Behavioral/Cognitive

Segmenting and Predicting Musical Phrase Structure Exploits Neural Gain Modulation and Phase Precession

Xiangbin Teng,[1] Pauline Larrouy-Maestri,[2,3] and David Poeppel[3,4,5,6]

[1]Department of Psychology, The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China, [2]Music Department, Max-Planck-Institute for Empirical Aesthetics, Frankfurt 60322, Germany, [3]Center for Language, Music, and Emotion (CLaME), New York, New York 10003, [4]Department of Psychology, New York University, New York, New York 10003, [5]Ernst Struengmann Institute for Neuroscience, Frankfurt 60528, Germany, and [6]Music and Audio Research Laboratory (MARL), New York, New York 11201

Music, like spoken language, is often characterized by hierarchically organized structure. Previous experiments have shown neural tracking of notes and beats, but little work touches on the more abstract question: how does the brain establish high-level musical structures in real time? We presented Bach chorales to participants (20 females and 9 males) undergoing electroencephalogram (EEG) recording to investigate how the brain tracks musical phrases. We removed the main temporal cues to phrasal structures, so that listeners could only rely on harmonic information to parse a continuous musical stream. Phrasal structures were disrupted by locally or globally reversing the harmonic progression, so that our observations on the original music could be controlled and compared. We first replicated the findings on neural tracking of musical notes and beats, substantiating the positive correlation between musical training and neural tracking. Critically, we discovered a neural signature in the frequency range ∼0.1 Hz (modulations of EEG power) that reliably tracks musical phrasal structure. Next, we developed an approach to quantify the phrasal phase precession of the EEG power, revealing that phrase tracking is indeed an operation of active segmentation involving predictive processes. We demonstrate that the brain establishes complex musical structures online over long timescales (>5 s) and actively segments continuous music streams in a manner comparable to language processing. These two neural signatures, phrase tracking and phrasal phase precession, provide new conceptual and technical tools to study the processes underpinning high-level structure building using noninvasive recording techniques.

Key words: **event segmentation; hierarchical structure; musical phrase; neural entrainment; phase precession; temporal prediction**

Significance Statement

Many music types are characterized by complex, hierarchical structures that evolve over time, requiring listeners to construct high-level musical structures, anticipate future content, and track notes and beats. There exists little evidence of how the brain performs online structural-level musical segmentation and prediction. This study reveals an ultralow-frequency neural component that modulates beat tracking and reliably correlates with parsing musical phrases. We further identified a phenomenon called "phrase phase precession," indicating that listeners use the ongoing listening experience to build structural predictions and track phrase boundaries. This study provides new conceptual and technical tools for studying the operation underlying structure building in various abstract musical features, using noninvasive recording techniques such as EEG or MEG.

Received July 17, 2023; revised May 29, 2024; accepted June 11, 2024.

Author contributions: X.T., P.L.-M., and D.P. designed research; X.T. and P.L.-M. performed research; X.T. and P.L.-M. analyzed data; X.T., P.L.-M., and D.P. wrote the paper.

We thank Johannes Messerschmidt, Dominik Thiele, Claudia Lehr, and Cornelius Abel for their technical support; Johannes Messerschmidt for his assistance with data collection; and Cecilia Musci and Lea Fink for their assistance in preparing music materials. We thank Yi Du, Molly Henry, and Andrew Chang for their comments on a previous version of the manuscript. This work was supported by the Max-Planck-Society and The Chinese University of Hong Kong.

The authors declare no competing financial interests.

Correspondence should be addressed to Xiangbin Teng at xiangbinteng@cuhk.edu.hk.

https://doi.org/10.1523/JNEUROSCI.1331-23.2024

Copyright © 2024 the authors

Introduction

The physical environment presents continuous streams of information, but we typically perceive discrete events (Newtson et al., 1977; Zacks and Swallow, 2007; Kurby and Zacks, 2008). This holds for spoken language, music, films, and motor sequences (Lashley, 1951; Lerdahl and Jackendoff, 1983; Baldassano et al., 2017). An intuitive example is that continuous speech is segmented into phonemes and syllables, which are grouped into words and phrases (Ghitza, 2012; Ding et al., 2016; Poeppel and Assaneo, 2020; Gwilliams et al., 2022). Likewise,

in music, units of different sizes are organized in nested structures (Lerdahl and Jackendoff, 1983; Larrouy-Maestri and Pfordresher, 2018). Accordingly, music listening involves segmenting hierarchical event structures, comparable to spoken language processing (Maess et al., 2001; Patel, 2003; Jackendoff, 2009; M. Rohrmeier, 2011; Koelsch et al., 2013). Evidence illuminating mechanisms of active structural segmentation in music has been sparse and circumstantial.

Music listeners internally form sequences based on musical cues including temporal and harmonic information. To perceive musical event structure in real time, listeners anticipate incoming musical sounds following past harmonic progressions (Huron, 2008; Vuust et al., 2009, 2022; M. A. Rohrmeier and Koelsch, 2012; Tillmann, 2012; Patel and Morgan, 2017). Musical phrase segmentation, measured behaviorally, has been described as natural and effortless for musicians as well as listeners without explicit musical knowledge (Trainor and Trehub, 1992; Kragness and Trainor, 2016; Hansen et al., 2021). Abstract online musical segmentation involves fast structural predictions, and behavioral evidence supporting that listeners process phrases is abundant. However, the processes underpinning online segmentation and prediction of musical phrases are unclear. We conjecture that neural dynamics should be observed that reflect simultaneously (1) parsing musical phrase structures and (2) predicting phrase boundaries. Previous research has not identified characteristic neural signatures of these concurrent operations.

Experiments studying online processing of music mainly focus on lower-level features, such as tracking notes or beats or rhythmicity (Nozaradan et al., 2011; Doelling and Poeppel, 2015; Fujioka et al., 2015; Lenc et al., 2018; Harding et al., 2019). Prediction is approached at this level using models based on information theory or transition probability, to estimate sequential predictability of notes and chords (Pearce, 2005; Di Liberto et al., 2020). Previous studies investigating musical syntax and structural prediction typically use deviation-detection paradigms: Neural responses to endings of well-formed musical segments (e.g., musical phrases) are compared with endings of manipulated sequences (Knosche et al., 2005; Koelsch et al., 2013, 2019; Silva et al., 2014). The observed differences, such as the Early Right Anterior Negativity (Koelsch et al., 2002) or the Closure Positive Shift (Neuhaus et al., 2006), are interpreted as evidence for processing and predicting musical syntax. It is not known how the brain, in real time and under natural listening conditions, segments high-level musical structures and establishes rapid predictions over a timescale of ∼1 s (e.g., note, beat levels) to ∼10 s (e.g., musical phrase levels).

Here we draw on concepts from other fields. First, it has been proposed that "gain modulation" in neural systems serves a fundamental role in scaffolding high-level computations (Salinas and Sejnowski, 2001; Martin, 2020). We hypothesize that musical phrasal segmentation can be implemented through modulating the gain of low-level neural processes, such as neural tracking of notes and beats. Second, we draw inspiration from mechanisms of spatial cognition that have been extended to language (Teng et al., 2020). The phenomenon of "phase precession" reveals that animals predict spatial positions of a future path during spatial navigation (Tolman, 1948; Jensen and Lisman, 1996; Buzsaki, 2005). We suggest that neural modulation components lock to musical phrasal boundaries and manifest phase precession, representing musical phrase segmentation with structural prediction (Fig. 1*B*,*C*).

To test these two hypotheses, we used four-part chorales composed by J.S. Bach (1685–1750), which follow strict harmonic rules (M. Rohrmeier, 2011; Fig. 1*A*,*B*), and we collected electrophysiological (EEG) data. We aim to identify neural signatures that reflect the active segmentation of musical phrases over a timescale of ∼5 to ∼10 s.

Materials and Methods

Data and code availability

The music materials and EEG data, and all the analysis codes have been deposited in the OSF folder and made publicly available (https://osf.io/vtgse/). Readers can contact the Lead Contact, Xiangbin Teng (xiangbinteng@cuhk.edu.hk), to request the raw EEG datasets.

Participants

Thirty-four native German speakers (age 18–34, 23 females) took part in the experiment. All the participants self-reported normal hearing and no neurological deficits. Twenty-nine participants were included (age 18–34, 20 females) in the analyses. Among the five participants excluded from EEG analyses: one participant did not undergo EEG recording; we encountered technical issues during EEG recording for two participants; and two participants could not finish the experiment. Written informed consent was obtained from each participant before the experiment, and monetary compensation was provided after the experiment. The experimental protocol was approved by the Ethics Council of the Max Planck Society.

Individual scores of musical training

In order to investigate the relation between the musical training of the participants and their neurophysiological responses to music listening, we used the Goldsmiths Musical Sophistication Index (GOLD-MSI; Mullensiefan et al., 2014). This self-report inventory was given to the participants in German before the EEG recording and includes six dimensions: Active Music Engagement, Self-reported Perceptual Abilities, Musical Training, Self-reported Singing Abilities, and Sophisticated Emotional Engagement with Music. We mainly focused on Musical Training here.

Experimental design

Stimuli. We selected 11 music pieces from the 371 four-part chorales by Johann Sebastian Bach (Breitkopf Edition, Nr. 8610). In the eighteenth century, chorales were congregational hymns performed in religious settings. The regularity and transparency of the harmonic structures of the Bach chorales facilitated our subsequent analyses in a frequency tagging paradigm to investigate musical phrase segmentation. In the selected music pieces, musical phrases are often marked by fermatas, indicating a pause or the prolongation of a note; the predefined phrasal structures of the selected pieces are shown in Table 1.

We then processed the original musical scores so that several confounds that may appear in EEG analyses could be avoided. We first ensured that each beat was always marked by a note onset, and hence the neural tracking of beat could be measured accordingly. Concretely, the whole and half notes were substituted by four or two quarter notes and the ties across notes were deleted so that they could be repeated. The material consisted thus mainly of quarter and eighth notes and a few sixteenth notes. More importantly, as fermatas in the original music pieces provided rhythmical or acoustic cues for phrasal structures (such as pauses between phrases and lengthened notes), we removed them so that temporal cues would be drastically limited for musical phrase segmentation.

Here, we investigate how listeners rely on chord progressions and abstract musical structures to parse music streams into phrases. In the original music pieces, there are two other nonstructural factors that potentially confound with structure-based musical segmentation: (1) The chords at the beginning and end of musical phrases, namely, onset and ending beats, are salient cues of phrasal boundaries; those boundary beats appear periodically as the phrasal structures are regular in the selected chorales (Table 1). Since listeners might lock to those salient boundary beats rhythmically to parse music streams, we need a control condition to tease out the effect of those boundary beats from the musical phrase segmentation based on harmonic progression. (2)

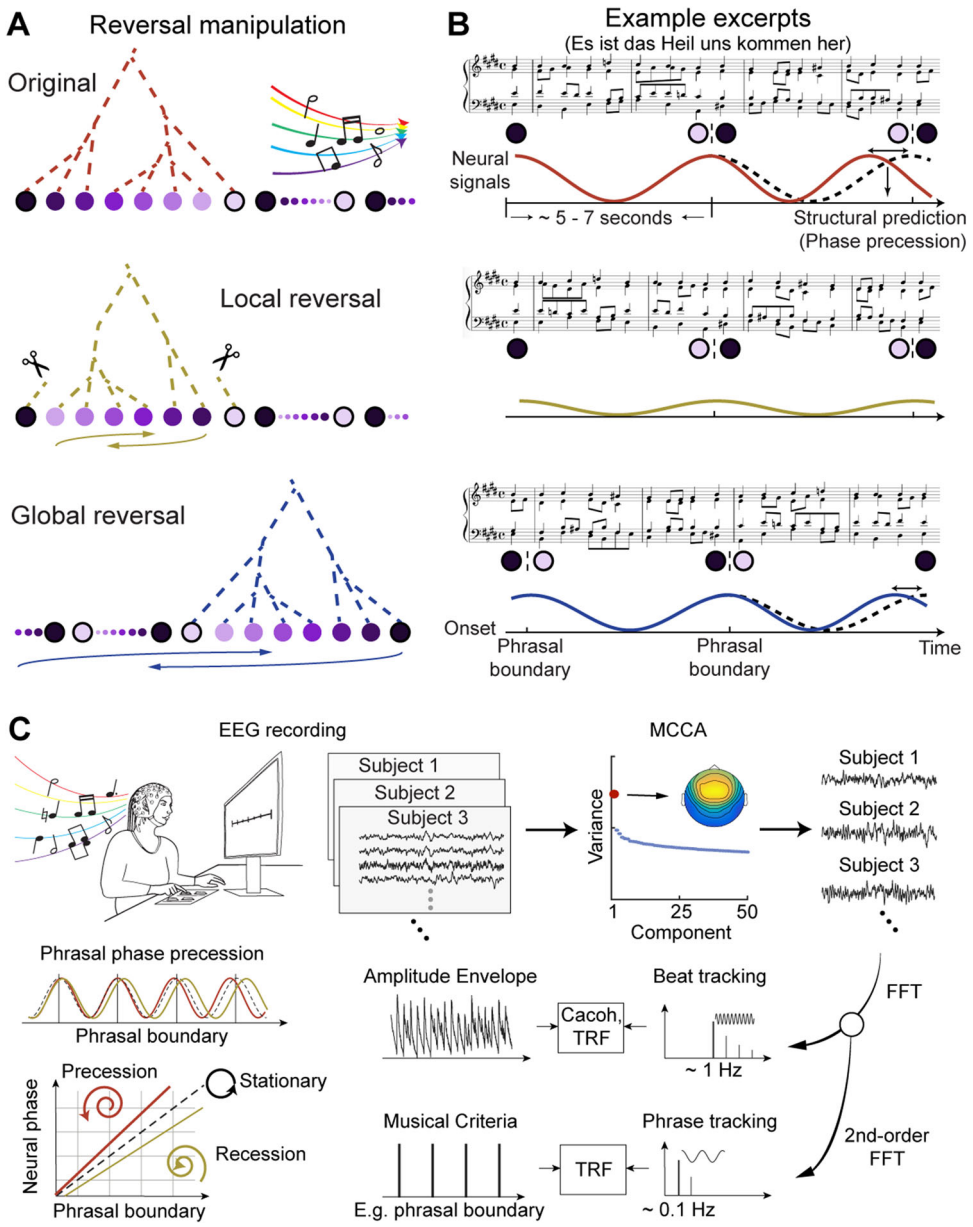


Figure 1. Stimulus manipulation and experimental paradigm. ***A***, Reversal manipulation. Ten Bach chorales (original, top) subjected to two reversal conditions: global reversal (bottom), the order of beats of an entire music piece was temporally reversed, and local reversal (middle)—the middle part of the musical phrases of each piece was reversed. The harmonic structure of musical phrases, suggested by the dashed tree structures, was hypothesized to contribute to musical phrase segmentation and prediction in the original and the global reversal conditions but not in the Local reversal condition. ***B***, Example excerpt. The first two phrases of a piece demonstrate the reversal manipulations. Neural signals are hypothesized to lock to the phrasal structures in the original (top) and global reversal (bottom) conditions but to a lesser degree in the local reversal (middle) condition, illustrated by the wave amplitude. Based on findings in speech segmentation, the harmonic progressions enable listeners to anticipate phrasal boundaries. This can be demonstrated by the phrase-segmenting neural signals advancing faster than the musical structures unfold physically—the phenomenon of phase precession. ***C***, Experimental paradigm and analysis. Participants listened to each piece while undergoing EEG recording and rated how much they liked each piece. We extracted shared neural components across the participants using MCCA (top right) and selected the component that explained the largest variance. We first conducted one Fourier decomposition to measure beat/note tracking ~1 Hz and derived temporal response function (TRF) and cerebro-acoustic coherence (Cacoh). We conducted the second Fourier decomposition that revealed how the power of neural signals was modulated by the phrasal structures ~0.1 Hz and derived TRFs using four different musical criteria. Lastly, we quantified phrasal phase precession in a neural-phase and phrasal-boundary pane: precession occurs when neural phase advances faster than phrasal boundaries; phase recession occurs otherwise.

