frontiers in Neuroscience

ORIGINAL RESEARCH
published: 16 July 2021
doi: 10.3389/fnins.2021.665767

An Analytical Framework of Tonal and Rhythmic Hierarchy in Natural Music Using the Multivariate Temporal Response Function

Jasmine Leahy[1†], *Seung-Goo Kim*[1*†], *Jie Wan*[1,2] *and Tobias Overath*[1,3,4*]

[1] Department of Psychology and Neuroscience, Duke University, Durham, NC, United States, [2] Department of Cognitive Sciences, University of California, Irvine, Irvine, CA, United States, [3] Duke Institute for Brain Sciences, Duke University, Durham, NC, United States, [4] Center for Cognitive Neuroscience, Duke University, Durham, NC, United States

Even without formal training, humans experience a wide range of emotions in response to changes in musical features, such as tonality and rhythm, during music listening. While many studies have investigated how isolated elements of tonal and rhythmic properties are processed in the human brain, it remains unclear whether these findings with such controlled stimuli are generalizable to complex stimuli in the real world. In the current study, we present an analytical framework of a linearized encoding analysis based on a set of music information retrieval features to investigate the rapid cortical encoding of tonal and rhythmic hierarchies in natural music. We applied this framework to a public domain EEG dataset (OpenMIIR) to deconvolve overlapping EEG responses to various musical features in continuous music. In particular, the proposed framework investigated the EEG encoding of the following features: *tonal stability*, *key clarity*, *beat*, and *meter*. This analysis revealed a differential spatiotemporal neural encoding of *beat* and *meter*, but not of *tonal stability* and *key clarity*. The results demonstrate that this framework can uncover associations of ongoing brain activity with relevant musical features, which could be further extended to other relevant measures such as time-resolved emotional responses in future studies.

Keywords: linearized encoding analysis, electroencephalography, tonal hierarchy, rhythmic hierarchy, naturalistic paradigm

Edited by:
Matthew Elliott Sachs,
Columbia University, United States

Reviewed by:
Dan Zhang,
Tsinghua University, China
Koichi Yokosawa,
Hokkaido University, Japan

Correspondence:
Seung-Goo Kim
solleo@gmail.com
Tobias Overath
t.overath@duke.edu

[†]These authors have contributed equally to this work and share first authorship

Specialty section:
This article was submitted to Auditory Cognitive Neuroscience, a section of the journal Frontiers in Neuroscience

Received: *08 February 2021*
Accepted: *24 June 2021*
Published: *16 July 2021*

Citation:
Leahy J, Kim S-G, Wan J and Overath T (2021) An Analytical Framework of Tonal and Rhythmic Hierarchy in Natural Music Using the Multivariate Temporal Response Function. Front. Neurosci. 15:665767. doi: 10.3389/fnins.2021.665767

INTRODUCTION

Music is a universal auditory experience known to evoke intense feelings. Even without musical training, humans not only connect to it on an emotional level but can also generate expectations as they listen to it (Koelsch et al., 2000). We gather clues from what we are listening to in real-time combined with internalized musical patterns, or schema, from our respective cultural settings to guess what will happen next, which ultimately results in a change in our emotions. Schemata consist of musical features, such as tonality (i.e., pitches and their relationship to one another) and rhythm. However, tonality has often been studied using heavily contrived chord progressions instead of more natural, original music in order to impose rigorous controls on the experiment (Fishman et al., 2001; Loui and Wessel, 2007; Koelsch and Jentschke, 2010). Likewise, beat perception studies have favored simplistic, isolated rhythms over complex patterns found in everyday music (Snyder and Large, 2005; Fujioka et al., 2009). Therefore, designs that take advantage of the multiple, complex features of natural music stimuli are needed to confirm the results of these experiments.

In order to devise a framework that can account for these complexities, we first considered how different musical features build up over the course of a piece of music. In everyday music, tonality and rhythm are constructed hierarchically, meaning some pitches in certain positions (e.g., in a bar) have more importance than others (Krumhansl and Shepard, 1979). One way that listeners assess this importance is *via* the temporal positions of pitches. Tones that occur at rhythmically critical moments in a piece allow us to more easily anticipate what we should hear next and when (Krumhansl, 1990; Krumhansl and Cuddy, 2010). This type of beat perception is considered hierarchical in the sense that it involves multiple layers of perception which interact with one another, namely beat and meter. Beat refers to the onset of every beat in a given measure, whereas meter refers to the importance of the beats relative to a given time signature (e.g., 4/4). Music listening has repeatedly been linked with activation of the motor cortex, in particular relating to anticipation of the beat (Zatorre et al., 2007; Chen et al., 2008; Gordon et al., 2018). The clarity of the beat matters during music perception as well; during moments of high beat saliency, functional connectivity increases from the basal ganglia and thalamus to the auditory and sensorimotor cortices and cerebellum, while low beat saliency correlates with increased connectivity between the auditory and motor cortices, indicating that we participate in an active search to find the beat when it becomes less predictable (Toiviainen et al., 2020). EEG studies, in particular, have shed light on how humans entrain beat and meter on both a micro-scale (e.g., milliseconds) (Snyder and Large, 2005; Fujioka et al., 2009) and macro-scale (e.g., years of genre-specific musical training) (Bianco et al., 2018). For example, it only requires a brief musical sequence to observe beta band activity (14–30 Hz) that increases after each tone, then decreases, creating beta oscillations synchronized to the beat of the music (Fujioka et al., 2009). Gamma band activity (∼30–60 Hz) also increases after each tone, even when a tone that was supposed to occur is omitted, suggesting that gamma oscillations represent an endogenous mechanism of beat anticipation (Fujioka et al., 2009). It was further found that phase-locked, evoked gamma band activity increases about 50 ms after tone onset and diminishes when tones are omitted, showing larger responses during accented beats vs. weak ones, which suggests a neural correlate for meter (Snyder and Large, 2005). Therefore, the aim of our study was to set up a continuous music framework that is not only able to detect encoding of beat and meter, but also able to distinguish between the two.

Tonality is another key component of real-life music listening. We learn what notes or chords will come next in a piece of music based, in part, on the statistical distribution, or frequency, of tones or sequences of tones (Krumhansl and Shepard, 1979). From these observations, Krumhansl and Cuddy (2010) derived the concept of tonal hierarchy, which describes the relative importance of tones in a musical context. By organizing tones in this way, humans assemble a psychological representation of the music based on tonality and rhythm. A few studies have attempted to develop multivariate frameworks that account for this prediction-driven, hierarchical nature of music. For example, Di Liberto et al. (2020) used EEG paired with continuous music stimuli to investigate the relative contributions of acoustic vs. melodic features of music to the cortical encoding of melodic expectation. However, they used monophonic melodies, rather than harmonic, complex music that we would hear in everyday life. They analyzed the EEG data with a useful tool for continuous stimuli, the Multivariate Temporal Response Function (mTRF) MATLAB toolbox, which maps stimulus features to EEG responses by estimating linear transfer functions (Crosse et al., 2016). Sturm et al. (2015) also used ridge regression with temporal embedding to calculate correlations between brain signal and music. Even though they used natural, complex piano music, they chose the power slope of the audio signal as a predictor, which is considered a basic acoustic measure that underlies more complex features such as beat and meter.

Building on the groundwork of these previous multivariate music analyses, we used the mTRF to analyze high-level tonal and rhythmic features of natural, continuous music stimuli extracted with the Music Information Retrieval (MIR) MATLAB toolbox (Lartillot and Toiviainen, 2007). The proposed framework aims to better understand how we process everyday music.

In an attempt to model Krumhansl's hierarchical organization of musical features, we also expanded on the features provided in the MIR toolbox to further enhance ecological validity. For example, *key clarity,* which measures how tonally similar a given frame of music is to a given key signature, has been used in several studies (Alluri et al., 2012; Sturm et al., 2015; Burunat et al., 2016), yet may not provide an accurate measurement of a musical event's tonality within the context of the entire musical excerpt. This motivated us to develop a novel feature called *tonal stability*, which contextualizes a particular musical event with respect to the tonal history thus far. *Tonal stability* quantifies the tonal hierarchy of a piece of music by taking the angular similarity between the *key strength* of a certain frame and the averaged *key strength* up until that frame. This allows us to determine how stable a musical event (or a frame) is within a given tonal hierarchy. In other words, it calculates how related the chord implied in an individual frame is to the overarching key, which is derived from a cumulative moving average. By continuously measuring local changes in tonal key centers with respect to the whole musical excerpt, we approximated the ongoing perception of tonal stability. To our knowledge, no prior study has developed such an analytical framework for combining MIR toolbox features with the mTRF to investigate how tonal and rhythmic features are encoded in the EEG signal during the listening of natural music.

We applied our framework to a public domain EEG dataset, the Open Music Imagery Information Retrieval dataset (Stober, 2017), to test the differential cortical encoding of tonal and rhythmic hierarchies. Using model comparisons, we inferred the contribution of individual features in EEG prediction. We show novel ecological evidence confirming and expanding Krumhansl's theory on how frequency and placement of musical features affect our perception and predictions (Krumhansl and Cuddy, 2010).

MATERIALS AND METHODS

The approach we used in the current study is known as linearized modeling of a sensory system (Wu et al., 2006), which has been successfully applied to M/EEG data (Lalor et al., 2006; Di Liberto et al., 2015; Brodbeck et al., 2018) as well as fMRI data (Kay et al., 2008; Huth et al., 2016) in response to naturalistic visual and auditory stimuli. The key idea of the approach is a linearization function (Wu et al., 2006), which captures the nonlinearity of stimulus-response mapping and provides an efficient parameterization of relevant aspects of a stimulus that can be linearly associated with its corresponding response. In this section, we will explain our linearization functions (i.e., musical features), the specifications of the analyzed public data, and practical details of the analysis, which was carried out using MATLAB (RRID:SCR_001622; R2020a) unless otherwise noted.

Musical Features

We considered a variety of features for the construction of our analytical models. The foundational feature in all models was the temporal *envelope* of the auditory stimulus, which contains low-level acoustic features such as amplitude variations. For tonal hierarchy, we used *key clarity* and *tonal stability* as our two additional features. For rhythmic hierarchy, we looked at the onset of every *beat* and their relative strengths within the given *meter*.

Acoustic Feature

A whole-spectrum *envelope* was calculated as the absolute value of the Hilbert transform of the musical signal. The envelope was down-sampled to the EEG's sampling rate after anti-aliasing high-pass filtering. This feature describing whole-spectrum acoustic energy served as a baseline for other models adding tonal and rhythmic features.

Tonal Features

As for high-level tonal features, we computed *key clarity* and *tonal stability*. Key clarity was derived from the MIR toolbox[1] (v1.7) function mirkeystrength, which computes a 24-dimensional vector of Pearson correlation coefficients corresponding to each of the 24 possible keys (12 major and 12 minor), which is called a *key strength* vector (Gómez, 2006). *Key clarity* is defined by the maximal correlation coefficient, which measures how strongly a certain key is implied in a given frame of interest (Lartillot and Toiviainen, 2007).

