to the fact that each of the four words have multiple dictionary definitions, the words were chosen based on the experimenters' subjective impression that they form semantic pairs (Table 1). The decision to restrict this stimulus set to only four words reflected the relative difficulty of constructing these couplets, though it limits direct comparison to the sentence stimuli.

2.3. Melodic stimuli

Melodic "progressions" were constructed that consisted of a pair of ascending triads in a particular musical key. Each triad consisted of three individual nonsense syllables with pitches corresponding to a major

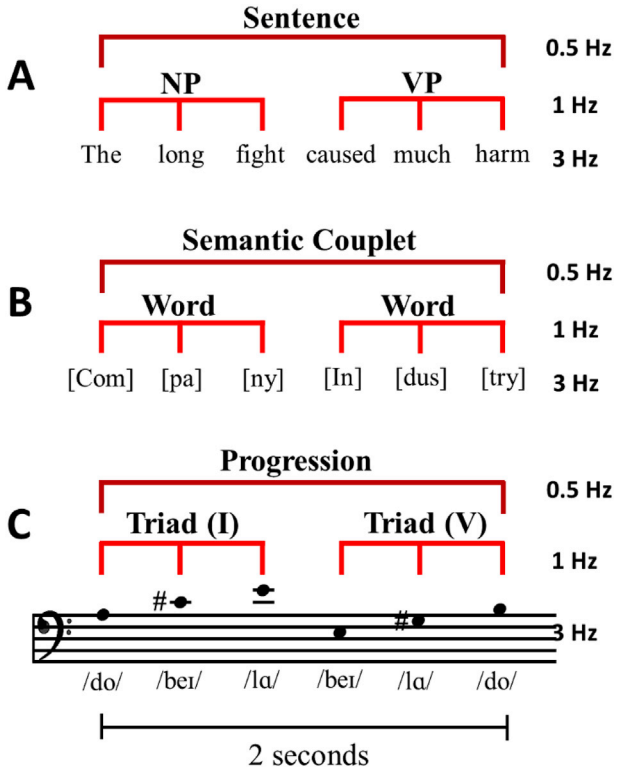
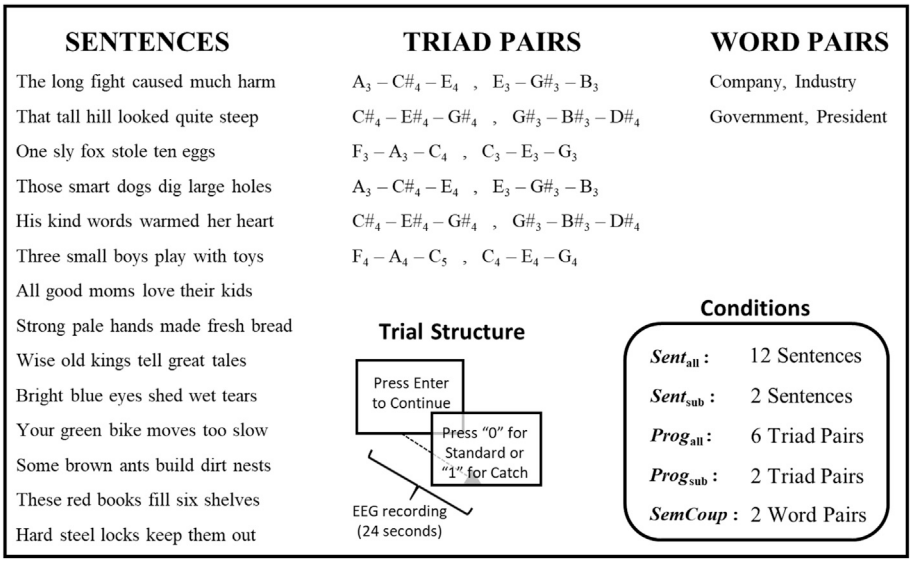


Fig. 1. Schematic of different stimulus types. (**A**) Individual words are presented at 3 Hz, and form a distinct noun phrase (NP) and verb phrase (VP) at 1 Hz. These 1-Hz phrases combine to form a complete sentence at 0.5 Hz. (**B**) Individual syllables are presented at 3 Hz, and form distinct words at 1 Hz. These two words are semantically related, and combine to form a semantic couplet at 0.5 Hz. (**C**) Individual sung syllables are presented at 3 Hz, and form distinct ascending triads at 1 Hz. These two triads combine to form a I–V chord progression at 0.5 Hz.

Table 1
List of stimuli used. A total of fourteen sentences were generated, with no words repeating. A sequence of six chord progressions was generated (subscript numbers refer to octave designation). Two word pairs were generated. The total number of novel stimuli included in each condition are indicated in the bottom right (see text for details).



triad. These "sung" syllables were generated with the same text-to-speech synthesizer used to generate the speech stimuli. The first triad in each pair consisted of the three pitches of the tonic (I) chord, while the second triad consisted of the three pitches of the dominant (V) chord, so each progression consisted of a total of six pitches. While these melodic stimuli consisted of sequentially presented sounds, they were designed to evoke an underlying harmonic structure. The word "harmonic", here and throughout, refers to the relationship between chords, rather than the relationship between components in the spectrum.

Each triad contained two intervals, an ascending major 3rd (four semitones) followed by an ascending minor 3rd (three semitones). The interval between the two triads within a progression was always a descending octave (12 semitones). Listeners therefore had two non-harmonic cues indicating the boundary between triads within a progression, in addition to the fact that each triad outlined a separate chord within the key. First, while the intervals within a triad were always ascending, the intervals between triads were always descending. Second, the size of the intervals within a triad were smaller than the intervals between triads.

Progressions were constructed in different keys and concatenated together such that the key of each progression was shifted up by a major 3rd relative to the previous progression. A major 3rd was chosen to (1) minimize the difference in mean pitch interval size within progressions and between progressions while (2) avoiding the tri-tone interval. According to this pattern, the original key is returned to every three key shifts. Since absolute pitch is extremely rare, even among highly-trained musicians (Deutsch et al., 2006), we did not control for the fact that progressions were repeated over the course of a trial. Furthermore, we chose a fixed-interval key shift instead of a random-interval key shift in order to avoid jarring or surprising transitions between progressions.

Pitch heights were constrained to fall within the normal vocal range for a male voice, resulting in the set of progressions shown in Table 1. While only six progressions are shown, two repetitions of this set were concatenated yielding a set of twelve progressions. Notice that for the six progressions shown, the pitch classes of sequences 4–6 are the same as in sequences 1–3, with the only difference between these sequences being found in the octave designation of the pitch classes.

Melodic stimuli consisted of individual "sung" nonsense syllables presented at 3 Hz, which means that individual triads were presented at 1 Hz, and progressions were presented at 0.5 Hz (Fig. 1c). The syllables/do/,/beɪ/, and/lɑ/were assigned such that the same syllable was never repeated sequentially.