In Western tonal music, the progressions between chords follow specific rules (i.e., the circle of fifths) that typically develop into cadences and lead to a closure feeling at the end of the musical phrase. The selected chorales follow those typical rules that are well known to listeners in Germany, surrounded by a Western tonal musical system. Therefore, identifying the phrasal structures of those chorales in their original condition does not present a challenging test on musical phrase segmentation. However, we can manipulate the selected materials in a way that, while basic musical contents (e.g., notes and tempi) are kept intact, a set of new harmonic progressions are generated, which are still plausible but less typical to Western listeners. If musical phrase segmentation can still be observed in those manipulated materials, this would provide additional evidence showing that listeners can parse music streams into phrases according to harmonic progressions, even when progressions are less familiar to listeners. For those two purposes, we employed a paradigm of temporal reversal that has been used in speech and language studies

Table 1. Selected chorales and their predefined phrasal structures marked by fermatas

No.	Piece code	Piece name	Phasal structure (beats per phrase)
0	BWV 255	Ach Gott und Herr	4 4 8 4 4 8
1	BWV153/1	Ach Gott, vom Himmel sieh darein	8 8 8 8 8 8
2	BWV 3/6	Ach Gott, wie manches Herzeleid	8 8 8 8
3	BWV267	An Wasserflüssen Babylon	8 8 8 8 8 8 8 8 8 12
4	BWV 6/6	Erhalt uns Herr, bei deinem Wort	8 8 8 8
5	BWV 248	Ermuntre dich, mein schwacher Geist	8 8 8 8 8 8 8 8
6	BWV 86/6	Es ist das Heil uns kommen her	8 8 8 8 8 8 8
7	BWV 308	Es spricht der Unweisen Mund wohl	8 8 8 8 8 8 8
8	BWV 20/7 11	O Ewigkeit, du Donnerwort	12 12 12 12 12 12
9	BWV 245/5	Vater unser im Himmelreich	8 8 8 8 8 8 8 8
10	BWV 110/7	Wir Christenleut	8 8 8 8 8 8

(Saberi and Perrott, 1999; Baldassano et al., 2017), although our reasons for using the reversal procedure here are different from such studies.

We created two reversal conditions. We first applied a local reversal procedure to the pieces. According to the phrasal structures defined in the musical scores, we kept the onset beat and the ending beat of each phrase intact but temporally reversed the order of the beats in the middle. For example, a phrase of eight beats [1 2 3 4 5 6 7 8] becomes a new phrase [1 7 6 5 4 3 2 8] after the local reversal. A schematic illustration of this procedure on the musical scores is illustrated in Figure 1*A*. The rationale for local reversal was as follows: as we kept the salient onset and ending beats intact but disrupted the harmonic progression, neural indices for tracking musical phrases and structural prediction should be lower in the condition of local reversal than in the original condition, supporting that intact harmonic progressions in the original pieces are used for phrasal segmentation and prediction. Alternatively, if the salient onset and ending beats are sufficient for listeners to parse continuous music streams into musical phrases, musical phrase segmentation should not differ between locally reversed pieces and the original pieces. To implement the local reversal procedure, we first used MuseScore 3 to generate audio files of the music pieces from the processed original scores. As the beats in the generated audio files had the exact same length at a certain tempo, we cut out each beat and then reversed the order of the beats while preserving the waveform of each beat (the note order in a beat was not reversed).

Second, we applied a global reversal procedure on the music pieces. For each piece, the order of all beats was temporally reversed. Figure 1*A* provides a schematic illustration of the procedure. This reversing procedure was implemented on the order of beats but not on the waveforms of beats. The chord progressions in the selected chorales follow the circle of fifths, a well-established structure in Western classical music and typical for listeners enculturated by Western tonal music (e.g., the participants recruited in Germany in the current study). In the reversed pieces, basic music content (e.g., tone scales and relationships between adjacent chords) is kept plausible, but the direction of chord progression is simply reversed. The reversed chord progressions, by no means, are random or lose any musical structures. The new (reversed) chord progressions, arguably, are less expected by listeners for this music genre but can still provide cues of phrasal opening and ending. In other words, the musical phrasal structures were preserved in a sense that the harmonic progressions in the reversed pieces can still be used to parse a phrasal structure. If musical phrase segmentation is still observed for globally reversed pieces, and if the neurophysiological tracking of musical phrases is comparable to the original pieces, it can be concluded that listeners do extract harmonic progressions to parse music streams into phrases and predict phrasal boundaries, even though the chord progressions are less familiar to the participants.

We need to emphasize that the reasons for using the reversal manipulations here differ from the reasons for the reversal procedures used in studies on spoken language and movies (Saberi and Perrott, 1999; Hasson et al., 2008; Lerner et al., 2011), where the reversal procedures of different timescales were used to completely disrupt the information within corresponding timescales. In the case of music, using random sequences by scrambling the order of beats cannot be used as a control condition, since compromising musical structure entirely leads to losing the "musical" character of the material (at least from this time period) and simply prevents direct comparison between conditions. Indeed, the current study focuses on the role of the harmonic progressions in musical segmentation and prediction and is specifically testing tonal material. Therefore, both control conditions contain musical structure while contrasting with the original condition. Note that the local reversal condition was designed to preserve the boundary beats and their temporal regularity, so the observation of certain neural responses locking to the boundary beats in the local reversal condition should not be taken as a failure to disrupt all musical structures.

We generated the music audio files in MuseScore 3 for three tempi, 66, 75, and 85 bpm (beats per minute), which arguably cover a reasonable range of tempi of actual recordings of the Bach chorales (the tempo range we observed in a set of 45 interpretations by voice, organ, or piano). Choosing three tempi, instead of only one, also enabled us to investigate how musical phrase segmentation varied with tempo. Furthermore, if phrase and beat tracking are observed in the music pieces of different tempi at different frequencies of EEG signals, this aids in validating the neural measurement—neural tracking of music structures is not constrained to a specific tempo or a frequency range of EEG signals but aligns with musical structures.

In summary, we had three reversal conditions—original (without reversal), global reversal, and local reversal—and three tempi: 66, 75, and 85 bpm. This generated a 3-by-3 experimental design. We used the music piece number 0, BWV 255 (Table 1), as the training piece before the formal experiment to familiarize participants with the experimental materials and the task. The 10 remaining pieces were included in the data analyses. In total, we presented 90 music pieces (3 conditions $*$ 3 tempi $*$ 10 pieces). The sampling rate of audio files was 44,100 Hz and the amplitude of the audio files was normalized to 70 dB SPL by referring the music materials to a 1 min white noise piece that had the same sampling rate and was measured beforehand to be 70 dB SPL at the experimental setting.

Acoustic analyses of the music material. To characterize temporal dynamics of the music pieces, we derived the amplitude envelopes of the musical materials. We filtered the audio files through a Gammatone filterbank of 64 bands logarithmically spanning from 50 to 4,000 Hz. The amplitude envelope of each cochlear band was extracted by applying the Hilbert transformation on each band and taking the absolute values (Glasberg and Moore, 1990). The amplitude envelopes of 64 bands were then averaged and downsampled from 44,100 to 100 Hz to match the sampling rate of the following processed EEG signals for further analyses [i.e., cerebral-acoustic coherence and temporal response function (TRF); see below].

We transformed the amplitude envelopes to modulation spectra so that the canonical temporal dynamics of the music pieces can be concisely shown in the spectral domain. We first calculated the modulation spectrum of the amplitude envelope of each music piece using FFT with zero-padding of 8,000 points and took the absolute value of each frequency point. As each piece had a different length and the zero-padding caused the modulation spectra of different music pieces to have different ranges of magnitude, we normalized the modulation spectrum of each piece by dividing the norm of its raw modulation spectrum. We then averaged the normalized modulation spectra across the 10 pieces for each condition at each tempo. At each tempo, the average modulation spectra across three conditions were highly similar, so we only showed the modulation spectra averaged across the three conditions for each tempo (Fig. 2*A*). Nonetheless, to demonstrate that the differences of the amplitude modulation spectra between the three conditions were not noteworthy, we calculated at each tempo the standard deviation over the modulation spectra of the three conditions (Fig. 2*B*). The music pieces of the three conditions showed the standard deviations close to zero below the beat rate. In addition, we further tested the acoustic

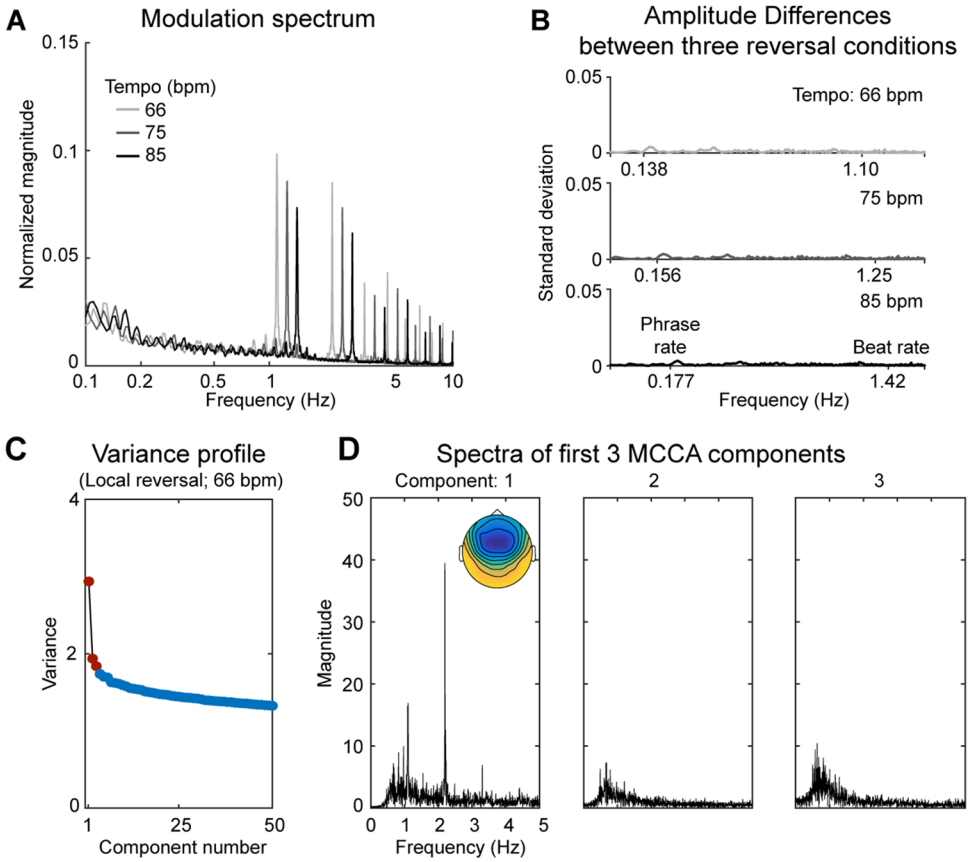


Figure 2. Acoustic analysis and MCCA. ***A***, Acoustic analyses. We extracted the amplitude envelopes of the music pieces and calculated the modulation spectra averaged over 10 music pieces of each tempo in each reversal condition. The modulation spectra of the three reversal conditions highly overlapped and were therefore plotted together. Note that there are no salient spectral peaks around the phrase rates (0.138, 0.156, and 0.177 Hz). ***B***, The standard deviation across the three reversal conditions at each tempo is tiny. In terms of the acoustic spectral components, the three conditions are comparable at each tempo. This shows that the acoustic properties of the music pieces do not differ and should not contribute to the difference in phrasal segmentation between the reversal conditions. ***C***, MCCA components of the local reversal condition at the tempo of 65 bpm. This provides an example of MCCA and shows that the first component of MCCA stands out and explains a considerable amount of variance. ***D***, Amplitude spectra of first three MCCA components of the local reversal condition at the tempo of 65 bpm. We calculated the amplitude spectrum of each MCCA component in ***C*** to examine the spectral component of each component. The spectrum of the first component shows amplitude peaks corresponding to the beat rate and the note rate (the first harmonic of the beat rate). This further supports our choice of analysis—specifically, the first MCCA component, but not other components, contained the neural signals induced by beat and note structures in the music pieces. Furthermore, this is consistent with the topography in the inset, showing that the topography reflects auditory responses to beats and notes. Similar results are shown for other tempi and other reversal conditions.

differences between the reversal conditions by conducting a one-way repeated-measures ANOVA at each frequency point for each tempo, with nine pieces of eight beats per phrase as "subjects" and the reversal condition as the main factor. For this analysis, we left out the piece of 12 beats per phrase, which has a different phrase rate than the other 9 pieces. Consistent with what is shown in Figure 2*B*, we found no significant effects at any frequency point for each tempo ($p > 0.05$), so the neural tracking of phrasal structures found later cannot be caused by the acoustic differences between the music stimuli of the different reversal conditions.

