Encoding and decoding analysis of music perception using intracranial EEG

Ludovic Bellier[1*], Anaïs Llorens[1], Déborah Marciano[1], Gerwin Schalk[4], Peter Brunner[2,3,4], Robert T. Knight[1,5], Brian N. Pasley[1]

[1]Helen Wills Neuroscience Institute, University of California, Berkeley, Berkeley, CA, USA
[2]Department of Neurosurgery, Washington University School of Medicine, St. Louis, MO, USA
[3]Department of Neurology, Albany Medical College, Albany, NY, USA
[4]National Center for Adaptive Neurotechnologies, Albany, NY, USA
[5]Department of Psychology, University of California, Berkeley, Berkeley, CA, USA

*Corresponding author: ludovic.bellier@berkeley.edu

Abstract

Music is core to human experience across cultures. Multiple brain regions are engaged in music perception, but the neural dynamics of music processing are not well defined. We applied predictive modeling tools to intracranial EEG data recorded from 29 patients who passively listened to a song (Another Brick in the Wall, Part 1, Pink Floyd), focusing on high-frequency activity (HFA; 70-150Hz) as a marker of local neural population activity. Encoding models characterized the spectrotemporal receptive fields (STRFs) of each electrode and decoding models investigated the population-level song representation. With the STRFs, we confirmed a central role of bilateral superior temporal gyri (STG) in music perception with additional involvement of bilateral sensory-motor cortices (SMC) and inferior frontal gyri (IFG). We also observed a right hemispheric preference for music perception. Using both an independent component analysis (ICA) and temporal modulations, we observed cortical regions tuned to specific musical elements including the vocals, lead guitar notes and rhythm guitar patterns. An ablation analysis selectively removed music-responsive electrodes from decoding models to assess the contribution of anatomical and functional regions to representing the song's acoustics. Ablating electrodes from either left and right hemispheres sites deteriorated decoding accuracy, and we confirmed a right STG preference for music perception. We also report, to our knowledge, the first attempt at reconstructing a song from neural activity using both linear and non-linear decoding models, and discuss methodological factors impacting decoding accuracy.

Introduction

Music is a universal experience across all ages and cultures and is a core part of our emotional, cognitive, and social lives (Merriam & Merriam, 1964; Peretz, 2006). A central goal in auditory neuroscience is understanding the neural code for music perception, i.e., how musical information is represented in the human brain. Multiple questions remain including which musical elements (e.g., melody, harmony, rhythm) are encoded in the brain and which brain areas and spectrotemporal dynamics support music processing. The last decades have seen tremendous progress in understanding the neural basis of music perception (Janata, 2015). Multiple studies have assessed the neural correlates of isolated musical elements, such as timbre (Deike et al., 2004; Goydke et al., 2004; Caclin et al., 2006, 2007; Alluri et al., 2012), pitch (Hyde et al., 2008; Kumar et al., 2011; Nan & Friederici, 2013; Plack et al., 2014), melody (Trainor et al., 2002; Baltzell et al., 2019), harmony (Janata et al., 2002; Brattico et al., 2006) and rhythm (Jongsma et al., 2004; Snyder & Large, 2005; Geiser et al., 2009; Grahn & Rowe, 2009; Abrams et al., 2011; Schaefer et al., 2011; Harding et al., 2019). These studies have established that music perception relies on a broad network of subcortical and cortical regions, including primary and secondary auditory cortices, sensorimotor areas, and inferior frontal gyri (Peretz & Zatorre, 2005; Limb, 2006; Koelsch, 2011; Zatorre & Salimpoor, 2013). Both hemispheres have been shown to be involved in music processing, with a relative preference for the right hemisphere (Warren, 2008; Ding et al., 2019).

These studies provide a foundation for understanding music perception. However, they mostly focus on isolated musical elements or cortical areas. In addition, they often rely on brain imaging methods with limited temporal or spatial resolution (fMRI and EEG, respectively; (Hall et al., 2014), and on standard trial-based paradigms and analytic methods. These studies paint a parcellated overview of the brain regions and neural dynamics involved in music perception. To address these limitations, we use a naturalistic auditory stimulus (the song *Another Brick in the Wall, Part 1*, by Pink Floyd) in a passive listening task, and apply encoding and decoding analyses to intracranial electroencephalography (iEEG) data, known for its unique spatiotemporal resolution.

Studies employing restricted or synthetic stimuli are useful to assess specific aspects of auditory processing but may miss brain regions involved in higher-order processing (Theunissen et al., 2000; David et al., 2007; Talebi & Baker, 2012; Sturm et al., 2014). Due to nonlinearities in the auditory pathway, probing the brain with isolated notes elicits neural activity in the primary auditory cortex (A1), but might fail to activate potential areas encoding key higher-order musical elements such as chords (i.e., at least three notes played together), harmony (i.e., the relationship between a system of chords), or rhythm (i.e., the temporal arrangement of notes). We adopted a naturalistic approach studying a real-world, popular rock song using a passive listening paradigm mimicking the everyday music-listening experience. The use of a rich and

complex auditory stimulus has the ability to elicit a robust and distributed neural response, allowing study of the neural network underlying music perception.

Music research participants are often asked to actively perform a task, such as detecting a target (Trainor et al., 2002; Goydke et al., 2004; Baltzell et al., 2019), focusing on a particular auditory object (Choi et al., 2013; Foldal et al., 2020), or judging a perceptual aspect (Brattico et al., 2006; Nan & Friederici, 2013). Such tasks are necessary to study key aspects of auditory cognition, such as attention, working memory or emotions. However, the dual task nature of these approaches requiring both listening and responding distracts participants from pure music listening and confounds neural processing of music with decision processes and motor activity. A naturalistic music listening experience provides an uninterrupted window for assessment of higher-order aspects of musical experience (e.g., sense of rhythm built over time) optimizing our chances at observing the full network underlying the perception of musical elements.

We used intracranial EEG (iEEG) data recorded directly from the cortical surface of neurosurgical patients (electrocorticography; ECoG). This unique window on cortical processing combines the temporal resolution of electrophysiological techniques, with the spatial resolution of fMRI (Lachaux et al., 2003). In addition, iEEG is the only brain imaging technique providing direct access to High-Frequency Broadband Activity (HFA; 70-150Hz). HFA provides an index of non-oscillatory neural activity reflecting information processing linked to local single unit firing in the infragranular cortical layers and dendritic potential in supraganular layers (Leszczyński et al., 2020) and to the BOLD signal in fMRI (Conner et al., 2011). Due to the direct contact between electrodes and brain tissue, iEEG benefits from an excellent signal-to-noise ratio. This is especially valuable in our naturalistic approach since it provides reliable HFA at the single-trial level enabling individual subject modeling.