2.4. Experimental design

The three stimulus types used in this experiment are shown in Fig. 1. To examine the effects of stimulus overexposure, and to establish a basis for comparison across all three stimulus types, "subset" versions of the sentence and progression stimuli were defined, yielding a total of five experimental conditions (Table 1, bottom right).

A total of 14 sentences were generated for this experiment, and are listed in Table 1 (left). In condition $Sent_{all}$, 12 of these sentences were selected (at random for each participant), and on each trial, were presented one after another in a (randomly) circularly-shifted order. Specifically, the initial order of the 12 sentences was random, and for each trial, the starting sentence was circularly shifted by a random number of sentences. This pseudo-randomization was done in order to match the $Prog_{all}$ condition, described in the subsequent paragraph. In condition $Sent_{sub}$, the remaining two sentences were selected, and on each trial, were presented six times in an alternating order (either ABAB … or BABA …), with the initial sentence being selected at random (either A or B). Following Ding et al. (2016), condition $Sent_{all}$ utilizes a large stimulus set such that we do not expect effects of stimulus overexposure. Condition $Sent_{sub}$ however, utilizes a limited stimulus set, allowing us to determine whether stimulus novelty has an effect on the SSR.

A total of six progressions were constructed for this experiment (Table 1, center). In condition $Prog_{all}$, two copies of these 6-pair sequences were concatenated to yield a 12-pair sequence. On each trial, the starting pair was circularly shifted by a random number of pairs so that the sequence was novel but the structure of the key shift was preserved. In condition $Prog_{sub}$, two adjacent triad pairs were randomly selected from the six available triad pairs and on each trial, these two triad pairs were each played six times in an alternating order with a random starting pair (either ABAB … or BABA …).

Two semantic couplets were constructed for this experiment (Table 1, top right). In condition *SemCoup*, these two semantic couplets were

presented six times in an alternating order with a random starting pair. For all conditions then, a trial contained twelve repetitions of the highest level of the hierarchy, and had a duration of 24 s. An example stimulus for each condition is provided as supplemental material.

2.5. Procedure

Each participant sat facing a computer monitor in a single-walled sound-attenuated booth with sound-treated interior walls. Stimuli were presented at 70 dB SPL over Final 500 electrostatic loudspeakers positioned at 45° and −45° degrees relative to the listener, at a distance of five feet from the listener's head. The order of the five experimental conditions (Table 1, bottom right) was randomized for each listener. Each condition contained 25 trials, and all conditions were tested in a single session. Each trial had a duration of 24 s, and contained 12 individual 2-s stimuli (sentences, word pairs, or triad pairs, see Fig. 1). Following Ding et al. (2016), listeners were instructed to detect catch trials. On a catch trial, the individual sound elements of a sentence/-word/triad were reversed in order (e.g. that tall hill looked quite steep → steep quite looked hill tall that, [in]-[dus]-[try] → [try]-[dus]-[in], E_3-G#$_3$-B_3 → B_3-G#$_3$-E_3, etc.). For sentence stimuli, a catch trial contained a single reversed sentence, while for word and triad stimuli, a catch trial contained two reversed words and two reversed triads, respectively. So as not to bias the participants towards grouping individual words into pairs (semantic couplets) and individual triads into pairs (progressions), the locations of two reversed words and triads were random, not sequential. At the end of each trial, listeners were asked to indicate whether or not it was a catch trial (Table 1, trial structure). For each condition five of the 25 trials were catch trails, which were excluded from the EEG analysis.

Prior to testing, listeners were asked to familiarize themselves with the stimuli on a self-paced GUI. As many of our listeners were unfamiliar with synthesized speech, we asked them to listen to the spoken speech materials (which were accompanied by text on the screen) until they were convinced they could understand the material without the aid of text. Listeners were also asked to familiarize themselves with the sung melodic stimuli, but were not directed to pay attention to the phonetic content.

2.6. EEG recording and Pre-processing

High-density EEG (128 channels) was recorded with equipment from Neuroscan. Electrodes were placed following the international 10/5 system (Oostenveld and Praamstra, 2001), and all channel impedances were kept below 10 kΩ. The EEG data was sampled at 1000 Hz, and filtered offline with a passband of .15–50 Hz. The filtered data were then segmented into individual trials which were 22 s long, beginning 2 s after the start of the sentence (yielding a frequency bin size of 1/22 Hz). This delay was incorporated to remove the onset response to the start of the stimulus. Artifacts were removed from the segmented EEG data using the Fast ICA algorithm (Hyvarinen and Oja, 1997).

Following Ding et al. (2016), the EEG responses were then denoised using the Denoising Source Separation (DSS) algorithm, which is a blind source separation technique that extracts neural response components that are consistent across trials (de Cheveigné and Simon, 2008). DSS computes a bias function based on the averaged neural data and applies a transformation to the unbiased (raw) neural data that maximizes this bias function. This bias function was computed across rather than within experimental conditions to avoid the artificial introduction of differences across these conditions.

2.7. Analysis

The denoised EEG data were analyzed in the frequency domain. For each participant, an average was taken over trials, and a Discrete Fourier Transform (DFT) was applied to the averaged data. To remove the 1/f trend in the denoised data, the magnitude of each Fourier coefficient was normalized by the median magnitude of neighboring coefficients while the phase spectrum was left intact (e.g. Srinivasan et al., 2006). Coefficient magnitudes at 0.5 Hz and above were normalized by the median magnitude of the coefficients at ± 11 bins (1/2 Hz). Coefficient magnitudes between 0.25 and 0.5 Hz were normalized by the median magnitude of the coefficients at ± 6 bins (1/4 Hz). Coefficients below 0.25 Hz were discarded. The median was used so that spectral extrema were not artificially boosted by the normalization procedure. Only broad trends in the spectrum were removed by this procedure, allowing us to optimize the response at each frequency before statistically evaluating the magnitude of individual frequency bins against neighboring bins.

To find optimal channel weights for each normalized frequency bin, a Singular Value Decomposition (SVD) was applied to local portions of the normalized (complex-valued) spectrum. These optimal channel weights can be thought of as a spatial filter that maximizes the EEG response at a particular frequency. A local portion was defined as 9 bins centered on the frequency bin of interest (±1/6 Hz). The local portion around each frequency bin was submitted to an SVD, and the first component was selected. The first component of the SVD reveals the spectrum that captures the most variance, along with the associated channel weights. Normalizing the spectrum as described above reduces the possibility that first component of the SVD captures a 1/f trend rather than a local spectral peak. The channel weights derived from the SVD refer to the relative distribution of activity across channels for a local frequency region. The magnitude of the Fourier coefficient at the bin of interest was normalized by the median of the magnitudes of the Fourier coefficients on either side (±1/6 Hz). The magnitude of the bin of interest was normalized to correct for unequal variance across SVD components across different portions of the spectrum.