Experimental protocol and EEG recording. EEG data were recorded using an actiCAP 64-channel, active electrode set (10–20 System, BrainVision Recorder, Brain Products, brainproducts.com), at a sampling rate of 500 Hz, with a 0.1 Hz online high-pass filter (12 dB/octave roll-off). There were 62 scalp electrodes, one electrode (originally, Oz) was placed on the tip of the nose. All impedances were kept below 5 kΩ, except for the nose electrode, which was kept below ∼10 kΩ. The auditory stimuli were delivered through plastic air tubes connected to foam earpieces (E-A-RTone Gold 3A Insert Earphones, Aearo Technologies Auditory Systems).

The experiment included a training session and a testing session. In the training session, we presented the three conditions of the piece BWV 255, at the three tempi, to the participants, who were instructed to answer a question after listening to each piece: "how do you like this music piece." The participants rated the music pieces using a 6-point scale from 1 to 6, with 6 being the most positive rating and 1 being the most negative rating. This behavioral task was designed primarily to keep the participants engaged and to make them pay attention to music pieces and is thus not reported here. In each trial, before each piece was presented, the participants were required to focus on a white cross in the center of a black screen. After 3–3.5 s of silence, a piece was presented, and the participants were instructed to keep their eyes open while listening to it. After each piece ended, the question appeared on the screen and the participants rated the piece. The next trial started right after the participants' responses. The order of music pieces was randomized between participants. The behavioral data and EEG signals were not recorded during the training session.

The testing session followed the training session. We presented the 90 music pieces in six blocks to the participants while they were undergoing EEG recording. Fifteen pieces were presented in each block. The order of the music pieces was randomized within and across the blocks, and a different order of pieces was presented to each participant. The trial structure was the same as in the training session. After each block, the participants could choose to take a short break of ∼1–3 min or to initiate the next block. Behavioral ratings are not analyzed in this study, as our main goal here was to investigate how the brain segments high-level musical structures.

EEG preprocessing and data analysis

EEG data analysis was conducted in MATLAB 2016b using the FieldTrip toolbox 20181024 (Oostenveld et al., 2011), the wavelet toolbox in MATLAB, NoiseTools (de Cheveigne et al., 2018, 2019), and the Multivariate Temporal Response Function (mTRF) Toolbox (Crosse et al., 2016).

EEG recordings were off-line referenced to the average of activity at all electrodes. Raw EEG data were first filtered through a bandpass filter from 0.5 to 35 Hz embedded in the FieldTrip toolbox (a FIR zero-phase forward and backward filter using MATLAB "fri1" function with an order of 4). Each trial (the recording for each music piece) was divided into an epoch of a length of each music piece plus a 3 s prestimulus period and a 3 s poststimulus period. Baseline was corrected for each trial by subtracting out the mean of each trial, which was necessary for the following multiway canonical correlation analysis (MCCA; de Cheveigne et al., 2018, 2019).

We applied an independent component analysis (ICA) on all the epoched trials, employing the FastICA algorithm (Hyvarinen, 1999) incorporated in the FieldTrip toolbox. This was used to remove artifacts resulting from eye blinks and movements. Following the epoching of trials, we amalgamated all the epoched trials from every condition and applied ICA to the EEG data, ensuring the subsequent step remains impartial to each condition. We then examined each of the initial 30 ICA components, eliminating those associated with eye blinks and both horizontal and vertical eye movements. These were chosen based on the topography of channel weights for each component and their temporal dynamics. Finally, we converted the remaining components back to the EEG channel space.

MCCA. We focused on the neurophysiological signals evoked by the musical pieces (auditory stimuli). To extract EEG signals that largely reflect auditory-related neural responses, instead of arbitrarily selecting certain electrodes (e.g., Cz or FCz), we deployed MCCA to extract shared EEG components across all the participants. The participants were listening to the same set of pieces and hence similar neural responses to the music pieces should be observed across participants, although each participant's EEG recording varied with his or her head size and EEG cap position and with different sources of noise. MCCA, briefly, first implements a principal component analysis (PCA) on each participant's EEG recordings and then pools all the PCA components across the participants to conduct another set of PCA. The components from the second PCA reflect shared components of neural responses across the participants that are invariant to EEG cap position and individual head sizes. Detailed procedures and further explanations can be found in de Cheveigne et al. (2019). This procedure of component extraction simplified further analyses and avoided biases introduced by arbitrary EEG channel selections and by differences of EEG cap positions and head sizes across participants.

We first sorted each participant' EEG recordings according to the music piece number and concatenated EEG recordings of each condition and each tempo after epoching to form one long trial. For example, the EEG recording for the piece 1 (BWV153/1) was put in the beginning and was followed by the EEG recordings of the piece number 2, 3, 4, 5, 6, 7, 8, 9, and 10, sequentially. The 3 s pre- and poststimulus periods were included for each trial. For each condition and each tempo, we applied MCCA to derive 50 components and selected the first component, as the first component explained the most variance (Figs. 1*C*, 2*C*). We checked MCCA weights over EEG channels and plotted them as topographies to check whether the first component is related to music-related auditory responses. We showed one condition (local reversal at 66 bpm) as an example in Figure 2*C*,*D*, from which it can be seen that the topography of MCCA weights resembled typical topographies of auditory EEG responses. To further validate our component selection, we calculated the power spectra for the first three components of MCCA of all conditions and showed an example in Figure 2*D* (local reversal, 66 bpm). We zero-padded the time series of each component to 8,000 points (8,000 points correspond to a stimulus duration of 80 s) and calculated power spectra using FFT. We calculated the absolute value for each frequency point in the amplitude spectrum obtained from the FFT and then normalized the amplitude spectrum by its norm. Figure 2*D* shows that only the first component showed amplitude peaks corresponding to the frequency of the beat rate of the corresponding tempo. This demonstrated that the first component of MCCA indeed represented neural responses to music pieces. All the other conditions showed similar patterns as the example in Figure 2*C*,*D*.

After conducting MCCA for each condition and each tempo and extracting the first component, we projected the component back to each participant and derived one long "trial" including the 10 music pieces for each participant. We then cut out the neural response for each piece from the long trial. As PCA sometimes reversed polarity of EEG signals, the polarity of the derived signals was manually checked and corrected for each participant.

Note that we performed the MCCA separately on the 10 pieces in each condition at each tempo because we hypothesized that the neural responses to music could be affected by both the reversal conditions and the tempo. If we performed the MCCA only once over all the conditions (all the 90 pieces), the MCCA components would reflect a mix of weights from all conditions. In other words, the MCCA components would reflect weights from all conditions—this would mean that the information about the reversal manipulations and tempi would be spread across all conditions and comparisons between conditions would be skewed (e.g., one condition may have very high magnitude of neural responses and this condition would drive the weights of the MCCA whereas other conditions contribute less to the MCCA). Therefore, we decided instead to conduct MCCA on each condition to prevent the extracted neural responses of each condition from being contaminated by other conditions. In this way, one MCCA component was independently derived from each condition without being biased by other conditions.

Cerebral-acoustic coherence (Cacoh). We calculated the coherence between the neural signals and the amplitude envelopes of music pieces in the spectral domain—cerebral-acoustic coherence (Cacoh)—to investigate how pieces of different conditions were tracked in the brain at different neural frequencies. The rationale is that, if the music pieces of one condition can be more reliably tracked by the auditory system than the pieces of the other conditions because of their specific musical structures or tempi, this difference can be observed in the magnitude of coherence between the neural signals and the amplitude envelopes. This measurement, Cacoh, has been used in neurophysiological studies on speech perception (Peelle et al., 2013; Doelling et al., 2014), which quantified how phase-locked neural responses correlate with acoustic signals—at what frequencies and how strong the neural signals and the acoustic signals correlate with each other. In essence, Cacoh calculates cross-spectrum coherence between neural signals and acoustic signals; here we calculated magnitude-squared coherence using the function "mscohere" in MATLAB 2016b (https://de.mathworks.com/help/signal/ref/mscohere.html). This allowed us to control the temporal window used to calculate the cross-spectrum coherence and the spectral resolution of the coherence values so that the temporal window sizes and the spectral resolution were matched across all the conditions. This helped avoid the biases of calculating spectral coherence over the entire music pieces introduced by different lengths of the music pieces. Nonetheless, we still referred to our calculation as "Cacoh" as this term has been well adopted in the literature.

We calculated Cacoh for each music piece and each participant using the neural data and the amplitude envelope derived between 1 s after stimulus onset and 1 s before offset, to minimize the influence of onset and offset neural responses of a whole musical piece. The temporal window used was a Hanning window of 1,024 points with an overlap of 512 points (50% of overlap between adjacent temporal windows). We constrained the frequency range from 0.1 to 20 Hz, with a step of 0.05 Hz. The inputs, the neural signals, and the amplitude envelope had a sampling rate of 100 Hz. We then averaged Cacoh values across the 10 pieces within a condition. Therefore, a Cacoh value was derived for each tempo and each condition at each frequency point.

Temporal response function. To investigate the temporal dynamics of the phase-locked responses to the music pieces, we calculated the reverse

correlation between the neural signals and the music amplitude envelopes in the temporal domain—TRFs. TRF has been commonly used to investigate how the auditory system codes acoustic and linguistic features in neurophysiological studies on speech perception (Di Liberto et al., 2015; Brodbeck et al., 2018; Broderick et al., 2018). The rationale for this analysis is similar to Cacoh: if music pieces of a condition robustly evoke auditory responses, such robust auditory responses should be reflected by TRF in the time domain. Furthermore, we could extract temporal information from TRFs on when the auditory responses are modulated by different conditions. For example, TRFs may reflect differences between tempi within the first 100 ms after the onset of an auditory event while TRFs may differ between the conditions after 100 ms because of different manipulations on the musical structures. Early auditory responses likely reflect sensory processes whereas late responses tend to reveal high-level cognitive processes in music perception (Koelsch et al., 2002, 2005; Koelsch and Siebel, 2005).

The TRF was derived from the amplitude envelopes of stimuli (S; for details see Acoustic analyses of the music material) and their corresponding EEG signals (R; for details see above, EEG preprocessing and data analysis) through ridge regression with a parameter (lambda) to control for overfitting and with M to enforce a smoothness constraint by penalizing the difference between adjacent time points (Lalor et al., 2006; Crosse et al., 2016; superscript t indicating transpose operation):

$$TRF = (S^t S + \lambda M)^{-1} S^t R$$

We calculated a TRF for each music piece for each participant using the neural data and the amplitude envelopes between 1 s from the stimulus onset to 1 s before the offset so that the influences of the onset and offset neural responses were avoided. The TRFs were calculated from 200 ms before onsets of auditory events and 500 ms after. As the highest tempo was 85 bpm, the shortest interbeat interval of the music pieces was ∼700 ms and the internote interval was <400 ms. In the statistical analysis, we focused on the period of TRFs from −100 to 400 ms but showed the TRFs from −200 to 500 ms. The EEG signals and the amplitude envelopes of music pieces were not filtered or decomposed into different frequency bands. We fixed the lambda at 0.1 for calculating TRFs across all the conditions and tempi, and hence the differences of TRFs across conditions should not be because of different lambda values. After calculating TRF for each music piece, we averaged TFRs over 10 pieces of one condition for each participant and conducted statistic tests on the mean TFRs across different conditions. The TRFs were calculated using the multivariate temporal response function (mTRF) toolbox (Crosse et al., 2016).

The fixed lambda value of 0.1 was decided empirically. The lambda value decides how signals are temporally smoothed during TRF calculation and makes sure that the encoding model does not overfit to noise: the larger the lambda value is, the signals are smoothed to a larger extent temporally. If zero is chosen for the lambda, the inversion of matrices cannot be implemented as sometimes the covariance of neural responses recorded here is not fully ranked. Therefore, we chose 0.1 as the smallest value of the lambda to make sure that the inversion can be implemented, and the signals were not smoothed seriously. We believe that the choice of the lambda value should not affect our results, as we compared TRFs across the conditions instead of drawing conclusions from the absolute TRF values and the choice of the lambda value exerted an equal influence on TRFs of all the conditions.

Temporal modulation of EEG power. Here we aim to investigate musical phrase segmentation beyond the level of the beat structure. The lengths of the music phrases ranged from ∼7 s (8 beats per music phrase at 65 bpm) to ∼5 s (8 beats per music phrase at 85 bpm), which corresponded to an ultralow EEG frequency range of ∼0.13 to ∼0.18 Hz. This ultralow EEG frequency range has proved to be challenging to analyze because of the 1/f shape of power spectrum of EEG recording and hence a low signal-to-noise ratio at the ultralow frequency range. Another fact is that the time-frequency analysis requires an ultra-long temporal window for such frequency ranges (i.e., two cycles for 0.13 Hz, which is ∼15 s). Previous studies have not yet shown a meaningful and robust neural signature within this range. Given the complexities mentioned earlier, it is indeed challenging to directly observe temporal dynamics of musical phrase segmentation and its corresponding frequency components in EEG signals.

