Title: Neural entrainment is strongest to the spectral flux of slow music and depends on familiarity and beat salience

Authors: Kristin Weineck[1,2], Olivia Xin Wen[1], Molly J. Henry[1,3]

[1]Research Group "Neural and Environmental Rhythms"
Max Planck Institute for Empirical Aesthetics
Grüneburgweg 14
60322 Frankfurt am Main, Germany

[2]Institute for Cell Biology and Neuroscience
Goethe University Frankfurt am Main
Max-von-Laue-Str. 13
60438, Frankfurt am Main, Germany

[3]Department of Psychology
Ryerson University
350 Victoria Street
Toronto, Ontario, Canada M5B 2K3

[*]Author to whom correspondence should be addressed. Electronic mail:
kristin.weineck@ae.mpg.de
molly.henry@ae.mpg.de

Abstract

Neural activity in the auditory system synchronizes to sound rhythms, and brain–environment synchronization is thought to be fundamental to successful auditory perception. Sound rhythms are often operationalized in terms of the sound's amplitude envelope. We hypothesized that – especially for music – the envelope might not best capture the complex spectro-temporal fluctuations that give rise to beat perception and synchronize neural activity. This study investigated 1) neural entrainment to different musical features, 2) tempo-dependence of neural entrainment, and 3) dependence of entrainment on familiarity, enjoyment, and ease of beat perception. In this electroencephalography study, 37 human participants listened to tempo-modulated music (1–4 Hz). Independent of whether the analysis approach was based on temporal response functions (TRFs) or reliable components analysis (RCA), the spectral flux of music – as opposed to the amplitude envelope – evoked strongest neural entrainment. Moreover, music with slower beat rates, high familiarity, and easy-to-perceive beats elicited the strongest neural response. Based on the TRFs, we could decode music stimulation tempo, but also perceived beat rate, even when the two differed. Our results demonstrate the importance of accurately characterizing musical acoustics in the context of studying neural entrainment, and demonstrate entrainment's sensitivity to musical tempo, familiarity, and beat salience.

Introduction

Neural activity synchronizes to different types of rhythmic sounds, such as speech and music (Doelling and Poeppel, 2015, Nicolaou et al., 2017, Ding et al., 2017, Kösem et al., 2018), over a wide range of rates. Neural oscillations are involved in the regulation of (patho-)physiological activity and are important for gating input during sensory perception and temporal processing (Giraud and Poeppel, 2012, Henry and Herrmann, 2014). For this reason, bringing neural oscillations into temporal alignment with a rhythmic stimulus – *neural entrainment* – can influence perception in the auditory, visual, and somatosensory modalities (Henry and Obleser, 2012, Spaak et al., 2014, Gundlach et al., 2016). In the auditory domain, neural oscillations entrained at syllabic, prosodic, and semantic rates in speech seem to play an important role in speech perception and intelligibility (Doelling et al., 2014, Peelle et al., 2013, Kösem et al., 2018). The current study examined neural entrainment to music.

Music is highly rhythmic, and neural oscillations can be entrained by the beat, the most prominent isochronous pulse in music, to which listeners would sway their bodies or tap their feet (Tierney and Kraus, 2015, Nozaradan et al., 2012, Large and Snyder, 2009, Doelling and Poeppel, 2015). Most studies that have examined cortical tracking of musical rhythm used simplified musical stimuli, such as MIDI melodies or click tracks (Kumagai et al., 2018, Nozaradan et al., 2012, Di Liberto et al., 2020, Nozaradan et al., 2011, Wollman et al., 2020) or monophonic melodies (Doelling and Poeppel, 2015); only a few studies have focused on naturalistic, polyphonic music (Tierney and Kraus, 2015, Madsen et al., 2019, Kaneshiro et al., 2020). Listeners show a strong preference for music at beat rates around 2 Hz (here, we use the term *tempo* to refer to the beat rate). The preference for 2 Hz coincides with the modal tempo of Western pop music (Moelants, 2002) and the most prominent frequency of natural adult body movements (MacDougall and Moore, 2005). Indeed, previous research showed that listeners perceive rhythmic sequences at beat rates around 2 Hz especially salient when they are able to track the beat by moving their bodies (Zalta et al., 2020). Despite the

perceptual and motor evidence, studies looking at tempo-dependence of neural entrainment are scarce (Doelling and Poeppel, 2015, Nicolaou et al., 2017) and we are not aware of any study using naturalistic polyphonic musical stimuli that are tempo-modulated. By examining entrainment across a relatively wide and finely spaced range of musical tempi (1–4 Hz, corresponding to the neural δ band), we aimed to test whether the preference for music with beat rates around 2 Hz is reflected in the strength of neural entrainment. In addition, a number of different musical, behavioral, and perceptual measures have been shown to modulate neural entrainment and influence music perception, including complexity, familiarity, repetition of the music, musical training of the listener, and attention to the stimulus (Kumagai et al., 2018, Madsen et al., 2019, Doelling and Poeppel, 2015). Thus, we investigated the effects of enjoyment, familiarity and the ease of beat perception on neural entrainment.

Most studies assessing neural entrainment to music have examined entrainment to either the stimulus amplitude envelope, which quantifies intensity fluctuations over time (Doelling and Poeppel, 2015, Kaneshiro et al., 2020, Wollman et al., 2020), or "higher order" musical features such as surprise and expectation (Di Liberto et al., 2020). This mimics approaches used for studying neural tracking of speech, where neural activity has been shown to be entrained by the amplitude envelope (Peelle and Davis, 2012), which roughly corresponds to syllabic fluctuations (Doelling et al., 2014), as well as by "higher order" semantic information (Broderick et al., 2019). "Higher order" musical features are difficult to compute for naturalistic music, which is typically polyphonic and has complex spectro-temporal properties (Zatorre et al., 2002). However, amplitude-envelope entrainment is well documented: neural activity synchronizes to the amplitude fluctuations in music between 1 Hz and 8 Hz, and entrainment is especially strong for listeners with musical expertise (Doelling and Poeppel, 2015).

Because of the complex nature of natural polyphonic music, we hypothesized that amplitude envelope might not be the only or most dominant feature by which neural activity would be entrained (Mller, 2015). Thus, the current study investigated neural responses to different musical features that evolve over time and capture different aspects of the stimulus dynamics. Here, we use the term *musical feature* to refer to time-varying aspects of music that fluctuate on time scales corresponding roughly to the neural δ band, as opposed to elements of music such as key, harmony or syncopation. We examined amplitude envelope, the first derivative of the amplitude envelope (usually more sensitive to sound onsets than the amplitude envelope), beat times, and *spectral flux*, which describes spectral changes of the signal on a frame-to-frame basis by computing the difference between the spectral vectors of subsequent frames (Mller, 2015). One distinct advantage of spectral flux over the envelope or its derivative is that spectral flux is sensitive to rhythmic information that is communicated by changes in pitch even when they are not accompanied by changes in amplitude.

The current study investigated neural entrainment to natural music by using two different analysis approaches: Reliable Components Analysis (RCA) (Kaneshiro et al., 2020) and temporal response functions (TRFs) (Di Liberto et al., 2020). RCA typically relies on stimulus–response correlation or stimulus–response coherence (Kaneshiro et al., 2020). These approaches have been criticized because of their potential susceptibility to autocorrelation, which is argued to be minimized in the TRF approach (Zuk et al., 2021). Thus, we tested the agreement between these two analysis approaches.

We aimed to answer four questions. 1) Does neural entrainment to natural music depend on tempo? 2) Which musical feature shows the strongest neural entrainment during natural music listening? 3) How compatible are RCA- and TRF-based methods with quantifying neural entrainment to natural music? 4) How do enjoyment, familiarity, and ease of beat perception affect neural entrainment? To answer these research questions, we recorded electroencephalography (EEG) data while participants listened to instrumental music

presented at different tempi (1–4 Hz). Strongest neural entrainment was observed in response to the spectral flux of music, for tempi between 1–2 Hz, to familiar songs, and to songs with an easy-to-perceive beat. Moreover, a classifier trained on the neural responses to each musical feature predicted the metrical level at which listeners tapped the beat. This indicates that the brain responded to *perceived* tempo, even when it was different from the stimulus tempo.

Results

Scalp EEG activity of 37 human participants was measured while they listened to instrumental segments of natural music from different genres (Supplementary Table 1). Music segments were presented at thirteen parametrically varied tempi (1–4 Hz in 0.25-Hz steps; see *Materials and Methods*). We assessed neural entrainment to four different musical features: amplitude envelope, first derivative of the amplitude envelope, beat times, and spectral flux. Neural entrainment was quantified using two different analysis pipelines and compared: 1) RCA combined with time- and frequency-domain analyses, and 2) TRFs (Crosse et al., 2016, Kaneshiro et al., 2020). As different behavioral and perceptual measures have been shown to influence neural entrainment to music (Madsen et al., 2019, Cameron et al., 2019), we investigated the effects of enjoyment, familiarity, and the ease with which a beat was perceived (Fig. 1A). To be able to use a large variety of musical stimuli on the group level, and to decrease any effects that may have arisen from individual stimuli occurring at certain tempi but not others, participants were divided into four subgroups that listened to different pools of stimuli (for more details please see *Materials and Methods*). The subgroups' stimulus pools overlapped, but the individual song stimuli were presented at different tempi for each subgroup.

Musical features

We examined neural synchronization to the time courses of four different musical features (Fig. 1B). First, we quantified energy fluctuations over time as the gammatone-filtered amplitude envelope (we report analyses on the full-band envelope in Supplementary Figures 2 and 4). Second, we computed the half-wave-rectified first derivative of the amplitude envelope, which is typically considered to be sensitive to the presence of onsets in the stimulus (Bello et al., 2005). Third, a percussionist drummed along with the musical segments to define beat times, which were here treated in a binary manner. Fourth, a spectral novelty function, referred to as spectral flux (Mller, 2015), was computed to capture changes in frequency content (as opposed to amplitude fluctuations) over time. In contrast to the first derivative, the spectral flux is better able to identify note onsets that are characterized by changes in spectral content (pitch or timbre), even if the energy level remains the same. To ensure that each musical feature possessed acoustic cues to the stimulation-tempo manipulation, we computed a fast Fourier transform (FFT) on the musical-feature time courses separately for each stimulation-tempo condition; the mean amplitude spectra are plotted in Figure 1C. Overall, amplitude peaks were observed at the intended stimulation tempo and at the harmonic rates for all stimulus features.

In order to assess the degree to which the different musical features might have been redundant, we calculated mutual information (MI) for all possible pairwise feature combinations and compared MI values to surrogate distributions calculated separately for each feature pair (Fig. 1D, E). MI quantifies the amount of information gained about one random variable by observing a second variable (Cover and Thomas, 2005). MI values were analyzed using separate three-way ANOVAs (MI data vs. MI surrogate × Tempo × Subgroup) for each musical feature. Spectral flux shared significant information with all other musical features; significant MI (relative to surrogate) was found between amplitude envelope and spectral flux ($F(1,102)=24.68$, $p_{FDR}=1.01e\text{-}5$, $\eta^2=0.18$), derivative and spectral flux

(F(1,102)=82.3, p_{FDR}=1.92e-13, η^2=0.45) and beat times and spectral flux (F(1,102)=23.05, p_{FDR}=1.3e-5, η^2=0.13). This demonstrates that spectral flux captures information from all three other musical features, and as such, we expected that spectral flux would be associated with strongest neural entrainment. Unsurprisingly, there was also significant shared information between the amplitude envelope and first derivative (F(1,102)=14.11, p_{FDR}=4.67e-4, η^2=0.09; other comparisons: ($F_{env\text{-}beat}$(1,102)=8.44, p_{FDR}=0.006, η^2=0.07; $F_{der\text{-}beat}$(1,102)=6.06, p_{FDR}=0.016, η^2=0.05).

There was a main effect of Tempo on MI shared between the amplitude envelope and derivative (F(12,91)=4, p_{FDR}=0.0002, η^2=0.32) and the spectral flux and beat times (F(12,91)=5.48, p_{FDR}=4.35e-6, η^2=0.37) (Supplementary Fig. 1). This is likely due to the presence of slightly different songs in the different tempo conditions, as the effect of tempo on MI was unsystematic for both feature pairs (see *Materials and Methods* and Supplementary Table 1). MI for the remaining feature pairs did not differ significantly across tempi.

No significant differences in MI were observed between subgroups, despite the subgroups hearing slightly different pools of musical stimuli: ($F_{env\text{-}der}$(3,100)=0.71, p_{FDR}=0.94, η^2=0.01; $F_{env\text{-}beat}$(3,100)=2.63, p_{FDR}=0.33, η^2=0.07; $F_{env\text{-}spec}$(3,100)=0.3, p_{FDR}=0.94, η^2=0.01; $F_{der\text{-}beat}$(3,100)=0.43, p_{FDR}=0.94, η^2=0.01; $F_{der\text{-}spec}$(3,100)=0.46, p_{FDR}=0.94, η^2=0.01; $F_{beat\text{-}spec}$(3,100)=0.13, p_{FDR}=0.94, η^2=0.002).