We employed predictive modeling tools to take advantage of the complexity of our naturalistic stimulus and the richness of iEEG data. Specifically, we used encoding models to characterize the spectrotemporal receptive fields (STRF) of each electrode and decoding models to reconstruct the song stimulus from population neural activity. Encoding models predict neural activity at one electrode from a representation of the stimulus (e.g., spectrogram, spectrotemporal modulations, onset of notes). When this representation is a spectrogram, encoding models are called spectro-temporal receptive fields (STRFs), and a plot of their trained coefficients (or weights) can be interpreted as the spectrogram of the ideal auditory stimulus needed to elicit an increase of neural activity at the observed electrode. These models have been successfully used to evidence key properties of the neural auditory system. This technique originated with action potential data recorded in animal models in response to artificial stimuli (Aertsen & Johannesma, 1980; Aertsen et al., 1980; Aertsen et al., 1981). Recent algorithmic and machine-learning developments expanded its use to human brain imaging data and naturalistic stimuli (Theunissen et al., 2000, 2001; Holdgraf et al., 2017). Within the last decade, STRFs have been used to quantitatively characterize the spectrotemporal tuning profile of neural

populations in response to speech or music. Notably, STRFs were used to evidence rapid plasticity of the human auditory cortex in speech perception (Holdgraf et al., 2016), an antero-posterior parcellation of the human superior temporal gyri (STG; Hullett et al., 2016), and a partial overlap between the neural activity underlying music imagery and music perception (Martin et al., 2017). By considering the full complexity of the auditory stimulus, as opposed to condition-based task design that often focuses on a single contrast dimension, and by revealing the tuning patterns of neural populations, STRFs constitute a tool of choice to investigate the neural coding supporting music perception.

Decoding models predict a representation of the stimulus from the elicited neural activity, often obtained from many electrodes and their implementation has exploded in the last decade for analyzing complex datasets without sacrificing potential dimensions of interest (for a review, see Holdgraf et al., 2017). In the music domain, most decoding models have been used in a classification approach, for example to identity a musical piece (Hoefle et al., 2018) or its genre (Casey, 2017; Nakai et al., 2021) from the elicited neural activity, or to estimate music-related aspects beyond the stimulus level, such as musical attention (Treder et al., 2014) or musicianship status of the listener (Saari et al., 2018). Another application of decoding models used for speech is stimulus reconstruction approach (Mesgarani et al., 2009; Pasley et al., 2012), where the auditory stimulus (i.e., the sound itself) is reconstructed from the elicited neural activity. Decoding performance informs on the nature of the information represented in the recorded neural activity. For example, if a musical element can be reconstructed, this means that it was represented within the set of electrodes used as input of the decoding model. Ablation analysis constitutes a method of making a virtual lesion on decoding models (Meyes et al., 2019; Kohoutová et al., 2020). It consists in removing (or ablating) sets of predictors (here, electrodes) to assess their impact on decoding accuracy. Moreover, comparing the impact of ablating different sets of electrodes provides insights on how information is uniquely or redundantly encoded between these sets.

Stimulus reconstruction has seen recent successes in the speech domain (Akbari et al., 2019; Anumanchipalli et al., 2019; Moses et al., 2021). Such studies have reconstructed intelligible speech from iEEG data, using nonlinear models (deep neural networks) combined with different representations of speech including speech kinematics or the movements of vocal articulators. Here we examine if a song can be reconstructed from direct brain recordings and quantify the factors impacting decoding accuracy including model type (linear vs nonlinear) and stimulus representation (low vs high level).

The dataset we analyzed has been the focus of previous studies, although not employing encoding and decoding models (Potes et al., 2012, 2014; Kubanek et al., 2013; Gupta et al., 2014; Sturm et al., 2014). These studies linked several musical elements, such as sound intensity or timber, to neural activity in the posterior superior temporal gyrus (STG) or sensorimotor areas. They constitute a foundation for comparison to the results of our modeling approach. Here, we use predictive modeling tools on iEEG data recorded from 2,668 electrodes across 29

neurological patients, who passively listened to a song (Another Brick in the Wall, Part 1; Pink Floyd). We used encoding models to identify responsive cortical areas and analyze their tuning patterns and decoding models both to analyze information processing through an ablation analysis and to reconstruct the song from the elicited neural activity.

Results

Distribution of song-responsive electrodes

To identify electrodes encoding information about the song, we fitted STRFs for all 2,379 artifact-free electrodes in the dataset, assessing how well the HFA recorded at these sites could be linearly predicted from the song's auditory spectrogram (see Fig. 1 for method). From a dense, bilateral, predominantly frontotemporal coverage (Fig. 2A), we identified 347 electrodes with a significant STRF (Fig. 2B). We found a higher proportion of song-responsive electrodes in the right hemisphere. There were 199 significant electrodes out of 1,479 total in the left hemisphere and 148 out of 900 in the right one (Fig. 2B, 13.5% against 16.4%, respectively; X^2 (1, N=2,379) = 4.01, p = .045).