To circumvent the challenge of analyzing the ultralow frequencies of EEG signals and to directly observe temporal dynamics of musical phrase segmentation, we resorted to a different strategy: The musical phrasal structures likely modulated the power of neural responses to each beat. The beat rates of the music pieces were above 1 Hz, at which neural signals can be well recorded by EEG. Therefore, we chose to measure the temporal modulation of EEG power at the beat rate to investigate musical phrasal tracking. If listeners do track musical phrases and the neural responses to beats are modulated by musical structures, we should be able to observe neural signatures reflecting musical phrase segmentation through temporal modulation of EEG power at the beat rates. This procedure is akin to recovering low-frequency amplitude envelopes from high-frequency carriers in speech signal processing (Teng et al., 2019). For instance, in speech signal processing, the amplitude envelopes of low-frequency ranges (e.g., <30 Hz) can be extracted from carrier frequencies above 100 Hz. Essentially, the temporal modulation of EEG power denotes the oscillatory pattern of EEG signal modulation. In signal processing terms, the bandpass filter operates on carrier signals (raw EEG signals), and the slow oscillatory pattern is the envelope of amplitude modulation on these carrier signals.

We first conducted a time-frequency analysis on EEG signals to derive power spectrograms of neural responses to the music pieces. The single-trial data derived from MCCA were transformed using the Morlet wavelets embedded in the FieldTrip toolbox, with a frequency range from 1 to 35 Hz in steps of 1 Hz and a temporal range from 3 s before the onset of music pieces to 3 s after the offset of music pieces in steps of 100 ms. To balance spectral and temporal resolution of the time-frequency transformation, from 1 to 35 Hz, the window length increased linearly from 3 cycles to 10 cycles. Power (squared absolute value) was extracted from the wavelet transform output at each time-frequency point. Moreover, the power values for each trial were normalized by dividing the mean power value over the baseline range (−1.5 to −0.5 s) and taking logarithms with base 10 and then was converted into values with unit of decibel by multiplying by 10.

We next quantified temporal modulation spectra of EEG power to examine at which frequency the EEG power was modulated and whether the salient modulation frequencies were related to the frequency range of musical phrasal structures. This procedure is similar to the calculations of acoustic modulation spectra of the music pieces (Fig. 2*A*). We calculated the modulation spectrum of EEG power at each frequency using FFT with zero-padding of 20,000 points and took the absolute value of each frequency point, so that the spectral resolution of the modulation spectra was 0.005 Hz. The high spectral resolution guaranteed to resolve different music phrase rates (8 beats per music phrase: 0.1375 Hz at 65 bpm; 0.1562 Hz at 75 bpm; 0.1771 Hz at 85 bpm). We then normalized the modulation spectra of each trial by dividing the modulation spectrum of each trial by the length of its corresponding music piece. As the music pieces of each condition had different lengths, they could not be summed together in the temporal domain, we chose to sum the modulation spectra (complex numbers) of 10 pieces within each condition in the spectral domain. This spectral summation on complex number kept phase information, and the shared frequency components with similar onset phases across the 10 pieces in a condition were emphasized. After the spectral summation, we took the absolute value of each frequency point to derive the modulation spectrum for each condition. The above procedure resulted in a two-dimensional modulation spectrum for each condition, with one dimension as the frequency of EEG power and the other dimension as the frequency of temporal modulation of EEG power. It should not matter for the following statistic tests in this case whether we summed the modulation spectra or averaged the modulation spectra over the music pieces, as both summation and averaging are linear operations.

To examine how the musical phrase segmentation varied with the tempo of music pieces, we averaged over three reversing conditions within each tempo and showed the modulation spectra of three tempi,

respectively (Fig. 4*A*). If the neural signatures of musical phrase segmentation varied with the tempo, this helped validate the observed neural signatures. We also checked whether other frequency ranges of EEG power (i.e., the alpha band, 8–12 Hz, and the beta band, 13–30 Hz) showed signatures of musical phrase segmentation, as the previous literature reported an important role of the neural beta band in music processing (Doelling and Poeppel, 2015; Fujioka et al., 2015).

We next focused on the frequencies of EEG power at the beat rates of three tempi, as we only observed salient modulations of EEG power ∼1 Hz (Fig. 4*A*), which was close to the beat rates of three selected tempi. We conducted once again the time-frequency transformations on the EEG signals, but only at the beat rates, 1.1001 Hz (65 bpm), 1.25 Hz (75 bpm), and 1.4165 Hz (85 bpm). The window length in the time-frequency transformations was three cycles. Other procedures and parameters were the same as the above analyses of calculating the EEG power and the modulation spectrum of each tempo and each reversing condition. We selected the third music piece of the original condition (BWV267), which had the longest duration, and showed the temporal dynamics of EEG power at the beat rate of each tempo (Fig. 4*B*). This directly showed how the EEG power was modulated by the musical phrasal structures. For each participant, we summed the modulation spectra of 10 pieces within a condition in the spectral domain and conducted statistics on the summed modulation spectrum of each condition (Fig. 4*C*).

Surrogate test on modulation spectra of EEG power. The modulation magnitudes of neural signals of musical phrase segmentation cannot be directly compared between different tempi, because the music pieces of different tempi have varying lengths, and the tempi bias estimation of the modulation spectra. As a result, frequency components in the lower frequency range tend to have greater amplitude due to the 1/f nature of neurophysiological signals. Therefore, individual baselines need to be constructed for the modulation spectra of different tempi; then the modulation spectra can be normalized according to the baselines within each tempo. Therefore, we conducted surrogate tests in the spectral domain to derive a null distribution of the modulation spectrum for each condition and each tempo. This served two purposes—we can directly test the significant frequency ranges of musical phrase segmentation using the null distributions and then normalize the modulation spectra to do comparisons across conditions.

The null hypothesis of the surrogate test here is that the neural signals did not track or phase-lock to the musical phrasal structure, and hence the sum of modulation spectra across 10 pieces of each condition should not differ from the surrogate modulation spectra derived from jittering the temporal alignment between each of the 10 music pieces and its neural recording temporally. However, it is not appropriate to conduct this surrogate test in the time domain. As the lengths of musical phases (>5 s) were much longer than the pre- and poststimulus periods (3 s), temporal surrogation cannot efficiently disrupt the temporal correspondence between neural signals and musical phrasal structures—the jittering can only happen within a cycle of musical phrase. More importantly, the tempi here still affect the procedure of surrogation—for example, jittering within a temporal range of 6 s at 85 bpm has a different jittering effect from the same jittering range at 65 bpm because of different lengths of musical phrases at the two tempi. Therefore, we conducted the surrogate test in the spectral domain after Fourier transformation of EEG power.

After calculating the modulation spectrum of the EEG power of each music piece, we multiplied the modulation spectrum (complex numbers) with a complex number that has a norm of 1 but a randomly generated phase from 0 to 2 ∗ pi. This multiplication reset the onset phase of the EEG power of each music piece so that the EEG power of each music piece now started with a random phase, whereas the amplitude magnitude of the modulation spectrum was kept intact. We then derived the summed modulation spectrum of the surrogated data for each tempo and each condition. By doing this, the 10 pieces within a condition all had different onset phases, and therefore the temporal components locked to the musical phrasal structures should be averaged out or severely smoothed. This surrogating procedure on phase in the spectral domain was invariant to different tempi, as the phase values were normalized to different tempi. We repeated this procedure 1,000 times and derived a null distribution of modulation spectra for each condition and each tempo and for each participant.

We averaged the modulation spectra of each surrogation across participants and derived a null distribution of the group-averaged modulation spectrum for each condition and each tempo. We chose a one-side alpha level of 0.01 as the significant threshold for each condition and each tempo (Fig. 4*C*, dashed lines). The frequency ranges where the empirical modulation spectra are above this threshold can be considered as the frequency ranges showing robust effects of musical phrase segmentation. To derive one frequency range across three conditions for one tempo so that the frequency range selected is unbiased to each condition for the following analysis, we chose the lower bound and the upper bound of the significant frequencies among all the three conditions and used them to define one frequency range for all the three conditions. The significant frequency ranges are as follows: at 65 bpm, 0.125–0.15 Hz; at 75 bpm, 0.14–0.175 Hz; at 85 bpm, 0.165–0.2 Hz.

Normalization of modulation spectra of EEG power. From the surrogate tests above, we obtained the null distributions for the modulation spectra of EEG power. As mentioned, using the null distributions we could normalize the raw modulation spectra that were biased because of different tempi of music pieces. Here, we calculated the mean of the null distribution for each condition and each tempo and subtracted out the mean of the null distribution from the raw modulation spectrum.

Temporal response function of EEG power envelope with phrasal boundaries as regressor. In the above spectral analysis, we characterized the frequency components reflecting musical phrase segmentation. What are the temporal trajectories of musical phrase segmentation? Is the musical phrase segmentation a result of neural signals locking to the onset beat or to the offset beat of a musical phrase? Spectral analyses cannot provide such temporal information. Additionally, the length of the chosen music pieces varies, resulting in differing precision in the estimation of fundamental frequency and harmonics in the spectrum. Normalizing the spectral magnitude between pieces of different lengths is also challenging. Conversely, TRF analysis, which is not affected by stimulus length and does not require normalization between stimuli, is more appropriate for this study. Therefore, we calculated TRFs of EEG power using the phrasal boundaries of music pieces as the regressors.

The TRFs of EEG power envelopes were calculated in a similar way as in the calculation of TRF of boundary beats of musical phrases. We marked the boundaries between two phrases in each music piece. For example, there are seven phrases in the music piece of BWV153/1 (Table 1, Piece Number 1), and hence there are six phrase boundaries; these six phrase boundaries served as a regressor for the EEG power envelope to derive a TRF for this piece. After deriving the EEG power envelopes at each tempo, we calculated TRF for each music piece from 1 s after stimulus onset to 1 s before stimulus offset. The length of TRF estimation was 6 s, with 3 s before the phrase boundary and 3 s after. We then averaged TRFs of 10 pieces within a condition of each tempo. The lambda was set as 0, which indicated that no smoothing was applied and the calculation of TRF was equal to a reverse correlation.

The TRFs of EEG power are shown in Figure 4*E*. It can be seen that the peak latencies for different conditions differ. To determine the latency of each condition and to find the peak time point, we fitted each group-averaged TRF using a one-term Gaussian model, whose center point determined the peak latency of each group-averaged TRF.

Phrasal phase precession and phrase segmentation at group-average level. The original music pieces are well structured, and a listener can readily extract the temporal regularity of the phrase structures and predict the unfolding of the musical pieces. In contrast, the reversed conditions may preserve the salient onset and offset beats of musical phrases or expected chord progressions but reduce the structural cues to the phrases, so listeners experience difficulties to predict the unfolding of the pieces. To test this conjecture, we tried to quantify phase precession

of neural responses on the level of musical phrases—phrasal phase precession (PPP), as phase precession of neural signals has been argued to reflect the prediction of future events (Jensen and Lisman, 1996; Lisman, 2005). The rationale here is that, if listeners can predict the incoming musical contents, the neural phase should proceed faster as the music unfolds. On the other hand, if listeners cannot predict the incoming music contents but passively respond to or are solely driven by the acoustic events in music, the neural phases would proceed at a constant speed, as the speed/tempo of the acoustic events is constant in each music piece.

To quantify phrasal phase precession, we need to conduct our analysis on the neural response to each music piece at a single-trial level. However, we conducted EEG recording here, and the EEG recording at a single-trial level often has a low signal-to-noise ratio; usually at least tens to hundreds of repetitions are needed for analyzing neural responses to a single stimulus. To resolve this issue, we took another strategy for quantifying phase precession for each music piece—we first averaged each trial across 29 participants and then conducted our analyses on the group-averaged trial. Although such a strategy prevented us from probing individual differences, the process of the group averaging did aid in emphasizing the neural components induced by the music pieces while averaging out neural signals unrelated to the music listening. If a music piece is predictable on the level of musical phrase segmentation, the phrasal phase precession should be observed for all the participants to various degrees. Therefore, the phase precession quantified from the group-averaged trials could collectively indicate that the participants predict the unfolding of each music piece as well as reflect the predictability of each music piece.

The above analyses extracted the EEG power envelopes reflecting the musical phrase segmentation and determined the significant frequency ranges of musical phrase segmentation at each tempo. We averaged the power envelopes of each music piece across 29 participants and employed a two-pass second-order Butterworth filter embedded in the FieldTrip toolbox to filter the group-averaged envelopes, using the significant frequency range of each tempo as the cutoff frequencies. The phases of the averaged power envelope of each music piece were extracted by applying the Hilbert transformation on the filtered signal and taking the angle at each time point. We showed in Figure 5*A* (left panel) the phase series of three conditions of the music piece, BWV267, at 75 bpm, as this piece has the largest length. We also plotted the phase values at the phrase boundaries in Figure 5*A* (right panel) so that it can be clearly seen that the proceeding speed of the neural phases was increasing in the original and global reversal conditions as the music piece unfolded.

We constructed an index, phrasal phase precession index (PPPi), to quantify phrasal phase precession over each music piece. If there is no phase precession, the neural phase series would proceed in a step of $2 * \text{pi}$ over every musical phrase. This is to say that, if we fit a line between the musical phrasal boundaries and their corresponding neural phase values, the slope will be exactly $2 * \text{pi}$ if there is no phase precession. In contrast, if the neural phase series is accelerating as the music piece is unfolding, this line will be steeper and the slope will be larger than $2 * \text{pi}$; if the phase series is slowing down, the slope will be smaller than $2 * \text{pi}$. Therefore, the differences between the slope of the fitted line and $2 * \text{pi}$ can represent whether there is phase precession and to what extent the phase series is accelerating or slowing down. The differences here were termed as PPPi. We provide an example in Figure 5*C*, where we show the lines fitted between the music phrase boundaries and their corresponding neural phase values in three reversing conditions at 75 bpm.

The peak frequencies of modulation spectra of EEG power correlate with PPPi, as the phase precession can be considered as frequency modulation or phase modulation—the frequency of musical phrase segmentation is modulated along the time—and the mean frequency over the whole music piece becomes larger than the phrase rate if the phrasal phase precession happens. Therefore, here, we also quantified the peak frequencies of the modulation spectrum of EEG power within the significant phrase segmentation range for each music piece, so that the results of PPPi can be further validated by a different index. In the sense of frequency modulation, PPPi reflects linear phase changes. For example, if the phrase rate is 0.1 Hz (10 s per circle) and the neural frequency is 0.105 Hz (~9.5 s per circle), the phase shift, measured by time, will be 0.5 s every phrase. If we use the phrase rate as the reference frequency (which was the case here), the phase shift will be $0.5/10 * 2 * \text{pi}$ every phrase ($0.1 * \text{pi}$). Admittedly, neural dynamics may not follow such a linear manner strictly. In future studies, neural recordings of high signal-to-noise ratios, such as intracranial recordings, can be used to measure the timing of each neural peak around the phrasal boundary.