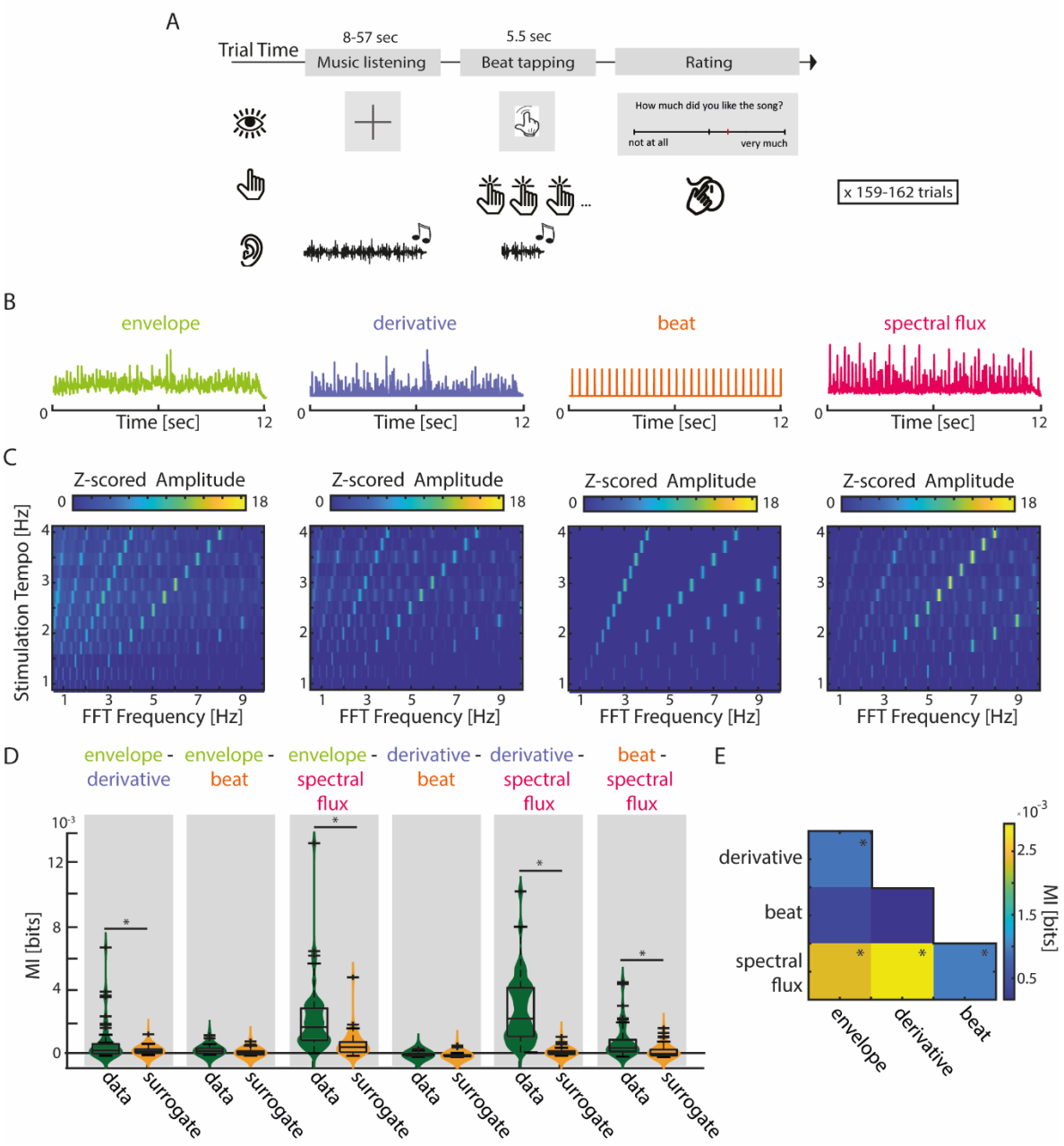

Figure 1. Experimental design and musical features. (A) Schematic of the experimental procedure. Each trial consisted of the presentation of one music segment, during which participants were instructed to listen attentively without moving. After a 1-s silence, the last 5.5 s of the music segment was repeated while participants tapped their finger along with the beat. At the end of each trial, participants rated their enjoyment and familiarity of the music segment, as well as the ease with which they were able to tap to the beat (Translated English example in Figure: "How much did you like the song?" rated from "not at all" to "very much"). **(B)** Exemplary traces of the four musical features of one music segment. **(C)** Z-scored mean amplitude spectrum of all 4 musical features. **(D)** Mutual information (MI) for all possible feature combinations (green) compared to a surrogate distribution (yellow, three-way ANOVA, *p_{FDR}<0.001, rest: p_{FDR}<0.05). Boxplots indicate the median, the 25th and 75th percentiles. **(E)** MI scores between all possible feature combinations (*p_{FDR}<0.001, rest: p_{FDR}<0.05).

190

Neural entrainment was strongest in response to slow music

Neural entrainment to music was investigated using two converging analysis pipelines based on (1) RCA followed by time- (stimulus-response correlation, SRCorr) and frequency- (stimulus-response coherence, SRCoh) domain analysis and (2) TRFs.

First, an RCA-based analysis approach was used to assess tempo effects on neural entrainment to music (Fig. 2, Supplementary Fig. 2). RCA involves estimating a spatial filter that maximizes correlation across data sets from multiple participants (for more details see *Materials and Methods*) (Kaneshiro et al., 2020, Parra et al., 2018). The resulting time course data from a single reliable component can then be assessed in terms of its correlation in the time domain (SRCorr) or coherence in the frequency domain (SRCoh) with different musical feature time courses. Our analyses focused on the first reliable component, which exhibited an auditory topography (Fig. 2A). SRCorrs were significantly tempo-dependent for all four musical features (repeated-measure ANOVAs with Greenhouse-Geiser correction where required: $F_{env}(12,408)=4.5$, p= 8.18e-7, $\eta^2=0.12$; $F_{der}(12,408)=2.5$, p=0.004, $\eta^2=0.07$; $F_{beat}(12,408)=2.5$, p=0.004, $\eta^2=0.07$; $F_{spec}(12,408)=5.84$, $p_{GG}=8.82e\text{-}6$, $\eta^2=0.15$). Highest correlations were found at slower tempi (~1-2 Hz). No significant differences were observed across subgroups ($F_{env}(3,32)=1.11$, $p_{FDR=}0.46$, $\eta^2=0.1$; $F_{der}(3,32)=0.88$, $p_{FDR}=0.46$, $\eta^2=0.08$; $F_{beat}(3,32)=1.5$, $p_{FDR}=0.46$, $\eta^2=0.12$; $F_{spec}(3,32)=2.05$, $p_{FDR}=0.26$, $\eta^2=0.16$). In the frequency domain, normalized SRCoh (Fig. 2E-H) showed clear peaks at the stimulation tempo and harmonics. Overall, SRCoh was stronger at the first harmonic of the stimulation tempo than at the stimulation tempo itself, regardless of the musical feature (Fig. 2E-I). This effect was significant for the envelope, derivative and spectral flux (Fig. 2I, paired-sample t-test, envelope: $t(12)=-4.21$, $p_{FDR}=0.005$, $r_e=0.65$; derivative: : $t(12)=-3.09$, $p_{FDR}=0.03$, $r_e=0.53$; beat: : $t(12)=-2.43$, $p_{FDR}=0.07$, $r_e=0.44$; spectral flux: : $t(12)=-8.26$, $p_{FDR}=2.25e\text{-}5$, $r_e=0.86$). The stimuli themselves mostly also contained highest FFT amplitudes at the first harmonic (Fig. 2J, envelope: $t(12)=-6.81$, $p_{FDR}=$ 5.23e-5, $r_e=0.81$; derivative: $t(12)=-6.88$, $p_{FDR}=$ 5.23e-

5, r_e=0.81; spectral flux: t(12)=-8.04, p_{FDR}= 2.98e-5, r_e=0.85), apart from the beat onsets (beat: t(12)=6.27, p_{FDR}= 8.56-5. r_e=0.79). For evaluating tempo-dependent effects, we averaged SRCoh across the stimulation tempo and first harmonic and submitted the average SRCoh values to repeated-measure ANOVAs for each musical feature. Highest SRCoh was found for slow music ($F_{env}(12,408)=2.58$, $p_{GG}=0.02$, $\eta^2=0.07$; $F_{der}(12,408)=2.76$, $p_{GG}=0.01$, $\eta^2=0.08$; $F_{beat}(12,408)=1.29$, $p_{GG}=0.25$, $\eta^2=0.04$; $F_{spec}(12,408)=3.86$, $p_{GG}=0.002$, $\eta^2=0.1$). No significant differences for the SRCoh were observed across subgroups ($F_{env}(3,32)=1.45$, $p_{FDR}=0.38$, $\eta^2=0.11$; $F_{der}(3,32)=1.39$, $p_{FDR}=0.38$, $\eta^2=0.1$; $F_{beat}(3,32)=1.2$, $p_{FDR}=0.38$, $\eta^2=0.09$; $F_{spec}(3,32)=1.07$, $p_{FDR}=0.38$, $\eta^2=0.08$). Individual data examples of the SRCorr and SRCoh can be found in Supplementary Figure 3.

Second, TRFs were calculated for each stimulation tempo. A TRF-based approach is a linear-system identification technique that serves as a filter describing the mapping of stimulus features onto the neural response (forward model) (Crosse et al., 2016). Using linear convolution and ridge regression to avoid overfitting, the TRF was computed based on mapping each musical feature to "training" EEG data. Using a leave-one-trial-out approach, the EEG response for the left-out trial was predicted based on the TRF and the stimulus feature of the same trial. The predicted EEG data were then correlated with the actual, unseen EEG data (we refer to this correlation value throughout as *TRF correlation*). We analyzed the two outputs of the TRF analysis: the filter at different time lags, which typically resembles evoked potentials, and the TRF correlations (Fig. 3, Supplementary Fig. 4). Again, strongest neural entrainment (here quantified as Pearson correlation coefficient between the predicted and actual EEG data) was observed for slower music (Fig. 3A). Repeated-measure ANOVAs showed that, significant effects of Tempo were observed for all musical features, with TRF correlations being strongest at slower tempi (~1-2 Hz) ($F_{env}(12,408)=2.35$, $p_{GG}=0.02$, $\eta^2=0.06$; $F_{der}(12,408)=1.82$, $p=0.04$, $\eta^2=0.05$; $F_{beat}(12,408)=2.29$, $p_{GG}=0.03$, $\eta^2=0.06$; $F_{spec}(12,408)=8.54$, $p_{GG}=2.36e-9$, $\eta^2=0.22$).

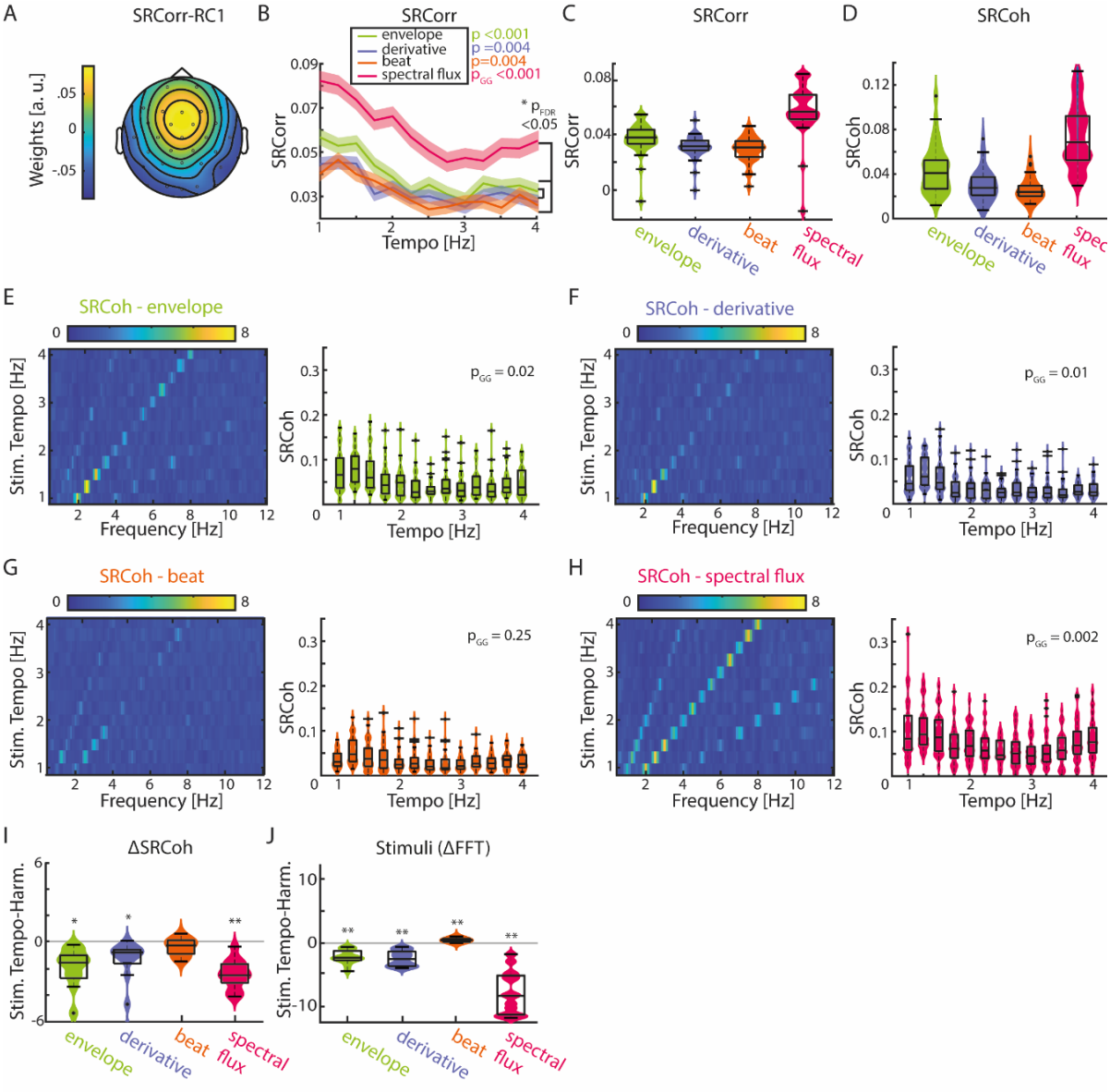

Figure 2. Stimulus–response correlation and stimulus–response coherence are tempo dependent for all musical features. (A) Projected topography of the first reliable component (RC1). **(B)** Average SRCorr across tempi for each musical feature (±SEM; shaded area). Highest correlations were found at slow tempi. Significant differences between tempi were assessed using a repeated-measure ANOVA (with Greenhouse-Geiser correction where applicable) and the slopes of regression models were used to compare the tempo-specificity between musical features. **(C)** Mean SRCorr across musical features. Highest correlations were found in response to spectral flux. There were significant differences between all possible feature combinations except between the derivative and beat onset features (derivative-beat: p_{FDR}=0.37; (repeated-measure ANOVA, Tukey's test, p_{FDR}<0.001). Boxplots illustrate the median, 25th and 75th percentiles. **(D)** Same as (C) for the frequency based SRCoh. All possible feature combinations were significantly different from each other apart from the derivative and beat onsets (derivative-beat: p_{FDR}=0.08; p_{FDR}<0.001). Coherence values were averaged over the stimulus tempo and first harmonic. Normalized SRCoh in response to the **(E)** amplitude envelope, **(F)** first derivative, **(G)** beat onsets and **(H)** spectral flux. Each panel depicts the stimulus response coherence as colorplot (left) and the pooled SRCoh values at the stimulation tempo and first harmonic (right). **(I)** Mean differences of SRCoh values at the stimulation tempo and the first

harmonic (negative values: higher SRCoh at harmonic, positive values: higher SRCoh at stimulation tempo, paired-sample t-test, *p_{FDR}<0.05; **p_{FDR}<0.001). **(J)** Same as (I) based on the FFT amplitudes of each musical feature.