Our novel *tonal stability* feature was designed to contextualize the *key strength* with respect to the overall *key strength* of a given musical piece. It is computed with an angular similarity between the *key strength* vector of a single frame and a cumulative average of *key strength* vectors up until the adjacent previous frame as:

$$\mathrm{s(t)} = 1 - \frac{\cos^{-1}\cos\theta}{\pi} = 1 - \frac{1}{\pi}\cos^{-1}\frac{\mathbf{v}(\mathrm{t})\cdot\bar{\mathbf{v}}(\mathrm{t}-1)}{||\mathbf{v}(\mathrm{t})||\cdot||\bar{\mathbf{v}}(\mathrm{t}-1)||} \quad (1)$$

where s(t) is *tonal stability* of the *t*-th frame, is an angle between the two *key strength* vectors, $\mathbf{v}$(t) is a *key strength* vector of the *t*-th frame, and $\bar{\mathbf{v}}(\mathrm{j}) = \sum_{i=1}^{j}\mathbf{v}(i)/j$ is a cumulative moving average of *key strength* vectors from the first to the *j*-th frame. The angular similarity is bounded between 0 and 1, inclusively (1 when two vectors are parallel, 0.5 when orthogonal, and 0 when opposite). Thus, the *tonal stability* is also bounded between 0 and 1: 0 when key strength vectors are in opposite directions (i.e., implied keys are most distant on the cycle of fifths; in other words, they share few common tones).

Using the tonal hierarchy profile (Krumhansl and Shepard, 1979) as an ideal chromagram, which yields a maximal *key strength* of one (**Figure 1A**), it can be shown that if a chromagram implies a distant key (e.g., C-major key in the F#-major key context), its *tonal stability* would be close to zero. A geometrical appreciation of the relations of *key strength* vectors can be made by a low-dimensional projection using principal component analysis (PCA). The first two principal components explained 65% of the total variance of all *key strength* vectors. When the *key strength* vectors of the 12 major keys are projected to the 2-dimensional plane of the first two principal components (**Figure 1B**), it becomes clear that the *key strength* vectors of C-major and F#-major are geometrically opposing. Therefore, the (high-dimensional) angular similarity between them would be close to zero (**Figure 1C**, marked by an arrow; not exactly zero because of higher dimensions that are not visualized), which is our definition of the *tonal stability* feature. On the other hand, the *key clarity* can be seen as the maximal projection to any of the 24 possible dimensions (i.e., maximal intensity projection). Therefore, it is constant regardless of context. In other words, the *tonal stability* quantifies how tonally stable a particular frame is within the context of the entire piece, whereas *key clarity* describes how strongly a tonal structure is implicated in an absolute sense (see **Supplementary Figure 1** for an example comparison).

The length of a time window to compute spectrograms should be long enough to cover the lower bound of pitch (i.e., 30 Hz; Pressnitzer et al., 2001) but also not too long to exceed the physiologically relevant spectral range. In the current work, we used a sparse encoding of tonal features based on the estimated beats and measures (see section "Rhythmic Features"). Specifically, at each beat (or measure), a time window was defined from the current to the next beat (or measure). For each time window, the spectrogram, cochleogram, and *key strength* vectors were estimated using the MIR function mirkeystrength, and the *key clarity* and *tonal stability* were calculated as described above. The approach of the sparse encoding is similar to assigning the "semantic dissimilarity" value of a word at the onset in a natural speech study, where N400-like temporal response functions (TRFs) were found (Broderick et al., 2018), and modeling the melodic entropy at the onset of a note (Di Liberto et al., 2020). Previous studies have found an early component (i.e., ERAN; Koelsch et al., 2003) in response to violations within local tonal contexts and a late component (i.e., N400; Zhang et al., 2018) during more global contexts. Therefore, *tonal stability* was expected to be encoded within these latencies.

[1] http://bit.ly/mirtoolbox

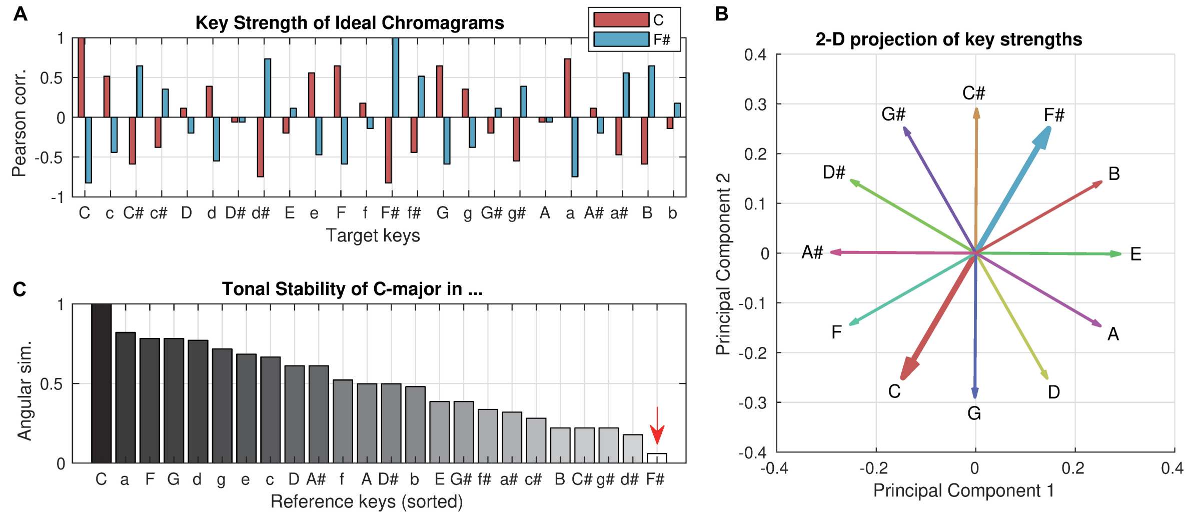


FIGURE 1 | Key strength and tonal stability. **(A)** *Key strength* of the C-major profile (red) and F#-major profile (blue) are computed for all 24 target keys (i.e., Pearson correlation between profiles). The profiles were used as ideal chromagrams that yield a maximal *key clarity* of one. Upper cases represent major keys, and lower cases represent minor keys. **(B)** For geometrical intuition of *tonal stability*, the *key strength* vectors of 12 major keys are projected on the 2-dimensional plane of the first two principal components, which together explained 65% of the total variance. **(C)** *Tonal stability* of the C-major profile with respect to all 24 reference keys (i.e., in all contexts) are computed (i.e., the angular similarity between *key strength* vectors) and sorted in descending order. Note that *key clarity* of the C-major profile is always one while its *tonal stability* varies depending on the reference key (i.e., context).

Rhythmic Features

As low-level rhythmic feature, we used *beats* (Grahn and McAuley, 2009; Stober, 2017). *Beats* were extracted using the dynamic beat tracker in the Librosa library[2] and included in the shared public data. We modeled *beats* using a unit impulse function (i.e., 1's at beats, 0's otherwise).

As high-level rhythmic feature, we used *meter,* which was based on *beats*. We weighted the strength of each beat in a musical excerpt, according to a beat accent system that is most prevalent in Western classical music, by separating beats into three tiers: strong, middle, and weak (Grahn and Rowe, 2009; Vuust and Witek, 2014). A separate unit impulse function was created for each of the three levels. Note that the tiers correspond to the strength of a beat, not the position (or phase) within a measure. The breakdown applies as follows:

4/4 meter signature: beat 1=strong; beat 2=weak; beat 3=middle; and beat 4=weak.

3/4 meter signature: beat 1=strong; beat 2=weak; and beat 3=weak.

OpenMIIR Dataset

We used the public domain Open Music Imagery Information Retrieval Dataset available on Github[3], which is designed to facilitate music cognition research involving EEG and the extraction of musical features. Given that we only analyzed a subset of the dataset, we will only summarize the relevant materials and methods. Complete details of the experimental procedure can be found in the original study (Stober, 2017).

Participants

Data was collected from ten participants. One participant was excluded from the dataset due to coughing and movement-related artifacts, resulting in a total of nine participants. Seven participants were female, and two were male. The average age of the participants was 23. Participants filled out a questionnaire asking about their musical playing and listening background. Seven out of the nine participants were musicians, which was defined as having engaged in a regular, daily practice of a musical instrument (including voice) for one or more years. The average number of years of daily musical practice was 5.4 years. The average number of formal years of musical training was 4.9 years.

Prior to the EEG recording, participants were asked to name and rate how familiar they were with the 12 stimuli of the experiment. Also, before the EEG experiment, they were asked to tap/clap along to the beat, which was then given a score by the researcher based on accuracy. Seven participants were given 100% on their ability to tap along to the beat, and two were given a 92%. All participants were familiar with 80% or more of the musical stimuli.

Stimuli

There were 12 different, highly familiar musical excerpts that ranged between 6.9 and 13.9 s, with an average duration of 10.5 s each. Exactly half of the songs had a 3/4 time signature, and the other songs had a 4/4 time signature. **Table 1** lists the popular songs that the stimuli were taken from. The tonal features of these stimuli are shown in **Figure 2**. The two features were not significantly correlated in any of the stimuli (minimum uncorrected-$p = 0.08$).

[2] https://github.com/bmcfee/librosa

[3] https://github.com/sstober/openmiir

TABLE 1 | Descriptive statistics of tonal features.

Stim#	Title	Duration (sec)	BPM	BPB	Key clarity (mean ± SD)	Tonal stability (mean ± SD)	Corr. (*p*-value)	Key clarity (mean ± SD)	Tonal stability (mean ± SD)	Corr. (*p*-value)
1	Chim Chim Cheree (lyrics)	13.3	213	3	0.53 ± 0.15	0.63 ± 0.19	0.17 (0.26)	0.52 ± 0.13	0.60 ± 0.24	0.16 (0.57)
2	Take me out to the ballgame (lyrics)	7.7	189	3	0.56 ± 0.14	0.63 ± 0.18	0.19 (0.42)	0.59 ± 0.15	0.60 ± 0.29	0.64 (0.12)
3	Jingle Bells (lyrics)	9.7	200	4	0.51 ± 0.16	0.65 ± 0.18	−0.09 (0.66)	0.55 ± 0.12	0.58 ± 0.28	0.14 (0.73)
4	Mary Had a Little Lamb (lyrics)	11.6	160	4	0.65 ± 0.11	0.68 ± 0.16	0.25 (0.19)	0.67 ± 0.12	0.64 ± 0.28	0.10 (0.81)
11	Chim Chim Cheree (no lyrics)	13.9	206	3	0.69 ± 0.10	0.72 ± 0.21	−0.00 (1.00)	0.70 ± 0.11	0.69 ± 0.27	−0.34 (0.22)
12	Take me out to the ballgame (no lyrics)	7.9	185	3	0.65 ± 0.12	0.69 ± 0.18	0.02 (0.95)	0.71 ± 0.04	0.67 ± 0.28	−0.08 (0.85)
13	Jingle Bells (no lyrics)	9.0	200	4	0.60 ± 0.12	0.72 ± 0.19	0.34 (0.08)	0.54 ± 0.10	0.63 ± 0.31	0.14 (0.76)
14	Mary Had a Little Lamb (no lyrics)	12.2	160	4	0.76 ± 0.09	0.81 ± 0.16	0.15 (0.44)	0.71 ± 0.10	0.78 ± 0.32	−0.33 (0.43)
21	Emperor Waltz	8.3	175	3	0.76 ± 0.11	0.76 ± 0.19	0.14 (0.54)	0.78 ± 0.14	0.71 ± 0.30	−0.15 (0.72)
22	Harry Potter theme	16.0	166	3	0.67 ± 0.16	0.68 ± 0.26	−0.03 (0.84)	0.72 ± 0.13	0.63 ± 0.31	−0.12 (0.69)
23	Star Wars theme	9.2	104	4	0.66 ± 0.16	0.70 ± 0.25	0.19 (0.50)	0.65 ± 0.11	0.68 ± 0.46	0.48 (0.52)
24	Eine Kleine Nachtmusik	6.9	140	4	0.64 ± 0.07	0.67 ± 0.23	0.10 (0.75)	0.69 ± 0.10	0.51 ± 0.36	−0.09 (0.91)

BPM, beats per minute; BPB, beats per bar; Corr., Pearson correlation coefficient between key clarity and tonal stability.