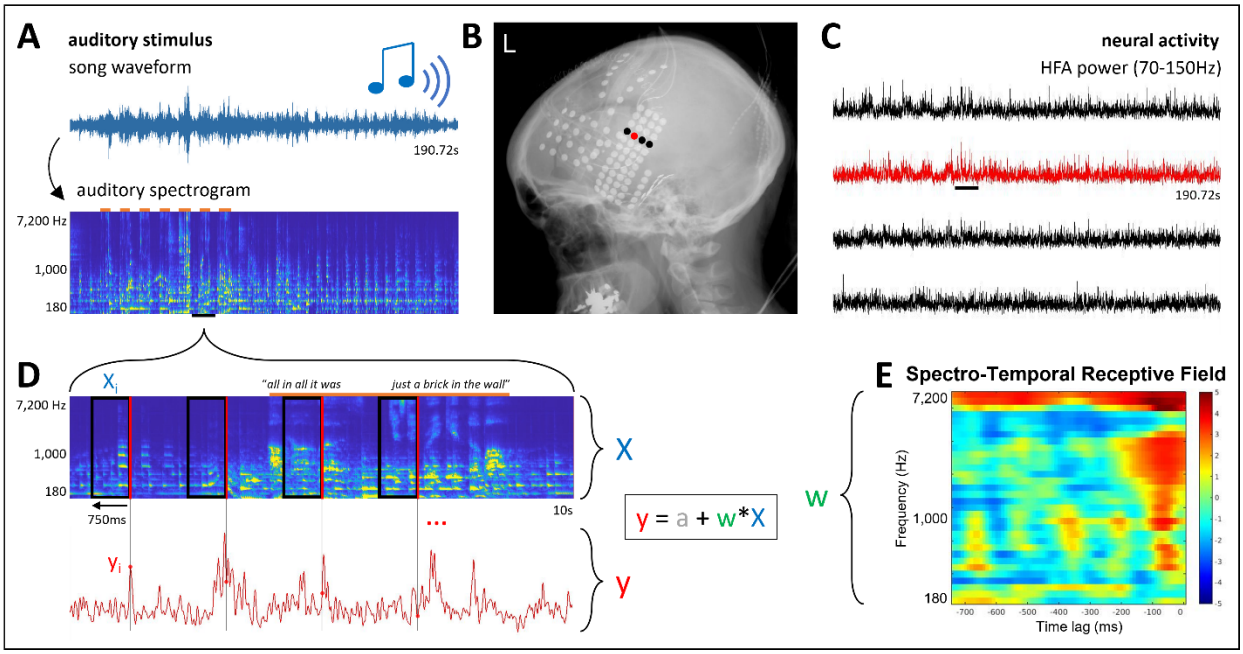

Fig 1. Protocol and data preparation for encoding models. **A.** Top. Waveform of the song. Participants listened to a 190.72-second-long rock song (Another Brick in the Wall, Part 1, Pink Floyd) using headphones. Bottom. Auditory spectrogram of the song. Orange dashes on top represent vocal parts. **B.** X-ray showing electrode coverage in one representative patient. Each dot is an electrode, and the signal of the four colored contacts is shown in C. **C.** High-Frequency Broadband (HFB; 70 to 150 Hz) power for four representative electrodes during the entire song duration. **D.** Zoom in on 10 seconds (see black dashes) of both the auditory spectrogram (in A) and the neural activity of an illustrative electrode (in C). Each time point of the neural estimate (y_i) is paired with a 750ms spectrotemporal chunk (X, black rectangle), ending at this time point (right, red edge of the rectangle). The set of all pairs (X, y_i) constitute the examples (or observations) used to train and evaluate the linear encoding models. Linear encoding model used here consists in predicting the neural activity (y) from the auditory spectrogram (X), by finding the optimal intercept (a) and coefficients (w). **E.** Spectro-temporal receptive field (STRF) for the depicted electrode. Note that 0 on the right with time lags preceding observed neural activity. Equivalent to W.

The majority of the 347 significant electrodes (87%) were concentrated in three regions: 68% in bilateral superior temporal gyri (STG), 14.4% in bilateral sensori-motor cortices (SMC, on the pre- and postcentral gyri), and 4.6% in bilateral inferior frontal gyri (IFG; Fig. 2C). The proportion of song-responsive electrodes per region was 55.7% for STG (236 out of 424 electrodes), 11.6% for SMC (45/389), and 7.4% for IFG (17/229). The remaining 13% of significant electrodes were distributed in the supramarginal gyri and other frontal and temporal regions.

Analysis of STRF prediction accuracies (Pearson's r) found a main effect of laterality (two-way ANOVA; $F(1, 346) = 7.48$, $p = 0.0065$; Fig. 2D), with higher correlation coefficients in the right hemisphere than in the left ($M_R = .203$, $SD_R = .012$; $M_L = .17$, $SD_L = .01$). We also found a main effect of cortical regions ($F(3, 346) = 25.09$, $p < .001$), with the highest prediction accuracies in STG (Tukey-Kramer post-hoc; $M_{STG} = .266$, $SD_{STG} = .007$; $M_{SMC} = .194$, $SD_{SMC} = .017$, $p_{STGvsSMC} < .001$; $M_{IFG} = .154$, $SD_{IFG} = .027$, $p_{STGvsSMC} < .001$; $M_{other} = .131$, $SD_{other} = .016$, $p_{STGvsSMC} < .001$). In addition, we found higher prediction accuracies in SMC compared to the group not including STG and IFG ($M_{SMC} = .194$, $SD_{SMC} = .017$; $M_{other} = .131$, $SD_{other} = .016$, $p_{SMCvsOther} = .035$).

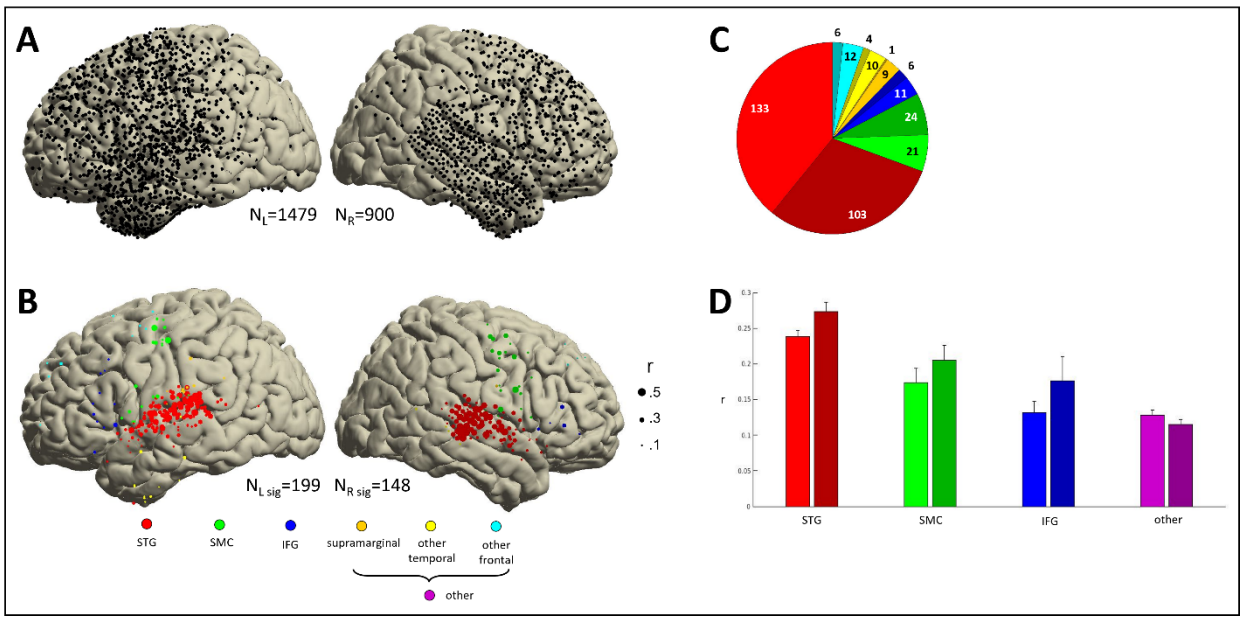

Fig. 2. Location of electrodes encoding the song. **A.** Electrode coverage across all 29 patients shown on the MNI template (N=2379). All presented electrodes are free of any artifactual or epileptic activity. Left hemisphere is plotted on the left. **B.** Location of electrodes significantly encoding the song (N_{sig}=347). Significance was determined by the STRF prediction accuracy bootstrapped over multiple resamples. Marker colors indicate the anatomical label as determined using the Freesurfer atlas, and marker size indicates the STRF's performance metrics (Pearson's R between actual and predicted HFB activity). Bright colors indicate a left location, while dark colors a right location. The same color code is used in the following panels. **C.** Number of significant electrodes per region. **D.** Average STRF performance metrics per region. Electrodes previously labelled as *supramarginal*, *other temporal* (other than STG) and *other frontal* (other than pre-central gyrus) are pooled together and labelled as *other*. Error bars indicate SEM.