Spectral flux drives strongest neural entrainment

As natural music is a complex, multi-layered auditory stimulus, we sought to explore the neural response to different musical features and to identify the stimulus feature or features that would evoke strongest neural entrainment. Regardless of the dependent measure (RCA-SRCorr, RCA-SRCoh, TRF correlation), strongest neural entrainment was found in response to the spectral flux (Fig. 2C-D, 3B). In particular, significant differences (as quantified with a repeated-measure ANOVA followed by Tukey's test) were observed between the spectral flux and all other musical features using the SRCorr ($F_{SRCorr}(3,140)=33.41$, p_{GG}= 4.01e-15, $\eta^2=0.5$), SRCoh ($F_{SRCoh}(3,140)=38.83$, p_{GG} =5.53e-10, $\eta^2=0.53$) and TRF correlations ($F_{TRF}(4,175)=24.56$, p_{GG}=2.2e-10, $\eta^2=0.42$).

As the TRF approach offers the possibility of running a multivariate analysis, all musical features were combined and compared to the single-feature TRF correlations (Fig. 3B). Although there was a significant increase in TRF correlations in comparison to the amplitude envelope (repeated-measure ANOVA with follow-up Tukey's test, p_{FDR}=1.66e-8), first derivative (p_{FDR} =1.66e-8) and beat onsets (p_{FDR}=1.66e-8), the spectral flux alone showed an advantage over the multi-featured TRF (p_{FDR}=2.18e-4). Thus, taking all stimulus features together is not a better descriptor of the neural response than the spectral flux alone, indicating together with the MI results from Figure 1 that spectral flux is a more complete representation of the rhythmic structure of the music than the other musical features.

To test how strongly modulated TRF correlations were by each musical feature, a regression line was fitted to single-participant TRF correlations as a function of tempo, and the slopes were compared across musical features (Fig. 3A). Linear slopes were significantly higher for the spectral flux and the multivariate model compared to the remaining three

musical features with the exception of the slopes of the multivariate model and envelope (one-way ANOVA, envelope-spectral flux: p_{FDR}= 0.005; envelope – all: p_{FDR}= 0.06; derivative-spectral flux: p_{FDR}= 3e-4; derivative – all: p_{FDR}= 0.005; beat-spectral flux: p_{FDR}= 3e-4; beat – all: p_{FDR}= 0.006). The results for SRCorr were qualitatively similar (envelope-spectral flux: p_{FDR}= 0.046; derivative-spectral flux: p_{FDR}= 3e-4; beat-spectral flux: p_{FDR}= 0.002; Fig. 2B).

We also examined the time courses of TRF weights (Fig. 3C–F) for time lags between 0 and 400 ms. Cluster-based permutation testing (1000 repetitions) was used to identify time windows in which TRF weights differed across tempi for each musical feature (see *Materials and Methods* for more details). Significant effects of tempo on TRF weights were observed for the beat times at time lags between 164–351 ms (p=0.036) and for the spectral flux between 101–242 ms (p<0.001) and 312–398 ms (p=0.033) (Fig. 3E-I). For these two musical features, the tempo specificity was observable in the amplitudes of the TRF weights, which were largest for slower music (Fig. 3G-I). The TRFs for the amplitude envelope and first derivative demonstrated similar patterns to each other, with strong deflections in time windows consistent with a canonical auditory P1–N1–P2 complex, but did not differ significantly between stimulation tempi (Fig. 3C-D). In contrast, the full-band (Hilbert) amplitude envelope and the corresponding first derivative (Supplementary Fig. 4) displayed tempo-specific effects at time lags of 172–305 ms (envelope, p=0.01) and 219–344 ms (derivative, p=0.01). Visual inspection suggested that TRF differences for these musical features were related to latency, as opposed to amplitude (Supplementary Fig. 4E-F, I-J). Therefore, we identified the latencies of the TRF-weight time courses within the time window of the N2, and fit a piece-wise linear regression to those mean latency values per musical feature (Supplementary Fig. 4G, K). In particular, TRF latency in the N2 time window decreased over the stimulation tempo conditions from 1–2.5 Hz and from 2.75–4 Hz for both stimulus features (envelope: $T_{1\text{-}2.5Hz}$=-0.86, p=0.43; $T_{2.75\text{-}4Hz}$=-2.04, p=0.11), but this was only significant for the derivative ($T_{1\text{-}2.5Hz}$=-4.44, p=0.007; $T_{2.75\text{-}4Hz}$=-4.05, p=0.016).

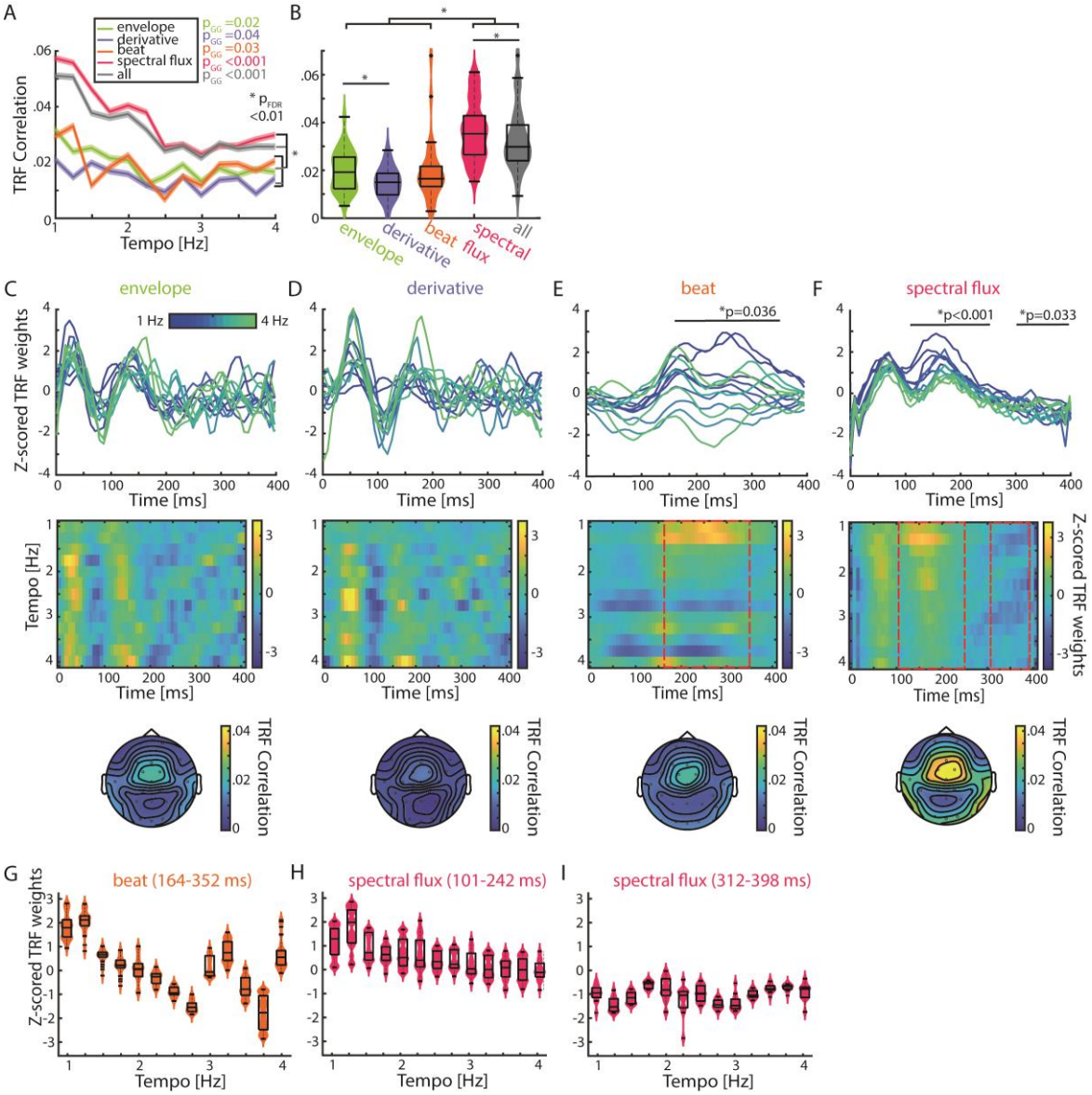

Figure 3. TRFs are tempo dependent. (A) Mean TRF (±SEM) correlations as a function of stimulation tempo per stimulus feature (p-values next to the legend correspond to a repeated-measure ANOVA across tempi for every musical feature and p-value below the legend to a one-way ANOVA based on the slopes of a linear regression model). TRF correlations were highest for spectral flux and combined musical features for slow tempi. **(B)** Violin plots of the TRF correlations across musical features. Boxplots illustrate the median, 25th and 75th percentiles (n=36). Significant pairwise musical feature comparisons were calculated using a repeated-measure ANOVA with follow-up Tukey's test, *p_{FDR}<0.001. **(C)** Top panel: TRF time lags (0-400 ms) of the amplitude envelope. Each line depicts one stimulation tempo (13 tempi between 1 Hz, blue and 4 Hz, green). Middle panel: Colormap of the normalized TRF weights of the envelope in the same time window across stimulation tempi. Lower panel: Topographies of the TRF correlations in response to the amplitude envelope. **(D)** Same as (C) for the first derivative, **(E)** beat onsets and **(F)** spectral flux. Cluster-based permutation testing was used to identify significant tempo-specific time windows (red dashed box, p<0.05). Mean TRF weights in response to **(G)** beat onsets from the significant time lag window (164-351 ms), **(H)** spectral flux for time lags between 101-242 ms and **(I)** between 312-398 ms.

293

Results of TRF and SRCorr/SRCoh converge

So far, we demonstrated that both RCA- and TRF-based measures of neural entrainment lead to similar results at the group level, and reveal strongest neural entrainment to spectral flux and at slow tempi. Next, we wanted to quantify the relationship between the SRCorr/SRCoh and TRF correlations across individuals (Fig. 4, Supplementary Fig. 3). This could have implications for the interpretation of studies focusing only on one method. To test this relationship, we predicted TRF correlations from SRCorr or SRCoh values (fixed effect) in separate linear mixed-effects models with Participant and Tempo as random effects (grouping variables). Each musical feature was modeled independently. For all four musical features, SRCorr significantly predicted TRF correlations ($t_{env}(466) = 6.46$, $\beta_{env} = 0.38$, $p_{FDR} = 5.37e\text{-}10$, $R^2 = 0.41$; $t_{der}(466) = 3.63$, $\beta_{der} = 0.22$, $p_{FDR} = 4e\text{-}4$, $R^2 = 0.16$; $t_{beat}(466) = 2.64$, $\beta_{beat} = 0.26$, $p_{FDR} = 0.009$, $R^2 = 0.41$; $t_{spec}(466) = 11.62$, $\beta_{spec} = 0.52$, $p_{FDR} = 5.78e\text{-}27$, $R^2 = 0.73$). The strongest correlations between neural entrainment measures were found for the spectral flux of music (Fig. 4D). In the frequency domain, we examined the SRCoh values at the stimulation tempo and first harmonic separately (Supplementary Fig. 5). SRCoh values at both the intended stimulation tempo and the first harmonic significantly predicted TRF correlations for all musical features. For all musical features, the intended stimulation tempo was a better predictor of TRF correlations than the first harmonic except for the spectral flux and derivative (intended tempo: $t_{env}(466) = 4.47$, $\beta_{env} = 0.15$, $p_{FDR} = 2.62e\text{-}05$, $R^2 = 0.32$; $t_{der}(466) = 2.03$, $\beta_{der} = 0.07$, $p_{FDR} = 0.04$, $R^2_r = 0.11$; $t_{beat}(466) = 2.37$, $\beta_{beat} = 0.12$, $p_{FDR} = 0.02$, $R^2 = 0.37$; $t_{spec}(466) = 3.65$, $\beta_{spec} = 0.1$, $p_{FDR} = 4e\text{-}4$, $R^2 = 0.62$; first harmonic: $t_{env}(466) = 5.89$, $\beta_{env} = 0.08$, $p_{FDR} = 2.96e\text{-}8$, $R^2 = 0.26$; $t_{der}(466) = 3.86$, $\beta_{der} = 0.06$, $p_{FDR} = 2e\text{-}4$, $R^2 = 0.1$; $t_{beat}(466) = 3.12$, $\beta_{beat} = 0.1$, $p_{FDR} = 0.003$, $R^2_t = 0.35$; $t_{spec}(466) = 6.39$, $\beta_{spec} = 0.09$, $p_{FDR} = 3.16e\text{-}9$, $R^2 = 0.62$). Overall, these results suggest that, despite their differences, TRF and RCA–SRCorr/RCA-SRCoh pick up on similar features of the neural response, but may potentially strengthen each other's explanatory power when used together.