2.8. Statistical analysis

This SVD procedure yielded a single magnitude spectrum with each frequency bin representing the summed activity over all channels with weights designed to optimally boost local activity at that frequency. Statistical analysis was performed on this optimized spectrum in order to determine which bins contained significant peaks (SSRs). Following Ding et al. (2016), one-tailed, two-sample t-tests were used to test whether the cortical response in a frequency bin was significantly stronger than the average of the neighboring four frequency bins (two bins on either side). A one-tailed *t*-test was performed for each frequency bin between 0.25 and 3.5 Hz, and an FDR correction for multiple comparisons (Benjamini and Hochberg, 1995) was applied ($\alpha = 0.01$).

In order to examine effects of musical training, a mixed effect model was fit to the data treating musical training (yes/no), condition ($Sent_{all}$, $Sent_{sub}$, *SemCoup*, $Prog_{all}$, $Prog_{sub}$), and frequency (0.5, 1, 3 Hz) as fixed effects and treating participant as a random effect. A priori expectations motivated the exclusive inclusion of the 0.5, 1, and 3 Hz frequency bins, and the design of this model was not contingent on the result of the SSR statistical analysis described in the previous paragraph. Since a Kolmogorov-Smirnov test revealed that the distribution of ages was not normal ($p < 0.001$), age was not included as a fixed effect.

3. Results

3.1. Behavior

To encourage listeners to remain attentive throughout the experiment, 1/5 of the trials in each condition were catch trials, and after each trial, listeners indicated whether they thought the trial was a standard or catch trial (see section 2.5). Percent correct responses for standard and catch trials are shown for each condition in Fig. 2. Listeners identified standard and catch trials above chance and with a high degree of accuracy, indicating that they were attending to the stimuli. Chance-level response rates were computed based on a hypothetical probability

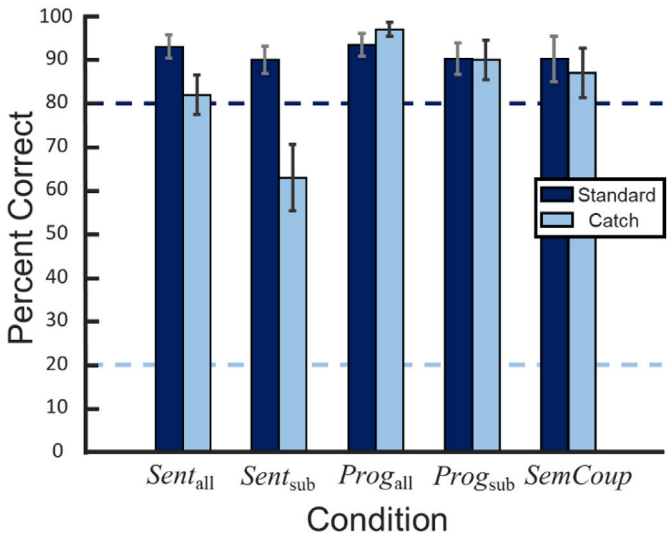


Fig. 2. Percent correct responses for each experimental condition, shown separately for standard and catch trials. Chance performance based on a probability-matching observer for standard (dark blue) and catch (light blue) trials are indicated with dotted lines.

matching observer who could not identify individual catch trials but had learned that 20% of trials were catch trials. Such an observer would randomly indicate "standard" on 20/25 trials and "catch" on 5/25 trials, and with this strategy would correctly identify 80% of standard trials and 20% of catch trials. It was somewhat surprising that performance for condition $Sent_{sub}$ was worse than condition $Sent_{all}$, since condition $Sent_{all}$ offered a more restrictive stimulus set, and may suggest that listeners may not have been attending as closely to the stimuli in $Sent_{sub}$ relative to $Sent_{all}$. Due to the differences in stimuli across conditions, and the fact that task difficulty was not controlled for, a statistical comparison of the behavioral data across conditions was not pursued.

3.2. Stimulus envelopes

Across conditions, individual sounds were presented at a rate of 3 Hz, but were hierarchically organized such that individual segments were presented at 1 Hz and 0.5 Hz. Since cortical entrainment to the acoustic envelope in the delta/theta range is known to be robust (Ding and Simon, 2012; Horton et al., 2013; Doelling et al., 2014), a potential concern is that 1-Hz and 0.5-Hz energy in the stimulus envelope could generate a cortical response at these frequencies. For each condition, 500 (25 trials x 20 participants) unique trial stimuli (24-s duration) were constructed, a Hilbert transform was applied, and the magnitude of the analytic signal was computed to extract the envelope from each example stimulus. The DFT of the stimulus envelope of each stimulus was taken, and magnitudes were averaged over examples, yielding average stimulus modulation spectra for each condition, shown in Fig. 3A.

The modulation spectra provide predictions for the entrainment response to the acoustic envelope. Across conditions there is a large peak at 3 Hz, which corresponds to the rate at which sounds were presented (Fig. 1). Also, the modulation spectra of $Sent_{all}$ and $Prog_{all}$ are flatter, respectively, then $Sent_{sub}$ and $Prog_{sub}$. This is because the longer the overall period, the smaller the fundamental frequency, and energy at the sidebands of the 3-Hz peak will be distributed to a greater number of bins.

Bins in the modulation spectrum were evaluated statistically following the same SSR procedure used to analyze the EEG response spectrum (section 2.8), and the analysis was similarly restricted to bins between 0.25 and 3.5 Hz. Shown in Fig. 3A, all visible peaks in the modulation spectra across conditions are significant. Importantly though, in no conditions is the energy at 0.5 and 1 Hz noticeably greater than the energy at other sideband peaks. This suggests that if a significant SSR is observed at these two frequencies (0.5 and 1 Hz), but not at other frequencies with greater or similar energy in the modulation spectrum, the responses cannot be interpreted as simply reflecting an entrainment response to the acoustic envelope.

In order to quantify the cortical response to significant sideband peaks in the stimulus modulation spectra, we sorted grand-averaged cortical responses (Fig. 4A, thick line) into three groups, pooled across conditions. The first group contained cortical responses at 0.5 and 1 Hz corresponding to significant peaks in the modulations spectrum. In other words, all significant peaks at 0.5 and 1 Hz were identified, across conditions, and the cortical responses corresponding to these peaks formed the first group. The second group contained cortical responses corresponding to all *other* significant peaks in the modulation spectra, with the exclusion of 3 Hz. The third group contained the cortical responses corresponding to all frequency bins that did not contain a significant peak in the modulation spectra. To the extent that the cortical responses at 0.5 and 1 Hz are driven by the acoustic envelope, we expect the response magnitudes in the first group to resemble the response magnitudes in second group. Furthermore, to the extent that the sideband peaks are *not* driving a cortical response, we expect the magnitudes in the second