Encoding of musical elements

We analyzed STRF coefficients (or weights) for all 347 significant electrodes to understand how different musical elements were encoded in different brain regions. This revealed a variety of spectrotemporal tuning patterns (Fig. 3A). To fully characterize the relationship between the song spectrogram and the neural activity, we performed an independent component analysis

(ICA) on all significant STRFs. We identified three components with distinct spectrotemporal tuning patterns, each explaining more than 5% variance (Fig. 3B).

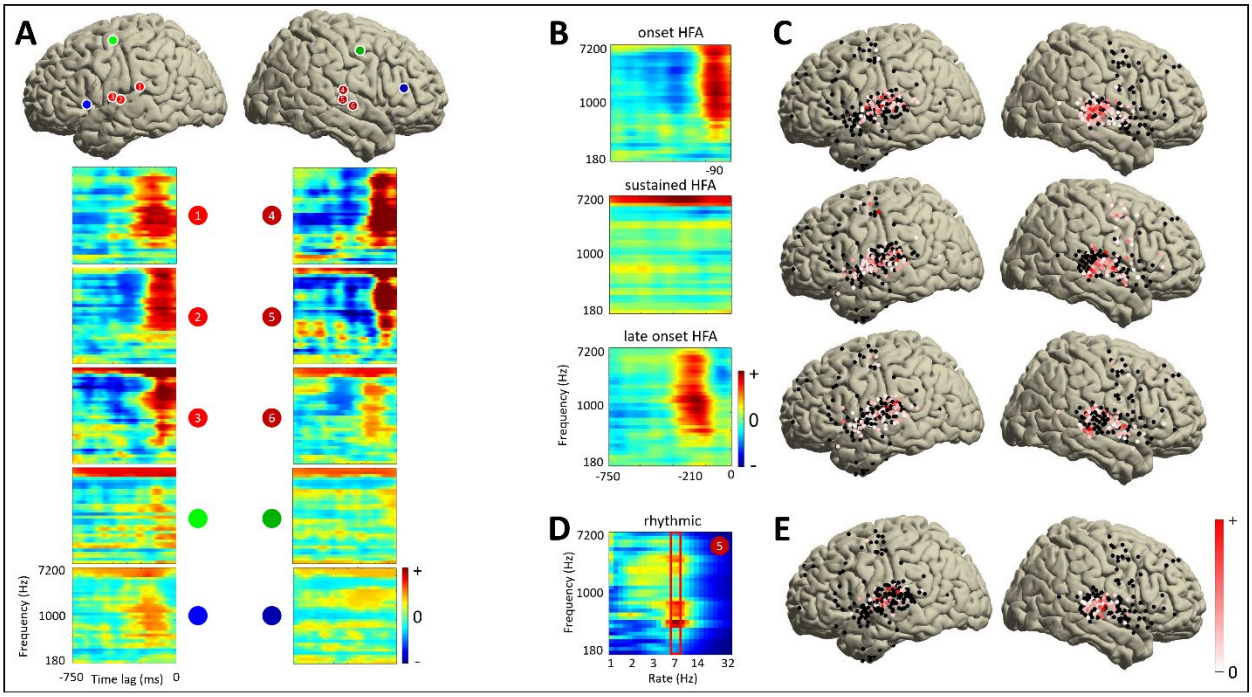

Fig. 3. Analysis of the STRF tuning patterns. **A.** Representative sample of 10 STRFs (out of the 347 available) with their respective locations on the MNI template. Electrode color code is identical to the one used in Fig. 1. **B.** Scores of the three main components from the ICA on all significant STRFs. These three components show *onset*, *sustained* and *late onset* activity. Percentages indicate the percentage of variance accounted for. **C.** ICA coefficients (color coded) of these three components for each electrode, plotted on the MNI template. Red indicates strong coefficients, i.e., electrodes which STRFs are built using this component. **D.** To capture the guitar struck rhythmicity (16th notes at 100 bpm, i.e., 6.66 Hz) pervasive throughout the song, we computed temporal modulation spectrums for each electrode's STRF. Example modulation spectrum is shown for STG electrode (5). The maximum value around 6.66 Hz of temporal modulation across all spectral frequencies (red frame) was extracted for each electrode. **E.** All extracted values are represented on the MNI template. Red indicates high modulation at the song's rhythmic rate.

The first component (28% explained variance) showed a cluster of positive coefficients (in red, in Fig. 3B, top row) spreading over a broad frequency range from about 500 Hz to 7 kHz, and over a narrow time window centered around 90 ms before the observed HFA (located at time lag = 0 ms, at the right edge of all STRFs). This temporally transient cluster revealed tuning to sound onsets. This component, referred to as the "onset component", was found exclusively in electrodes located in bilateral posterior STG (Fig. 3C, top row, electrodes depicted in red). Fig. 4C, top row showed in red the parts of the song eliciting the highest HFA increase in electrodes possessing this onset component. These parts corresponded to onsets of lead guitar or synthesizer motifs (Fig. 4A, blue and purple lines, respectively; see Fig. 4E for a zoom-in) played every two bars (green lines), and to onsets of syllable nuclei in the vocals (orange lines; see Fig. 4D for a zoom-in).

The second component (14.7% explained variance) showed a cluster of positive coefficients (in red, in Fig. 3B, middle row) spreading over the whole 750ms time window, and over a narrow frequency range from about 4.8 to 7 kHz. This component, referred to as the "sustained component", was found in electrodes located in bilateral mid- and anterior STG, and

in bilateral SMC (Fig. 3C, middle row). It was most correlated with parts of the song containing vocals, thus suggesting tuning to speech (Fig. 4C, middle row, in red; see Fig. 4D for a zoom-in).