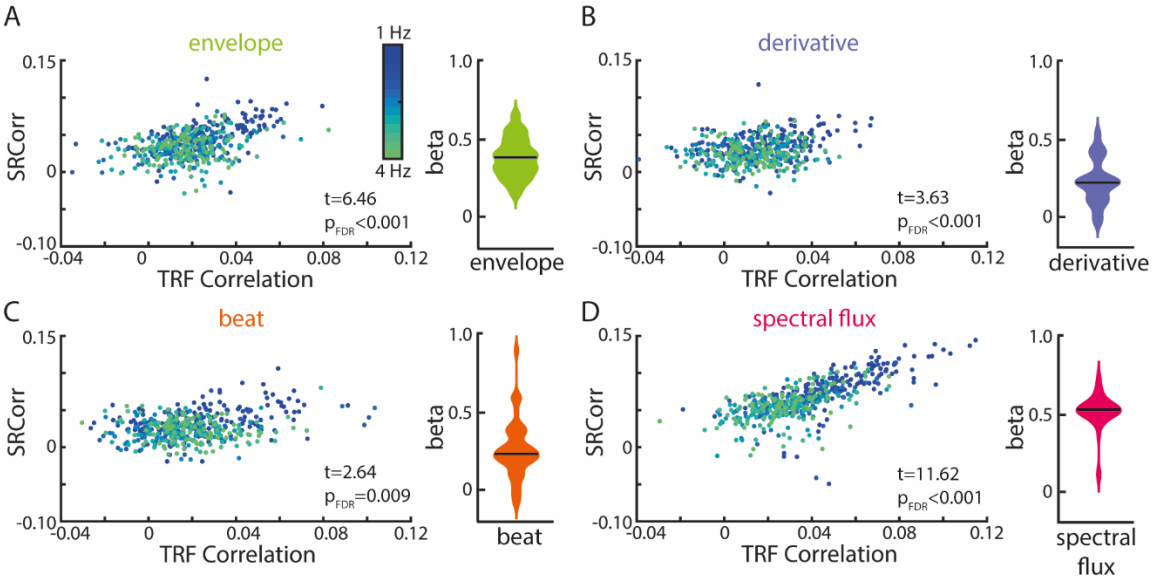

Figure 4. Significant relationships between SRCorr and TRF correlations for all musical features. (A) Linear-mixed effects models of the SRCorr (predictor variable) and TRF correlations (response variable) in response to the amplitude envelope. Each dot represents the mean correlation of one participant (n=36) at one stimulation tempo (n=13) (=grouping variables; blue, 1 Hz-green, 4 Hz). Violin plots illustrate fixed effects coefficients (β). **(B)-(D)** same as (A) for the first derivative, beat onsets and spectral flux. For all musical features, the fixed effects were significant.

Familiar songs and songs with an easy-to-tap beat drive strongest neural entrainment

Next, we tested whether neural entrainment to music depended on 1) how much the song was enjoyed, 2) the familiarity of the song, and 3) how easy it was to tap the beat of the song; each of these characteristics was rated on a scale ranging between –100 and +100. We hypothesized that difficulty to perceive and tap to the beat in particular would be associated with weaker neural entrainment. Ratings on all three dimensions are shown in Figure 5A. To evaluate the effects of tempo on the individual's ratings, separate repeated-measure ANOVAs were conducted for each behavioral rating. Although enjoyment ($F(12,408)=1.2$, $p_{GG}=0.31$, $\eta^2=0.03$) and familiarity ($F(12,408)=1.93$, $p_{GG}=0.09$, $\eta^2=0.05$) were unaffected by tempo, participants indicated that it was more difficult to tap to the beat of faster than slower stimuli ($F(12,408)=6.3$, $p_{GG}=4.71e\text{-}06$, $\eta^2=0.17$).

To assess the effects of familiarity, enjoyment, and beat-tapping ease on neural entrainment, TRFs in response to spectral flux were calculated for the 15 trials with the

highest and the 15 trials with the lowest ratings per participant per behavioral rating condition (Fig. 5B-F). TRF correlations were not significantly different for less enjoyed compared to more enjoyed music (paired-sample t-test, $t(35)=1.22$, $p_{FDR}=0.23$, $r_e=0.24$; Fig. 5C). In contrast, significantly higher TRF correlations were observed for familiar vs. unfamiliar songs ($t(35)=-2.88$, $p_{FDR}=0.02$, $r_e=0.51$), and there was a trend for stronger neural entrainment to songs with an easier-to-perceive beat ($t(35)=-1.94$, $p_{FDR}=0.09$, $r_e=0.37$). These results were reflected in the TRFs at time lags between 0-400 ms (Fig. 5D-F).

Next, we wanted to entertain the possibility that musical training could modulate neural entrainment to music. Therefore, participants with less than 2 years of regular, daily music training were assigned to a "non-musician" group (n=17) and participants with over 6 years of regular music training were labelled as "musicians" (n=14). Although there is little agreement about the specific criterion that should be used to defined musician and non-musician participants, this division had the advantages that it ignored participants with medium amounts of training and it roughly equally divided our sample. Subsequently, TRF correlations were compared between groups (Supplementary Fig. 6). Regardless of the stimulus feature, no significant differences were detected between participants with different levels of musical expertise (paired-sample t-test, envelope: $p_{FDR}=0.93$; derivative: $p_{FDR}=0.93$; beats: $p_{FDR}=0.85$; spectral flux: $p_{FDR}=0.93$). Moreover, the Goldsmith's Musical Sophistication Index (Gold-MSI) was used to quantify musical "sophistication" (referring not only to the years of musical training, but also e. g. musical engagement or self-reported perceptual abilities (Müllensiefen et al., 2014)), which we then correlated with neural entrainment. No significant correlations were observed between musical sophistication and TRF correlations (Pearson correlation, envelope: $R=-0.22$, $p_{FDR}=0.26$; derivative: $R=-0.16$, $p_{FDR}=0.26$; beats: $R=-0.25$, $p_{FDR}=0.34$; spectral flux: $R=-0.29$, $p_{FDR}=0.26$; Supplementary Fig. 6).

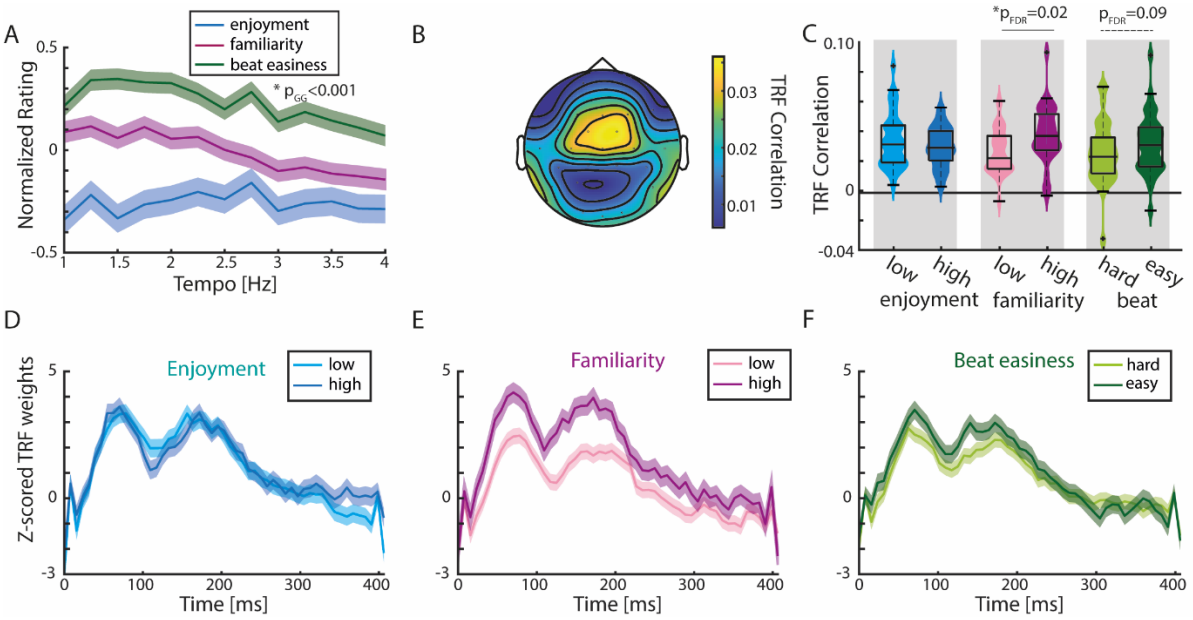

Figure 5. TRF correlations are highest in response to familiar songs. (A) Normalized (to the maximum value per rating/participant), averaged behavioral ratings of enjoyment, familiarity and easiness to tap to the beat (±SEM). Significant differences across tempo conditions were observed (repeated-measure ANOVA with Greenhouse-Geiser correction). **(B)** Mean TRF correlations topography across all ratings (based on the analysis of 15 trials with highest and lowest ratings per behavioral measure). **(C)** Violin plots of TRF correlations comparing low vs. highly enjoyed, low vs. highly familiar, and subjectively difficult vs. easy beat trials. Strongest TRF correlations were found in response to familiar music (paired-sample t-test, p_{FDR}=0.02). **(D)** Mean TRFs (±SEM) for time lags between 0-400 ms of more and less enjoyable music songs. **(E)-(F)** Same as (D) for trials with low vs. high familiarity and difficult vs. easy beat ratings.

Brain responses to musical features predicts perceived beat rate

In natural music, the beat can be perceived at multiple metrical levels. For that reason, it was possible that listeners did not perceive the beat at the tempo we intended (the stimulation tempo), but may have instead perceived the beat at double or half that rate. Thus, we wanted to explore whether our TRF-based measures of neural entrainment simply reflected the stimulus tempo that we presented, or whether they might be sensitive to perceived beat rate when that differed from the stimulation tempo, i.e., the intended beat rate. For this analysis, we made use of the tapping data that were collected in the final part of each trial, during which participants finger-tapped to the beat for 5.5 s. Trials with at least three consistent taps were assigned to a perceived tempo condition (1-4 Hz in steps of 0.25-Hz, see *Materials and Methods* for more details). In this study, we will use the term "stimulation tempo" to refer to

the predominant beat frequency in each music segment, whereas we will use the term "tapped beat rate" when referring to the tapped frequency. The preferred tapped beat rate on the group level was ~1.55 Hz (Supplementary Fig. 7C, mode of skewed Gaussian fitted to mean histograms of the relative number of trials per tapped beat rate).

We wanted to test if we could identify the stimulation tempo (chosen by us) or the tapped beat rate (rate the participant tapped to) based on the neural data, in particular when the stimulation tempo and the tapped beat rate were different. We used a support vector machine (SVM) classifier to first, predict the stimulation tempo (Fig. 6A-B) and second, to predict the perceived (tapped) rate based on the neural response to different musical features (Fig. 6C-D). For predicting the stimulation tempo, we identified two sets of 6 trials (per participant) each, one set where the participants tapped the intended stimulation tempo and the other set where they tapped the same rate, but the intended stimulation tempo was twice as fast as what the participants tapped, i.e., participants tapped the subharmonic of the stimulation tempo. We were able to do this for 19 of our 36 participants. Next, TRFs were computed in response to each musical feature for each set of trials (tapped rate = intended stimulation tempo vs. same tapped rate = 2*stimulation tempo). The SVMs were computed using bootstrapping (100 repetitions) and a leave-one-out approach. The mean SVM prediction accuracies for each musical feature were compared to a surrogate distribution generated by randomly shuffling the tempo labels (tapped rate = intended stimulation tempo vs. same tapped rate = 2*stimulation tempo) when training the SVM classifier. We observed significantly higher prediction accuracies in comparison to the surrogate data for all musical features (paired-sample t-test, envelope: $t(18)=51.89$, $p_{FDR}<1e\text{-}15$, $r_e=0.996$; derivative: $t(18)=124.4$, $p_{FDR}< 1e\text{-}15$, $r_e=0.999$; beat onsets: $t(18)=78.91$, $p_{FDR}<1e\text{-}15$, $r_e=0.998$; spectral flux: $t(18)=99.92$, $p_{FDR}<1e\text{-}15$, $r_e=0.998$; Fig. 6A). This shows that even if the perceived tempo of two musical pieces is the same, the intended (acoustic) stimulation tempo evokes varying levels of neural entrainment. For comparing the prediction accuracies across musical

features, an accuracy index($(Accuracy_{Data}-Accuracy_{Surr})/(Accuracy_{Data}+Accuracy_{Surr})$) was submitted to a repeated-measure ANOVA. No significant differences between musical features were observed ($F(3,72)=0.83$, $p=0.49$, $\eta^2=0.05$; Fig.6B).