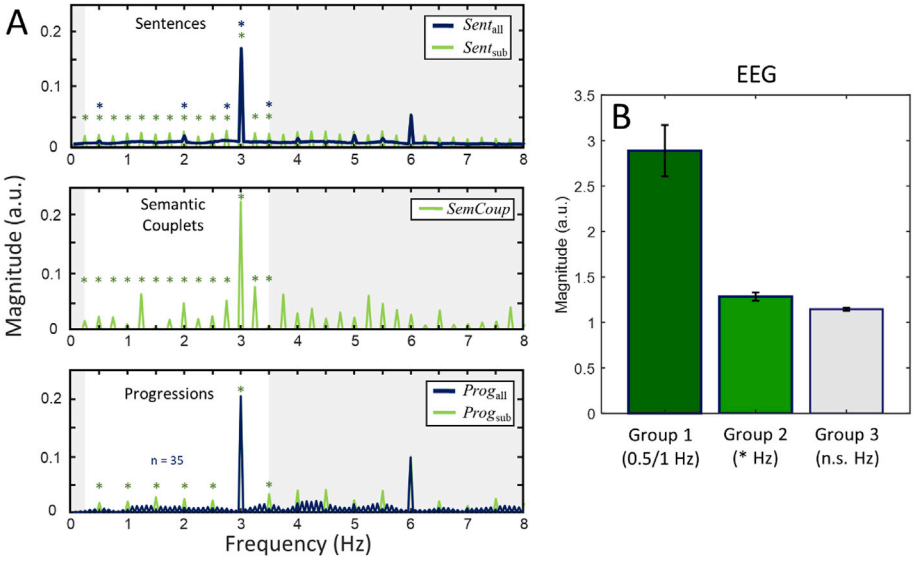


Fig. 3. (**A**) Modulation spectra of the stimuli in different conditions. Modulation spectra were computed by taking the DFT of the absolute value of the Hilbert transform of example trial stimuli. Significant peaks in the 0.25–3.5 Hz range are indicated by asterisks, except in the $Prog_{all}$ condition, where 35 individual peaks were significant. (**B**) Magnitude of cortical responses sorted according to the stimulus modulation spectra (0.25–3.5 Hz), pooled across conditions. The left bar indicates the mean cortical response corresponding to significant peaks in the modulation spectra at 0.5 and 1 Hz. The middle bar indicates the mean cortical response corresponding to all other significant peaks (excluding 3 Hz). The right bar indicates the mean cortical response corresponding to all frequency bins that did not contain a significant peak. Error bars indicate standard error of the mean.

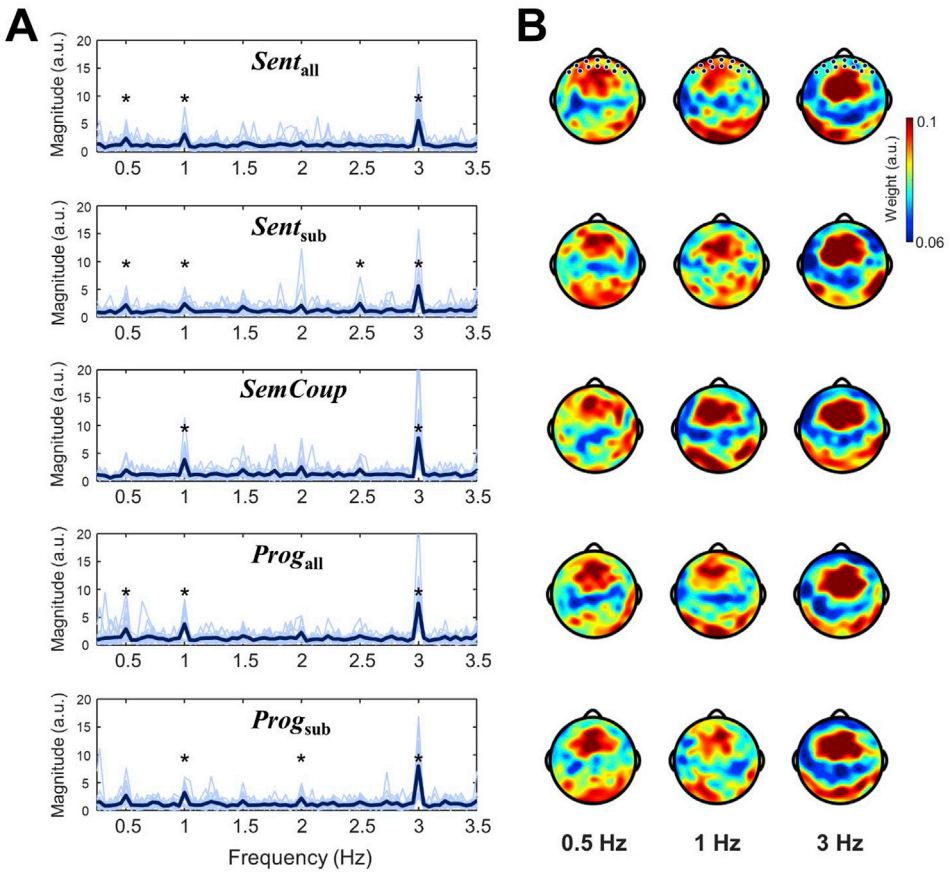


Fig. 4. Weighted average of cortical responses across EEG sensors. (**A**) Cortical response spectra for the different conditions. Individual participants are shown in light blue traces, and the grand average is shown in dark blue. Cortical encoding of hierarchical sequences in the stimulus are reflected in spectral peaks. Frequency bins with significantly (FDR corrected) more energy than neighboring bins (±2 bins) are marked with an asterisk. (**B**) Scalp-topographic maps of channel weights derived from the SVD procedure. Channel weights can be interpreted as the relative cortical activity across channels associated with each local frequency region centered at 0.5 Hz, 1 Hz, and 3 Hz. Channels included in the frontal ROI analysis (Fig. 5) are indicated with blue dots corresponding to electrode locations (top row).

group to resemble the response magnitudes in the third group. Shown in Fig. 3B, we see that response magnitudes in the first group are dramatically larger than the response magnitudes in the second group, suggesting that the responses at 0.5 and 1 Hz are not reducible to an envelope following response. We also see that the response magnitudes in the second group are only marginally larger than the response magnitudes in the third group, suggesting that for the most part, the sidebands are not driving a cortical response.

3.3. Cortical encoding of hierarchical boundaries

Based on the hierarchical structure of our stimuli (Fig. 1), we expected to observe cortical encoding of the stimulus at three distinct frequencies, corresponding to this structure. The cortical response spectrum was analyzed from 0.25 to 3.5 Hz for each condition (72 bins x 5 conditions), and an FDR correction for multiple comparisons was applied across conditions with a cutoff $\alpha = 0.01$ (yielding a p-value criterion of 0.0004).

Across conditions, we expected an SSR at 3 Hz, corresponding to the rate at which sounds were presented. Shown in Fig. 4a, a significant response at 3 Hz was observed for all conditions. For conditions $Sent_{all}$ and $Sent_{sub}$, we expect an SSR at 1 Hz, corresponding to the phrase presentation rate, and at 0.5 Hz, corresponding to the sentence presentation rate. We observed a significant response at both 1 Hz and 0.5 Hz, for both $Sent_{all}$ and $Sent_{sub}$. For conditions $Prog_{all}$ and $Prog_{sub}$, we expected an SSR at 1 Hz, corresponding to the triad presentation rate, and at 0.5 Hz, corresponding to the progression presentation rate. We observed a significant response at both 1 Hz and 0.5 Hz for $Prog_{all}$. We observed a significant response at 1 Hz for $Prog_{sub}$, but did not observe a significant response at 0.5 Hz ($p = 0.0019$). For condition *SemCoup*, we expected an SSR at 1 Hz, corresponding to the word presentation rate, and at 0.5 Hz, corresponding to the semantic couplet presentation rate. We observed a significant response at 1 Hz, but did not observe a significant response at 0.5 Hz ($p = 0.0021$).