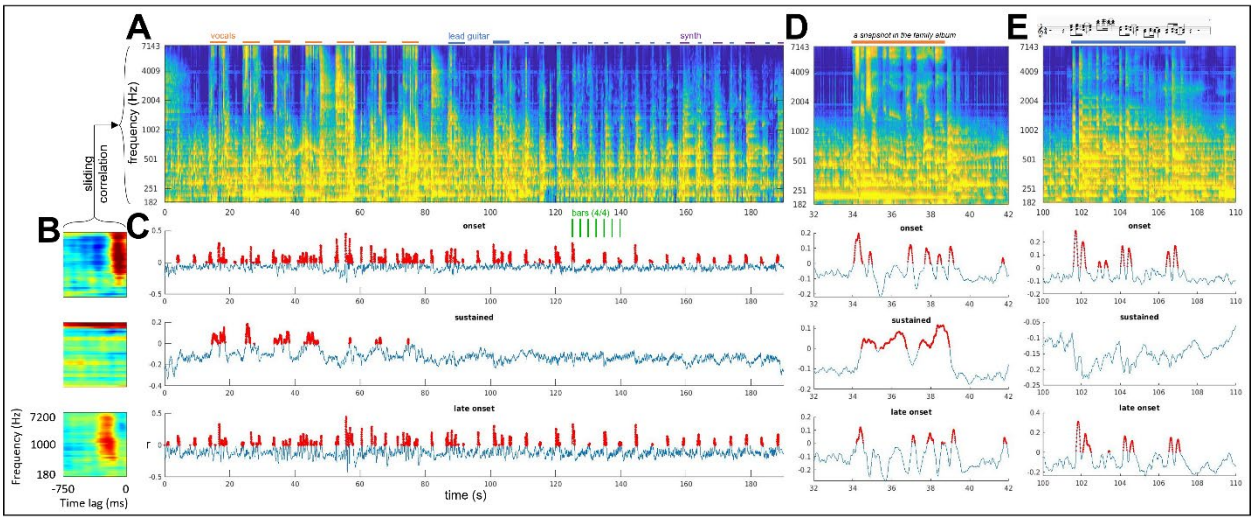

Fig. 4. Encoding of musical elements. **A.** Auditory spectrogram of the whole song. Orange lines above the spectrogram mark all vocals and purple lines all synthesizer motifs. Blue lines mark a subset (for the sake of clarity) of lead guitar motifs. Green vertical lines below the spectrogram mark a series of seven 4/4 bars. Bold orange and blue lines mark locations of the zoom-in presented in D and E, respectively. **B.** Three main STRF components as presented in Fig. 3B, namely onset HFA (top), sustained HFA (middle) and late onset HFA (bottom). **C.** Output of the sliding correlation between the song spectrogram in A and each of the three STRF components in B. Positive Pearson's r are shown in red, marking parts of the song that elicit an increase of HFA in electrodes exhibiting the given component. Note that for the sustained plot (middle), positive correlation coefficients are specifically observed during vocals, suggesting that the sustained component is a marker for speech perception. Also, note for both the onset and late onset plots (top and bottom, respectively), positive r values in the second half of the song corresponds to lead guitar and synthesizer motifs, occurring every two bars. **D.** Zoom in on the third vocals. Lyrics are presented above the spectrogram. Most syllables trigger an HFA increase in both onset and late onset plots, while a sustained increase of HFA is observed during the whole vocals. **E.** Zoom in on a lead guitar motif. Sheet music is presented above the spectrogram. Most notes trigger an HFA increase in both onset and late onset plots, while there is no HFA increase in the sustained component.

The third component (9.8% explained variance) showed a similar tuning pattern as the onset component, only with a longer latency of about 210 ms before the observed HFA (Fig. 3B, bottom row). This component, referred from now on as the "late onset component", was found in bilateral posterior and anterior STG, neighboring the electrodes representing the onset component, and in bilateral SMC (Fig. 3C, bottom row). As with the onset component, this late onset component was most correlated with onsets of lead guitar and synthesizer motifs and of syllable nuclei in the vocals, only with a longer latency (Fig. 4C, bottom row; see Fig. 4D and 4E for zoom-ins).

A fourth component was found by computing the temporal modulations and extracting the maximum coefficient around a rate of 6.66 Hz for all 347 STRFs (Fig. 3D, red rectangle). This rate corresponded to the 16th notes of the rhythm guitar, pervasive throughout the song, at the song tempo of 99 bpm (beats per minute). It was translated in the STRFs as small clusters of positive coefficients spaced by 150 ms (1 / 6.66 Hz) from each other (e.g., Fig. 3A, electrode 5). This component, referred from now on as the "rhythmic component", was found in electrodes located in bilateral mid STG (Fig. 3E).

Anatomo-functional distribution of the song's acoustic information

To assess the role of these different cortical regions and functional components in representing musical features, we performed an ablation analysis using linear decoding models. We first computed linear decoding models for each of the 32 frequency bins of the song spectrogram, using the HFA of all 347 significant electrodes as predictors. This yielded an average prediction accuracy of .62 (Pearson's r; min .27 - max .81). We then removed (or *ablated*) anatomically- or functionally defined sets of electrodes and computed a new series of decoding models, to assess how each ablation would impact the decoding accuracy. We used prediction accuracies of the full, 347-electrode models as baseline values (Fig. 5). We found a significant main effect of electrode sets (one-way ANOVA; $F(1, 24) = 78.4$, $p < .001$). We then ran a series of post-hoc analyses to examine the impact of each set on prediction accuracy.

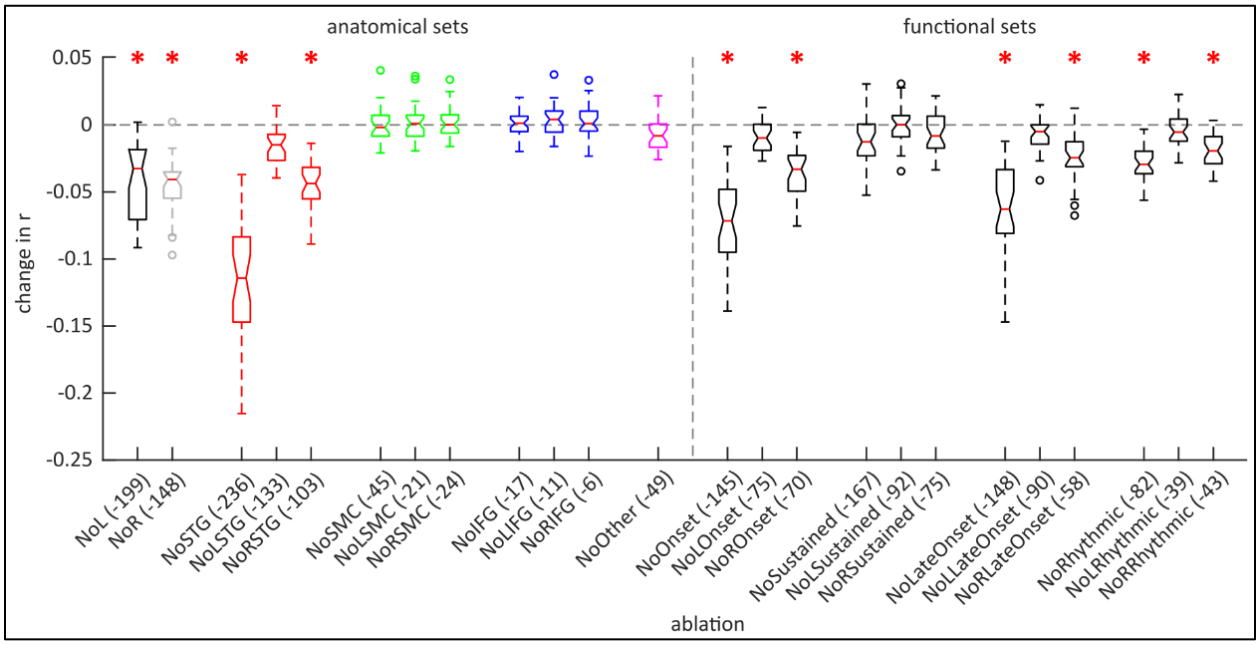

Fig. 5. Ablation analysis for linear decoding. Sets of electrodes are either anatomical or functional. Ablations are shown on the horizontal axis, and their impacts on the linear decoding performance metrics (Pearson's r), as compared to a linear decoding model using all 347 significant electrodes, are shown on the vertical axis. For each ablation, a notched box plot represents the distribution of the 32 decoding models (one per frequency bin of the auditory spectrogram) change in performance metrics.