Next, the neural response to different musical features were used to predict the tapped beat rate for sets of trials with the same stimulation tempo (intended stimulation tempo = tapped rate vs. same stimulation tempo = 2*tapped rate). Analogous to the previously described analysis pipeline, 13 individual datasets from different tempo conditions (this time from only 9 participants with each one dataset and two participants with each two datasets to increase the sample size) were identified that met the criterion. All SVM classifier prediction accuracies yielded significant differences in comparison to the surrogate data (paired-samples t-test, envelope: $t(12)=52.02$, $p_{FDR}= 2.22e-15$, $r_e=0.996$; derivative: $t(12)=122.61$, $p_{FDR}< 1e-15$, $r_e=0.999$; beat onsets: $t(12)=44.57$, $p_{FDR}= 1.07e-14$, $r_e=0.994$; spectral flux: $t(12)=54.31$, $p_{FDR}=2e-15$, $r_e=0.996$; Fig. 6C), suggesting that entrained neural responses also possess unique signatures of the perceived beat rate, even when it is different from the stimulation tempo. No significant differences in predicting the tapped beat rate between musical features were observed ($F(3,48)=0.17$, $p=0.92$, $\eta^2=0.02$; Fig. 6D).

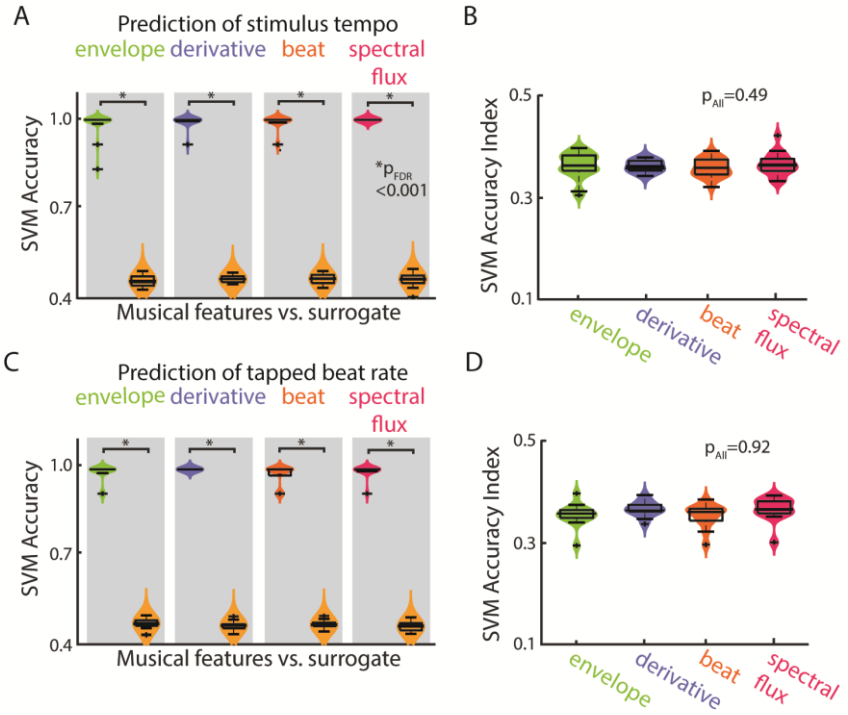

Fig. 6. Tapped beat rate can be predicted based on the neural response to musical features. (A) Violin plot of the mean accuracies of a support vector machine (SVM) classifier predicting the stimulation tempo (n=19; tapped rate = intended stimulation tempo vs. same tapped rate = 2*stimulation tempo). Based on the TRFs to all musical features, significant differences in prediction accuracies were computed in comparison to a surrogate (paired-sample t-test, $*p_{FDR}<0.001$). **(B)** Comparison of SVM classifier accuracies $((Accuracy_{Data}-Accuracy_{Surr})/(Accuracy_{Data}+Accuracy_{Surr}))$ across musical features revealed no significant differences in predicting the stimulation tempo (repeated-measure ANOVA, p=0.92). **(C)-(D)** Same as (A)-(B), but here the SVM classifier predicted the tapped rate based on the TRFs (n=13; intended stimulation tempo = tapped rate vs. same stimulation tempo = 2*tapped rate) (paired-sample t-test, $p_{FDR}<0.001$). No differences were observed in SVM prediction accuracies across musical features (repeated-measure ANOVA, p=0.92).

Discussion

We investigated neural entrainment to naturalistic, polyphonic music presented at different tempi. The music stimuli varied along a number of dimensions in idiosyncratic ways, including the familiarity and enjoyment of the music, and the ease with which the beat was perceived. The current study demonstrates that neural entrainment is strongest to 1) music with beat rates between 1 and 2 Hz, 2) spectral flux of music, 3) familiar music and music with an easy-to-perceive beat. In addition, 4) brain responses to the music stimuli were

informative regarding the listeners' perceived metrical level of the beat, and 5) analysis approaches based on TRF and RCA revealed converging results.

Neural entrainment was strongest to music with beat rates in the 1–2 Hz range

Strongest neural entrainment was found in response to stimulation tempi between 1 and 2 Hz in terms of SRCorr (Fig. 2B), TRF correlations (Fig. 3A), and TRF weights (Fig. 3C-F). Moreover, we observed a behavioral preference to tap to the beat in this frequency range, as the group preference for music tapping was at 1.55 Hz (Supplementary Fig. 7C). Previous studies have shown a preference to listen to music with beat rates around 2 Hz (Bauer et al., 2015), which is moreover the modal beat rate in Western pop music (Moelants, 2002) and the rate at which the modulation spectrum of natural music peaks (Ding et al., 2017). Even in nonmusical contexts, spontaneous adult human locomotion is characterized by strong energy around 2 Hz (MacDougall and Moore, 2005). Moreover, when asked to rhythmically move their bodies at a comfortable rate, adults will spontaneously move at rates around 2 Hz (McAuley et al., 2006) regardless whether they use their hands or feet (Rose et al., 2020). Thus, there is a tight link between preferred rates of human body movement and preferred rates for the music we make and listen to that was moreover reflected in our neural data. This is perhaps not surprising, as musical rhythm perception activates motor areas of the brain, such as the basal ganglia and supplementary motor area (Grahn and Brett, 2007), and is further associated with increased auditory–motor functional connectivity (Chen et al., 2008). In turn, involving the motor system in rhythm perception tasks improves temporal acuity (Morillon et al., 2014), but only for beat rates in the 1–2 Hz range (Zalta et al., 2020).

In the frequency domain, SRCoh was strongest at the stimulation tempo and its harmonics (Fig. 2E-I). In fact, highest coherence was observed at the first harmonic and not at the stimulation tempo itself (Fig. 2I). This replicates previous work that also showed higher coherence (Kaneshiro et al., 2020) and spectral amplitude (Tierney and Kraus, 2015) at the first harmonic than at the musical beat rate. There are several potential reasons for this

finding. One reason could be that the stimulation tempo that we defined for each musical stimulus was based on beat rate, but natural music can be subdivided into smaller units (e.g., notes) that can occur at faster time scales. A recent MEG study demonstrated inter-trial phase coherence for note rates up to 8 Hz (Doelling and Poeppel, 2015). Hence, the neural responses to the music stimuli in the current experiment likely tracked not only the beat rate, but also faster elements such as notes. In line with this hypothesis, FFTs conducted on the stimulus features themselves showed higher amplitudes at the first harmonic than the stimulation tempo for all musical features except the beat onsets (Fig. 2J). Moreover, there are other explanations for higher coherence at the first harmonic than at the beat rate. For example, the low-frequency beat-rate neural responses fall into a steeper part of the 1/f slope, and as such may simply suffer from worse signal-to-noise ratio than their harmonics.

Regardless of the reason, since frequency-domain analyses separate the neural response into individual frequency-specific peaks, it is easy to interpret neural tracking (SRCoh) or stimulus spectral amplitude at the beat rate and the note rate – or at the beat rate and its harmonics – as independent (Keitel et al., 2021). However, music is characterized by a nested, hierarchical rhythmic structure, and it is unlikely that neural tracking at different metrical levels goes on independently and in parallel. One potential advantage of TRF-based analyses is that they operate on relatively wide-band data compared to Fourier-based approaches, and as such are more likely to preserve nested neural activity and perhaps less likely to lead to over- or misinterpretation of frequency-specific effects.

Neural entrainment is driven by spectral flux

Neural entrainment was strongest in response to the spectral flux of music, regardless whether the analysis was based on TRFs or RCA. Similar to speech-tracking studies, music-tracking studies typically use the amplitude envelope of the sound to characterize the stimulus rhythm (Vanden Bosch der Nederlanden et al., 2020, Kumagai et al., 2018, Doelling and Poeppel,

2015, Decruy et al., 2019, Reetzke et al., 2021). Although speech and music share features such as amplitude fluctuations over time and hierarchical grouping (Patel, 2003), there are differences in their spectro-temporal composition that make spectral information especially important for music perception. For example, while successful speech recognition requires 4-8 spectral channels, successful recognition of musical melodies requires at least 16 spectral channels (Shannon, 2005) – the flipside of this is that music is more difficult than speech to understand based only on amplitude-envelope information. Moreover, increasing spectral complexity of a music stimulus enhances neural entrainment (Wollman et al., 2020). Critically, both temporal and spectral information influence the perceived accent structure in music (Pfordresher, 2003).

A recent study claimed that neuronal activity synchronizes less strongly to music than to speech (Zuk et al., 2021); notably they focused specifically on amplitude envelope to characterize the stimulus rhythms. We argue that the amplitude envelope – even when passed through a model of the peripheral auditory system – is a suboptimal measure to approximate individual note onsets that convey rhythmic structure in music and to which neural activity can be entrained (Mller, 2015). Imagine listening to a melody played in a *glissando* fashion on a violin. There might never be a clear onset that would be represented by the amplitude envelope – all of the rhythmic structure is communicated by spectral changes. Thus, in this study we wanted to compare neural entrainment by the amplitude envelope to neural entrainment by spectral flux, which compares spectral content, i.e., power spectra, on a frame-to-frame basis, and which is arguably a more appropriate measure of rhythmic and metrical structure in music. Indeed, many automated tools for extracting the beat in music used in the musical information retrieval (MIR) literature rely on spectral flux information (Oliveira et al., 2010). Also in the context of body movement, spectral flux has been associated with the type and temporal acuity of synchronization between the body and music at the beat rate (Burger et al., 2018) to a greater extent than other acoustic characterizations of musical

499 rhythmic structure. As such, we found that spectral flux drove stronger entrainment than the
500 amplitude envelope.

501 Using TRF analysis, we found that not only was neural entrainment to spectral flux
502 stronger than to any other musical feature, it was also stronger than to the response to a mutli-
503 variate predictor that combined all musical features. For this reason, we calculated the shared
504 information (MI) between each pair of musical features, and found that spectral flux shared
505 significant information with all other musical features (Fig. 1). Hence, spectral flux seems to
506 capture information also contained in, for example, the amplitude envelope, but contains
507 unique information about rhythmic structure that cannot be gleaned from the other acoustic
508 features (Fig. 3). This finding has potentially important implications for direct comparisons of
509 neural tracking of music and speech, or music and natural sounds (Zuk et al., 2021). We
510 would caution that conclusions about differences in how neural activity entrains to different
511 categories of sounds should be sure to characterize stimuli as fairly as possible rather than
512 relying on the amplitude envelope as a one-size-fits-all summary of rhythmic structure.
513
514 *Neural entrainment was strongest to familiar songs and songs with an easy beat*
515 We found that the strength of neural entrainment depended on the familiarity of music and, to
516 a lesser extent, the ease with which a beat could be perceived (Fig. 5). This is in line with a
517 previous study showing stronger neural entrainment to familiar music (Madsen et al., 2019). It
518 is likely that songs a person knows – familiar songs – increase engagement. We note that we
519 did not have a measure of engagement, though engagement has been shown to be a major
520 driver of neural entrainment during film viewing (Dmochowski et al., 2014).

521 There was also a trend for higher neural entrainment to music with subjectively "easy-to-
522 tap-to" beats. However, both neural entrainment and ease of beat tapping were highest for
523 slow stimulation tempi; faster songs were associated with weaker entrainment and were rated
524 as more difficult to tap to. Thus, in the current study, it is not possible to separate the

influences of stimulation tempo and beat salience on neural entrainment. Here, we chose music stimuli with salient, easy-to-perceive beats. However, a design including more "weakly metrical" or syncopated rhythms may have more success in doing so. Overall, we interpret our results as indicating that stronger neural entrainment is evoked in response to music that is more predictable: familiar music and with easy-to-track beat structure.

Musical training did not affect the degree of neural entrainment in response to tempo-modulated music (Supplementary Fig. 6). This contrasts with previous music research showing that musicians' neural activity was entrained more strongly by music than non-musicians' (Madsen et al., 2019, Doelling and Poeppel, 2015, Di Liberto et al., 2020). There are several possible reasons for this discrepancy. One is that our study recruited participants with varying level of musical expertise and did not aim for a specific target group; our study was not intended to examine the role of musical training in neural entrainment. Furthermore, most studies that have done so have focused on classical music (Doelling and Poeppel, 2015, Madsen et al., 2019, Di Liberto et al., 2020), whereas we incorporated music stimuli with different instruments and from different genre (e. g. Rock, Pop, Techno, Western, Hip Hop or Jazz). We suspect that musicians are more likely to be familiar with, in particular, classical music, and as we have shown that familiarity with the individual piece increases neural entrainment, these studies may have inadvertently confounded musical training with familiarity.