PPPi, as stated above, reflects the value of a shifted phrase every phrase (phase shift per phrase). As the music is unfolding, the amount of shifted phase accumulates. The accumulated phase shift correlates with the length of the music piece, so it is not a good measurement and is affected by the music piece length, whereas PPPi can be viewed as an index normalized by the music piece length, which also directly reflects how the prediction is built up as the music is unfolding. For the longest music piece at 75 bpm, the PPPi value is ~0.1 (Fig. 5*E*); the accumulated phase shift is 0.9 after nine phrases, which is $\sim 0.29 * \text{pi}$ and is close to the length of one beat in the phase domain $(0.25 * \text{pi}) - 2 * \text{pi}$ covers one phrase, which includes eight beats; a PPPi value of $0.25 * \text{pi}$ represents that neural signals shift forward by one beat. Asymptotically, the accumulated phase shift will amount to $2 * \text{pi}$ (eight beats, a whole phrase), if the wrapped phase is used and its value range is constrained to be between 0 and $2 * \text{pi}$. For example, if PPPi for an extremely long music piece is 0.1, after 62 stationary phrases (~8 min), the accumulated phase shift will reach $2 * \text{pi}$. It would be interesting to test the maximum value of PPPi using longer music pieces; it is likely that PPPi values would asymptote to a few beats per phrase.

Statistical analysis

One-way, two-way, and three-way repeated-measures ANOVAs (rmANOVAs) were employed to examine differences across conditions, consistent with our within-group experimental design. The Greenhouse–Geisser correction was applied when the assumption of sphericity was potentially violated, ensuring the validity of our analyses; however, none of the tests exhibited violations of this assumption. Mauchly's test of sphericity was rigorously used to assess these assumptions. All ANOVA analyses were performed using SPSS software (version 24, IBM). For instances requiring post hoc analyses, the false discovery rate (FDR) method (Benjamini and Hochberg, 1995) was utilized to adjust for multiple comparisons. The level of significance for all tests was set at $p < 0.05$. We explicitly noted cases where the adjustment was not applied in post hoc analyses.

Correlation analyses involved calculating Pearson's correlation coefficient (r) to evaluate the strength and direction of linear relationships between variables. The significance of these correlations was assessed, maintaining a significance level of $p < 0.05$.

A cluster-based rmANOVA test identified temporal regions with significant effects under reversal conditions (Fig. 3*D*; Maris and Oostenveld, 2007). The null hypothesis posited no difference between reversal conditions. To construct the null distribution, condition labels for data points were shuffled, and a one-way rmANOVA with reversal conditions as the main factor was performed at each time point. Significant time points were identified at a significance threshold of 0.05, grouped into clusters, and cluster-level F values calculated. The cluster with the highest F value was used to form a distribution across 1,000 permutations. An alpha level of 0.01 was set as the significance threshold. Clusters in empirical data with F values exceeding this threshold, determined from the permutation-derived null distribution, were deemed significant.

A surrogate statistical test was conducted to identify spectral regions significantly affected by reversal conditions in EEG power modulation spectra. This test followed a clustering and thresholding procedure similar to the cluster-based rmANOVA, but with a null distribution creation method tailored to the modulation of EEG power calculations. Detailed procedures are provided above, Surrogate test on modulation spectra of EEG power.

Results

We first examined how the reversal manipulations modulated low-level (note and beat) musical tracking, hence providing a replication of previous research using new stimuli. Next, we explored musical phrase segmentation and compared the three reversal conditions. Lastly, we discovered a neural mechanism, "phrasal phase precession," that supports the process of predicting musical phrase boundaries.

Neural tracking of notes and beats is not modulated by high-level musical structure

The results are depicted in Figure 3*A*. Peak Cacoh values were observed around the beat rate (B), the note rate (N), and their harmonics, across all tempi and reversal conditions. We selected the peak Cacoh values around the beat and note rates (Fig. 3*B*) and conducted a three-way repeated-measure ANOVA (rmANOVA) (condition × tempo × frequency, beat rate vs note rate). We found a main effect of Frequency, with Cacoh values of the note rate significantly larger than the beat rate ($F_{(1,28)} = 108.09$; $p < 0.001$; $\eta_p^2 = 0.794$). The main effect of tempo was significant ($F_{(2,56)} = 3.68$; $p = 0.032$; $\eta_p^2 = 0.116$), and the interaction effect between tempo and frequency was significant ($F_{(2,56)} = 124.47$; $p < 0.001$; $\eta_p^2 = 0.816$). A two-way rmANOVA (tempo × frequency) showed that Cacoh increased with the tempo at the beat rate ($F_{(2,56)} = 26.33$; $p < 0.001$; $\eta_p^2 = 0.485$) but decreased at the note rate ($F_{(2,56)} = 100.76$; $p < 0.001$; $\eta_p^2 = 0.783$). Tempo-modulated music tracking can be explained by the resolution-integration mechanism in auditory perception (Teng et al., 2016): shortening intervals between acoustic elements facilitates global integration of acoustic information over a beat;

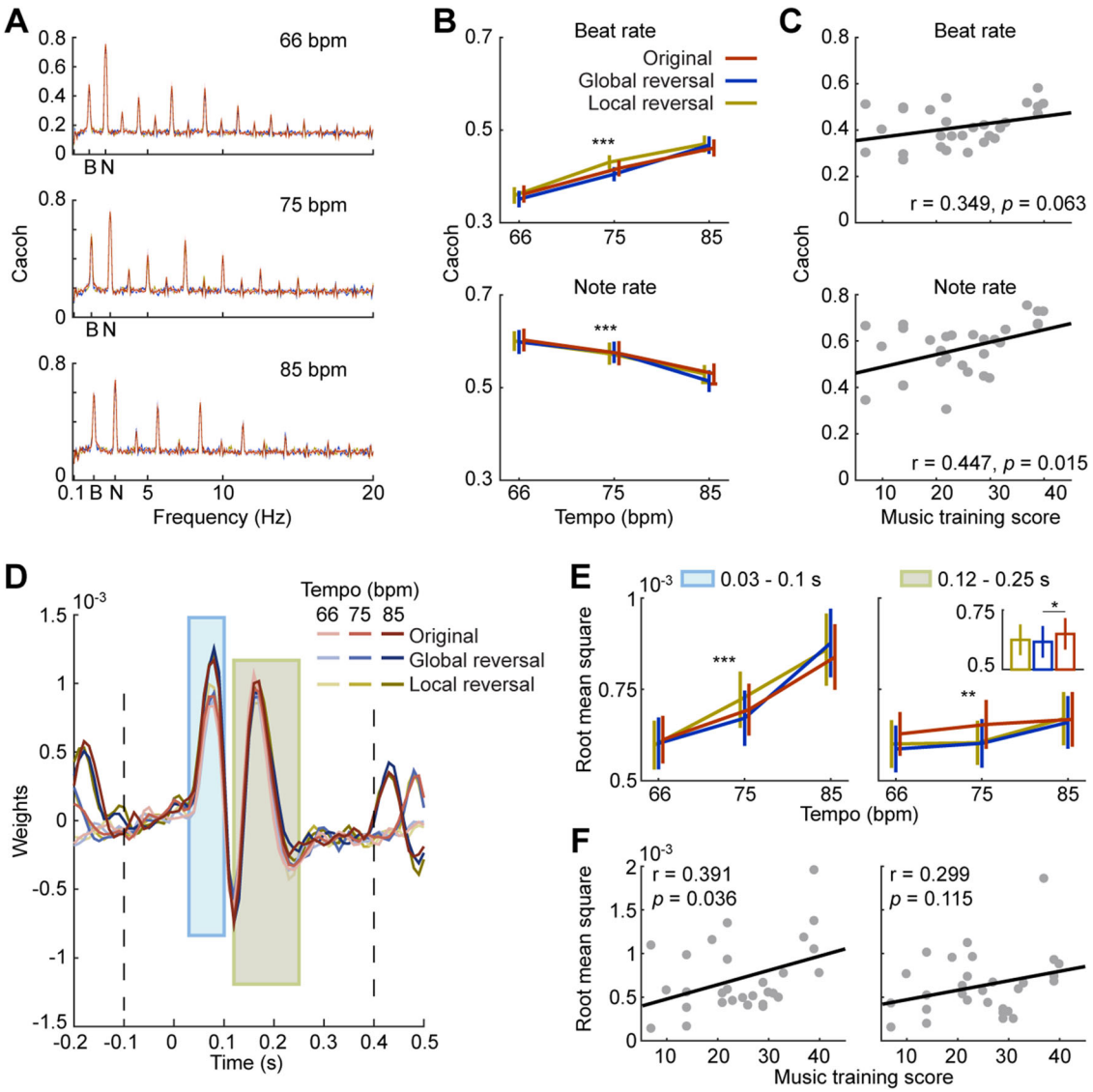


Figure 3. Neural tracking of beats and notes. ***A***, Cerebro-acoustic coherence (Cacoh) for notes (N) and beats (B) at each tempo. Line color indicates the reversal condition. The three conditions original, local, and global are shown in each panel, but only the original condition can be seen because the three conditions overlap. ***B***, Tempo modulates neural tracking of beats and notes. Increasing tempo positively modulates beat tracking ($p < 0.05$) but negatively modulates note tracking ($p < 0.05$). ***C***, Correlation between Cacoh and music training score. We used the GOLD-MSI questionnaire to quantify how much musical training each participant received and correlated the score of this subscale with Cacoh. ***D***, TRF for each condition and tempo. We identified two periods (shaded boxes) showing significant differences ($p < 0.01$) across all conditions and tempi. Line shade codes for the tempo. The dashed lines indicate the boundaries of the permutation test. ***E***, RMS of TRF within each period. We calculated RMS within each significant period and found that the tempo positively modulated both periods ($p < 0.05$). In the late period, we found a main effect of condition ($p < 0.05$); the RMS of the original condition is larger than the global reversal ($p < 0.05$). ***F***, Correlation between RMS and music training score. We correlated RMS of each period with the music training score and found a significant positive correlation in the early period ($p < 0.05$).

lengthening the intervals facilitates perceptual analyses of local information to extract notes within a beat. Note that the Cacoh of note tracking is stronger than beat tracking, probably because the first harmonic of the beat note overlaps with the note rate, increasing the neural signal strength around the note rate.

We next correlated the beat and note tracking averaged over all reversal conditions with participants' musical training scores. In line with previous research on neural tracking of musical beats and notes (Nozaradan et al., 2011, 2012; Doelling and Poeppel, 2015; Lenc et al., 2018), we show significant correlations between music training and note tracking (Koelsch et al., 2002; Fujioka et al., 2004; Doelling and Poeppel, 2015; Harding et al., 2019). We found a significant positive correlation between training and note tracking ($r_{(27)} = 0.447$; $p = 0.015$) but not beat tracking ($r_{(27)} = 0.349$; $p = 0.063$; Fig. 2*C*).

Surprisingly—and crucially—the main effect of condition (original, local, global) was not significant ($p > 0.05$) across all analyses. This suggests that previous findings of note and beat tracking may have less to do with high-level musical structures but rather reflect lower-level auditory processing of acoustic contents and tempi in musical materials.

Reverse correlation shows temporal dynamics of music tracking

Neural tracking of notes and beats has often been investigated in the spectral domain, forgoing information on temporal dynamics. To characterize the time domain, we calculated a TRF over 10 music pieces in their nine presented versions (Fig. 3*D*). To reveal the modulation effects of the tempi and reversal conditions, we conducted a cluster-based rmANOVA analysis over the nine conditions from 100 ms before the zero point to 400 ms after (see Materials and Methods). We found two significant periods ($p < 0.01$): one early period, 30–100 ms, and one late period, 120–250 ms. We then calculated root mean square (RMS) representing overall signal strength over the TRF weights within each significant period, depicted in Figure 3*E*. We next conducted a two-way rmANOVA (condition × tempo) in each period. In the early period, we found a significant main effect of tempo ($F_{(2,56)} = 33.29$; $p < 0.001$; $\eta_p^2 = 0.543$), with the higher tempo showing higher RMS values, but not for condition ($F_{(2,56)} = 0.24$; $p = 0.788$; $\eta_p^2 = 0.008$). This finding echoes the result of the beat tracking analysis in Figure 2*B* and shows that tempo modulates early auditory responses and consequently modulates beat tracking. In the late period, we found a significant main effect of tempo ($F_{(2,56)} = 6.34$; $p = 0.003$; $\eta_p^2 = 0.185$) as well as, importantly, of condition ($F_{(2,56)} = 4.01$; $p = 0.024$; $\eta_p^2 = 0.125$). We compared the RMS of the late period between conditions and found that, after FDR correction, the original condition had higher RMS values than the global reversal ($t_{(28)} = 3.09$; $p = 0.021$; (Fig. 3*E*, inset). We then correlated the averaged RMS across all the versions in each period with the music training score (Fig. 3*F*). A significant positive correlation was shown in the early period ($r_{(27)} = 0.391$; $p = 0.036$), but not in the late period ($r_{(27)} = 0.299$; $p = 0.115$).

Summarizing these first analyses (Fig. 3), the data illustrate the basic temporal dynamics of neural responses during music listening: an early response period is primarily modulated by the tempo of music; a later period is likely related to putative musical structures. The correlations between the musical training and the neural signals are constrained to the early period (30–100 ms), suggesting that the observed correlation between neural tracking of beats and notes with music training in the spectral domain was mostly driven by early cortical responses, arguably related to basic auditory processing, instead of by processes underpinning musical structure tracking (Koelsch et al., 2002, 2005; Mankel and Bidelman, 2018). This is consistent with the view that music training can increase the efficiency of processing low-level acoustic contents in musical materials.