Anatomical ablations. Removing all STG or all right STG electrodes impacted prediction accuracy ($p < .001$; Fig. 5), with removal of all STG electrodes having the highest impact compared to all other electrode sets ($p < .001$). Removal of right STG electrodes had higher impact than left STG removal ($p < .001$), and no impact of removing left STG electrodes was found ($p = .156$). Together, this suggests that: 1) bilateral STG represented unique musical information compared to other regions, 2) right STG had unique information compared to left STG, and 3) part of the musical information in left STG was redundantly encoded in right STG. Ablating SMC, IFG or all other regions did not impact prediction accuracy ($p > .998$). Removing either all left or all right electrodes significantly reduced the prediction accuracy ($p < .001$), with no significant difference between all left and all right ablations ($p = 1$). These results suggest that both hemispheres represent unique information and contribute to song decoding. Furthermore, the fact that

removing single regions in the left hemisphere had no impact but removing all left electrodes did, suggests redundancy within the left hemisphere, with musical information being spatially distributed across left regions.

Functional ablations. Removing all onset electrodes and right onset electrodes both impacted prediction accuracy ($p < .001$), with a highest impact for all onset ($p < .001$). No impact of removing left onset electrodes was found ($p = .994$). This suggests that right onset electrodes had unique information compared to left onset electrodes, and that at least part of the musical information in left onset electrodes was redundantly encoded in right onset electrodes. A similar pattern of higher right hemisphere involvement was observed with the late onset component ($p < .001$). Removing all rhythmic and right rhythmic electrodes both significantly impacted the decoding accuracy ($p < .001$ and $p = .007$, respectively), while we found no impact of removing left rhythmic electrodes ($p = 1$). We found no difference between removing all rhythmic and right rhythmic electrodes ($p = .973$). This suggests that right rhythmic electrodes had unique information, none of which was redundantly encoded in left rhythmic electrodes. Despite the large number of sustained electrodes, no impact of removing any set was found ($p > .745$). Note that as opposed to anatomical sets, functional sets of electrodes partially overlapped. This impeded our ability to reach conclusions regarding the uniqueness or redundancy of information *between* functional sets.

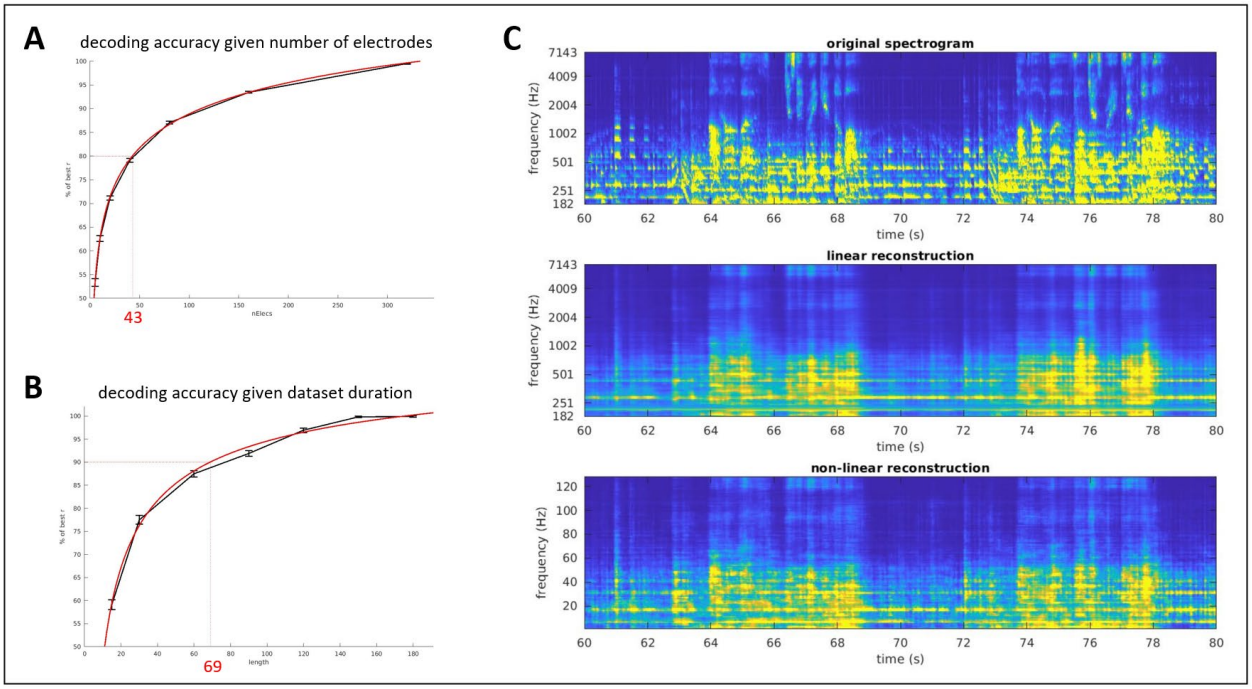

Fig. 6. Methodological factors impacting decoding accuracy. **A.** Decoding accuracy as a function of the number of electrodes included as predictors in the decoding model. On the Y axis, 100% represents the maximum decoding accuracy, obtained using all 347 significant electrodes. Black curve shows data points obtained from bootstrapping analyses; red curve shows a fit line. **B.** Decoding accuracy as a function of dataset duration. See A for plotting details. **C.** Auditory spectrograms of the original song (top) and of the reconstructed song using either linear (middle) or non-linear models (bottom).

Song reconstruction and methodological factors impacting decoding accuracy

Finally, we tested if we could reconstruct the song from neural activity, and how methodological factors such as the number of electrodes included in the model, the dataset duration or the model type at use would impact decoding accuracy. A bootstrap analysis revealed a logarithmic relationship between how many electrodes were used as predictors in the decoding model and the resulting prediction accuracy (Fig. 6A). For example, 80% of the best prediction accuracy (using all 347 significant electrodes) was obtained with 43 (or 12.4%) electrodes. A similar relationship was observed between dataset duration and prediction accuracy (Fig. 6B). For example, 90% of the best performance (using the whole 191s song) was obtained using 69 seconds (or 36.1%) of data. Regarding model type, linear decoding provided an average decoding accuracy of .237 (r-squared), while non-linear decoding using a fully connected neural network (multi-layer perceptron) yielded an average decoding accuracy of .23. Despite similar decoding accuracies, the reconstructed spectrograms looked different, with the linear reconstruction showing smoothed envelopes with only coarse spectrotemporal changes, and the non-linear reconstruction showing more detailed spectro-temporal information with fine-grained modulations (Fig. 6C, middle and bottom row, respectively).