Neural responses predicted tempo perception

One interesting yet difficult aspect of music, when it comes to studying entrainment, is that music has metrical structure; that is, there are several levels at which nested periodicities can be perceived. Here, we asked participants to tap along with short sections of each musical stimulus so that we could confirm that their perceived (tapped) beat rate matched our intended stimulation tempo. Although participants mostly tapped at the rate we intended, they

551 sometimes tapped at half or double the intended stimulation tempo, especially when the
552 stimulation tempo was particularly fast or slow, respectively. Here, we applied a classification
553 approach to demonstrate that entrained neural responses to music can predict a) whether
554 participants tapped at double-time or half-time to stimuli with the same stimulation tempo, or
555 b) whether stimuli to which participants tapped identically belonged to the double-time or
556 half-time stimulation-tempo condition. Importantly, neural activity was measured in response
557 to auditory stimulation (without movement) and the perceived metrical level was based on the
558 beat tapping rate established in a separate part of each trial after the listening portion was
559 over. To our knowledge, this study constitutes the first to successfully identify the specific
560 metrical level at which individuals perceived a beat in the absence of overt movement.
561 Nonetheless, there are a few caveats to mention. First, we chose musical stimuli that all had a
562 relatively easy-to-perceive beat. As a result, only 11 participants had enough trials with
563 metrically ambiguous tapping behaviour to stimuli belonging to the same intended stimulation
564 tempo condition for conducting TRF analysis. Moreover, we initially only included the beat-
565 tapping section of each trial as a verification of the validity of our tempo manipulation. As
566 such, we only collected tapping responses for 5.5 s per trial, and tapping behavior was quite
567 difficult to analyze due to the short tapping epochs, which resulted in many tapping trials
568 being discarded.
569
570 *TRF- and RCA-based measures show converging results*
571 In the present study, we used the TRF and RCA analysis approaches to quantify neural
572 entrainment. Here, we have purposefully avoided the debate about whether these metrics
573 measure entrainment "in the narrow sense" (Obleser and Kayser, 2019), meaning phase-
574 locked and (mainly) unidirectional coupling between a rhythmic input and neural activity
575 generated by a neural oscillator (Lakatos et al., 2019) or whether neural tracking reflects
576 convolution with an evoked response (Zuk et al., 2021). Here, we prefer to remain agnostic,

and refer rather to "entrainment in the broad sense" (Obleser and Kayser, 2019), that is neural tracking of music independent of the underlying physiological mechanism.

RCA and TRF approaches share their ability to characterize neural responses to single-trial, ongoing, naturalistic stimuli. As such, both techniques afford something that is challenging or impossible to accomplish with "classic" ERP analysis. However, we made use of two techniques in parallel in order to leverage their unique advantages. RCA allows for frequency-domain analysis such as SRCoh, which can be useful for identifying neural tracking responses specifically at the beat rate, for example. Past music studies often used a "frequency-tagging" approach for this, which is based on averaging over trials in the time domain (so requires repetition of stimuli) rather than relating the neural response to the stimulus time course, and moreover operates in electrode as opposed to component space (Nozaradan et al., 2012, Nozaradan et al., 2011). TRFs rather take into account wider-band neural data, which may better capture the tracking of nested metrical structure as in music. Moreover, TRFs offer a univariate and multivariate analysis approach that allowed us to show that adding other musical features to the model did not improve the correspondence to the neural data over and above spectral flux alone. Despite their differences, we found strong correspondence between the dependent variables from the two approaches. Specifically, TRF correlations were strongly correlated with stimulation-tempo SRCoh, and this correlation was higher than for SRCoh at the first harmonic of the stimulation tempo for the amplitude envelope, derivative and beat onsets (Supplementary Fig. 5). Thus, despite being computed on a relatively broad range of frequencies, the TRF seems to be correlate with frequency-specific measures at the stimulation tempo.

Conclusions

This study presented new insights into neural entrainment to natural music. We compared neural entrainment to different musical features and showed strongest neural responses to the

603 spectral flux. This has important implications for research on neural entrainment to music
604 research, which has so far often quantified stimulus rhythm with what we would argue is a
605 subpar acoustic feature – the amplitude envelope. Moreover, our findings demonstrate that
606 neural entrainment is strongest for slower beat rates, and for predictable stimuli, namely
607 familiar music with an easy-to-perceive beat.
608

Materials and Methods

Participants

Thirty-seven participants completed the study (26 female, 11 male, mean age = 25.7 years, SD = 4.33 years, age range = 19-36 years); data for 36 were included in the final analysis (see *EEG data preprocessing*). The sample-sizes for all projects funded by the ERC Starting Grant (ERC-STG-804029 BRAINSYNC) were pre-calculated with 24 + 4 individuals for between condition- comparisons experiments and with 32 + 4 individuals for built-in correlational experiments to obtain 80% power for a significant medium-sized effect while allowing to discard ~ 15% of the recorded data (G*Power3). Since the current experiment was designed to have both types of comparisons, we defaulted to the larger sample size. Prior to the EEG experiment, all participants filled out an online survey about their demographic and musical background using LimeSurvey (LimeSurvey GmbH, Hamburg, Germany, http://www.limesurvey.org). All participants self-identified as German speakers. Most participants self-reported normal hearing (7 participants reported occasional ringing in one or both ears). Thirty-four participants were right- and three were left-handed. Seventeen participants reported having no musical background (0-2 years of daily music training, here termed as “non-musicians”) and 14 reported at least 6 years of musical training (“musicians”). Musical expertise was assessed using the Goldsmith Music Sophistication Index (Gold-MSI;(Müllensiefen et al., 2014)). Participants received financial compensation for participating (Online: 2.50 €, EEG: 7€ per 30 min). All participants signed the informed consent before starting the experiment. The study was approved by the Ethics Council of the Max Planck Society Ethics Council in compliance with the Declaration of Helsinki.

Stimuli

The stimulus set started from 39 instrumental versions of musical pieces from different genres, including techno, rock, blues, and hip-hop. The musical pieces were available in a

635 **.wav* format on Qobuz Downloadstore (https://www.qobuz.com/de-de/shop). Each musical
636 piece was segmented manually using Audacity (Version 2.3.3, Audacity Team,
637 https://www.audacityteam.org) at musical phrase boundaries (e.g., between chorus and verse),
638 leading to a pool of 93 musical segments with varying lengths between 14.4 – 38 s. We did
639 not use the beat count from any publicly available beat-tracking softwares, because they did
640 not track beats reliably across genres. Due to the first Covid-19 lockdown, we assessed the
641 original tempo of each musical segment using an online method. Eight tappers, including the
642 authors, listened to and tapped to each segment on their computer keyboard for a minimum of
643 17 taps; the tempo was recorded using an online BPM estimation tool
644 (https://www.all8.com/tools/bpm.htm). In order to select stimuli with unambiguous strong
645 beats that are easy to tap to, we excluded 21 segments due to high variability in tapped
646 metrical levels (if more than 2 tappers tapped different from the others) or bad sound quality.
647 The remaining 72 segments were then tempo-manipulated using a custom-written
648 MAX patch (Max 8.1.0, Cycling ’74, San Francisco, CA, USA). Each segment was shifted to
649 tempi between 1–4 Hz in steps of 0.25 Hz. Subsequently, the authors screened all of the
650 tempo-shifted music and eliminated versions where the tempo manipulation led to acoustic
651 distortions, made individual notes indistinguishable, or excessively repetitive. Overall, 703
652 music stimuli with durations of 8.3–56.6 sec remained. All stimuli had a sampling rate of
653 44,100 Hz, were converted from stereo to mono, linearly ramped with 500-ms fade-in and
654 fade-out and root-mean-square normalized using Matlab (R2018a; The MathWorks, Natick,
655 MA, USA). A full overview of the stimulus segments can be found in the Supplementary
656 Material (Supplementary Table 1).
657 Each participant was assigned to one of four pseudo-randomly generated stimulus
658 lists. Each list comprised 4–4.6 min of musical stimulation per tempo condition (Kaneshiro et
659 al., 2020), resulting in 7–17 different musical segments per tempo and a total of 159–162
660 segments (trials) per participant. Each segment was repeated only once per tempo but was

allowed to occur for up to three times at different tempi within one experimental session (tempo difference between two presentations of the same segment was 0.5 Hz minimum). The presentation order of the musical segment was randomly generated for each participant prior to the experiment. The music stimuli were played at 50 dB sensation level (SL), based on individual hearing thresholds that were determined using the method of limits (Leek, 2001).

Experimental design

After attaching the EEG electrodes and seating the participant in an acoustically and electrically shielded booth, the participant was asked to follow the instructions on the computer screen (BenQ Monitor XL2420Z, 144Hz, 24", 1920x1080, Windows 7 Pro (64-bit)). The auditory and visual stimulus presentation was achieved using custom-written Matlab scripts using Psychtoolbox (PTB-3, (Brainard, 1997)) in Matlab (R2017a; The MathWorks, Natick, MA, USA).

The overall experimental flow for each participant can be found in Figure 1A. First, each participant conducted a self-paced spontaneous motor tempo task (SMT; (Fraisse, 1982)) which is a commonly used technique to assess individual's preferred tapping rate (Rimoldi, 1951, McAuley, 2010). To obtain SMT, each participant tapped for thirty seconds (3 repetitions) at a comfortable rate with a finger on the table close to a contact microphone (Oyster S/P 1605, Schaller GmbH, Postbauer-Heng, Germany). Second, we estimated individual's hearing threshold using the method of limits. All sounds in this study were delivered by a Fireface soundcard (RME Fireface UCX Audiointerface, Audio AG, Haimhausen, Germany) via on-ear headphones (Beyerdynamics DT-770 Pro, Beyerdynamic GmbH & Co. KG, Heilbronn, Germany). After a short three-trial training, the main task was performed. The music stimuli in the main task were grouped into eight blocks with approximately 20 trials per block and the possibility to take a break in between.

Each trial comprised two parts: attentive listening (music stimulation without movement) and tapping (music stimulation + finger tapping; Fig. 1A). During attentive listening, one music stimulus was presented (8.3–56.6 s) while the participant looked at a fixation cross on the screen; the participant was instructed to mentally locate the beat without moving. Tapping began after a 1-s interval; the last 5.5 s of the previously listened musical segment were repeated, and participants were instructed to tap a finger to the beat of the musical segment (as indicated by the replacement of the fixation cross by a hand on the computer screen). Note that 5.5 s of tapping data is not sufficient to conduct standard analyses of sensorimotor synchronization; rather, our goal was to confirm that the participants tapped at the intended beat rate based on our tempo manipulation. After each trial, participants were asked to rate the segment based on *enjoyment*/*pleasure*, *familiarity* and *ease of tapping to the beat* with the computer mouse on a visual analogue scale ranging from -100 to +100. At the end of the experiment, the participant performed the SMT task again for three repetitions.

EEG data acquisition

EEG data were acquired using BrainVision Recorder (v.1.21.0303, Brain Products GmbH, Gilching, Germany) and a Brain Products actiCap system with 32 active electrodes attached to an elastic cap based on the international 10-20 location system (actiCAP 64Ch Standard-2 Layout Ch1-32, Brain Products GmbH, Gilching, Germany). The signal was referenced to the FCz electrode and grounded at the AFz position. Electrode impedances were kept below 10 kOhm. The brain activity was acquired using a sampling rate of 1000 Hz via a BrainAmp DC amplifier (BrainAmp ExG, Brain Products GmbH, Gilching, Germany). To ensure correct timing between the recorded EEG data and the auditory stimulation, a TTL trigger pulse over a parallel port was sent at the onset and offset of each musical segment and the stimulus envelope was recorded to an additional channel using a StimTrak (StimTrak, Brain Products GmbH, Gilching, Germany).

712

713 *Data Analysis*

714 *Behavioral data.* Tapping data was processed offline with a custom-written Matlab script. To
715 extract the taps, the **.wav* files were imported and downsampled (from 44.1 kHz to 2205 Hz).
716 The threshold for extracting the taps was adjusted for each trial manually (SMT and music
717 tapping) and trials with irregular tap intervals were rejected. The SMT results were not
718 analyzed as part of this study and will not be discussed further. For the music tapping, only
719 trials with at least three taps (two intervals) were included for further analysis. Five
720 participants were excluded from the music tapping analysis due to irregular and inconsistent
721 taps within a trial (if > 40% of the trials were excluded).

722 One of our goals was to test whether we could identify trials based on the neural data
723 where the perceived tempo differed from the intended stimulation rate (see *Brain responses to*
724 *musical features can predict the produced beat tapping rate*). For this analysis, we identified
725 two subsets of participants: those that tapped the same tempo to two sets of stimuli with
726 different intended stimulation tempi, and those that tapped the intended stimulation tempo on
727 some trials and a different tempo than what was intended (the harmonic or subharmonic) on
728 other trials. We identified 19 participants that tapped for at least 6 trials at the intended
729 stimulation tempo and tapped for at least 6 trials at the same tempo when the stimulation
730 tempo was something different (double the tapped tempo; i.e., participants tapped at half the
731 intended stimulation tempo). In contrast, we identified 11 participants that tapped for at least
732 6 trials at the intended stimulation tempo and for at least 6 trials at half-/double the
733 stimulation tempo. TRFs were submitted to a SVM classifier (see section *EEG – Temporal*
734 *Response Function*).