EEG power modulations at ultralow frequencies reflect musical phrase segmentation

By testing how the power of electrophysiological responses is modulated over the entire pieces, we sought to reveal genuine musical phrase segmentation and expected to observe differences between three conditions: original, local reversal, and global reversal. We first aimed to determine whether EEG power was modulated. We show the modulation spectrum for each tempo in Figure 4*A*. EEG power between 1 and 2 Hz (precisely around the beat rate; *y*-axis) was modulated. Importantly, the largest modulations fell around the phrase rate (*x*-axis) at each tempo.

We extracted EEG power at the beat rate for each tempo and show the EEG power of different tempi, averaged over 29 participants, for the longest original piece in Figure 4*B*. EEG power is indeed quasiperiodically modulated by the phrasal structure, which can be seen clearly from peaks of EEG power locked to the phrasal boundaries. We next quantified the modulation spectrum of each condition for the nine music pieces having eight beats per musical phrase. It is important to note that the music piece with 12 beats per phrase was excluded from this analysis because its phrasal rate differs from the other nine pieces.

Remarkably, the power modulations of the reversal conditions were found to be significantly above the threshold around the corresponding phrase rate (Fig. 4*C*). We corrected the power modulations by subtracting the mean of the null distribution of each condition derived from the surrogate tests and then identified the significant frequency ranges around the phrase rates (*F*0) and their first harmonic (*F*1). The analyses in the spectral domain—by examining how the neural signals fluctuated over entire music pieces—were primarily deployed to locate the neural correlates of the phrasal segmentation, the power modulation of the neural beat tracking at the phrase rate (Fig. 4*A*–*C*). But we have not yet shown how individual musical phases were segmented. The power modulations tracking the phrasal structures are certainly complex waveforms in the time domain (Fig. 4*B*). Accordingly, between the three reversal conditions, the differences of the neural correlates of the phrasal segmentation were not necessarily reflected only at *F*0 but in the shape of temporal waveforms that was captured by all the relevant harmonics (i.e., *F*0 and *F*1). By simply estimating the magnitude of *F*0 as the neural signature of phrase segmentation, we may not robustly estimate musical phrase segmentation. A more accurate estimation of phrase segmentation can be achieved by estimating the TRF around phrasal boundaries in the time domain (refer to Materials and Method). This approach can capture the complex waveforms in the time domain without fragmenting the neural signature into multiple frequency components.

Nevertheless, we conducted statistical tests on the effects of condition and tempo in the spectral domain, but we urge readers to consider this finding together with the following analyses in the temporal domain (Fig. 4*E*,*F*). We conducted a two-way rmANOVA (tempo × condition) on the corrected modulation amplitude within the significant range around the phrase rate (the fundamental frequency of the phrase segmentation, *F*0). We did not find any significant main effect or interactions ($p > 0.05$). We then averaged the magnitude over the significant frequency ranges around *F*0 of the phrase segmentation and its first harmonic (*F*1). The two-way rmANOVA showed a significant

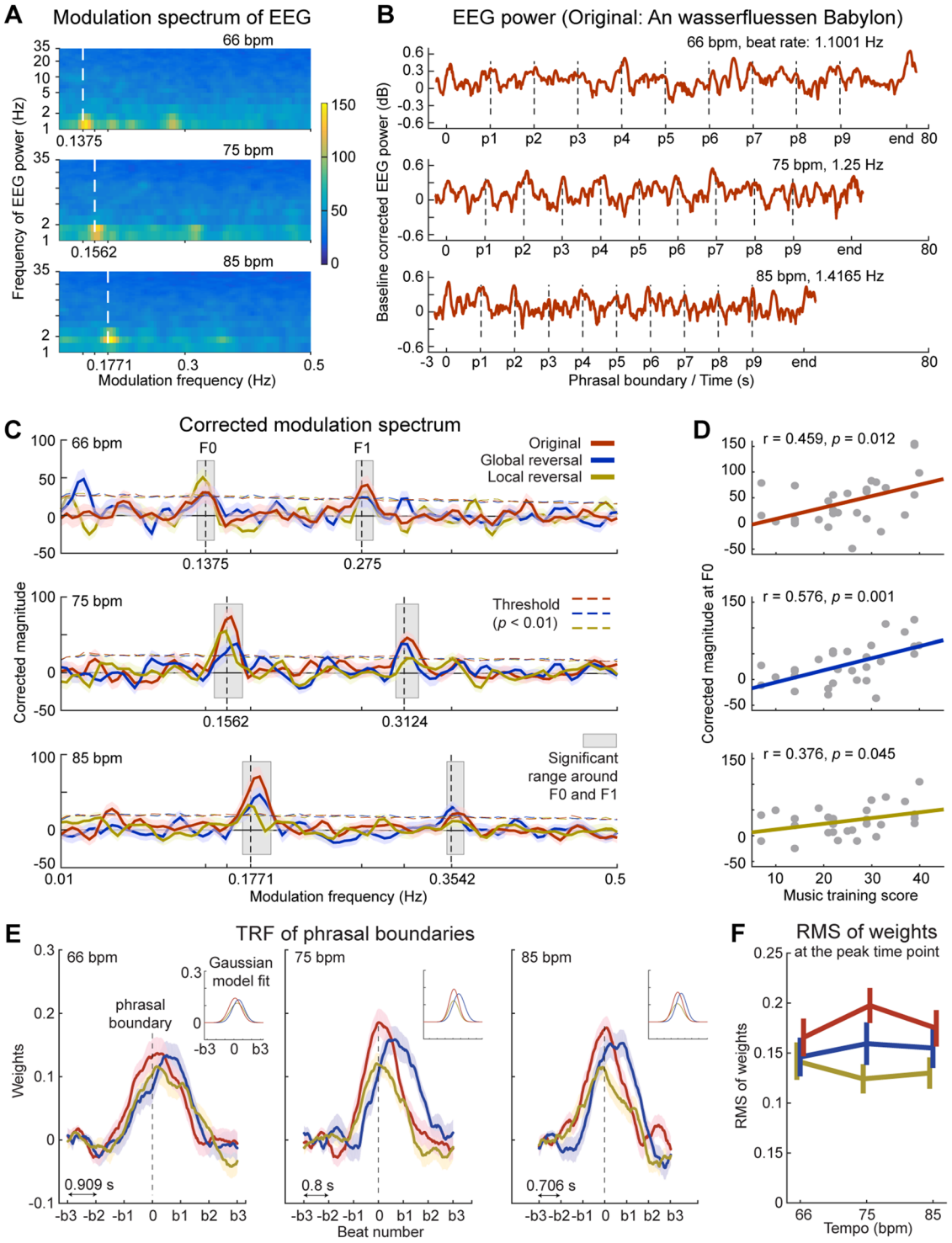


Figure 4. Spectral and temporal analyses of musical phrase segmentation. ***A***, Modulation spectra of EEG power. The modulation spectrum for each frequency of EEG power at each tempo was computed, averaged over the three conditions. The *y*-axis indicates the frequency of EEG power; *x*-axis the frequency of EEG power modulation. ***B***, EEG power of the longest piece modulated by phrasal structure. The dashed vertical lines mark the phrasal boundaries. The fluctuations of EEG power are locked to the phrase boundaries. ***C***, Modulation spectra at the beat rate. We show the modulation spectrum at the beat rate for each condition (color code as in Fig. 1) and each tempo. The horizontal dashed lines indicate the threshold derived from the surrogate tests in the spectral domain. The gray shaded areas represent frequency ranges with modulation amplitude above the thresholds around the fundamental frequency (*F*0) and the first harmonic (*F*1) of the phrase rate. ***D***, Correlation between musical training score and EEG power modulation of each reversal condition around *F*0. ***E***, TRFs of phrasal boundaries and the Gaussian model fits. We calculated the TRF of EEG power using the phrasal boundaries as the regressor and we fitted Gaussian models to the group averaged TRFs (inset in each panel). The *x*-axis marks the number of beats; the double-arrow line indicates the beat length at each tempo. ***F***, RMS of TRFs at the peak time points determined by the Gaussian fits. The shaded areas of color represent ±1 standard error of the mean over participants.

main effect of condition ($F_{(2,56)} = 3.66$; $p = 0.032$; $\eta_p^2 = 0.115$), but not of tempo ($F_{(2,56)} = 0.66$; $p = 0.521$; $\eta_p^2 = 0.023$) or an interaction ($F_{(4,112)} = 1.48$; $p = 0.214$; $\eta_p^2 = 0.050$). In the post hoc test, after FDR correction, we did not find differences between the conditions (before FDR correction, the original is significantly larger than the local reversal, $t_{(28)} = 2.36$; $p = 0.026$).

We averaged the corrected modulation amplitude across tempi and correlated the corrected modulation magnitude of each condition around the phrase rate with the music training score (Fig. 4*D*). We found significant positive correlations for all the three conditions (original: $r_{(27)} = 0.459$, $p = 0.012$; global reversal: $r_{(27)} = 0.576$, $p = 0.001$; local reversal: $r_{(27)} = 0.376$, $p = 0.045$). Note that the correlation between the musical training score and the magnitude of neural tracking of beats is not significant (Fig. 3*C*), but the power modulation of beat tracking significantly correlated with the musical training. Although the strength of correlations varied across different conditions, all the correlations are significant, even in the local reversal condition. This is probably because limited tonal cues (i.e., end and start of phrases) remained despite the harmonic manipulation. The amount of music training exerts a general influence on music listening, benefiting not only extracting harmonic progressions but also following the regular boundary beat pattern (Mankel and Bidelman, 2018).

TRFs of EEG power reveal different temporal dynamics of phrase segmentation

To capture the temporal dynamics of the phrase segmentation, we calculated TRFs of EEG power at the beat rate of each tempo using the phrasal boundaries as the regressor (Fig. 4*E*). Note that all 10 music pieces were included in this analysis. The TRFs of all three conditions show peaks around the phrasal boundaries, but their weights (TRF magnitude) differed. To summarize the TRFs quantitatively, we fitted the group-averaged TRFs with Gaussian models (Fig. 4*E*, insets) and tested the significance of the differences of RMS centered around the TRF peak point between the conditions (Fig. 4*F*). We performed a two-way rmANOVA (tempo × condition) and found a significant main effect of condition ($F_{(2,56)} = 5.73$; $p = 0.005$; $\eta_p^2 = 0.170$), but not of tempo ($F_{(2,56)} = 0.31$; $p = 0.732$; $\eta_p^2 = 0.011$) or an interaction ($F_{(4,112)} = 0.68$; $p = 0.608$; $\eta_p^2 = 0.024$). The post hoc test shows that the TRFs of the original condition is significantly larger than the TRFs of the local reversal ($t_{(28)} = 3.19$; $p = 0.009$; FDR corrected). Similar results were observed when we calculated RMS within ±1 standard deviation or within ±2 standard deviation of the Gaussian model fits.

The TRFs of the original and the local reversal mostly peaked with long tails around the boundaries of the predefined phrasal structure (Fig. 4*E*). The phrasal boundaries were defined at a single time point in musical scores, but the "mental phrase structuring" process during natural listening is built up over chord progressions. The fact that the TRFs started to peak above baseline around the last chord of the musical phrase probably suggests that the segmentation process starts around the offset beat.

Summarizing to this point, the findings of musical phrase segmentation in both the spectral and temporal domains (Fig. 4) demonstrate that we can capture mechanisms that support how listeners parse musical phrases online. The local reversal manipulation was designed to keep the regular temporal structures formed by the boundary beats and was used as a control with matched temporal patterns (i.e., eight beats per segment) for the original condition (Fig. 1*A*). Rhythmic power modulations locking to the boundary beats can still be observed in the Local reversal condition (Fig. 4*C*,*E*), suggesting that the salient boundary beats and the regular temporal structures indeed aid in segmenting continuous music streams. Note that such temporal regularities (facilitating phrasal segmentation) suit the original lyric structure and thus contribute to the compositional goal of such pieces. However, since the phrase segmentation was better in the original condition than that in the local reversal condition (Fig. 4*E*), online phrase segmentation cannot solely be explained by boundary beats of musical phrases and their regular temporal structures; the harmonic progressions were extracted by listeners for musical phrasal segmentation (Hołubowska et al., 2023). On the other hand, the global reversal condition showed musical phrasal segmentation between the local reversal condition and the original condition, suggesting that the harmonic progressions, even in the reversed direction, were partly exploited by listeners for segmenting musical streams.

The last question is, since the boundary beats already presented cues, though limited, for segmenting musical streams (as shown in the Local reversal condition), how do the harmonic progressions (original and reversed) contribute to segmenting musical streams?

Phrasal phase precession reveals predictive processes of musical phrase segmentation

Music listening is not a passive process but involves listeners actively predicting incoming musical contents based on previous musical information (Vuust et al., 2022). In the case of musical phrase segmentation, listeners may exploit the harmonic progressions to predict the phrasal boundaries. This predictive process further distinguishes the local reversal condition from the other two conditions in musical phrase segmentation: boundary beats should drive, passively, boundary segmentation in the local reversal condition, whereas an active segmenting process should be observed in the original and global reversal conditions, as listeners can exploit the relationship between the harmonic progressions and the phrasal boundaries. To explore our conjecture, the predictive nature of musical phrase segmentation, we quantified phase precession of EEG power modulations. The important phenomenon of phase precession has been shown in spatial navigation (Buzsaki, 2005) and related to prediction of further events (Jensen and Lisman, 1996; Lisman, 2005; Qasim et al., 2021). At the beginning of each musical piece, the neural responses will track the phrasal structure that is constructed internally from the physical stimuli (like the cognitive map constructed in the spatial domain); as the music stimuli unfold, predictions for the coming phrasal structures are established and the neural phases advance faster. Because phase precession is difficult to show in continuous naturalistic stimuli, this is an opportunity to explore a fundamental neurobiological mechanism in the context of complex perceptual processing.