Procedure

We analyzed the "Perception" condition, which was the first out of four experimental conditions. The rest of the conditions involved musical imagery tasks, which we did not include in our analysis. Each condition consisted of five blocks. All 12 stimuli were played in a randomized order once per block. This resulted in a total of 60 trials for each condition (i.e., five repetitions per stimulus). In each trial, a stimulus was preceded by two measures of cue beats.

Data Acquisition and Preprocessing

Neural signals were measured during the experiment using a BioSemi Active-Two EEG system in 64 channels at a sampling rate of 512 Hz. Independent components associated with ocular and cardiac artifacts were detected using the MNE-python[4] (RRID:SCR_005972; v0.20.7) (Gramfort et al., 2013), of which demixing matrices were also included in the open dataset. After projecting out the artifact-related components using mne.preprocessing.ica.apply, the EEG signal was converted to handle in EEGLAB[5] (RRID:SCR_007292; v14.1.2). Then, the data was bandpass-filtered between 1 and 8 Hz using Hamming windowed sinc finite impulse response (FIR) filter using pop_eegfiltnew as the low-frequency activity was previously found to encode music-related information (Di Liberto et al., 2020). Trials were epoched using pop_epoch between 100 ms after music onset (i.e., after beat cues) and 100 ms before music offset with a window length of 200 ms for tonal feature extraction. The EEG signal was then down-sampled to 128 Hz (pop_resample) and normalized by Z-scoring each trial.

mTRF Analysis

Model Prediction

The linearized encoding analysis was carried out using the mTRF MATLAB Toolbox[6] (v2.1) created by Crosse et al. (2016). In a FIR model, we fit a set of lagged stimulus features to response timeseries to estimate time-varying causal impacts of features to the response timeseries:

$$\mathrm{y(t)} = \sum_{d=0}^{D} x(t-d)b(d)$$

where y(t) and x(t) are a response and a feature at a time point t, respectively, and $b(d)$ is a weight that represents the impact of a feature at a delay d. A timeseries of these weights (i.e., a transfer function or a kernel of a linear filter) is called a TRF. In the mTRF encoding analysis, we use a regularized regression (e.g., ridge) to estimate TRFs where multicollinearity exists among multiple features. The encoding analysis is performed at each channel at a time (i.e., multiple independent variables and a univariate dependent variable). The validity of the estimated TRFs is often tested *via* cross-validation (i.e., convolving test features with a kernel estimated from a training set to predict test responses).

When considering multiple features, the FIR model can be expressed in a matrix form:

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon \tag{2}$$

where $\mathbf{y} \in \mathbb{R}^{T \times 1}$ is an EEG response vector from a given channel over T time points, $\mathbf{X} \in \mathbb{R}^{T \times PD}$ is a feature matrix of which columns are P features lagged over D time points (i.e., a Toeplitz matrix), $\beta \in \mathbb{R}^{1 \times PD}$ is a vector of unknown weights, and $\varepsilon \in \mathbb{R}^{T \times 1}$ is a vector of Gaussian noise with unknown serial correlation. Note that a feature set could consist of multiple sub-features (e.g., 16-channel cochleogram and 3-channel meter). The vector β is concatenated weights over D delays for P features. In the current analysis, we column-wise normalized **y** and **X** by taking Z-scores per trial.

A ridge solution of Eq. 2 is given (Hoerl and Kennard, 1970) as:

$$\widehat{\beta}(\lambda) = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X} + \lambda\mathbf{I}\right)^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\mathbf{y}, \tag{3}$$

[4] https://mne.tools/stable/index.html

[5] https://sccn.ucsd.edu/eeglab/index.php

[6] https://github.com/mickcrosse/mTRF-Toolbox

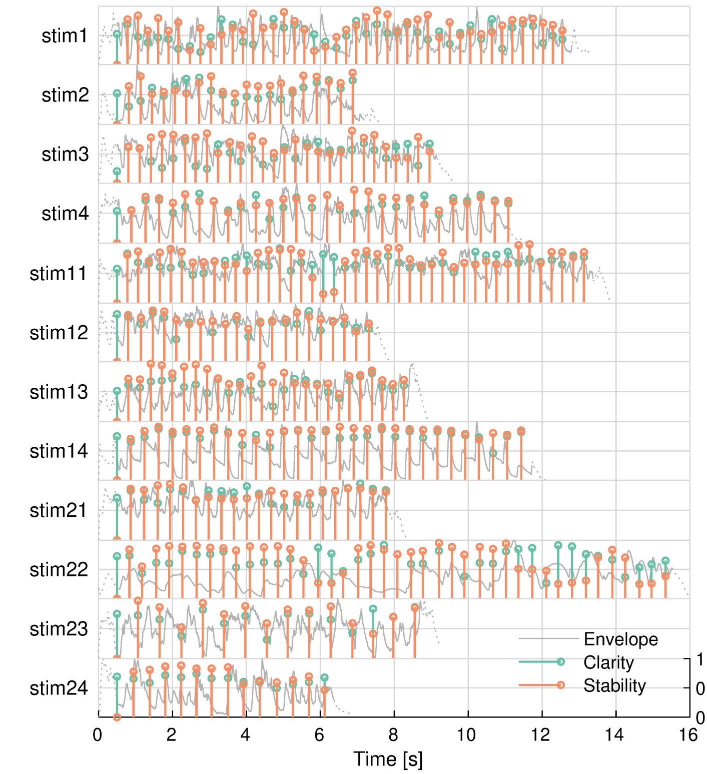


FIGURE 2 | Tonal features of musical stimuli. *Envelope* (gray), *key clarity* (green), and *tonal stability* (orange) are shown. Envelopes outside the analysis window are shown in dashed lines. Stimulus IDs are noted on the left.

where $\mathbf{I} \in \mathbb{R}^{PD \times PD}$ is an identity matrix and λ 0 is a regularization parameter that penalizes (i.e., shrinks) estimates. That is, the ridge estimates are dependent on the selection of regularization. The lambda was optimized on training data (i.e., a lambda that yields the maximal prediction accuracy for each channel), and the validity of this model was tested on testing data (i.e., predicting EEG response based on given features) through the leave-one-out cross-validation scheme using mTRFcrossval, mTRFtrain, and mTRFevalute. The prediction accuracy was measured by the Pearson correlation coefficient. Specifically, we used 79 delays from −150 ms to 450 ms and 21 loglinearly spaced lambda values from 2^{-10} to 2^{10}. We discarded time points where the kernel exceeded trial boundaries (i.e., valid boundary condition) to avoid zero-padding artifacts (e.g., high peaks at zero-lag from short trials).

Model Comparison

We created multiple models with varying terms and compared prediction accuracies to infer the significance of encoding of a specific feature in the responses. The families of models were:

$$\mathbf{y} = [\mathbf{X}_{env}]\,[\beta_{env}] + \varepsilon \tag{4}$$

$$\mathbf{y} = \begin{bmatrix} \mathbf{X}_{env}\ \mathbf{X}_{beat} \end{bmatrix} \begin{bmatrix} \beta_{env} \\ \beta_{beat} \end{bmatrix} + \varepsilon \tag{5-1}$$

$$\mathbf{y} = \begin{bmatrix} \mathbf{X}_{env}\ \mathbf{X}_{meter} \end{bmatrix} \begin{bmatrix} \beta_{env} \\ \beta_{meter} \end{bmatrix} + \varepsilon \tag{5-2}$$

$$\mathbf{y} = \begin{bmatrix} \mathbf{X}_{env}\ \mathbf{X}_{meter}\ \mathbf{X}_{clarity} \end{bmatrix} \begin{bmatrix} \beta_{env} \\ \beta_{meter} \\ \beta_{clarity} \end{bmatrix} + \varepsilon \tag{6-1}$$

$$\mathbf{y} = \begin{bmatrix} \mathbf{X}_{env}\ \mathbf{X}_{meter}\ \mathbf{X}_{stability} \end{bmatrix} \begin{bmatrix} \beta_{env} \\ \beta_{meter} \\ \beta_{stability} \end{bmatrix} + \varepsilon \tag{6-2}$$

where $\mathbf{X}_\mathbf{i}$ and $\beta_\mathbf{i}$ are a Toeplitz matrix and a weight vector for the *i*-th feature, respectively. Equation 4 served as a baseline model and Eq. 5 are rhythmic models and Eq. 6 are tonal models while covarying rhythmic features. Comparisons of interest were: (a) Eq. 5-1 vs. Eq. 4, (b) Eq. 5-2 vs. Eq. 5-1, (c) Eq. 6-1 vs. Eq. 5-2, and (d) Eq. 6-2 vs. Eq. 5-2. Note that the comparisons were made to infer the effect of the addition of each feature despite their multicollinearity. That is, if there is no uniquely explained variance by the last term, the full model (with the last term) cannot yield greater prediction accuracy than the reduced model (without the last term).

Cluster-based Monte Carlo permutation test (Maris and Oostenveld, 2007), using ft_statistics_montecarlo in FieldTrip[7]

[7] https://www.fieldtriptoolbox.org/

(RRID:SCR_004849; v20180903), was used to calculate cluster-wise p-values of one paired t-test on differences in prediction accuracies across all channels with the summed t-statistics as a cluster statistic. 10,000 permutations with replacement were made to generate null distributions. In permutation tests, a cluster-forming threshold does not affect the family-wise error rate (FWER) but only sensitivity (see Maris, 2019 for formal proof). Thus, clusters were defined at an arbitrary threshold of the alpha-level of 0.05, and the cluster-wise p-values are thresholded at the alpha-level of 0.05 to control the FWER to 0.05.

To estimate the variation of the point estimate of a prediction accuracy difference, we bootstrapped cluster-mean prediction accuracies for 10,000 times to compute 95% confidence intervals. For the visualization of results, modified versions of topoplot in EEGLAB and cat_plot_boxplot in CAT12[8] are used.

Control Analysis

To demonstrate the false positive control and the sensitivity of the current procedure, we randomized the phases of envelopes (Menon and Levitin, 2005; Abrams et al., 2013; Farbood et al., 2015; Kaneshiro et al., 2020) to create control features with disrupted temporal structure, but with identical spectra. If the prediction is not due to the encoding of temporal information, this control feature (i.e., phase-randomized envelope) would be expected to explain the EEG data as well as the original envelope. Specifically, the phases of envelopes were randomized *via* fast Fourier transform (FFT) and inverse FFT for each stimulus. That is, within each randomization, the randomized envelope was identical throughout repeated representations over trials. MATLAB's fft and ifft were used. The phase randomization, model optimization, and model evaluation processes were repeated 50 times across all participants. Then, the prediction accuracies averaged across phase-randomizations were compared with the prediction accuracies with the actual envelopes using the cluster-based Monte Carlo permutation test with the same alpha-levels as in the main analysis.