735 On each trial, participants were asked to rate the musical segments based on
736 *enjoyment/pleasure*, *familiarity* and *ease to tap to the beat.* The rating scores were normalized
737 to the maximum absolute rating per participant and per category. For the group analysis the

738 mean and standard error of the mean (SEM) were calculated. For assessing the effects of each
739 subjective dimension on neural entrainment, the 15 trials with the highest and lowest ratings
740 (regardless of the tempo) per participant were further analyzed (see *EEG – Temporal*
741 *Response Function*).
742
743 *Audio Analysis.* We assessed neural entrainment to four different musical features (Fig. 1B-
744 C). Note that the term "musical feature" is used to describe time-varying features of music
745 that operate on a similar time-scale as neural entrainment as opposed to the classical musical
746 elements such as syncopation or harmony; 1) Amplitude envelope – gammatone filtered
747 amplitude envelope in the main manuscript and absolute value of the full-band Hilbert
748 envelope in the Supplementary Material; the gammatone filterbank consisted of 128 channels
749 linearly spaced between 60-6000 Hz. 2) Half-wave rectified, first derivative of the amplitude
750 envelope, which detects energy changes over time and is typically more sensitive to onsets
751 (Daube et al., 2019, Di Liberto et al., 2020). 3) Binary-coded beat onsets (0= no beat; 1=beat);
752 a professionally trained percussionist tapped with a wooden drumstick on a MIDI drum pad to
753 the beat of each musical segment at the original tempo (3 trials per piece). After latency
754 correction, the final beat times were taken as the average of the two takes with the smallest
755 difference (Harrison and Müllensiefen, 2018). 4) Spectral novelty ("spectral flux") (Mller,
756 2015) was computed using a custom-written Python script (Python 3.6, Spyder 4.2.0) using
757 the packages *numpy* and *librosa*. For computing the spectral flux of each sound, the
758 spectrogram across frequencies of consecutive frames (frame length = 344 samples) was
759 compared. All stimulus features were z-scored and downsampled to 128 Hz for computing the
760 stimulus-brain synchrony.
761 To validate that each musical feature contained acoustic cues to our tempo
762 manipulation, we conducted a discrete Fourier transform using a Hamming window on each

763 musical segment (resulting frequency resolution of 0.0025 Hz), averaged and z-scored the
764 amplitude spectra per tempo and per musical feature (Fig. 1C).
765 To assess how much information the different musical features share, a mutual
766 information (MI) score was computed between each pair of musical features (Fig. 1D). MI (in
767 bits) is a time-sensitive measure that quantifies the reduction of uncertainty for one variable
768 after observing a second variable (Cover and Thomas, 2005). MI was computed using
769 *quickMI* from the Neuroscience Information Theory Toolbox with 4 bins, no delay, and a p-
770 value cut-off of 0.001 (Timme and Lapish, 2018). For each stimulus feature, all trials were
771 concatenated in the same order for each tempo condition and stimulation subgroup (Time x 13
772 Tempi x 4 Subgroups). MI values for pairs of musical features were compared to surrogate
773 datasets in which one musical feature was time reversed (Fig. 1D). To statistically asses the
774 shared information between musical features, a three-way ANOVA test was performed (with
775 first factor: data-surrogate comparison; second factor: tempo and third factor: stimulation
776 subgroup).
777
778 *EEG data preprocessing.* Unless stated otherwise, all EEG data were analyzed offline using
779 custom-written Matlab code (R2019b; The MathWorks, Natick, MA, USA) combined with
780 the Fieldtrip toolbox (Oostenveld et al., 2011). The continuous EEG data were bandpass
781 filtered between 0.5-30 Hz (Butterworth filter), re-referenced to the average reference,
782 downsampled to 500 Hz, and epoched between 1 s after stimulus onset (to remove onset
783 responses to the start of the music stimulus) until the end of the initial musical segment
784 presentation (attentive listening part of the trial). Single trials and channels containing large
785 artefacts were removed based on an initial visual inspection. Missing channels were
786 interpolated based on neighbouring channels with a maximum distance of 3 cm
787 (*ft_prepare_neighbours*). Subsequently, Independent Component Analysis (ICA) was applied
788 to remove artefacts and eye movements semi-automatically. After transforming the data back

from component to electrode space, electrodes that exceeded 4 standard deviations of the mean (of the squared data) for at least 10% of the recording time were excluded. If bad electrodes were identified, pre-processing for that recording was repeated after removing the identified electrode (Kaneshiro et al., 2020). Next, noisy segments of the single-trial, single-electrode recordings were rejected. For the RCA analysis, the data points were replaced by NaNs when the segment exceeded a threshold of two standard deviations of the single-trial, single-electrode mean amplitude. For the TRF analysis, which does not operate on NaNs, noisy transients were replaced by estimates using spherical spline interpolation with a pre- and post- window length of 0.5 s. This step was repeated four times to ensure that all artefacts were removed (Kaneshiro et al., 2020). The dataset of one participant was discarded because of large artefacts in the EEG signal and for not following the experimental instructions. The behavioral and neural data of the remaining 36 participants were utilized for further analysis.

Next, the data were restructured to match the requirements of the RCA or TRF (see sections *EEG – Temporal Response Function* and *EEG – Reliable Component Analysis*), downsampled to 128 Hz and z-scored. For the RCA analysis approach, the trials in each tempo condition were concatenated resulting in a time-by-electrode matrix (Time x 33 Electrodes; with Time varying across tempo condition). Subsequently the data of participants in the same subgroup were pooled together in a time-by-electrode-by-participant matrix (Time x 33 Electrodes x 9 or 10 Participants depending on the subgroup). In contrast to the RCA, for TRF analysis, trials in the same stimulation condition were not concatenated in time, but grouped into cell arrays per participant according to the stimulus condition (Tempo x Trials x Electrodes x Time).

EEG – Reliable Component Analysis. To reduce data dimensionality and enhance the signal-to-noise ratio, we performed RCA (reliable components analysis, also correlated components analysis) (Dmochowski et al., 2012). RCA is designed to capture the maximum correlation

between datasets of different participants by combining electrodes linearly into a vector space. One important feature of this technique is that it maximizes the correlation between electrodes across participants (which differentiates it from the similar canonical correlation analysis) (Madsen et al., 2019). Using the *rcaRun* Matlab function (Dmochowski et al., 2012, Kaneshiro et al., 2020), the time-by-electrode matrix was transformed to a time-by-component matrix with the maximum across-trial correlation in the first reliable component (RC1), followed by components with correlation values in descending order. For each RCA calculation, for each tempo condition and subgroup, the first three RCs were retained, together with forward-model projections for visualizing the scalp topographies. The next analysis steps in the time and frequency-domain were conducted on the maximally correlated RC1 component.

To examine the correlation between the neural signal and stimulus over time, the stimulus-response correlation (SRCorr) was calculated for every musical feature. This analysis procedure was adopted from (Kaneshiro et al., 2020). In brief, every stimulus feature was concatenated in time with trials of the same tempo condition and subgroup to match the neural component-by-time matrix. The stimulus features were temporally filtered to account for the stimulus–brain time lag, and the stimulus features and neural time-courses were correlated. To create a temporal filter, every stimulus feature was transformed into a Toeplitz matrix, where every column repeats the stimulus-feature time course, shifted by one sample up to a maximum shift of 1 s, plus an additional intercept column. The Moore-Penrose pseudoinverse of the Toeplitz matrix and temporal filter was used to calculate the SRCorr. To report the SRCorr, the mean (± SEM) correlation coefficient across tempo conditions for every stimulus feature was calculated. For comparing tempo-specificity between musical features, a linear regression was fit to SRCorr values (and TRF correlations) as a function of tempo for every participant and for every musical feature (using *fitlm*). We compared the resulting slopes across musical features with a one-way ANOVA

Stimulus-response coherence (SRCoh) is a measure that quantifies the consistency of phase and amplitude of two signals in a specific frequency band and ranges from 0 (no coherence) to 1 (perfect coherence) (Srinivasan et al., 2007). Here, the magnitude-squared coherence between different stimulus features and neural data was computed using the function *mscohere* with a Hamming window of 5 s and 50% overlap, resulting in a frequency range 0–64 Hz with a 0.125 Hz resolution. For visualizing the mean frequency response per musical feature, the coherence values at each stimulation tempo were normalized by dividing by the mean coherence across all other stimulation tempi per frequency bin (Fig. 2E-H) (van Bree et al., 2021). As strong coherence was found at the stimulation tempo and the first harmonic, the SRCoh values of each frequency vector were compared between musical features.

EEG – Temporal Response Function. The TRF is a system identification technique, which computes a filter that optimally describes the relationship between the brain response and stimulus features (Ding and Simon, 2012, Crosse et al., 2016). Via linear convolution, the filter delineates how the stimulus features map onto the neural response (forward model), using ridge regression to avoid overfitting. All computations of the TRF used the Matlab toolbox "The multivariate Temporal Response Function (mTRF) Toolbox" (Crosse et al., 2016). The TRF was calculated in a leave-one-out cross-validation manner for all trials per stimulation tempo; this procedure was repeated for each musical feature separately, and additionally for all musical features together in a multivariate model (using *mTRFcrossval* and *mTRFtrain*) using time lags 0–400 ms (Di Liberto et al., 2020). Using *mTRFpredict*, the neural time course of the left-out trial was predicted based on the time course of the corresponding musical feature of that trial. The quality of the predicted neural data was assessed by computing Pearson correlations between the predicted and actual EEG data separately for each electrode (TRF correlations). We averaged over the eight electrodes with

867 the highest TRF correlations that also corresponded to a canonical auditory topography. To
868 quantify differences in the TRFs, the mean TRF correlation across stimulation tempo and/or
869 musical feature was calculated per participant. The TRF weights across time lags were Fisher-
870 z-scored (Fig. 3C-F) (Crosse et al., 2016).

871 The assessment of TRF weights across time lags was accomplished by using a
872 clustering approach for each musical feature and comparing significant data clusters to
873 clusters from a random distribution (Fig. 3C-F). To extract significant time windows in which
874 the TRF weights were able to differentiate the different tempo conditions, a one-way ANOVA
875 was performed at each time point. Clusters (consecutive time windows) were identified if the
876 p-value was below a significance level of 0.05 and the size and F-statistic of those clusters
877 were retained. Next, the clusters were compared to a surrogate dataset, which followed the
878 same procedure, but had the labels of the tempo conditions randomly shuffled before entering
879 it to the ANOVA. This step was repeated for 1000 times (permutation testing). At the end, the
880 significance of clusters was evaluated by subtracting the proportion of times the summed F-
881 values of each clusters exceeded the summed F-values of the surrogate clusters from 1. A p-
882 value below 0.05 was considered significant (Fig. 3G-I). This approach yielded significant
883 regions for the full-band (Hilbert) envelope and derivative (Supplementary Fig. 4). As these
884 clusters did not show differences across amplitudes but rather in time, a latency analysis was
885 conducted. Therefore, local minima around the grand average minimum within the significant
886 time lag window were identified for every participant/tempo condition and the latencies
887 retained. As there was no significant correlation between latencies and tempo conditions, the
888 stimulation tempi were split upon visual inspection into two groups (1-2.5 Hz and 2.75-4 Hz).
889 Subsequently, a piecewise linear regression was fitted to the data and the root mean square
890 error (RMSE) and p-value calculated (Supplementary Fig. 4G, K).

891 TRFs were evaluated based on participant ratings of enjoyment, familiarity, and ease
892 to tap to the beat. Two TRFs were calculated per participant based on the 15 highest and 15

lowest ratings on each measure (ignoring tempo condition and subgroup), and the TRF correlations and time lags were compared between the two groups of trials (Fig. 5). Significant differences between the groups were evaluated based on paired-sample t-tests.

The effect of musical sophistication was analyzed by computing the Pearson correlation coefficients between the maximum TRF correlation across tempi per participant and the general musical sophistication (Gold-MSI) per participant (Supplementary Fig. 6).

A support vector machine (SVM) classifier tested whether TRFs captured information about the intended stimulation tempo, the perceived beat rate, or both (Fig. 6). As described previously (see *Behavioral Analysis*), individual tempo conditions were identified in which participants tapped the same rate for two sets of trials that had different intended stimulation tempi, and conditions were also identified in which participants tapped two different rates in response to the same intended stimulation tempo. TRF analysis was performed separately for those two groups of trials, and the z-scored TRF weights were fed into the SVM classifier. First, the SVM classifier was trained to predict the stimulation tempo based on the TRF weights for trials on which the stimulation tempo corresponded to the tapped rate versus trials when the same tapped rate was twice the stimulation tempo (tapped rate = intended stimulation tempo vs. same tapped rate = 2*stimulation tempo; n=19). In comparison, we next identified participants that tapped for 6 trials at the intended tempo and for 6 trials at the harmonic of that intended tempo (intended stimulation tempo = tapped rate vs. same stimulation tempo = 2*tapped rate, n=13). The resulting TRFs were used to predict the tapped rate of the participants. Overall, the classifier was trained to find the optimal hyperplane that separates the data (*fitcsvm*) and was validated in with a leave-one-out cross-validation method (*crossval*). Classification error (quantified with *kfoldLoss*) was compared to a surrogate condition in which the labels of the classifier were randomly shuffled during the training step. The SVM was computed for 100 iterations of the surrogate data. An SVM-accuracy metric was quantified as:

$$\frac{Data\ Accuracy - Surrogate\ Accuracy}{Data\ Accuracy + Surrogate\ Accuracy} \quad (1)$$

leading to a matrix of 4 Musical Features x 13 or 19 Tempo conditions x 100 SVM repetitions.

EEG – Comparison of TRF and RCA measures. The relationship between the TRF analysis approach and the SRCorr was calculated using a linear-mixed effects model (using *fitlme*). Participant and tempo were random (grouping) effects; SRCorr the fixed (predictor) effect and TRF correlations the response variable. To examine the underlying model assumption, the residuals of the linear-mixed effects model were plotted and checked for consistency. The best predictors of the random effects and the fixed-effects coefficients (beta) were computed for every musical feature and illustrated as violin plot (Fig. 4).

Statistical Analysis.

For each analysis, we assessed the overall difference between multiple subgroups using a one-way ANOVA. To test for significant differences across tempo conditions and musical features (TRF Correlation, SRCorr and SRCoh), repeated-measure ANOVAs were conducted coupled to Tukey's test and Greenhouse-Geiser correction was applied when the assumption of sphericity was violated (as calculated with the Mauchly's test). As effect size measures, we report partial η^2 for repeated-measures ANOVAs and $r_{equivalent}$ for paired sample t-test (Rosenthal and Rubin, 2003). Where applicable, the p-values were corrected using the False Discovery Rate (FDR).