We quantified phrase-level phase precession for each musical piece, as the different pieces were different in terms of their musical structures, which potentially leads to different degrees of predictability among the individual pieces. Note that the music piece with 12 beats per phrase (Piece 8) was not included here because its phrasal rate is different from the other nine pieces and cannot be averaged together with the other nine pieces in the spectral domain under the frequency tagging paradigm. Figure 5*A* (left panel) shows the wrapped phase series of the group-averaged EEG power, for the longest music piece in the three reversal conditions at 75 bpm, plotting the phase values at the phrasal boundaries (Fig. 5*A*, right panel). The phase series at the phrasal boundaries in the original and the global

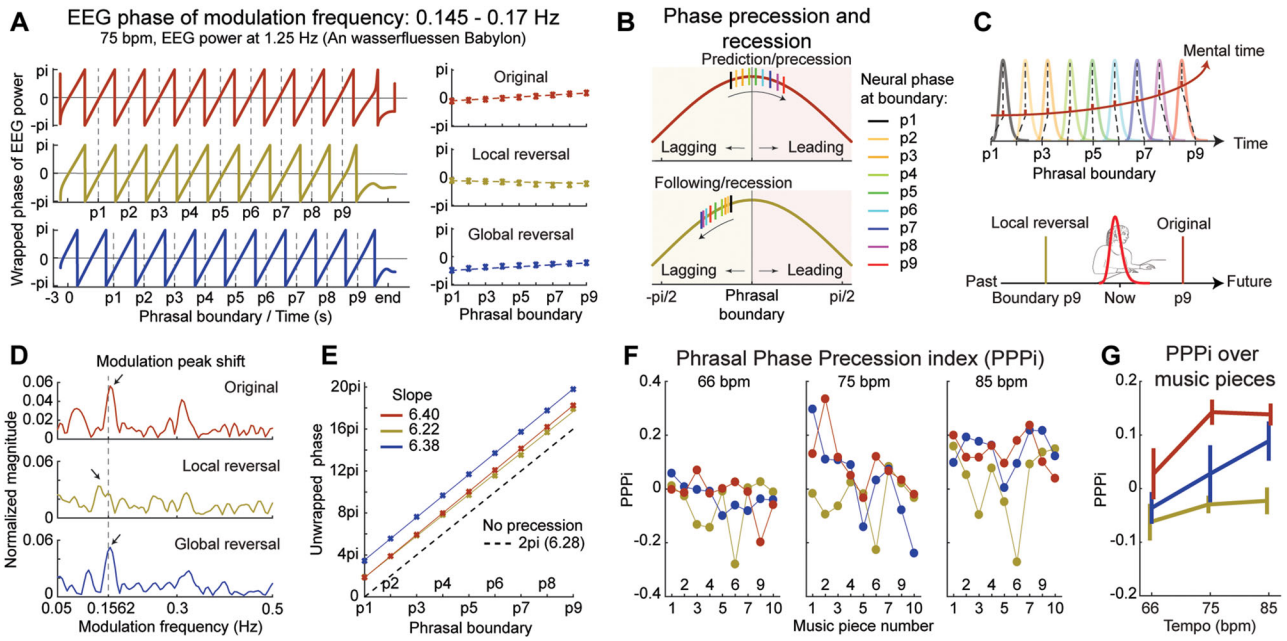


Figure 5. Phrasal phase precession. ***A***, Phase series of group-averaged EEG power of an example music piece (left panel). The phase of zero indicates where the peaks of cosine waves should be. The neural phases at the phrasal boundaries in the original and the global reversal conditions are advancing (tilting upward) as the phrasal structures unfold (right panel). ***B***, Distribution of neural phases around phrasal boundaries. Cosine waves were plotted in both panels with the phase of zero aligning with the phrasal boundary. The bars indicate neural phases at the phrasal boundaries in the right panel of ***A***. In the original condition, neural peaks lagged behind phrasal boundaries in the beginning but predicted phrasal boundaries after the fourth phrase; an opposite pattern was observed in the local reversal condition. See Figure S5 for the global reversal. ***C***, Schematic indication of phase precession. As the phrasal phase precession occurs, time is warped mentally (top panel). Phrase-segmenting neural components followed the phrasal boundaries in the local reversal condition but predicted the incoming phrasal boundaries in the original condition by the end of the music piece (bottom panel). The phrasal boundaries in the local reversal and original conditions are at the same physical time. ***D***, Shifted modulation spectral peak. The peak frequencies of neural signals in the original and global reversal conditions are higher than the phrase rate. ***E***, PPPi. We fit a line between the phrasal boundary number and the unwrapped boundary phase series. If no phase precession occurs, the slope is 2×pi (dashed line). The difference between the slope of each condition and 2×pi is indicated as PPPi and represents the degree and the direction of phase precession. ***F***, PPPi for each music piece. We calculated PPPi within the significant frequency ranges determined in Figure 4*C*. Note that PPPi values correspond to different time shifts at different tempi, which is represented by the *y*-axis on the right. ***G***, Averaged PPPi for each condition. The reversal manipulation modulated the phase precession (see main text). The error bars represented ±1 standard error of the mean over nine music pieces.

reversal conditions is gradually accelerating, but not in the local reversal.

We grouped the neural phases of the music piece at the phrasal boundaries in the original and the local reversal conditions and visualize them in Figure 5*B*, since both the conditions have the same boundary beats. As the music was unfolding, the neural phases for the original condition were increasingly moving forward; in the local reversal condition, the neural phases were lagging. This can be interpreted as the phrasal boundaries constructed from the physical stimuli in the brain are shifted forward, and the mental intervals between the phrasal boundaries are decreasing—i.e., time is warped mentally (Fig. 5*C*, top panel). By the end of the music piece, the neural signal indicating musical phrase segmentation is ahead of the phrasal boundaries defined physically in the original condition and predicts the future boundaries. In contrast, in the local reversal condition the neural signal lags behind or is passively driven by the phrasal boundaries. Concretely, the timing of the phrasal boundaries is different even though the physical final boundaries in the two conditions are defined at the same time point (Fig. 5*C*, bottom panel). Hence, the peak frequency of phrase-segmenting neural signals should be higher than the phrase rate of the music piece in the original condition (warped time) but lower than the phrase rate in the local reversal condition, which is exactly what the data show in Figure 5*D*.

We then constructed an index to summarize the phrasal phase precession by fitting a line between the order number of phrase boundaries and the unwrapped phase values for the example music piece (Fig. 5*E*). If no phase precession occurs, the slope of the fitted line would be 2×pi (a full cycle of neural signal for each phrase). A slope larger than 2×pi would be observed if the phase series accelerates, and vice versa (see Materials and Methods for more details on linearity and asymptote). We subtracted 2×pi from the slope fitted for each piece, and the index was termed the PPPi. Note that PPPi values between 0.1 and 0.2 represent a timescale of hundreds of milliseconds. For example, a PPPi value of 0.1 at 66 bpm refers to ∼110 ms advancement of neural signals per phrase; for the longest piece, after nine phrases, the advancement of the neural signals amounts to ∼990 ms, a considerable time shift.

The PPPi values showed positive values for the original version and the global reversal of the music piece, but a negative value for the local reversal. We quantified the effect of tempo and reversal condition on PPPi (shown for each piece in Fig. 5*F*) by conducting a two-way rmANOVA (tempo × condition). Each "participant" here is a music piece and the rmANOVA only included nine "participants." We found a significant main effect of tempo ($F_{(2,16)} = 21.65$; $p < 0.001$; $\eta_p^2 = 0.730$), but no significant effect was shown for condition ($F_{(2,16)} = 3.44$; $p = 0.057$; $\eta_p^2 = 0.301$) or an interaction ($F_{(4,32)} = 1.44$; $p = 0.243$; $\eta_p^2 = 0.153$). Note that after removing the piece number 10, which was picked because this piece showed a pattern of PPPi opposite to the mean PPPi pattern, we found a significant main effect for tempo ($F_{(2,14)} = 26.21$; $p < 0.001$; $\eta_p^2 = 0.789$) and condition ($F_{(2,14)} = 4.78$; $p = 0.026$; $\eta_p^2 = 0.406$),

but the interaction remained nonsignificant ($F_{(4,32)} = 1.33$; $p = 0.284$; $\eta_p^2 = 0.160$).

The significant main effect of tempo on PPPi may be caused by the normalization according to tempo in the phase domain. For example, a PPPi value of 0.1 * pi/phrase at 66 bpm means that, every phrase, there is an increase of 0.1 * pi in the phase domain, which is a decrease of ~0.364 s in the latency of neural responses to the phrasal boundaries; in contrast, a decrease of ~0.364 s at 85 bpm corresponds to a PPPi value of 0.128 * pi/phrase. Subsequently, we converted the PPPi values to absolute time and reconducted the statistical test but still found a significant main effect for tempo ($p < 0.001$). One reason for the significant effect of tempo can be simply that the neural correlates of musical phrase segmentation were not robustly estimated due to low signal-to-noise ratio at the low frequency. As the magnitude of beat tracking increased with tempo (Fig. 3*B*), PPPi was less robustly estimated for the phrase rate at 66 bpm than the other two tempi. The other reason could be that, when the music is slow (i.e., at 66 bpm), each chord tends to stand out and to less likely be grouped together, challenging the integration of events in phrasal structures. When the tempo increased and the phrase length decreased, the PPPi value increased quickly and, in the original condition, reached to an asymptote. Tempo may be critical to establish prediction in music listening.

In summary, phrasal phase precession demonstrates that listeners establish structural predictions at the phrasal level from the preserved harmonic structure in the original and global reversal conditions during music listening and predict future event (phrasal) boundaries. The phrase segmentation in the local reversal condition can be viewed as a passive process of phrase segmentation, as the phrasal boundaries drive neural signals, but predictive processes are only involved in the phrase segmentation in the original and global reversal conditions—a manifestation of musical expectation at the structural level. Nevertheless, this estimation of the PPPi value, which relies on the group-level method, hinders us from demonstrating how this index is related to music experience. Future studies employing neuroimaging methods with superior signal quality than EEG, such as intracranial EEG, should be utilized to investigate phrase phase precession at an individual level, thereby further validating the findings presented here.

Discussion

One of the remarkable features of our perceptual and cognitive systems is that we can track not just surface properties of naturalistic stimuli but also extract and track higher-order structural features. This capacity is especially prominent when we process continuously varying, complex stimuli such as speech and music. Investigating these internal, inferential processes has proved to be challenging. We discovered a neural signature that segments musical phrases online, in a manner that exploits predictive processes demonstrated by phase precession.

We first analyzed neural tracking of notes and beats in the spectral domain (applying novel analytic approaches to a well-studied question). Importantly, we found that note and beat tracking did not interact with the presence or integrity of phrasal musical structures (Fig. 3*A–C*). The subsequent temporal analysis showed that the neural tracking revealed an early period modulated by tempo and a late period sensitive to phrasal structure in music (Fig. 3*D–F*). We then demonstrated that EEG power of beat tracking was modulated by musical structure, reflecting musical phrase segmentation (Fig. 4). The analyses culminate in the quantification of phrasal phase precession. We show that musical phrase segmentation is a predictive process rather than being passively evoked by the phrasal boundaries (Fig. 5). Our findings support the view that listeners actively segment music streams online, conforming to high-level musical structures.

The approach shows that it is possible to study abstract musical structures at long timescales (>5 s), using new analysis approaches to noninvasive EEG recording. Previous studies have typically shied away from analyzing EEG signals at ultralow frequencies (i.e., ~0.1 Hz) due to the low signal-to-noise ratio. However, the neural signatures of phrase segmentation were reliable and could be estimated from a single music piece, allowing the possibility to study differences between individual music pieces (Figs. 4*B*, 5*A*,*F*). Although we were motivated by previous research showing characteristic responses to the endings of manipulated musical structures, such as ERAN (Koelsch et al., 2002) and CPS (Neuhaus et al., 2006), our new approach reveals neurobiological insights about music listening in natural conditions as well as neural signatures of phrase segmentation that differ in kind from ERAN and CPS. (1) The neural signature of processing the musical syntax can be directly observed (Fig. 4*B*,*C*), instead of by comparing different conditions of musical materials, as in ERAN and CPS. (2) The phrase segmentation derived from the low-level neural tracking of beat, modulated by high-level musical structures. More importantly and generally, we show that segmenting sensory (music) streams based on high-level grouping structures is a dynamic operation that exploits predictive processes (Fig. 5).

The phenomenon of phase precession has been argued to reflect predictions of future events (Jensen and Lisman, 1996; Lisman, 2005). The phrasal phase precession we show—the discovery that neural phase accelerates as music unfolds—illustrates that the structure-dependent segmentation of music streams is not solely based on the musical stimuli but driven by listeners' internal construction of schematic knowledge of segmentation (Newtson et al., 1977; Tillmann et al., 1998; Tillmann and Bigand, 2001). Listeners potentially extract structural regularities in music from the beginning of a musical piece and gradually internalize the pattern of structure of the musical pieces, which allows listeners to construct a template to segment the musical piece. We provide evidence that segmentation is an interplay between stimulus-driven processes and the construction of future events. Admittedly, the phenomenon of phrasal phase precession and its corresponding quantification require further exploration.

The musical phrase segmentation we observed probably reveals a neural signature of segmenting high-level musical structures. The original music pieces we used were selected so that they had clear and simple musical phrasal structures and facilitated the successful identification of the neural signature of musical phrase segmentation, but contribution of various music theoretical criteria, such as voice leading, group-changing rule, and cadence, should be investigated in the future. The approach applied here, particularly the TRF procedure, can be extended to genres with varying or irregular phrasal structures.

Musical phrase segmentation occurs at a timescale that can exceed ~5 s, requiring recruitment of brain areas capable of processing information over a long timescale. The core auditory system does not typically represent such long time constants in studies specifically testing timescales of acoustic processing (Overath et al., 2015; Teng et al., 2017; Donhauser and Baillet, 2020; Teng and Poeppel, 2020), though several fMRI studies on scrambled spoken language suggest representations of long time constants in the posterior temporal regions (Hasson et al.,

2008; Lerner et al., 2011). It is also possible that other brain regions send signals to modulate neural activities in the auditory system. Other candidate brain regions with long time constants are frontal areas, which have been proposed to integrate information over a long time period in language processing (Hagoort, 2005; Hickok and Poeppel, 2007; Lerner et al., 2011; Hagoort and Indefrey, 2014) and movie processing (Hasson et al., 2008) and have been demonstrated to be involved with processing musical syntax (Maess et al., 2001; Knosche et al., 2005; Koelsch et al., 2005). It is likely that the frontal areas (e.g., potentially right inferior frontal areas) establish the phrasal structures online and convey signals to modulate the neural responses of beat tracking. With regard to predicting phrasal event boundaries (Fig. 5), the hippocampus likely aids in encoding past musical structural information and fast retrieval of structural knowledge before phrasal boundaries for prediction (Baldassano et al., 2017; Ben-Yakov and Henson, 2018). Lastly, predicting the timing of auditory sequences has been shown to involve motor areas, including premotor cortex, supplementary motor area, and the basal ganglia (Grahn and Rowe, 2013; Morillon and Baillet, 2017; Assaneo and Poeppel, 2018; Rimmele et al., 2018; Assaneo et al., 2019). Therefore, we conjecture that a fast interplay between the frontal areas and hippocampus during music listening supports extracting event structures and establishing structural predictions online, with assistance from motor areas for predicting the precise timing of musical events. It would be interesting to test these hypotheses by employing MEG and/or functional MRI to pinpoint the brain areas involved in the musical phrase segmentation for a more comprehensive understanding.