Discussion

We applied predictive modeling analyses on iEEG data obtained from patients listening to a song. Encoding models showed a central role of bilateral STG and a right-hemisphere preference in music perception. The results revealed partially overlapping cortical areas that encoded different musical elements. An ablation analysis for decoding models showed that both the left and right hemispheres contained unique musical information, and that part of the information between left and right STG was redundant. Moreover, in the left hemisphere, we observed that musical information was spatially distributed between regions, beyond STG. On a methodological side, we quantified the impact of the number of electrodes, dataset duration and model type (linear vs non-linear) on decoding accuracy.

We observed a right hemisphere preference for music perception, with a higher proportion of electrodes with significant STRFs, higher STRF prediction accuracies, and a higher impact of ablating right electrode sets (both anatomical and functional) from the decoding models. While there was a statistical preference for the right hemisphere, left hemisphere electrodes also exhibited significant STRFs and a significantly reduced prediction accuracy with ablation. These results are in accord with prior research, showing that music perception relies on a bilateral network, with a relative right lateralization (for a review, see Warren, 2008; Ding et al., 2019; Albouy et al., 2020).

We also found that the spatial distribution of musical information differed between hemispheres, as suggested by the ablation results. Redundant musical information was distributed between STG, SMC and IFG in the left hemisphere, whereas unique musical information was focused in STG in the right hemisphere. Such spatial distribution is reminiscent of the dual-stream model of speech processing (Hickok & Poeppel, 2007). However, the absence

of right SMC or IFG involvement in the ablation analysis was surprising given their role in music processing (Albouy et al., 2013; Zatorre & Salimpoor, 2013). Still, we observed significant STRFs in bilateral SMC and IFG, with possible roles in encoding vocals-related information and speech or melodic syntaxis, respectively (Zatorre et al., 2007; Gordon et al., 2018; Anumanchipalli et al., 2019).

We found a critical role of bilateral STG in representing musical information, in line with prior human studies (Albouy et al., 2013; Toiviainen et al., 2014; Martin et al., 2017; Nakai et al., 2021). As observed in other studies, STRFs obtained from the STG had rich, complex tuning patterns. To assess the anatomo-functional organization of music perception in the STG, we employed an ICA analysis on all STRFs, which revealed four components: onset, sustained, late onset and rhythmic.

The onset and sustained components were similar to those observed for speech in prior work (Hullett et al., 2016; Hamilton et al., 2018). Specifically, the onset component was tuned for high temporal/low spectral modulations while the sustained component was tuned to low temporal/high spectral modulations.

The onset component was tuned to a broad range of frequencies but to a narrow time window peaking at 90ms. This latency is similar to the lag time between HFA and auditory stimulus intensity reported by Ding et al., 2019. We found that the onset component was activated by both vocals (that is, syllables) and instrumental onsets (or notes). This confirms that the onset component is not speech specific, consistent with prior work (Hamilton et al., 2018) showing that reversed and spectrally rotated speech also elicited onset responses.

In contrast to the onset component, we found that the sustained component (tuned to a narrow high-frequency band but to a wide time window) was activated by the vocals only. As seen in prior work (Hullett et al., 2016; Hamilton et al., 2018) we observed these components in anatomically distinct STG subregions, with the onset component in posterior STG and the sustained component in mid- and anterior STG. Interestingly, we observed single electrodes representing both the onset and the sustained components, which were mostly located in mid STG. This was not found in previous studies, likely due to the use of different data-driven approaches (clustering vs ICA). Surprisingly, in our functional ablation analysis, removing all electrodes representing the sustained component did not impact decoding accuracy, despite their substantial number (167 out of 347). This might be due to the fact that as the song is dominated by instrumentals, removing a component related to vocals has negligible impact on the decoding accuracy.

In addition to the onset and sustained component, we found evidence for two other distinct components: late onset and rhythmic. The late onset component was found in electrodes neighboring the onset component in STG and had similar tuning properties as the onset component, only peaking at a later latency of 210ms. This is in line with the findings of Nourski et al., 2014, who, using click trains and a speech syllable, observed a concentric spatial gradient of HFA onset latencies in STG, with shorter latencies in post-/mid-STG and longer latencies in surrounding tissue. Further studies are needed to understand better the relationship between the onset and late onset components, as their similar functional behavior despite such different latencies appears as a discrepancy.

The rhythmic component, tuned to the 6.66 Hz sixteenth notes of the rhythmic guitar pervasive throughout the song, was observed in mid STG, especially in electrodes representing both the onset and the sustained components. Here, this is the first time to our knowledge that we evidence a link between HFA and a specific rhythm guitar signature in a subregion of STG. This extends the knowledge brought by past studies that have found an involvement of STG in a range of rhythmic processes, i.e., beat perception (Grahn & McAuley, 2009), omissions (Vikene et al., 2019), periodicity (Herff et al., 2020). Altogether, these four components paint a rich picture of the anatomo-functional organization of complex sound processing in the human STG.

On the methodological side, the observed logarithmic relationship between decoding accuracy and the number of electrodes (a proxy for electrode density) or dataset duration is in line with previous literature for speech stimuli (e.g., Akbari et al., 2019). We showed that 80% of the maximum observed decoding accuracy is achieved with 43 electrodes or in 37 seconds, which supports the feasibility of using a predictive modeling approach in relatively small datasets. Moreover, so far, we have not evidenced any advantages of using non-linear models over linear models to reconstruct the song. This might in part due to the fact that our electrodes (and electrode coverages in papers mentioned above) were located in non-primary auditory areas, known to represent auditory information as abstractions from the stimulus acoustics (Kumar et al., 2007). Decoding the song spectrogram from electrodes in primary auditory areas (A1, accessible with stereotactic EEG/depth electrodes) might enable successful decoding. In addition, further tuning of the non-linear models might lead to better decoding accuracy.

An important open question is whether there exist brain regions and networks that are specific to music, or whether music-related information is processed in general auditory pathways (Zatorre & Gandour, 2008; Albouy et al., 2020; Norman-Haignere et al., 2020). While this study links musical elements to STRF components and thus to precise anatomical locations, it is unlikely that these regions respond specifically to music. Rather our findings suggest non-music-specific encoding of musical elements. The fact that onset and late onset components respond to syllables, lead guitar and synthesizer (Fig. 4) suggests that subparts of STG process both vocals and music. Although one could argue that the rhythmic component (Fig. 3D and E) is music specific as it is clearly related to the 6.66 Hz sixteenth notes of the rhythm guitar, this same rhythmic component shows diffuse energy between 2 and 8 Hz in the temporal modulation spectrum (Fig. 3D), compatible with syllabic rhythm (Giraud et al., 2007). On the other hand, a specificity for speech is suggested by the sustained component, as it is only activated by vocals (Fig. 4C, D and E).