Acknowledgments

This work was funded by the ERC Starting Grant awarded to Molly Henry (ERC-STG-804029 BRAINSYNC). We would like to thank the lab staff of the Max Planck Institute for

Empirical Aesthetics for technical support during data acquisition and Lauren Fink for valuable input during data analysis and stimulus feature design.

Author Contributions

KW: Conceptualization, Methodology, Software, Investigation, Formal Analysis, Visualization, Writing - Original draft preparation. **OXW**: Software. Writing - Reviewing and Editing. **MJH**: Conceptualization, Methodology, Software, Formal Analysis, Supervision, Writing - Original draft preparation, Writing- Reviewing and Editing, Funding acquisition.

Competing financial interests

There are no competing financial interests in relation to the work described in this manuscript.

References

BAUER, A.-K. R., KREUTZ, G. & HERRMANN, C. S. 2015. Individual musical tempo preference correlates with EEG beta rhythm. *Psychophysiology,* 52**,** 600-604.

BELLO, J. P., DAUDET, L., ABDALLAH, S., DUXBURY, C., DAVIES, M. & SANDLER, M. B. 2005. A tutorial on onset detection in music signals. *IEEE Transactions on Speech and Audio Processing,* 13**,** 1035-1047.

BRAINARD, D. H. 1997. The Psychophysics Toolbox. *Spatial Vision,* 10**,** 433-436.

BRODERICK, M. P., ANDERSON, A. J. & LALOR, E. C. 2019. Semantic Context Enhances the Early Auditory Encoding of Natural Speech. *The Journal of Neuroscience,* 39**,** 7564.

BURGER, B., LONDON, J., THOMPSON, M. R. & TOIVIAINEN, P. 2018. Synchronization to metrical levels in music depends on low-frequency spectral components and tempo. *Psychological Research,* 82**,** 1195-1211.

CAMERON, D. J., ZIOGA, I., LINDSEN, J. P., PEARCE, M. T., WIGGINS, G. A., POTTER, K. & BHATTACHARYA, J. 2019. Neural entrainment is associated with subjective groove and complexity for performed but not mechanical musical rhythms. *Experimental Brain Research,* 237**,** 1981-1991.

CHEN, J. L., PENHUNE, V. B. & ZATORRE, R. J. 2008. Listening to musical rhythms recruits motor regions of the brain. *Cereb Cortex,* 18**,** 2844-54.

COVER, T. & THOMAS, J. 2005. Entropy, Relative Entropy, and Mutual Information.

CROSSE, M. J., DI LIBERTO, G. M., BEDNAR, A. & LALOR, E. C. 2016. The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience,* 10.

DAUBE, C., INCE, R. A. A. & GROSS, J. 2019. Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. *Current Biology,* 29**,** 1924-1937.e9.

DECRUY, L., VANTHORNHOUT, J. & FRANCART, T. 2019. Evidence for enhanced neural tracking of the speech envelope underlying age-related speech-in-noise difficulties. *Journal of Neurophysiology,* 122**,** 601-615.

DI LIBERTO, G. M., PELOFI, C., BIANCO, R., PATEL, P., MEHTA, A. D., HERRERO, J. L., DE CHEVEIGNÉ, A., SHAMMA, S. & MESGARANI, N. 2020. Cortical encoding of melodic expectations in human temporal cortex. *eLife,* 9**,** e51784.

DING, N., PATEL, A. D., CHEN, L., BUTLER, H., LUO, C. & POEPPEL, D. 2017. Temporal modulations in speech and music. *Neuroscience & Biobehavioral Reviews,* 81**,** 181-187.

DING, N. & SIMON, J. Z. 2012. Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of Neurophysiology,* 107**,** 78-89.

DMOCHOWSKI, J. P., BEZDEK, M. A., ABELSON, B. P., JOHNSON, J. S., SCHUMACHER, E. H. & PARRA, L. C. 2014. Audience preferences are predicted by temporal reliability of neural processing. *Nature Communications,* 5**,** 4567.

DMOCHOWSKI, J. P., SAJDA, P., DIAS, J. & PARRA, L. C. 2012. Correlated components of ongoing EEG point to emotionally laden attention - a possible marker of engagement? *Frontiers in human neuroscience,* 6**,** 112-112.

DOELLING, K. B., ARNAL, L. H., GHITZA, O. & POEPPEL, D. 2014. Acoustic landmarks drive delta–theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage,* 85**,** 761-768.

DOELLING, K. B. & POEPPEL, D. 2015. Cortical entrainment to music and its modulation by expertise. *Proceedings of the National Academy of Sciences,* 112**,** E6233.

FRAISSE, P. 1982. 6 - Rhythm and Tempo. *In:* DEUTSCH, D. (ed.) *Psychology of Music.* San Diego: Academic Press.

GIRAUD, A.-L. & POEPPEL, D. 2012. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature Neuroscience,* 15**,** 511-517.

GRAHN, J. A. & BRETT, M. 2007. Rhythm and Beat Perception in Motor Areas of the Brain. *Journal of Cognitive Neuroscience,* 19**,** 893-906.

GUNDLACH, C., MÜLLER, M. M., NIERHAUS, T., VILLRINGER, A. & SEHM, B. 2016. Phasic Modulation of Human Somatosensory Perception by Transcranially Applied Oscillating Currents. *Brain Stimulation,* 9**,** 712-719.

HARRISON, P. M. C. & MÜLLENSIEFEN, D. 2018. Development and Validation of the Computerised Adaptive Beat Alignment Test (CA-BAT). *Scientific Reports,* 8**,** 12395.

HENRY, M. J. & HERRMANN, B. 2014. Low-Frequency Neural Oscillations Support Dynamic Attending in Temporal Context. *Timing & Time Perception,* 2**,** 62-86.

HENRY, M. J. & OBLESER, J. 2012. Frequency modulation entrains slow neural oscillations and optimizes human listening behavior. *Proceedings of the National Academy of Sciences,* 109**,** 20095.

KANESHIRO, B., NGUYEN, D. T., NORCIA, A. M., DMOCHOWSKI, J. P. & BERGER, J. 2020. Natural music evokes correlated EEG responses reflecting temporal structure and beat. *NeuroImage,* 214**,** 116559.

KEITEL, C., OBLESER, J., JESSEN, S. & HENRY, M. J. 2021. Frequency-Specific Effects in Infant Electroencephalograms Do Not Require Entrained Neural Oscillations: A Commentary on Köster et al. (2019). *Psychological Science,* 32**,** 966-971.

KÖSEM, A., BOSKER, H. R., TAKASHIMA, A., MEYER, A., JENSEN, O. & HAGOORT, P. 2018. Neural Entrainment Determines the Words We Hear. *Current Biology,* 28**,** 2867-2875.e3.

KUMAGAI, Y., MATSUI, R. & TANAKA, T. 2018. Music Familiarity Affects EEG Entrainment When Little Attention Is Paid. *Front Hum Neurosci,* 12**,** 444.

LAKATOS, P., GROSS, J. & THUT, G. 2019. A New Unifying Account of the Roles of Neuronal Entrainment. *Current Biology,* 29**,** R890-R905.

LARGE, E. W. & SNYDER, J. S. 2009. Pulse and meter as neural resonance. *Ann N Y Acad Sci,* 1169**,** 46-57.

LEEK, M. R. 2001. Adaptive procedures in psychophysical research. *Perception & Psychophysics,* 63**,** 1279-1292.

MACDOUGALL, H. G. & MOORE, S. T. 2005. Marching to the beat of the same drummer: the spontaneous tempo of human locomotion. *J Appl Physiol (1985),* 99**,** 1164-73.

MADSEN, J., MARGULIS, E. H., SIMCHY-GROSS, R. & PARRA, L. C. 2019. Music synchronizes brainwaves across listeners with strong effects of repetition, familiarity and training. *Scientific Reports,* 9**,** 3576.

MCAULEY, J. D. 2010. Tempo and rhythm. *Music perception.* New York, NY, US: Springer Science + Business Media.

MCAULEY, J. D., JONES, M. R., HOLUB, S., JOHNSTON, H. M. & MILLER, N. S. 2006. The time of our lives: life span development of timing and event tracking. *J Exp Psychol Gen,* 135**,** 348-67.

MLLER, M. 2015. *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*, Springer Publishing Company, Incorporated.

MOELANTS, D. Preferred tempo reconsidered. Proceedings of the 7th international conference on music perception and cognition, 2002. Citeseer, 1-4.

MORILLON, B., SCHROEDER, C. E. & WYART, V. 2014. Motor contributions to the temporal precision of auditory attention. *Nature Communications,* 5**,** 5255.

MÜLLENSIEFEN, 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.

NICOLAOU, N., MALIK, A., DALY, I., WEAVER, J., HWANG, F., KIRKE, A., ROESCH, E. B., WILLIAMS, D., MIRANDA, E. R. & NASUTO, S. J. 2017. Directed Motor-Auditory EEG Connectivity Is Modulated by Music Tempo. *Frontiers in Human Neuroscience,* 11**,** 502.

NOZARADAN, S., PERETZ, I., MISSAL, M. & MOURAUX, A. 2011. Tagging the Neuronal Entrainment to Beat and Meter. *The Journal of Neuroscience,* 31**,** 10234.

NOZARADAN, S., PERETZ, I. & MOURAUX, A. 2012. Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. *J Neurosci,* 32**,** 17572-81.

OBLESER, J. & KAYSER, C. 2019. Neural Entrainment and Attentional Selection in the Listening Brain. *Trends Cogn Sci,* 23**,** 913-926.

OLIVEIRA, J., GOUYON, F., MARTINS, L. & REIS, L. 2010. *IBT: A Real-time Tempo and Beat Tracking System*.

OOSTENVELD, R., FRIES, P., MARIS, E. & SCHOFFELEN, J.-M. 2011. FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Computational Intelligence and Neuroscience,* 2011**,** 156869.

PARRA, L., HAUFE, S. & DMOCHOWSKI, J. 2018. Correlated Components Analysis --- Extracting Reliable Dimensions in Multivariate Data.

PATEL, A. D. 2003. Rhythm in language and music: parallels and differences. *Ann N Y Acad Sci,* 999**,** 140-3.

PEELLE, J. & DAVIS, M. 2012. Neural Oscillations Carry Speech Rhythm through to Comprehension. *Frontiers in Psychology,* 3.

PEELLE, J. E., GROSS, J. & DAVIS, M. H. 2013. Phase-Locked Responses to Speech in Human Auditory Cortex are Enhanced During Comprehension. *Cerebral Cortex,* 23**,** 1378-1387.

PFORDRESHER, P. Q. 2003. The Role of Melodic and Rhythmic Accents in Musical Structure. *Music Perception,* 20**,** 431-464.

REETZKE, R., GNANATEJA, G. N. & CHANDRASEKARAN, B. 2021. Neural tracking of the speech envelope is differentially modulated by attention and language experience. *Brain and Language,* 213**,** 104891.

RIMOLDI, H. J. 1951. Personal tempo. *J Abnorm Psychol,* 46**,** 283-303.

ROSE, D., CAMERON, D. J., LOVATT, P. J., GRAHN, J. A. & ANNETT, L. E. 2020. Comparison of Spontaneous Motor Tempo during Finger Tapping, Toe Tapping and Stepping on the Spot in People with and without Parkinson's Disease. *Journal of movement disorders,* 13**,** 47-56.

ROSENTHAL, R. & RUBIN, D. B. 2003. requivalent: A simple effect size indicator. *Psychological Methods,* 8**,** 492-496.

SHANNON, R. V. 2005. Speech and Music Have Different Requirements for Spectral Resolution. *International Review of Neurobiology.* Academic Press.

SPAAK, E., DE LANGE, F. P. & JENSEN, O. 2014. Local entrainment of α oscillations by visual stimuli causes cyclic modulation of perception. *J Neurosci,* 34**,** 3536-44.

SRINIVASAN, R., WINTER, W. R., DING, J. & NUNEZ, P. L. 2007. EEG and MEG coherence: measures of functional connectivity at distinct spatial scales of neocortical dynamics. *Journal of neuroscience methods,* 166**,** 41-52.

TIERNEY, A. & KRAUS, N. 2015. Neural entrainment to the rhythmic structure of music. *J Cogn Neurosci,* 27**,** 400-8.

TIMME, N. M. & LAPISH, C. 2018. A Tutorial for Information Theory in Neuroscience. *eneuro,* 5**,** ENEURO.0052-18.2018.

VAN BREE, S., SOHOGLU, E., DAVIS, M. H. & ZOEFEL, B. 2021. Sustained neural rhythms reveal endogenous oscillations supporting speech perception. *PLOS Biology,* 19**,** e3001142.

VANDEN BOSCH DER NEDERLANDEN, C. M., JOANISSE, M. F. & GRAHN, J. A. 2020. Music as a scaffold for listening to speech: Better neural phase-locking to song than speech. *NeuroImage,* 214**,** 116767.

WOLLMAN, I., ARIAS, P., AUCOUTURIER, J.-J. & MORILLON, B. 2020. Neural entrainment to music is sensitive to melodic spectral complexity. *Journal of Neurophysiology,* 123**,** 1063-1071.

ZALTA, A., PETKOSKI, S. & MORILLON, B. 2020. Natural rhythms of periodic temporal attention. *Nature Communications,* 11**,** 1051.

ZATORRE, R. J., BELIN, P. & PENHUNE, V. B. 2002. Structure and function of auditory cortex: music and speech. *Trends in Cognitive Sciences,* 6**,** 37-46.

ZUK, N. J., MURPHY, J. W., REILLY, R. B. & LALOR, E. C. 2021. Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies. *PLOS Computational Biology,* 17**,** e1009358.