RESULTS

Envelope Tracking

The control analysis revealed that the mTRF analysis sensitively detects envelope tracking compared to models with phase-randomized envelopes (**Figure 3**). In a cluster with 38 channels over the central and frontal scalp regions, the prediction accuracy with the observed envelopes was significantly higher than randomized envelopes [cluster-mean $r_{\text{rand}} = 0.0517$; $r_{\text{obs}} = 0.0670$; $r_{\text{obs}} - r_{\text{rand}} = 0.0153$, 95% CI = (0.0091, 0.0217); summary statistics $\Sigma T = 143.7$; cluster-$p = 0.0001$]. As discussed above (see section "Model Comparison"), the higher prediction accuracy of the full model than that of the reduced model (or the null model) indicates that the term of the full model that differs from the reduced (or null) model adds a unique contribution to the prediction, reflecting the neural encoding of the corresponding information. Here, the results suggest that the sound envelope is encoded in the cluster. Note that the peaks at the zero-lag in the TRFs (**Figure 3E**) are due to the free boundary condition (zero-padding at the boundaries of trials; note that the "condition" here refers to a mathematical constraint and not relevant to experimental conditions), which predicted trial-onset responses in phase-randomized models. When a weaker null model without the trial-onset was compared (i.e., valid boundary condition), the testing revealed increased prediction accuracy in 56 electrodes (cluster-$p = 0.0001$), presumably reflecting the widespread auditory activity *via* volume conduction (figure not shown).

Rhythmic Hierarchy

With respect to the low-level rhythmic feature, the analysis revealed significant encoding of *beat* (Eq. 5-1 vs. Eq. 4; **Figure 4**) in a cluster of 20 central channels [cluster-mean $r_{\text{reduced}} = 0.0314$; $r_{\text{full}} = 0.0341$; $r_{\text{full}} - r_{\text{reduced}} = 0.0027$, 95% CI = (0.0013, 0.0042); $\Sigma T = 52.4$; cluster-$p = 0.0234$]. Similarly to the envelope tracking results, a significant increase of prediction accuracy indicates a unique contribution of *beat* in addition to *envelope*.

With respect to the high-level rhythmic feature, the analysis revealed significant encoding of *meter* (Eq. 5-2 vs. Eq. 5-1; **Figure 5**) in a cluster of 16 frontal and central channels [cluster-mean $r_{\text{reduced}} = 0.0337$; $r_{\text{full}} = 0.0398$; $r_{\text{full}} - r_{\text{reduced}} = 0.0062$, (0.0023, 0.0099); $\Sigma T = 34.6$; cluster-$p = 0.0137$]. Likewise, a significant increase of prediction accuracy indicates a unique contribution of *meter* in addition to *envelope* and *beat*. The TRFs for *meter* showed different patterns by accents.

Tonal Hierarchy

We did not find a significant increase of prediction accuracy for either *key clarity* or *tonal stability* calculated on each beat or measure (Eq. 6-1 vs. Eq. 5-2, minimum cluster-$p = 0.1211$; Eq. 6-2 vs. Eq. 5-2, minimum cluster-$p = 0.0762$; **Supplementary Figures 2–5**).

DISCUSSION

Validity of the Proposed Framework

The view that individual elements in natural music may not produce the same effect as they do in isolation is not new. It has been claimed that music is not an objective entity but rather something that is experienced and perceived, raising the need for a dynamic, event-based processing framework (Reybrouck, 2005). The fundamental issue, however, is that often diverse musical features covary to maximize emotional effect (e.g., slow and elegiac melodies in Niccolò Paganini's Caprice for Solo Violin Op. 1 No. 3 in E minor and energetic arpeggios and triple stops in No. 1 in E major; or subdued, low vocals in Nirvana's melancholic "Something In The Way" and loud, angry drums in "Smells Like Teen Spirit"). Unless hundreds (if not thousands) of natural stimuli are used (Eerola et al., 2009; Cowen et al., 2020), it is impossible to tease out the effect of one element (or an independent component of elements) from another with a small number of stimuli. For this reason, it has been an established tradition to isolate and

[8] http://www.neuro.uni-jena.de/cat/

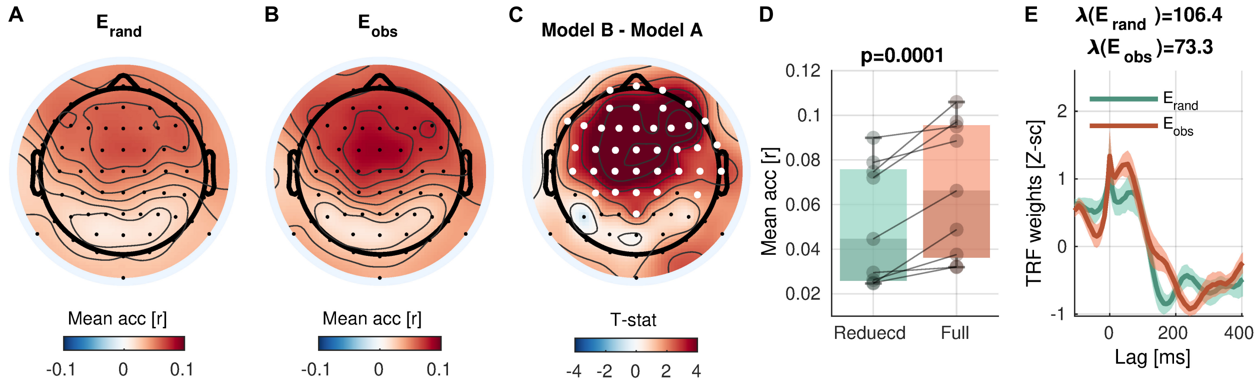


FIGURE 3 | Envelope encoding. **(A,B)** Mean prediction accuracies averaged across subjects are shown in topoplots for a null model (E_{rand}, phase-randomized envelope) and a faithful model (E_{obs}, observed envelope), respectively. **(C)** *t*-statistics comparing differences in prediction accuracies are shown. Channels included in significant clusters (cluster-$p < 0.05$) are marked in white. **(D)** Prediction accuracies averaged within the cluster with the smallest *p*-value are plotted for each participant. **(E)** Temporal response functions averaged within the cluster are shown. Shades mark one standard error of the mean across participants.

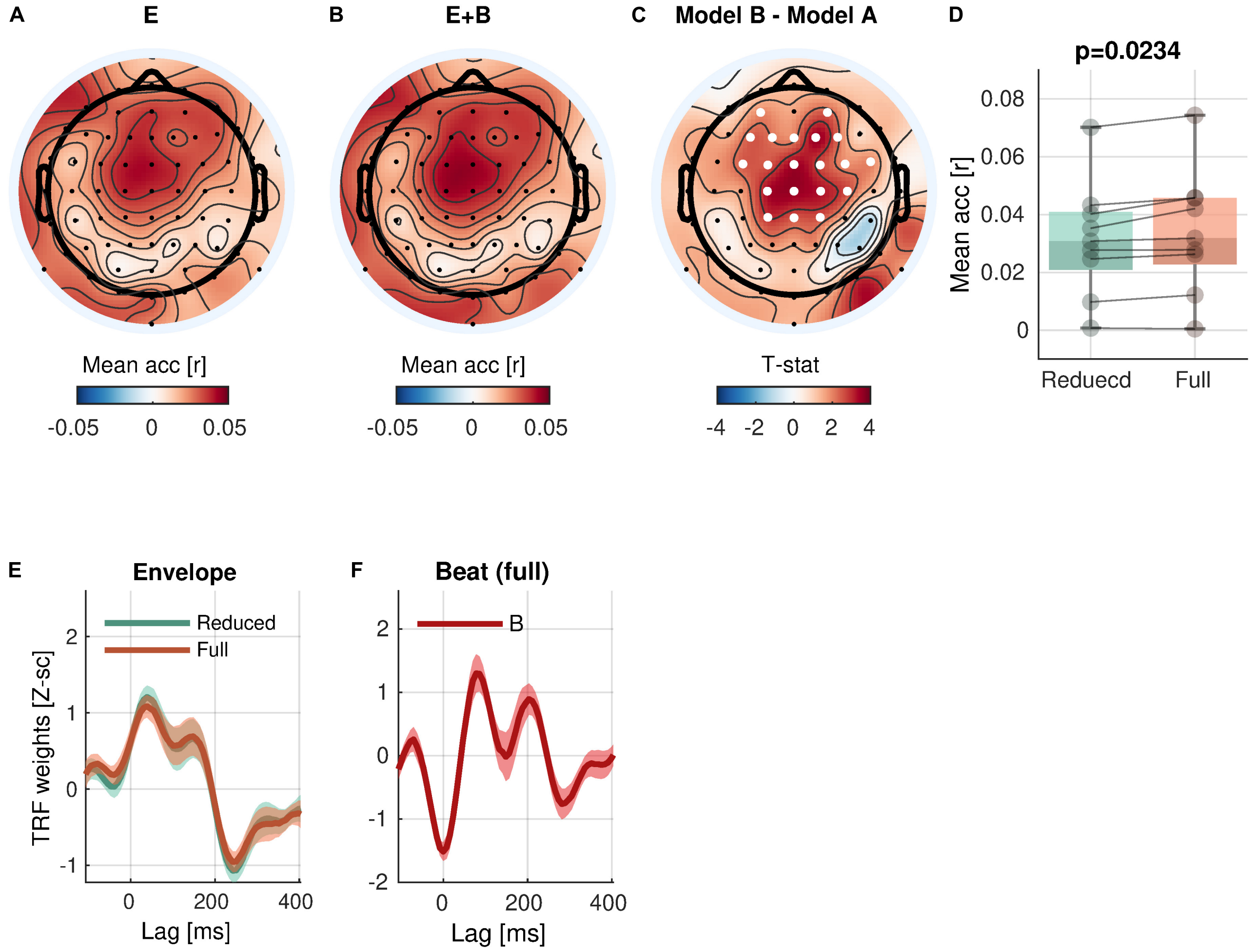


FIGURE 4 | Beat encoding. **(A,B)** Mean prediction accuracies of a reduced model (E, Envelope) and a full model (E + B, Envelope + Beat), respectively. **(C)** *t*-statistics comparing differences in prediction accuracies are shown. Channels included in significant clusters (cluster-$p < 0.05$) are marked in white. **(D)** Prediction accuracies averaged within the cluster with the smallest *p*-value are plotted for each subject. **(E,F)** Temporal response functions of features averaged across electrodes within the cluster are shown. TRFs are *Z*-scored across lags for different regularizations across electrodes/participants.