We also observe two significant SSRs at frequencies that do not correspond to hierarchical structures in the stimulus; at 2 Hz in the $Prog_{sub}$ condition, and at 2.5 Hz in the $Sent_{sub}$ condition. Considering that there isn't substantial energy in the acoustic envelope at these frequencies, it is unlikely that they reflect an entrainment response to the acoustic envelope. Instead, these responses likely reflect distortion products generated by the cortical encoding of the hierarchical stimuli.

Channel weights derived from the SVD are shown in Fig. 4b for 0.5 Hz, 1 Hz, and 3 Hz frequency bins. These weights represent the relative activity across the scalp that contributed to the optimized response at a particular frequency. These weight patterns are broadly consistent with scalp topographies of an entrainment response to the acoustic envelope measured with EEG (Nozaradan et al., 2012; Baltzell et al., 2017), a response typified by a dominant fronto-central to posterior dipole. Seen in Fig. 3b, this pattern is clearer at 3 Hz than 1 Hz and 0.5 Hz, suggesting that while the 3 Hz response likely reflects an entrainment response to the acoustic envelope, the 1 Hz and 0.5 Hz response may reflect contributions from additional sources.

Since frontal sources are more active during the processing of linguistic information as opposed to strictly acoustic information (DéMonet et al., 1992; for review, see Friederici, 2002), we anticipated differences in frontal channel weights across frequencies (0.5 Hz, 1, Hz, 3 Hz). In an attempt to focus on frontal sources, we defined a frontal ROI (channels included in this ROI are shown in the top row of topographic maps in Fig. 4b) so as not to overlap with the main fronto-central to posterior dipole based on visual inspection. Since the optimized topographies we are analyzing contain relative weights rather than scalp voltages, source analysis was not pursued. Our analysis of the frontal ROI can therefore

only determine if the relative activity of frontal electrodes (included in the ROI) was different across frequencies. Relative weights in this frontal ROI is shown in Fig. 5. A mixed-effects model was fit treating Frequency and Condition as fixed effects and treating participant as a random effect. An ANOVA revealed a significant main effect of Frequency ($p = 0.033$) and a significant main effect of Condition ($p < 0.001$). The interaction was not significant ($p = 0.26$).

Post-hoc planned constrasts of Frequency revealed a significant (Bonferroni corrected) difference between 0.5 Hz and 3 Hz ($p < 0.001$) and between 1 Hz and 3 Hz ($p < 0.001$), but not between 0.5 Hz and 1 Hz ($p = 0.64$). In other words, relative frontal activity is significantly higher at 0.5 Hz and 1 Hz than at 3 Hz (Fig. 5a). Post-hoc contrasts of Condition only revealed a significant difference between $Sent_{all}$ and $Prog_{sub}$. Since stimulus overexposure may have created unequal attentional demands, we compared relative frontal activity across full stimulus set conditions $Sent_{all}$ and $Prog_{all}$ to relative frontal activity across stimulus subset conditions $Sent_{sub}$, $Prog_{sub}$, and *SemCoup*. Shown in Fig. 5b, this comparison (paired *t*-test) was not significant ($p = 0.37$).

3.4. Effects of musical training

In order to examine effects of musical training, a mixed effect model was fit to the data treating musical training, condition, and frequency (0.5, 1, 3 Hz) as fixed effects and treating participant as a random effect. An ANOVA revealed a significant main effect of frequency ($p < 0.001$) and condition ($p = 0.035$), but the main effect of musical training was not significant ($p = 0.38$). Furthermore, there were no significant interactions (all $p > 0.35$), suggesting that musical training did not significantly modulate cortical responses in any condition or frequency bin.

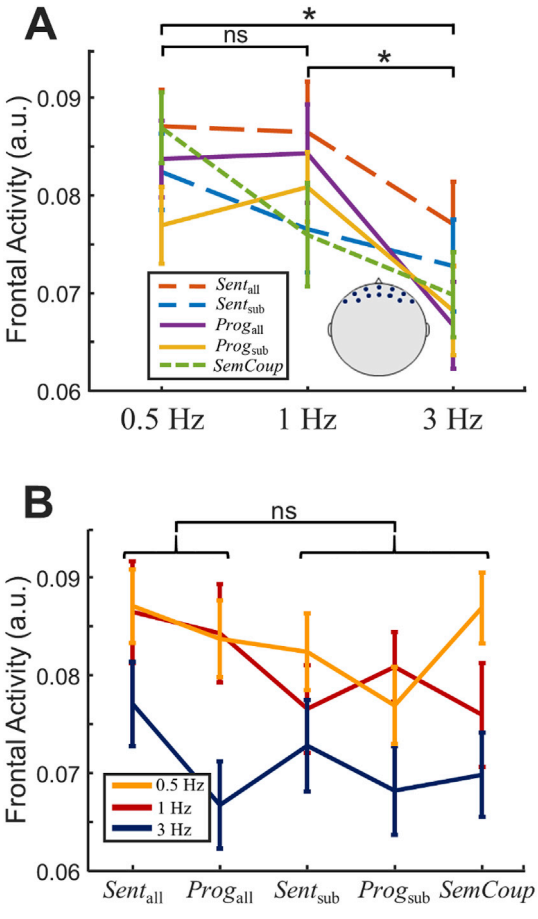


Fig. 5. (**A**) Activity in the frontal ROI (blue dots on gray scalp diagram) for each conditions as a function of frequency. Error bars refer to the standard error of the mean. Asterisks indicate significant post-hoc paired comparisons. (**B**) The same data from (A) but shown as a function of condition. Error bars refer to the standard error of the mean. The comparison between the *all* conditions ($Sent_{all}$ and $Prog_{all}$) and the *sub* conditions ($Sent_{sub}$, $Prog_{sub}$, *SemCoup*) was not significant.

3.5. Individual differences in cortical responses to musical hierarchies

In the $Prog_{all}$ condition, we observed a significant SSR at 1 Hz and at 0.5 Hz, suggesting that particiapants simultaneously encoded the melodic stimuli at multiple timescales. Crucially, it was not simply the case that this pattern in the group average reflected the fact that some participants showed a large response at 1 Hz, while the rest showed a large response at 0.5 Hz. In Fig. 6, we show the cortical responses of three representative participants. One participant shows a relatively larger response at 0.5 Hz, one participant shows a relatively larger response at 1 Hz, and one participant shows an equally large response at 0.5 Hz and 1 Hz. However, all of these participants show a response at both 0.5 Hz and 1 Hz.