We identified a robust neural signature that captures the online segmentation of music according to its high-level phrasal structure. The novel quantification of phrasal phase precession further demonstrates a predictive process of phrase segmentation during music listening. The neural signatures we highlight (musical phrase segmentation and phrasal phase precession) and the analysis procedures (EEG power modulation and PPPi) provide novel directions for studying cognitive processes of high-level structures using noninvasive recording techniques such as EEG or MEG.

References

Assaneo MF, Poeppel D (2018) The coupling between auditory and motor cortices is rate-restricted: evidence for an intrinsic speech-motor rhythm. Sci Adv 4:eaao3842.

Assaneo MF, Ripolles P, Orpella J, Lin WM, de Diego-Balaguer R, Poeppell D (2019) Spontaneous synchronization to speech reveals neural mechanisms facilitating language learning. Nat Neurosci 22:627–632.

Baldassano C, Chen J, Zadbood A, Pillow JW, Hasson U, Norman KA (2017) Discovering event structure in continuous narrative perception and memory. Neuron 95:709–721.e5.

Ben-Yakov A, Henson RN (2018) The hippocampal film editor: sensitivity and specificity to event boundaries in continuous experience. J Neurosci 38:10057–10068.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Series B Stat Methodol 57:289–300.

Brodbeck C, Hong LE, Simon JZ (2018) Rapid transformation from auditory to linguistic representations of continuous speech. Curr Biol 28:3976–3983.e5.

Broderick MP, Anderson AJ, Di Liberto GM, Crosse MJ, Lalor EC (2018) Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. Curr Biol 28:803–809.e3.

Buzsaki G (2005) Theta rhythm of navigation: link between path integration and landmark navigation, episodic and semantic memory. Hippocampus 15:827–840.

Crosse MJ, Di Liberto GM, Bednar A, Lalor EC (2016) The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. Front Hum Neurosci 10:604.

de Cheveigne A, Di Liberto GM, Arzounian D, Wong DDE, Hjortkjaer J, Fuglsang S, Parra LC (2019) Multiway canonical correlation analysis of brain data. Neuroimage 186:728–740.

de Cheveigne A, Wong DDE, Di Liberto GM, Hjortkjaer J, Slaney M, Lalor E (2018) Decoding the auditory brain with canonical component analysis. Neuroimage 172:206–216.

Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-frequency cortical entrainment to speech reflects phoneme-level processing. Curr Biol 25:2457–2465.

Di Liberto GM, Pelofi C, Bianco R, Patel P, Mehta AD, Herrero JL, de Cheveigne A, Shamma S, Mesgarani N (2020) Cortical encoding of melodic expectations in human temporal cortex. Elife 9:e51784.

Ding N, Melloni L, Zhang H, Tian X, Poeppel D (2016) Cortical tracking of hierarchical linguistic structures in connected speech. Nat Neurosci 19:158–164.

Doelling KB, Arnal LH, Ghitza O, Poeppel D (2014) Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. Neuroimage 85:761–768.

Doelling KB, Poeppel D (2015) Cortical entrainment to music and its modulation by expertise. Proc Natl Acad Sci U S A 112:E6233–E6242.

Donhauser PW, Baillet S (2020) Two distinct neural timescales for predictive speech processing. Neuron 105:385–393.e9.

Fujioka T, Ross B, Trainor LJ (2015) Beta-band oscillations represent auditory beat and its metrical hierarchy in perception and imagery. J Neurosci 35:15187–15198.

Fujioka T, Trainor LJ, Ross B, Kakigi R, Pantev C (2004) Musical training enhances automatic encoding of melodic contour and interval structure. J Cogn Neurosci 16:1010–1021.

Ghitza O (2012) On the role of theta-driven syllabic parsing in decoding speech: intelligibility of speech with a manipulated modulation spectrum. Front Psychol 3:238.

Glasberg BR, Moore BCJ (1990) Derivation of auditory filter shapes from notched-noise data. Hear Res 47:103–138.

Grahn JA, Rowe JB (2013) Finding and feeling the musical beat: striatal dissociations between detection and prediction of regularity. Cereb Cortex 23:913–921.

Gwilliams L, King JR, Marantz A, Poeppel D (2022) Neural dynamics of phoneme sequences reveal position-invariant code for content and order. Nat Commun 13:6606.

Hagoort P (2005) On Broca, brain, and binding: a new framework. Trends Cogn Sci 9:416–423.

Hagoort P, Indefrey P (2014) The neurobiology of language beyond single words. Annu Rev Neurosci 37:347–362.

Hansen NC, Kragness HE, Vuust P, Trainor L, Pearce MT (2021) Predictive uncertainty underlies auditory boundary perception. Psychol Sci 32:1416–1425.

Harding EE, Sammler D, Henry MJ, Large EW, Kotz SA (2019) Cortical tracking of rhythm in music and speech. Neuroimage 185:96–101.

Hasson U, Yang E, Vallines I, Heeger DJ, Rubin N (2008) A hierarchy of temporal receptive windows in human cortex. J Neurosci 28:2539–2550.

Hickok G, Poeppel D (2007) Opinion - the cortical organization of speech processing. Nat Rev Neurosci 8:393–402.

Hołubowska Z, Teng X, Larrouy-Maestri P (2023) The effect of temporal regularity on tracking musical phrases. In Proceedings of the international conference on music perception and cognition (ICMPC), Tokyo, Japan.

Huron D (2008) *Sweet anticipation: music and the psychology of expectation*. Cambridge, MA: MIT press.

Hyvarinen A (1999) Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw 10:626–634.

Jackendoff R (2009) Parallels and nonparallels between language and music. Music Percept 26:195–204.

Jensen O, Lisman JE (1996) Hippocampal CA3 region predicts memory sequences: accounting for the phase precession of place cells. Learn Mem 3:279–287.

Knosche TR, Neuhaus C, Haueisen J, Alter K, Maess B, Witte OW, Friederici AD (2005) Perception of phrase structure in music. Hum Brain Mapp 24:259–273.

Koelsch S, Gunter TC, Wittfoth M, Sammler D (2005) Interaction between syntax processing in language and in music: an ERP study. J Cogn Neurosci 17:1565–1577.

Koelsch S, Rohrmeier M, Torrecuso R, Jentschke S (2013) Processing of hierarchical syntactic structure in music. Proc Natl Acad Sci U S A 110:15443–15448.
Koelsch S, Schmidt BH, Kansok J (2002) Effects of musical expertise on the early right anterior negativity: an event-related brain potential study. Psychophysiology 39:657–663.
Koelsch S, Siebel WA (2005) Towards a neural basis of music perception. Trends Cogn Sci 9:578–584.
Koelsch S, Vuust P, Friston K (2019) Predictive processes and the peculiar case of music. Trends Cogn Sci 23:63–77.
Kragness HE, Trainor LJ (2016) Listeners lengthen phrase boundaries in self-paced music. J Exp Psychol Hum Percept Perform 42:1676–1686.
Kurby CA, Zacks JM (2008) Segmentation in the perception and memory of events. Trends Cogn Sci 12:72–79.
Lalor EC, Pearlmutter BA, Reilly RB, McDarby G, Foxe JJ (2006) The VESPA: a method for the rapid estimation of a visual evoked potential. Neuroimage 32:1549–1561.
Larrouy-Maestri P, Pfordresher PQ (2018) Pitch perception in music: do scoops matter? J Exp Psychol Hum Percept Perform 44:1523–1541.
Lashley KS (1951) The problem of serial order in behavior. In: *Cerebral mechanisms in behavior; the Hixon symposium*, pp 112–146. Oxford: Wiley.
Lenc T, Keller PE, Varlet M, Nozaradan S (2018) Neural tracking of the musical beat is enhanced by low-frequency sounds. Proc Natl Acad Sci U S A 115:8221–8226.
Lerdahl F, Jackendoff R (1983) An overview of hierarchical structure in music. Music Percept 1:229–252.
Lerner Y, Honey CJ, Silbert LJ, Hasson U (2011) Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. J Neurosci 31:2906–2915.
Lisman J (2005) The theta/gamma discrete phase code occuring during the hippocampal phase precession may be a more general brain coding scheme. Hippocampus 15:913–922.
Maess B, Koelsch S, Gunter TC, Friederici AD (2001) Musical syntax is processed in Broca's area: an MEG study. Nat Neurosci 4:540–545.
Mankel K, Bidelman GM (2018) Inherent auditory skills rather than formal music training shape the neural encoding of speech. Proc Natl Acad Sci U S A 115:13129–13134.
Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG-and MEG-data. J Neurosci Methods 164:177–190.
Martin AE (2020) A compositional neural architecture for language. J Cogn Neurosci 32:1407–1427.
Morillon B, Baillet S (2017) Motor origin of temporal predictions in auditory attention. Proc Natl Acad Sci U S A 114:E8913–E8921.
Mullensiefan D, Gingras B, Musil J, Stewart L (2014) The musicality of non-musicians: an index for assessing musical sophistication in the general population. PLoS One 9:e89642.
Neuhaus C, Knosche TR, Friederici AD (2006) Effects of musical expertise and boundary markers on phrase perception in music. J Cogn Neurosci 18:472–493.
Newtson D, Engquist GA, Bois J (1977) The objective basis of behavior units. J Pers Soc Psychol 35:847–862.
Nozaradan S, Peretz I, Missal M, Mouraux A (2011) Tagging the neuronal entrainment to beat and meter. J Neurosci 31:10234–10240.
Nozaradan S, Peretz I, Mouraux A (2012) Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. J Neurosci 32:17572–17581.
Oostenveld R, Fries P, Maris E, Schoffelen JM (2011) FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci 2011:156869.
Overath T, McDermott JH, Zarate JM, Poeppel D (2015) The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat Neurosci 18:903–911.
Patel AD (2003) Language, music, syntax and the brain. Nat Neurosci 6:674–681.
Patel AD, Morgan E (2017) Exploring cognitive relations between prediction in language and music. Cogn Sci 41:303–320.
Pearce MT (2005) The construction and evaluation of statistical models of melodic structure in music perception and composition. City University London.
Peelle JE, Gross J, Davis MH (2013) Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cereb Cortex 23:1378–1387.
Poeppel D, Assaneo MF (2020) Speech rhythms and their neural foundations. Nat Rev Neurosci 21:322–334.
Qasim SE, Fried I, Jacobs J (2021) Phase precession in the human hippocampus and entorhinal cortex. Cell 184:3242–3255.e10.
Rimmele JM, Morillon B, Poeppel D, Arnal LH (2018) Proactive sensing of periodic and aperiodic auditory patterns. Trends Cogn Sci 22:870–882.
Rohrmeier M (2011) Towards a generative syntax of tonal harmony. J Math Music 5:35–53.
Rohrmeier MA, Koelsch S (2012) Predictive information processing in music cognition. A critical review. Int J Psychophysiol 83:164–175.
Saberi K, Perrott DR (1999) Cognitive restoration of reversed speech. Nature 398:760–760.
Salinas E, Sejnowski TJ (2001) Correlated neuronal activity and the flow of neural information. Nat Rev Neurosci 2:539–550.
Silva S, Barbosa F, Marques-Teixeira J, Petersson KM, Castro SL (2014) You know when: event-related potentials and theta/beta power indicate boundary prediction in music. J Integr Neurosci 13:19–34.
Teng X, Cogan GB, Poeppel D (2019) Speech fine structure contains critical temporal cues to support speech segmentation. Neuroimage 202:116152.
Teng X, Ma M, Yang J, Blohm S, Cai Q, Tian X (2020) Constrained structure of ancient Chinese poetry facilitates speech content grouping. Curr Biol 30:1299–1305.e97.
Teng X, Poeppel D (2020) Theta and gamma bands encode acoustic dynamics over wide-ranging timescales. Cereb Cortex 30:2600–2614.
Teng X, Tian X, Poeppel D (2016) Testing multi-scale processing in the auditory system. Sci Rep 6:34390.
Teng X, Tian X, Rowland J, Poeppel D (2017) Concurrent temporal channels for auditory processing: oscillatory neural entrainment reveals segregation of function at different scales. PLoS Biol 15:e2000812.
Tillmann B (2012) Music and language perception: expectations, structural integration, and cognitive sequencing. Top Cogn Sci 4:568–584.
Tillmann B, Bigand E (2001) Global context effect in normal and scrambled musical sequences. J Exp Psychol Hum Percept Perform 27:1185–1196.
Tillmann B, Bigand E, Pineau M (1998) Effects of global and local contexts on harmonic expectancy. Music Percept 16:99–117.
Tolman EC (1948) Cognitive maps in rats and men. Psychol Rev 55:189–208.
Trainor LJ, Trehub SE (1992) A comparison of infants and adults sensitivity to western musical structure. J Exp Psychol Hum Percept Perform 18:394–402.
Vuust P, Heggli OA, Friston KJ, Kringelbach ML (2022) Music in the brain. Nat Rev Neurosci 23:287–305.
Vuust P, Ostergaard L, Pallesen KJ, Bailey C, Roepstorff A (2009) Predictive coding of music - brain responses to rhythmic incongruity. Cortex 45:80–92.
Zacks JM, Swallow KM (2007) Event segmentation. Curr Dir Psychol Sci 16:80–84.