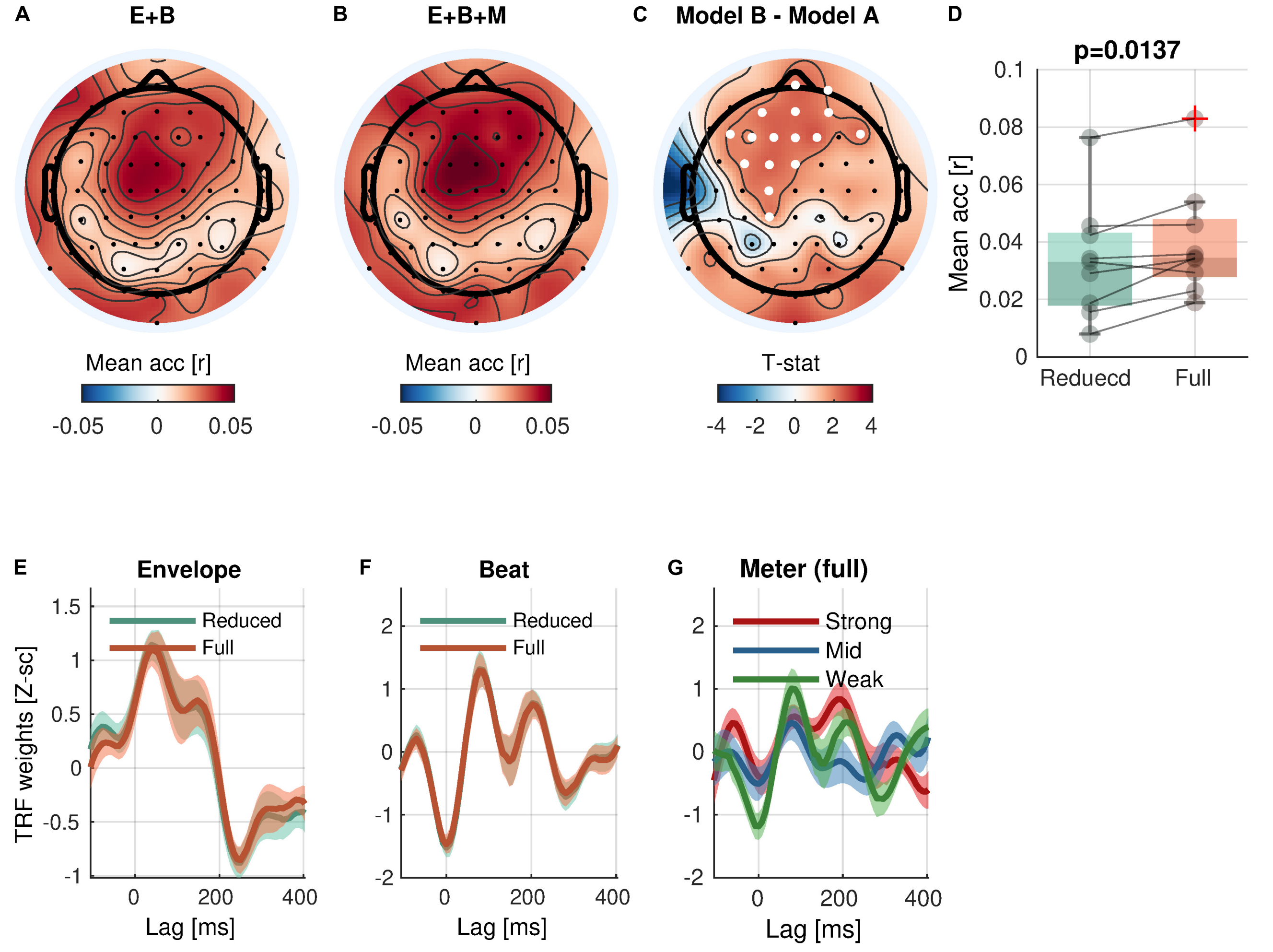


FIGURE 5 | Meter encoding. **(A,B)** Mean prediction accuracies of a reduced model (E + B, Envelope + Beat) and a full model (E + B + M, Envelope + Beat + Meter), respectively. **(C)** *t*-statistics comparing differences in prediction accuracies are shown. Channels included in significant clusters (cluster-$p < 0.05$) are marked in white. **(D)** Prediction accuracies averaged within the cluster with the smallest *p*-value are plotted for each subject. **(E,F)** Temporal response functions of features averaged within the cluster are shown. **(G)** Temporal response functions of the meter features (strong beat, middle beat, and weak beat; e.g., 4/4/: strong-weak-middle-weak) averaged within the cluster are shown. TRFs are *Z*-scored across lags for different regularizations across electrodes/participants.

orthogonalize musical features or acoustic properties to study their effects in music psychology and cognitive neuroscience of music. However, now that computational models can translate naturalistic stimuli into relevant features (i.e., linearizing functions), recent human neuroimaging studies have shown that it is possible to analyze complex interactions among natural features while taking advantage of the salience of naturalistic stimuli to evoke intense emotions and provide ecologically valid contexts (Goldberg et al., 2014; Sonkusare et al., 2019; Jääskeläinen et al., 2021).

In the current study, we demonstrated a simple yet powerful framework of a linearized encoding analysis by combining the MIR toolbox (a battery of model-based features) and mTRF Toolbox (FIR modeling with ridge regression). First, we showed that ridge regression successfully predicted envelope-triggered cortical responses in the ongoing EEG signal in comparison to null models with a phase-randomized envelope. Furthermore, our proposed framework detected cortical encoding of rhythmic, but not tonal, features while listening to naturalistic music. In addition, the estimated transfer functions and the spatial distribution of the prediction accuracies made neuroscientific interpretations readily available. These findings differentiate themselves from previous studies using similar regression analyses that only used either monophonic music or simple, low-level acoustic features, such as note onset (Sturm et al., 2015; Di Liberto et al., 2020).

Cortical Encoding of Musical Features

We showed cortical encoding of beats and meter during the listening of every day, continuous musical examples. This was observed most strongly over frontal and central EEG channels, which have long been implicated as markers of auditory processing activity (Näätänen and Picton, 1987; Zouridakis et al., 1998; Stropahl et al., 2018). However, *key clarity* and *tonal stability* were not conclusively represented in the cortical signal in our models.

Unlike the tonal features, both of the rhythmic features (*beat* and *meter*) were encoded strongly in the neural signal. The TRF for *beat* showed a steady periodic signal, consistent with the finding in the original OpenMIIR dataset publication by Stober (2017) that the peaks of the event-related potentials (ERPs) corresponded to the beat of the music. This means that both the ERPs in the study by Stober (2017) and our *beat* TRFs displayed large peaks at zero-lag, implying that beats may be anticipated. The possibility of an anticipatory mechanism of beats is consistent with the view that humans may possess an endogenous mechanism of beat anticipation that is active even when tones are unexpectedly omitted (Fujioka et al., 2009). The relatively early latency of the additional TRF peaks between 100 and 200 ms suggests that beats may be processed in a bottom-up fashion as well. Humans engage in an active search for the beat when it becomes less predictable by adaptively shifting their predictions based on the saliency of the beats in the music, suggesting that beats also provide useful exogenous cues (Toiviainen et al., 2020). The use of continuous music and EEG in the proposed framework lends itself particularly well to determining these various mechanisms of beat perception.

It has also been shown that different populations of neurons entrain to beats and meter (Nozaradan et al., 2017). Moreover, phase-locked gamma band activity has further suggested a unique neural correlate to meter (Snyder and Large, 2005). Extending these previous findings, the current results in the low frequency band (1–8 Hz) revealed this dichotomy between beats and meter through their different topologies. *Beat* was encoded over a tight cluster of central channels, but *meter* was encoded over a large cluster of frontal channels. The significant increase in prediction accuracy observed over widespread frontal channels for *meter* might suggest a distant source although it is not possible to uniquely determine the source location only from the sensor topography (i.e., inverse problem). That is, the topography also could be due to widely spread but synchronized cortical sources. However, there is evidence based on deep brain stimulation and scalp recording that EEG is sensitive to subcortical sources (Seeber et al., 2019). The putamen, in particular, has been proposed as a region of meter entrainment, while the cortical supplementary motor area is more associated with beats (Nozaradan et al., 2017; Li et al., 2019). The distinct topologies observed between the beats and meter features are especially intriguing given the relatively short duration of each stimulus (10.5 s on average).

It was unexpected that neither of the tonal features was significantly correlated with the EEG signal, given that previous studies suggested that information about these tonal structures is reflected in non-invasive neural recordings. For instance, previous ERP studies showed stronger responses to deviant harmonies than normative ones (Besson and Faïta, 1995; Janata, 1995; Koelsch et al., 2003). Additionally, in a recent MEG study (Sankaran et al., 2020), a representational similarity analysis revealed that distinctive cortical activity patterns at the early stage (around 200 ms) reflected the absolute pitch (i.e., fundamental frequencies) of presented tones, whereas late stages (after 200 ms onward) reflected their relative pitch with respect to the established tonal context (i.e., tonal hierarchy) during the listening of isolated chord sequences and probe tones played by a synthesized piano. In a study with more naturalistic musical stimuli (Di Liberto et al., 2020), the cortical encoding of melodic expectation, which is defined by how surprising a pitch or note onset is within a given melody, was shown using EEG and the TRF during the listening of monophonic MIDI piano excerpts generated from J. S. Bach Chorales. With respect to *key clarity*, it was shown that *key clarity* correlates significantly with behavioral ratings (Eerola, 2012) and is anti-correlated with the fMRI signal timeseries in specific brain regions, including the Rolandic Operculum, insula, and precentral gyrus, while listening to modern Argentine Tango (Alluri et al., 2012). In a replication study with identical stimuli (Burunat et al., 2016), *key clarity* showed scattered encoding patterns across all brain regions with weaker magnitudes of correlations, although such an association with evoked EEG responses (or the absence thereof) has not been previously reported. One possibility for the current negative finding with respect to tonal features is that the musical stimuli in the current dataset might not have been optimal for our interest in the tonal analysis given their tonal simplicity (see section "Limitations" for further discussion).

Limitations

The stimuli were relatively short in duration (10-s long on average) and often repetitious in nature. These stimulus characteristics limited the ability to observe the response to larger changes in *key clarity* and *tonal stability*. For instance, the ranges of standard deviation of *key clarity* and *tonal stability* were (0.0667, 0.1638) and (0.1570, 0.2607), respectively, when calculated on beats. These were narrower than typical musical stimulus sets [e.g., 360 emotional soundtrack 15-s excerpts (Eerola and Vuoskoski, 2011); (0.0423, 0.2303) and (0.11882, 0.3441) for *key clarity* and *tonal stability*, respectively]. These limitations (short lengths and limited variation in tonality) might have contributed to negative findings in the current study. Another limitation in the dataset was the small number of participants (n = 9), which limited statistical power. Future neuro-music public datasets (e.g., the one developed by Grahn et al., 2018) may want to consider using longer, more dynamic musical excerpts, especially ones that have increased dramatic shifts in tonality with more participants. The dataset also did not contain simultaneous behavioral ratings of the music, which resulted in us being unable to analyze our data alongside measures such as emotion.

One limitation in our analysis is that we used a single regularization parameter for all features, as currently implemented in the mTRF Toolbox. However, it has been shown that using independent regularization for each feature set ("banded ridge") can improve the prediction and interpretability of joint modeling in fMRI encoding analysis (Nunez-Elizalde et al., 2019). Thus, it is expected that a systematic investigation on the merits of banded ridge regression in mTRF analysis on M/EEG data would benefit the community.

FUTURE DIRECTIONS AND CONCLUSION

Ultimately, we hope that this framework can serve two broad purposes. The first is for it to enhance the ecological validity of future music experiments. The second is for it to be used as a tool that can be paired with other metrics of interest. Emotion is perhaps the most fitting application of this framework, given the special ability of music to make us experience intense feelings. Combining the current analytic framework with behavioral measures like emotion will be especially useful because it could shed light on what factors interact with our anticipation of tonality and rhythm during music listening. In particular, when combined with continuous behavioral measures, such as emotion or tension, this might 1 day be used to elucidate how changes in certain musical features make us happy or sad, which could deepen our knowledge of how music can be used therapeutically or clinically. Furthermore, some current limitations of the *tonal stability* measure provide future researchers with opportunities for innovation. Looking forward, it would be useful to create a *tonal stability* measure that can account for multiple (shifting) tonal centers within a single piece of music.

In summary, we presented an analytical framework to investigate tonal and rhythmic hierarchy encoded in neural signals while listening to homophonic music. Though the model did not demonstrate the presence of the proposed *tonal stability* measure, it did successfully capture cortical encoding of rhythmic hierarchy. Moreover, the framework was able to differentiate the spatial encoding of low/high-level features, as represented by the separate encoding of beat and meter, suggesting distinct neural processes. The current framework is applicable to any form of music by directly feeding audio signals into the linearizing model. In addition, it has the possibility of including other time-resolved measures to appropriately address the complexity and multivariate nature of music and other affective naturalistic stimuli. This will bring us to a more complete understanding of how tonality and rhythm are processed over time and why the anticipation and perception of these features can induce a variety of emotional responses within us.