4. Discussion

Our results show that cortical activity simultaneously encodes melodic structure at multiple temporal scales corresponding to its hierarchical organization, paralleling the hierarchical processing of syntactic structure in speech. This suggests that cortical encoding of hierarchical structure may not be unique to speech.

4.1. Cortical responses to speech

Cortical responses to speech were measured in three conditions: $Sent_{all}$, $Sent_{sub}$, and *SemCoup* (Table 1). In condition $Sent_{all}$, using a relatively large set of sentences (n = 12), we found that cortical responses encode hierarchically-organized changes in linguistic structure (phrases and sentences), replicating the main result of Ding et al. (2016). In condition $Sent_{sub}$, using only a pair of sentences, we found a similar result, suggesting that overexposure to the stimulus does not eliminate the cortical response observed in $Sent_{all}$. In condition *SemCoup*, using a pair of semantically coupled tri-syllabic words, we observed a significant response to the word presentation rate. While Ding et al. (2016) found a significant response to the word presentation rate, it was unclear whether this response reflected an acoustic envelope-following response or an internal computation of word boundaries. Here, we provide evidence of cortical encoding of internally computed word boundaries.

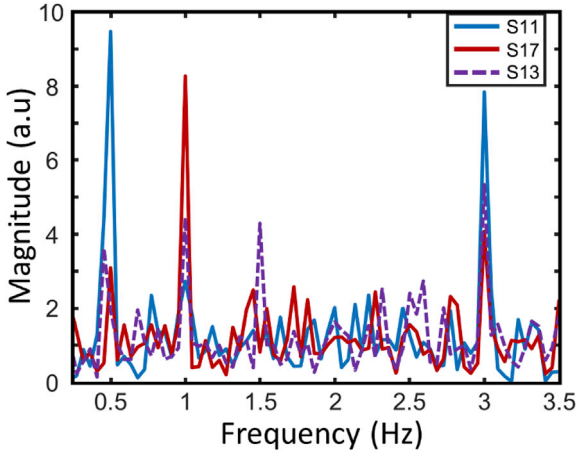


Fig. 6. Cortical responses from the $Prog_{all}$ condition for three representative participants. In general participants show a response at 0.5 Hz and 1 Hz, though some participants show a relatively larger response at 0.5 Hz or at 1 Hz.

4.2. Cortical responses to melodies

Cortical responses to melodies were measured in two conditions: $Prog_{all}$ and $Prog_{sub}$ (Table 1). In condition $Prog_{all}$, melodies were constructed based on a repeating 6-note sequence, composed of two ascending triads, and outlining a tonic-dominant progression in a particular musical key. The key of each progression shifted throughout the melody, giving rise to a hierarchical melodic structure based on chord changes and key changes (Fig. 1). We observed a significant response to the presentation rate of individual triads and to the presentation rate of individual progressions, suggesting that listeners simultaneously segment triads into discrete melodic sections, and segment pairs of triads into discrete melodic sections. However, in condition $Prog_{sub}$, where the same two progressions were repeated over the course of a trial, we observed only a marginally significant response to the progression presentation rate, suggesting that the cues for grouping progressions are not as robust with a limited set of repeated progressions.

While the boundaries between individual triads and individual progressions were not contained in the acoustic envelope (Fig. 3), changes in internally-computed pitch may have contributed to the cortical encoding of these boundaries. Because the melodic stimuli contained regular pitch changes, it is possible that in the absence of any harmonic (chordal) understanding of the relationship between pitches, listeners may have parsed the melodic sequences based on modulations in pitch. Shown in Fig. 7, there are observable peaks at both 1 Hz and 0.5 Hz in the f_0 modulation spectrum.

While prosody can certainly have an effect linguistic meaning in conversational contexts, linguistic meaning can, for the most part, be transmitted to the listener irrespective of changes in pitch. Musical "meaning", on the other hand, essentially depends on pitch. Patterson et al. (2002) showed differential activation in auditory cortex to diatonic compared to random melodic sequences consistent with hierarchical processing of musical pitch, suggesting that in the context of music, pitch is not simply a low-level sound feature computed peripherally. While it is possible to construct melodic stimuli without regular changes in pitch, it is difficult to evoke regular harmonic changes without correspondingly regular changes in pitch. Further studies are required to tease apart the contribution of harmonic (chordal) cues and pitch cues to the hierarchical cortical responses we observed.

Effects of musical training on the cortical responses to melodic sequences were not observed. While it may be the case that with more participants an effect of musical training could be observed, our data do not support the conclusion that neural signatures of melodic segmentation, at least for the stimuli used in the present study, are stronger for individuals with musical training compared to individuals passively exposed to Western music. While Doelling and Poeppel (2015) observed a significant effect of musical expertise on entrainment to acoustic envelope, their stimuli consisted of natural piano recordings, which may have differentially recruited attention and/or musical knowledge for those listeners with greater expertise.

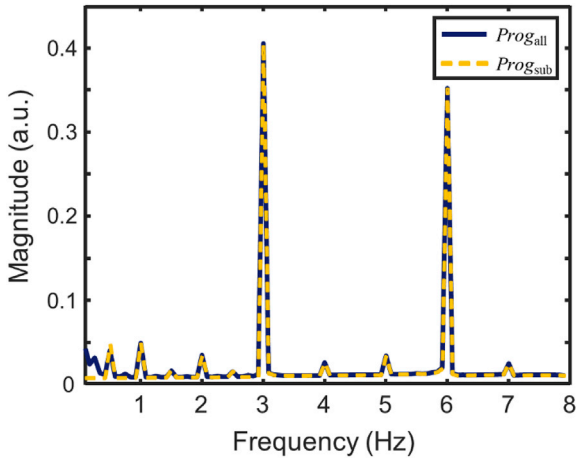


Fig. 7. Pitch modulation spectrum for melodic stimuli. The fundamental frequency (f_0) was extracted from all variations of the $Prog_{all}$ and $Prog_{sub}$ stimuli used in the experiment. F_0 values were log-transformed and the DFT was computed over these values. The magnitude spectrum is shown here, and we can see distinct peaks at 3 Hz, 1 Hz, and 0.5 Hz, corresponding to the regular pitch changes in our melodic stimuli.

4.3. Metrical organization of isochronous stimuli

Nozaradan et al. (2012) showed that listeners can impose spontaneous metrical organization (or segmentation) to repeating sequences, and that this organization is evidenced by an SSR. It is therefore possible that listeners imposed a metrical organization on the isochronous stimuli used in this experiment, although the imposing of such an organization would not predict simultaneous encoding at multiple time-scales to the hierarchical structure of the stimuli. However, while music is typically organized according to some strict metrical structure, speech is typically not, and it is possible that the mechanisms involved in the processing of isochronous speech may be partially distinct from those involved in the processing of natural speech.

4.4. Relative frontal activity

Across conditions, the relative activity at each frequency mainly resembles the potential distribution due to a dipole source in auditory cortex that typifies the entrainment response to the acoustic envelope (Fig. 4b) as well as the N1 auditory evoked potential (Giard et al., 1994). However, relative frontal activity is significantly higher at frequencies related to linguistic and musical sequences (0.5 Hz and 1 Hz) than at the 3-Hz stimulus envelope rate (Fig. 4a). This is consistent with the fact that frontal areas are recruited for the processing of linguistic and musical information rather than acoustic information (Friederici, 2002; Koelsch et al., 2002), and suggests that these regions may be relatively more active at frequencies related to linguistic/musical sequences. Interestingly, the scalp topography at 1 Hz in the *SemCoup* condition (corresponding to the presentation rate of individual words) strongly resembles the topography at 3 Hz in the $Sent_{all}$ and $Sent_{sub}$ conditions (Fig. 4b), consistent with the suggestion that the higher-order linguistic computations of phrases and sentences recruit frontal sources more heavily than lower-order linguistic computation of words (Friederici, 2003; Just et al., 1996).

However, strong conclusions cannot be made here about the involvement of frontal sources for three reasons. First, because our analysis focused on the relative channel weights corresponding to the optimized spectrum, source analysis was not possible, so it may be the case that non-frontal sources contributed to the activity in the frontal ROI we defined. Second, because the weights are relative, an increased activity in frontal cortex cannot be distinguished from a decreased activity in auditory cortex, rendering it impossible to conclude that frontal sources are more active at 0.5 Hz and 1 Hz compared to 3 Hz. Lastly, the spatial resolution of EEG is relatively poor, especially when searching for secondary neural sources, and more sensitive methods are required to make strong claims about the loci of neural sources.

4.5. Cortical distortion products

We observed significant cortical responses at frequencies unrelated to sequences in the stimulus (Fig. 4a). These responses are not straightforwardly attributable to an entrainment response to the acoustic envelope. Instead, they likely reflect intermodulation distortion products generated by cortical sources that process the stimulus at multiple time scales. The fact that these responses tend to appear at 0.5 Hz intervals is consistent with this interpretation, as this pattern could emerge from a cortical source that is oscillating at both 0.5 Hz and 1 Hz (or 0.5 Hz and 3 Hz).

4.6. Hierarchical vs. simultaneous encoding of boundaries

Throughout this discussion, we have suggested that the simultaneous cortical encoding of multiple boundaries in the stimulus implies hierarchical encoding. For speech stimuli, the implication is that in order for word boundaries to be meaningful, individual syllables must be parsed, and in order for phrase boundaries to be meaningful, individual words must be parsed, etc. For the melodic stimuli, the implication is that in order for triad boundaries to be meaningful, the individual pitches must be extracted and the relationship between the pitches computed, and in order for the progression boundaries to be meaningful, the relationship between the pairs of triads must be computed.

In this sense, hierarchical refers more explicitly to the stimulus rather to the cortical response, and our data do not explicitly suggest that the simultaneous cortical responses we observe are hierarchically coupled.

4.7. Limitations

There are a number of limitations in the current study that could be improved upon in future work. First, despite the fact that we did not observe a significant effect of musical training on the cortical responses to our melodic stimuli, it is possible that significant effects could be observed with more subjects and/or a more systematic quantification of musical training. Second, because the age distribution in our sample was not normally distributed, the effect of age was not considered in our analysis. However, it is possible that significant effects could be observed if both older and younger listeners were explicitly recruited. Third, the fact that the repetition rates of our structural boundaries were harmonically related raises the possibility of cortical distortion products confounding our results. While this is a natural consequence of building hierarchical isochronous stimuli, it is a consequence worth considering carefully, and the inclusion of control conditions where the hierarchical structure was removed would help mitigate (though not entirely alleviate) these concerns. Lastly, the fact that peaks in the f_0 modulation spectrum of the melodic stimuli were observed at 0.5 and 1 Hz raises the possibility that the encoding of pitch, rather than melodic structure per se, is contributing to the cortical response. It is perhaps possible to construct hierarchically-organized melodic stimuli whose structure remains perceptually salient while also eliminating peaks in the f_0 modulation spectrum corresponding to that structure.

4.8. Conclusions

Previous literature has revealed cortical entrainment to internally computed structural boundaries in both speech and music. While cortical entrainment to hierarchically-organized structures have been demonstrated for speech, we show entrainment to hierarchically-organized melodic structures in music. The fact that this response can be observed across domains suggests that hierarchical cortical entrainment may reflect a general property of the auditory system, or at least one that is not unique to speech.

Further studies are required to understand the neural mechanisms that underlie the cortical responses we observe, and the extent to which these mechanisms are specific to speech and music.

Acknowledgements

Research reported in this publication was supported by National Institute of Mental Health 2RO1-MH-68004 and National Institutes of Health NIDCD R21 DC013406.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2019.06.054.