DATA AVAILABILITY STATEMENT

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found in the article/**Supplementary Material**.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Ethics Board at the University of Western Ontario. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JL and S-GK conceived the ideas, developed the analytic framework, analyzed the public data, and wrote the first draft together. S-GK formulated models and wrote code for analysis and visualization. JW and TO contributed to conceiving ideas, interpreting results, and writing the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by Duke University startup funds to TO.

ACKNOWLEDGMENTS

We thank Sebastian Stober for sharing the OpenMIIR dataset, code, and documentation. We also thank Michael Broderick for technical discussion on the mTRF analysis. The editor and reviewers helped us in improving the earlier version of the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fnins.2021.665767/full#supplementary-material

REFERENCES

Abrams, D. A., Ryali, S., Chen, T., Chordia, P., Khouzam, A., Levitin, D. J., et al. (2013). Inter-subject synchronization of brain responses during natural music listening. *Eur. J. Neurosci.* 37, 1458–1469. doi: 10.1111/ejn.12173

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., and Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage* 59, 3677–3689. doi: 10.1016/j.neuroimage.2011.11.019

Besson, M., and Faïta, F. (1995). An event-related potential (ERP) study of musical expectancy: comparison of musicians with nonmusicians. *J. Exp. Psychol.* 21:1278. doi: 10.1037/0096-1523.21.6.1278

Bianco, R., Novembre, G., Keller, P. E., Villringer, A., and Sammler, D. (2018). Musical genre-dependent behavioural and EEG signatures of action planning. A comparison between classical and jazz pianists. *Neuroimage* 169, 383–394. doi: 10.1016/j.neuroimage.2017.12.058

Brodbeck, C., Presacco, A., and Simon, J. Z. (2018). Neural source dynamics of brain responses to continuous stimuli: speech processing from acoustics to comprehension. *NeuroImage* 172, 162–174. doi: 10.1016/j.neuroimage.2018.01.042

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., and Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Curr. Biol.* 28, 803–809.e3.

Burunat, I., Toiviainen, P., Alluri, V., Bogert, B., Ristaniemi, T., Sams, M., et al. (2016). The reliability of continuous brain responses during naturalistic listening to music. *Neuroimage* 124, 224–231. doi: 10.1016/j.neuroimage.2015.09.005

Chen, J. L., Penhune, V. B., and Zatorre, R. J. (2008). Listening to musical rhythms recruits motor regions of the brain. *Cereb. Cortex* 18, 2844–2854. doi: 10.1093/cercor/bhn042

Cowen, A. S., Fang, X., Sauter, D., and Keltner, D. (2020). What music makes us feel: at least 13 dimensions organize subjective experiences associated with music across different cultures. *Proc. Natl. Acad. Sci. U.S.A.* 117, 1924–1934. doi: 10.1073/pnas.1910704117

Crosse, M. J., Di Liberto, G. M., Bednar, A., and Lalor, E. C. (2016). The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli. *Front. Hum. Neurosci.* 10:604. doi: 10.3389/fnhum.2016.00604

Di Liberto, G. M., O'sullivan, J. A., and Lalor, E. C. (2015). Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Curr. Biol.* 25, 2457–2465. doi: 10.1016/j.cub.2015.08.030

Di Liberto, G. M., Pelofi, C., Bianco, R., Patel, P., Mehta, A. D., Herrero, J. L., et al. (2020). Cortical encoding of melodic expectations in human temporal cortex. *eLife* 9:e51784.

Eerola, T. (2012). Modeling listeners' emotional response to music. *Top. Cogn. Sci.* 4, 607–624. doi: 10.1111/j.1756-8765.2012.01188.x

Eerola, T., and Vuoskoski, J. K. (2011). A comparison of the discrete and dimensional models of emotion in music. *Psychol. Music* 39, 18–49. doi: 10.1177/0305735610362821

Eerola, T., Lartillot, O., and Toiviainen, P. (2009). "Prediction of multidimensional emotional ratings in music from audio using multivariate regression models," in *Proceedings of the International Society for Music Information Retrieval (ISMIR)* (Kobe), 621–626.

Farbood, M. M., Heeger, D. J., Marcus, G., Hasson, U., and Lerner, Y. (2015). The neural processing of hierarchical structure in music and speech at different timescales. *Front. Neurosci.* 9:157. doi: 10.3389/fnins.2015.00157

Fishman, Y. I., Volkov, I. O., Noh, M. D., Garell, P. C., Bakken, H., Arezzo, J. C., et al. (2001). Consonance and dissonance of musical chords: neural correlates in auditory cortex of monkeys and humans. *J. Neurophysiol.* 86, 2761–2788. doi: 10.1152/jn.2001.86.6.2761

Fujioka, T., Trainor, L. J., Large, E. W., and Ross, B. (2009). Beta and gamma rhythms in human auditory cortex during musical beat processing. *Ann. N. Y. Acad. Sci.* 1169, 89–92. doi: 10.1111/j.1749-6632.2009.04779.x

Goldberg, H., Preminger, S., and Malach, R. (2014). The emotion–action link? Naturalistic emotional stimuli preferentially activate the human dorsal visual stream. *NeuroImage* 84, 254–264. doi: 10.1016/j.neuroimage.2013.08.032

Gómez, E. (2006). Tonal description of polyphonic audio for music content processing. *Informs J. Comput.* 18, 294–304. doi: 10.1287/ijoc.1040.0126

Gordon, C. L., Cobb, P. R., and Balasubramaniam, R. (2018). Recruitment of the motor system during music listening: an ALE meta-analysis of fMRI data. *PLoS One* 13:e0207213. doi: 10.1371/journal.pone.0207213

Grahn, J. A., and McAuley, J. D. (2009). Neural bases of individual differences in beat perception. *NeuroImage* 47, 1894–1903. doi: 10.1016/j.neuroimage.2009.04.039

Grahn, J. A., and Rowe, J. B. (2009). Feeling the beat: premotor and striatal interactions in musicians and nonmusicians during beat perception. *J. Neurosci.* 29, 7540–7548. doi: 10.1523/jneurosci.2018-08.2009

Grahn, J., Diedrichsen, J., Gati, J., Henry, M., Zatorre, R., Poline, J.-B., et al. (2018). *OMMABA: The Open Multimodal Music and Auditory Brain Archive Project Summaries*. London, ON: Western University.

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., et al. (2013). MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* 7:267. doi: 10.3389/fnins.2013.00267

Hoerl, A. E., and Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12, 55–67. doi: 10.1080/00401706.1970.10488634

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532:453. doi: 10.1038/nature17637

Jääskeläinen, I. P., Sams, M., Glerean, E., and Ahveninen, J. (2021). Movies and narratives as naturalistic stimuli in neuroimaging. *NeuroImage* 224:117445. doi: 10.1016/j.neuroimage.2020.117445

Janata, P. (1995). ERP measures assay the degree of expectancy violation of harmonic contexts in music. *J. Cogn. Neurosci.* 7, 153–164. doi: 10.1162/jocn.1995.7.2.153

Kaneshiro, B., Nguyen, D. T., Norcia, A. M., Dmochowski, J. P., and Berger, J. (2020). Natural music evokes correlated EEG responses reflecting temporal structure and beat. *NeuroImage* 214:116559. doi: 10.1016/j.neuroimage.2020.116559

Kay, K. N., Naselaris, T., Prenger, R. J., and Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature* 452, 352–355. doi: 10.1038/nature06713

Koelsch, S., and Jentschke, S. (2010). Differences in electric brain responses to melodies and chords. *J. Cogn. Neurosci.* 22, 2251–2262. doi: 10.1162/jocn.2009.21338

Koelsch, S., Gunter, T., Friederici, A. D., and Schröger, E. (2000). Brain indices of music processing: "nonmusicians" are musical. *J. Cogn. Neurosci.* 12, 520–541. doi: 10.1162/089892900562183

Koelsch, S., Gunter, T., Schröger, E., and Friederici, A. D. (2003). Processing tonal modulations: an ERP study. *J. Cogn. Neurosci.* 15, 1149–1159. doi: 10.1162/089892903322598111

Krumhansl, C. L. (1990). Tonal hierarchies and rare intervals in music cognition. *Music Percept.* 7, 309–324. doi: 10.2307/40285467

Krumhansl, C. L., and Cuddy, L. L. (2010). "A theory of tonal hierarchies in music," in *Music Perception*, eds M. Riess Jones, R. R. Fay, and A. N. Popper (New York, NY: Springer), 51–87. doi: 10.1007/978-1-4419-6114-3_3

Krumhansl, C. L., and Shepard, R. N. (1979). Quantification of the hierarchy of tonal functions within a diatonic context. *J. Exp. Psychol.* 5:579. doi: 10.1037/0096-1523.5.4.579

Lalor, E. C., Pearlmutter, B. A., Reilly, R. B., Mcdarby, G., and Foxe, J. J. (2006). The VESPA: a method for the rapid estimation of a visual evoked potential. *NeuroImage* 32, 1549–1561. doi: 10.1016/j.neuroimage.2006.05.054

Lartillot, O., and Toiviainen, P. (2007). "A Matlab toolbox for musical feature extraction from audio," in *Proceedings of the International Conference on Digital Audio Effects (DAFx)* (Bordeaux), 237–244.

Li, Q., Liu, G., Wei, D., Liu, Y., Yuan, G., and Wang, G. (2019). Distinct neuronal entrainment to beat and meter: revealed by simultaneous EEG-fMRI. *NeuroImage* 194, 128–135. doi: 10.1016/j.neuroimage.2019.03.039

Loui, P., and Wessel, D. (2007). Harmonic expectation and affect in Western music: effects of attention and training. *Percept. Psychophys.* 69, 1084–1092. doi: 10.3758/bf03193946

Maris, E. (2019). Enlarging the scope of randomization and permutation tests in neuroimaging and neuroscience. *bioRxiv* [Preprint] doi: 10.1101/685560v4

Maris, E., and Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* 164, 177–190. doi: 10.1016/j.jneumeth.2007.03.024

Menon, V., and Levitin, D. J. (2005). The rewards of music listening: response and physiological connectivity of the mesolimbic system. *Neuroimage* 28, 175–184. doi: 10.1016/j.neuroimage.2005.05.053

Näätänen, R., and Picton, T. (1987). The N1 wave of the human electric and magnetic response to sound: a review and an analysis of the component structure. *Psychophysiology* 24, 375–425. doi: 10.1111/j.1469-8986.1987.tb00311.x

Nozaradan, S., Schwartze, M., Obermeier, C., and Kotz, S. A. (2017). Specific contributions of basal ganglia and cerebellum to the neural tracking of rhythm. *Cortex* 95, 156–168. doi: 10.1016/j.cortex.2017.08.015

Nunez-Elizalde, A. O., Huth, A. G., and Gallant, J. L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage* 197, 482–492. doi: 10.1016/j.neuroimage.2019.04.012

Pressnitzer, D., Patterson, R. D., and Krumbholz, K. (2001). The lower limit of melodic pitch. *J. Acoust. Soc. Am.* 109, 2074–2084. doi: 10.1121/1.1359797

Reybrouck, M. (2005). A biosemiotic and ecological approach to music cognition: event perception between auditory listening and cognitive economy. *Axiomathes* 15, 229–266. doi: 10.1007/s10516-004-6679-4

Sankaran, N., Carlson, T. A., and Thompson, W. F. (2020). The rapid emergence of musical pitch structure in human cortex. *J. Neurosci.* 40:2108. doi: 10.1523/jneurosci.1399-19.2020

Seeber, M., Cantonas, L.-M., Hoevels, M., Sesia, T., Visser-Vandewalle, V., and Michel, C. M. (2019). Subcortical electrophysiological activity is detectable with high-density EEG source imaging. *Nat. Commun.* 10:753.