References

Abla, D., Katahira, K., Okanoya, K., 2008. On-line assessment of statistical learning by event-related potentials. J. Cogn. Neurosci. 20 (6), 952–964.
Baltzell, L.S., Srinivasan, R., Richards, V.M., 2017. The effect of prior knowledge and intelligibility on the cortical entrainment response to speech. J. Neurophysiol. 118 (6), 3144–3151.
Batterink, L.J., Paller, K.A., 2017. Online neural monitoring of statistical learning. Cortex 90, 31–45.
Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Ser. B 57 (1), 289–300.
Brochard, R., Abecasis, D., Potter, D., Ragot, R., Drake, C., 2003. The "Ticktock" of our internal clock: direct brain evidence of subjective accents in isochronous sequences. Psychol. Sci. 14 (4), 362–366.
Buiatti, M., Pena, M., Dehaenelambertz, G., 2009. Investigating the neural correlates of continuous speech computation with frequency-tagged neuroelectric responses. Neuroimage 44 (2), 509–519.
Buzsaki, G., Draguhn, A., 2004. Neuronal oscillations in cortical networks. Science 304 (5679), 1926–1929.
Canolty, R.T., Edwards, E., Dalal, S.S., Soltani, M., Nagarajan, S.S., Kirsch, H.E., Knight, R.T., 2006. High gamma power is phase-locked to theta oscillations in human neocortex. Science 313 (5793), 1626–1628.
Cunillera, T., Càmara, E., Toro, J.M., Marco-Pallares, J., Sebastián-Galles, N., Ortiz, H., Rodríguez-Fornells, A., 2009. Time course and functional neuroanatomy of speech segmentation in adults. Neuroimage 48 (3), 541–553.
de Cheveigné, A., Simon, J.Z., 2008. Denoising based on spatial filtering. J. Neurosci. Methods 171 (2), 331–339.
De Diego Balaguer, R., Toro, J.M., Rodriguez-Fornells, A., Bachoud-Lévi, A.-C., 2007. Different neurophysiological mechanisms underlying word and rule extraction from speech. PLoS One 2 (11), e1175.
DéMonet, J.-F., Chollet, F., Ramsay, S., Cardebat, D., Nespoulous, J.-L., Wise, R., Frackowiak, R., 1992. The anatomy of phonological and semantic processing in normal subjects. Brain 115 (6), 1753–1768.
Deutsch, D., Henthorn, T., Marvin, E., Xu, H., 2006. Absolute pitch among American and Chinese conservatory students: prevalence differences, and evidence for a speech-related critical period. J. Acoust. Soc. Am. 119 (2), 719.
Ding, N., Simon, J.Z., 2012. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J. Neurophysiol. 107 (1), 78–89.
Ding, N., Melloni, L., Zhang, H., Tian, X., Poeppel, D., 2016. Cortical tracking of hierarchical linguistic structures in connected speech. Nat. Neurosci. 19 (1), 158–164.
Ding, N., Patel, A.D., Chen, L., Butler, H., Luo, C., Poeppel, D., 2017. Temporal modulations in speech and music. Neurosci. Biobehav. Rev. 81, 181–187.
Doelling, K.B., Arnal, L.H., Ghitza, O., Poeppel, D., 2014. Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. Neuroimage 85, 761–768.
Doelling, K.B., Poeppel, D., 2015. Cortical entrainment to music and its modulation by expertise. Proc. Natl. Acad. Sci. Unit. States Am. 112 (45), E6233–E6242.
Francois, C., Schön, D., 2010. Learning of musical and linguistic structures: comparing event-related potentials and behavior. Neuroreport 21 (14), 928–932.
Friederici, A.D., 2002. Towards a neural basis of auditory sentence processing. Trends Cognit. Sci. 6 (2), 78–84.
Friederici, A.D., 2003. The role of left inferior frontal and superior temporal cortex in sentence comprehension: localizing syntactic and semantic processes. Cerebr. Cortex 13 (2), 170–177.
Fujioka, T., Trainor, L.J., Large, E.W., Ross, B., 2012. Internalized timing of isochronous sounds is represented in neuromagnetic beta oscillations. J. Neurosci. 32 (5), 1791–1802.
Ghitza, O., 2011. Linking speech perception and neurophysiology: speech decoding guided by cascaded oscillators locked to the input rhythm. Front. Psychol. 2.
Giard, M.H., Perrin, F., Echallier, J.F., Thévenet, M., Froment, J.C., Pernier, J., 1994. Dissociation of temporal and frontal components in the human auditory N1 wave: a scalp current density and dipole model analysis. Electroencephalogr. Clin. Neurophysiology Evoked Potentials Sect. 92 (3), 238–252.
Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T.E., Frackowiak, R.S.J., Laufs, H., 2007. Endogenous cortical rhythms determine cerebral specialization for speech perception and production. Neuron 56 (6), 1127–1134.
Giraud, A.-L., Poeppel, D., 2012. Cortical oscillations and speech processing: emerging computational principles and operations. Nat. Neurosci. 15 (4), 511–517.
Horton, C., D'Zmura, M., Srinivasan, R., 2013. Suppression of competing speech through entrainment of cortical oscillations. J. Neurophysiol. 109 (12), 3082–3093.
Hyvarinen, A., Oja, E., 1997. A fast fixed-point algorithm for independent component analysis. Neural Comput. 9, 1483–1492.
Jackendoff, R., Lerdahl, F., 2006. The capacity for music: what is it, and what's special about it? Cognition 100 (1), 33–72.
Jackendoff, R., 2009. Parallels and nonparallels between language and music. Music Percept. 26 (3), 195–204.
Just, M.A., Carpenter, P.A., Keller, T.A., Eddy, W.F., Thulborn, K.R., 1996. Brain activation modulated by sentence comprehension. Science 274 (5284), 114–116.
Koelsch, S., Gunter, T., Friederici, A.D., Schröger, E., 2000. Brain indices of music processing: "nonmusicians" are musical. J. Cogn. Neurosci. 12 (3), 520–541.
Koelsch, S., Gunter, T.C., v.Cramon, D.Y., Zysset, S., Lohmann, G., Friederici, A.D., 2002. Bach speaks: a cortical "Language-Network" serves the processing of music. Neuroimage 17 (2), 956–966.

Lakatos, P., Shah, A.S., Knuth, K.H., Ulbert, I., Karmos, G., Schroeder, C.E., 2005. An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. J. Neurophysiol. 94, 1904–1911.
Large, E.W., Palmer, C., 2002. Perceiving temporal regularity in music. Cogn. Sci. 26, 1–37.
Lerdahl, F., Jackendoff, R., 1985. A Generative Theory of Tonal Music. MIT Press, Cambridge, MA.
Nozaradan, S., Peretz, I., Missal, M., Mouraux, A., 2011. Tagging the neuronal entrainment to beat and meter. J. Neurosci. 31 (28), 10234–10240.
Nozaradan, S., Peretz, I., Mouraux, A., 2012. Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. J. Neurosci. 32 (49), 17572–17581.
Oostenveld, R., Praamstra, P., 2001. The five percent electrode system for high-resolution EEG and ERP measurements. Clin. Neurophysiol. 112 (4), 713–719.
Patterson, R.D., Uppenkamp, S., Johnsrude, I.S., Griffiths, T.D., 2002. The processing of temporal pitch and melody information in auditory cortex. Neuron 36 (4), 767–776.
Repp, B.H., 2005. Sensorimotor synchronization: a review of the tapping literature. Psychonomic Bull. Rev. 12 (6), 969–992.
Saffran, J.R., Aslin, R.N., Newport, E.L., 1996. Statistical learning by 8-month-old infants. Science 274 (5294), 1926–1928.
Saffran, J.R., Johnson, E.K., Aslin, R.N., Newport, E.L., 1999. Statistical learning of tone sequences by human infants and adults. Cognition 70 (1), 27–52.
Sanders, L.D., Newport, E.L., Neville, H.J., 2002. Segmenting nonsense: an event-related potential index of perceived onsets in continuous speech. Nat. Neurosci. 5 (7), 700–703.
Sanders, L.D., Ameral, V., Sayles, K., 2009. Event-related potentials index segmentation of nonsense sounds. Neuropsychologia 47 (4), 1183–1186.
Schön, D., François, C., 2011. Musical expertise and statistical learning of musical and linguistic structures. Front. Psychol. 2.
Schroeder, C.E., Lakatos, P., 2009. Low-frequency neuronal oscillations as instruments of sensory selection. Trends Neurosci. 32 (1), 9–18.
Srinivasan, R., Bibi, F.A., Nunez, P.L., 2006. Steady-state visual evoked potentials: distributed local sources and wave-like dynamics are sensitive to flicker frequency. Brain Topogr. 18 (3), 167–187.
Wilsch, A., Neuling, T., Obleser, J., Herrmann, C.S., 2018. Transcranial alternating current stimulation with speech envelopes modulates speech comprehension. Neuroimage 172, 766–774.
Zoefel, B., VanRullen, R., 2016. EEG oscillations entrain their phase to high-level features of speech sound. Neuroimage 124, 16–23.