Snyder, J. S., and Large, E. W. (2005). Gamma-band activity reflects the metric structure of rhythmic tone sequences. *Cogn. Brain Res.* 24, 117–126. doi: 10.1016/j.cogbrainres.2004.12.014

Sonkusare, S., Breakspear, M., and Guo, C. (2019). Naturalistic stimuli in neuroscience: critically acclaimed. *Trends Cogn. Sci.* 23, 699–714. doi: 10.1016/j.tics.2019.05.004

Stober, S. (2017). Toward studying music cognition with information retrieval techniques: lessons learned from the OpenMIIR initiative. *Front. Psychol.* 8:1255. doi: 10.3389/fpsyg.2017.01255

Stropahl, M., Bauer, A.-K. R., Debener, S., and Bleichner, M. G. (2018). Source-modeling auditory processes of EEG data using EEGLAB and brainstorm. *Front. Neurosci.* 12:309. doi: 10.3389/fnins.2018.00309

Sturm, I., Dähne, S., Blankertz, B., and Curio, G. (2015). Multi-variate EEG analysis as a novel tool to examine brain responses to naturalistic music stimuli. *PLoS One* 10:e0141281. doi: 10.1371/journal.pone.0141281

Toiviainen, P., Burunat, I., Brattico, E., Vuust, P., and Alluri, V. (2020). The chronnectome of musical beat. *Neuroimage* 216:116191. doi: 10.1016/j.neuroimage.2019.116191

Vuust, P., and Witek, M. A. (2014). Rhythmic complexity and predictive coding: a novel approach to modeling rhythm and meter perception in music. *Front. Psychol.* 5:1111. doi: 10.3389/fpsyg.2014.01111

Wu, M. C.-K., David, S. V., and Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.* 29, 477–505. doi: 10.1146/annurev.neuro.29.051605.113024

Zatorre, R. J., Chen, J. L., and Penhune, V. B. (2007). When the brain plays music: auditory-motor interactions in music perception and production. *Nat. Rev. Neurosci.* 8, 547–558. doi: 10.1038/nrn2152

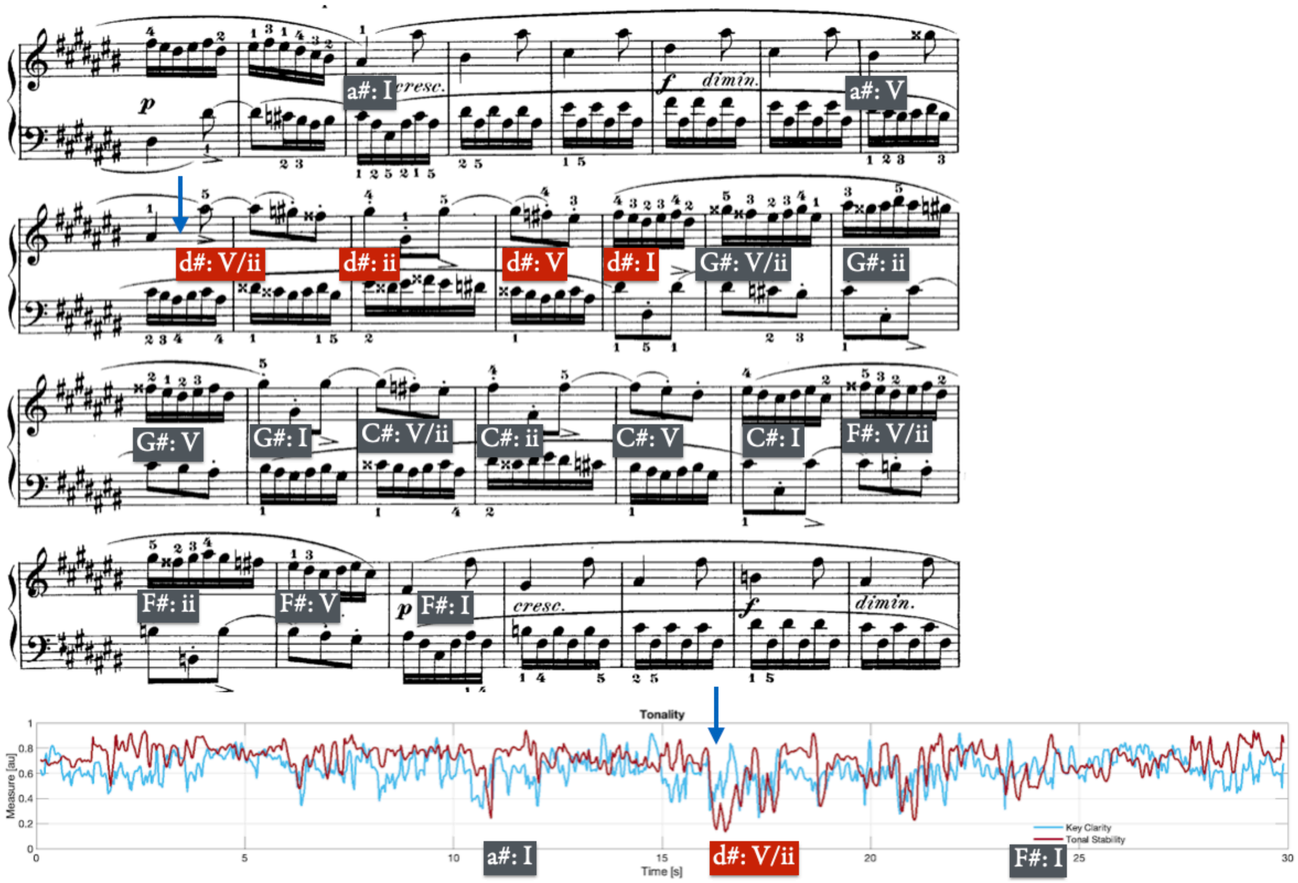
Zhang, J., Zhou, X., Chang, R., and Yang, Y. (2018). Effects of global and local contexts on chord processing: an ERP study. *Neuropsychologia* 109, 149–154. doi: 10.1016/j.neuropsychologia.2017.12.016

Zouridakis, G., Simos, P. G., and Papanicolaou, A. C. (1998). Multiple bilaterally asymmetric cortical sources account for the auditory N1m component. *Brain Topogr.* 10, 183–189.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

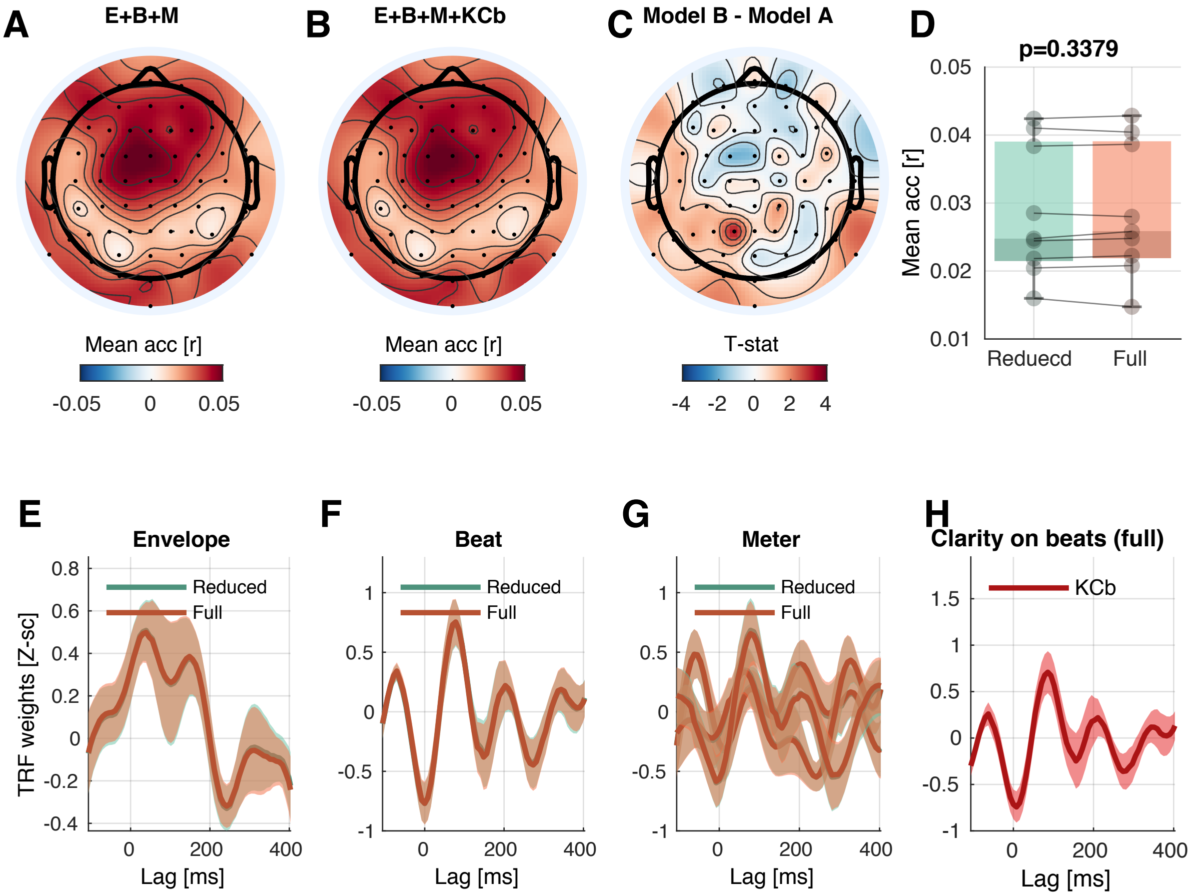
Supplementary Material

1 Supplementary Figures

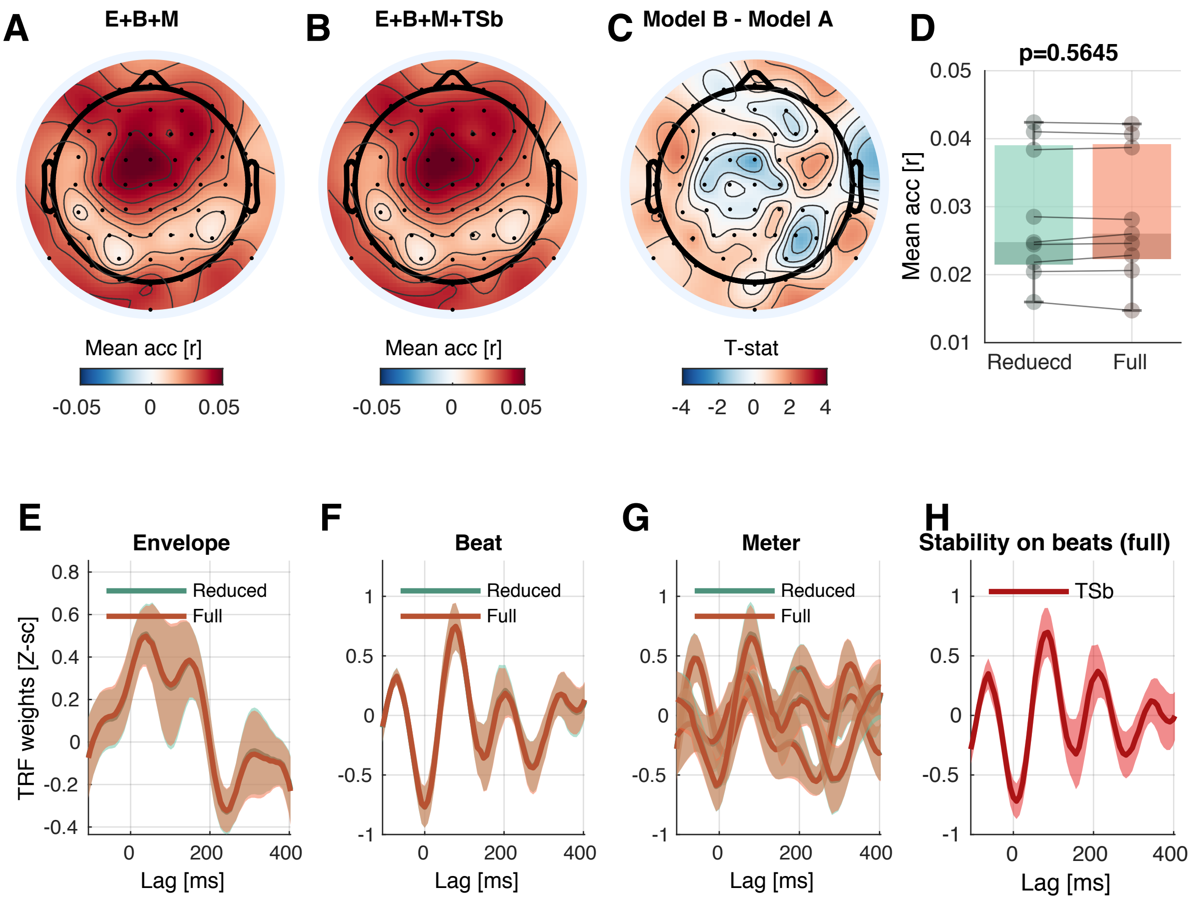


Supplementary Figure S1. **Key clarity and tonal stability**. As an example comparison, key clarity (blue) and tonal stability (red) were calculated for a 30-s excerpt from J. S. Bach's Prelude and Fugue in C# major, BWV 848 with 50% overlapping 200-ms frames. A modulation to a key (D# minor) that is distant from the overall key of the excerpt (C# major) was detected by a sudden decrease in tonal stability (marked with blue arrows), whereas key clarity was insensitive to such tonal relationships. The musical score is in the public domain[1].

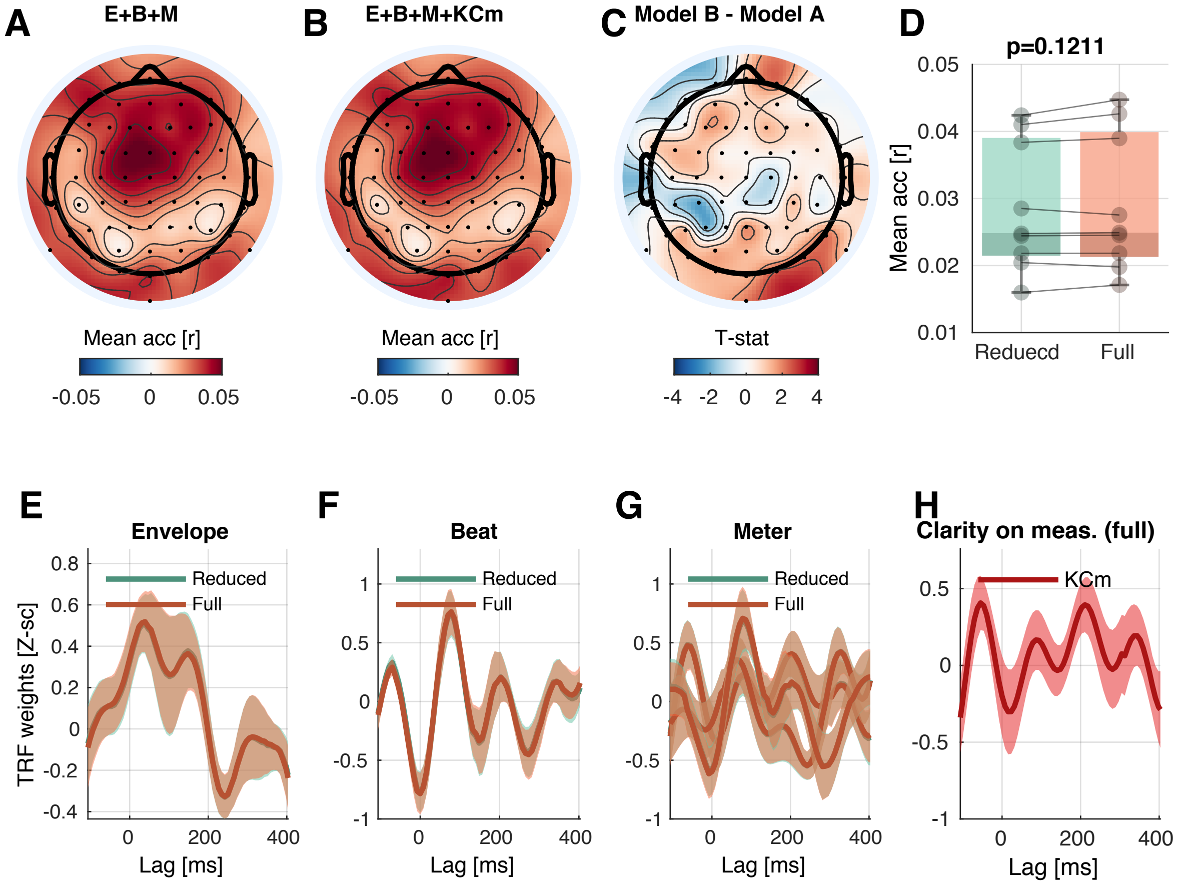
[1] https://imslp.org/wiki/Prelude_and_Fugue_in_C-sharp_major,_BWV_848_(Bach,_Johann_Sebastian)



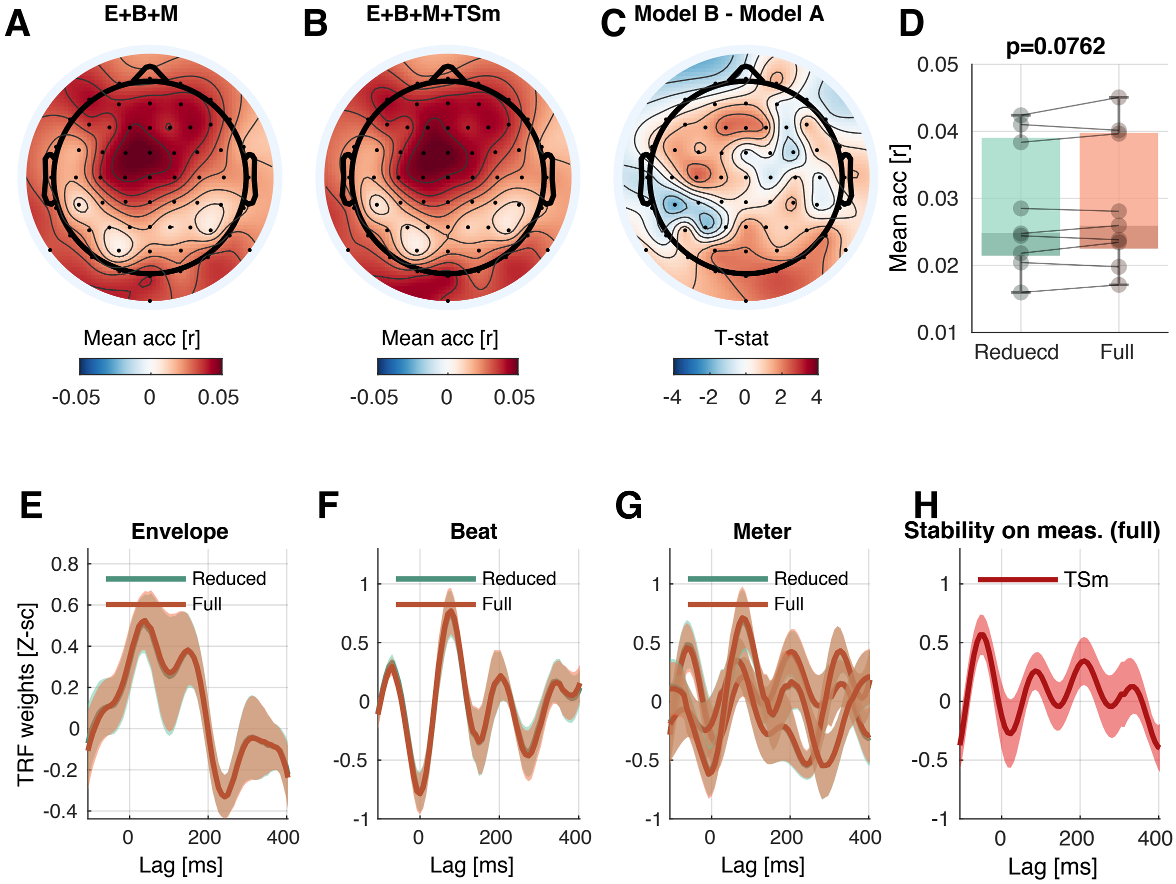
Supplementary Figure S2. **Key-clarity-on-beat encoding**. (A, B) Mean prediction accuracies of a reduced model (E+B+M: Envelope + Beat + Meter) and a full model (E+B+M+KCb: Envelope + Beat + Meter + Key clarity on beats), respectively. (C) T-statistics comparing differences in prediction accuracies are shown. (D) Prediction accuracies averaged across all channels are plotted for each subject. (E–H) Temporal response functions of features averaged across all channels are shown. TRFs are Z-scored across lags for different regularizations across electrodes/subjects.



Supplementary Figure S3. **Tonal-stability-on-beat encoding**. (A, B) Mean prediction accuracies of a reduced model (E+B+M: Envelope + Beat + Meter) and a full model (E+B+M+TSb: Envelope + Beat + Meter + Tonal stability on beats), respectively. (C) T-statistics comparing differences in prediction accuracies are shown. (D) Prediction accuracies averaged across all channels are plotted for each subject. (E–H) Temporal response functions of features averaged across all channels are shown. TRFs are Z-scored across lags for different regularizations across electrodes/subjects.



Supplementary Figure S4. **Key-clarity-on-measure encoding**. (A, B) Mean prediction accuracies of a reduced model (E+B+M: Envelope + Beat + Meter) and a full model (E+B+M+KCm: Envelope + Beat + Meter + Key clarity on measures), respectively. (C) T-statistics comparing differences in prediction accuracies are shown. (D) Prediction accuracies averaged across all channels are plotted for each subject. (E–H) Temporal response functions of features averaged across all channels are shown. TRFs are Z-scored across lags for different regularizations across electrodes/subjects.



Supplementary Figure S5. **Tonal-stability-on-measure encoding**. (A, B) Mean prediction accuracies of a reduced model (E+B+M: Envelope + Beat + Meter) and a full model (E+B+M+TSm: Envelope + Beat + Meter + Tonal stability on measures), respectively. (C) T-statistics comparing differences in prediction accuracies are shown. (D) Prediction accuracies averaged across all channels are plotted for each subject. (E–H) Temporal response functions of features averaged across all channels are shown. TRFs are Z-scored across lags for different regularizations across electrodes/subjects.