Our study had several limitations. Importantly, the models we used in this study to investigate the neural dynamics of music perception estimate the linear relationship between the song's acoustics and the HFA. It is possible that regions not highlighted by our study respond to the song, either in other neural frequency bands, or encoding higher-order musical information. Another limitation was the short duration of the song (three minutes) with relatively limited spectrotemporal variability. More data would enhance statistical power and enable the use of more complex non-linear models. Another limitation is the lack of patient-related information about musicianship status or degree of familiarity with the song. Finally, we used a passive

listening task requiring the models to be trained using an acoustic representation of the song (the auditory spectrogram). However, neural activity recorded in higher-order regions also reflects intrinsic, goal-directed perceptual processes. Further studies could assess encoding differences between different listening conditions (e.g., focus on the bass line, vs on the vocals), to extend the analysis outside of passively presented physical properties of the song.

Combining a naturalistic paradigm, unique iEEG data and novel modeling-based analyses, this study extends musical knowledge at two levels. At the brain level, we observed a right-hemisphere preference and a preponderant role of bilateral STG in representing the song's acoustics. Within bilateral STG, we observed partially overlapping neural populations tuned to distinct musical elements. The ablation analysis revealed the presence of unique musical information in both hemispheres, which was spatially distributed within the left hemisphere between STG, SMC and IFG. At a technical level, we showed the feasibility of applying predictive modeling on a relatively short dataset and quantified the impact of different methodological factors on the prediction accuracy of decoding models. To our knowledge, this is also the first attempt at reconstructing the acoustics of a song from direct brain recordings. Together, these two levels pave the way to describing in greater detail the neural dynamics of music perception. Follow-up studies could investigate different representations of the song (i.e., notes, chords, sheet music) and different neural frequency bands (e.g., theta, alpha, beta power), and will add another brick in the wall of our understanding of music processing in the human brain.

Methods

Participants. Twenty-nine patients with pharmacoresistant epilepsy participated in the study. All had intracranial grids or strips of electrodes (electrocorticography, ECoG) surgically implanted to localize their epileptic foci, and electrode location was solely guided by clinical concern. Recordings took place at the Albany Medical Center (Albany, NY). All patients volunteered and gave their informed consent prior to participating in the study. The experimental protocol has been approved by the Institutional Review Boards of both the Albany Medical Center and the University of California, Berkeley. All patients had self-declared normal hearing.

Task. Patients passively listened to the song *Another Brick in the Wall* (Part 1) from Pink Floyd (released on the album The Wall, Harvest Records/Columbia Records, 1979). They were instructed to listen attentively to the music, without focusing on any special detail. Total song duration was 190.72 seconds. The auditory stimulus was digitized at 44.1 kHz, and delivered through in-ear monitor headphones (bandwidth 12Hz-23.5kHz, 20dB isolation from surrounding noise) at a comfortable sound level adjusted for each patient (50 to 60 dB SL). Eight patients had more than one recording of the present task, in which cases we selected the cleanest one (i.e., containing the least epileptic activity or noisy electrodes).

Intracranial recordings. Direct cortical recordings were obtained through grids or strips of platinum-iridium electrodes (Ad-Tech Medical, Oak Creek, WI), with center-to-center distances of 10 mm for 21 patients, 6 mm for 4, 4 mm for 3 or 3 mm for 1. We recruited patients in the study if their implantation map covered at least partially the superior temporal gyri (left or right). The cohort consists of 28 unilateral cases (18 left, 10 right) and one bilateral case. Total number of electrodes across all 29 patients was 2,668 (range 36-250, mean 92 electrodes). ECoG activity was recorded at a sampling rate of 1,200 Hz using g.USBamp biosignal acquisition devices (g.tec, Graz, Austria) and BCI2000 (Schalk et al., 2004).

Preprocessing – Auditory stimulus. To study the relationship between the acoustics of the auditory stimulus and the ECoG-recorded neural activity, the song waveform was transformed into a magnitude-only auditory spectrogram using the NSL Matlab Toolbox (Chi et al., 2005). This transformation mimics the processing steps of early stages of the auditory pathways, from the cochlea's spectral filter bank to the midbrain's reduced upper limit of phase-locking ability, and outputs a psychoacoustic-, neurophysiologic-based spectrotemporal representation of the song. The resulting auditory spectrogram has 128 frequency bins from 180 to 7,246 Hz, with characteristic frequencies uniformly distributed along a logarithmic frequency axis (24 channels per octave), and a sampling rate of 100 Hz. This full-resolution, 128-frequency-bin spectrogram is used in the song reconstruction analysis. For all other analyses, to decrease the computational load and the number of features, we outputted a reduced spectrogram with 32 frequency bins from 188 to 6,745 Hz.

Preprocessing – ECoG data. We used the High-Frequency Activity (HFA; 70 to 150 Hz) as an estimate of local neural activity (Rich & Wallis, 2017). For each dataset, we visually inspected raw recorded signals and removed electrodes exhibiting noisy or epileptic activity, with the help of a

neurologist (RTK). We then extracted data aligned with the song stimulus, adding 10 seconds of data padding before and after the song (to prevent filtering-induced edge artifacts). We filtered out power-line noise, using a range of notch filters centered at 60 Hz and harmonics up to 300 Hz (Butterworth, 4th order, 2 Hz bandwidth), and removed slow drifts with a 1 Hz high-pass filter (Butterworth, 4th order). We used a bandpass-Hilbert approach (Bruns, 2004) to extract HFA, with 20-Hz-wide sub-bands spanning from 70 to 150 Hz in 5 Hz steps (70 to 90, 75 to 95, … up to 130 to 150 Hz). We chose a 20 Hz bandwidth to enable the observation of temporal modulations up to 10 Hz (Dvorak & Fenton, 2014), encompassing the 6.66 Hz sixteenth-note rhythm guitar pattern, pervasive throughout the song. This constitutes a crucial methodological point, enabling the observation of the rhythmic component (Fig. 3D). For each sub-band, we first bandpass-filtered the signal (Butterworth, 4th order), then performed median-based Common Average Reference (CAR; Liu et al., 2015), and computed the Hilbert transform to obtain the envelope. We standardized each sub-band envelope using robust scaling on the whole time period (subtracting the median and dividing by the interquartile range between the 10th and 90th percentiles), and average them together to yield the HFA estimate. We performed CAR separately for electrodes plugged on different splitter boxes to optimize denoising in 14 participants. Finally, we removed the 10-second pads, down-sampled data to 100 Hz to match the stimulus spectrogram's sampling rate, and tagged outlier time samples exceeding 7 standard deviations for later removal in the modeling preprocessing. We used Fieldtrip (version from May 11, 2021; Oostenveld et al., 2011) and homemade scripts to perform all above preprocessing steps. Unless specified otherwise, all further analyses and computations were implemented in MATLAB (The MathWorks, Natick, MA, USA; version 2021a). Code is available upon request.

Preprocessing – Anatomical data. We followed the anatomical data processing pipeline presented in Stolk et al., 2018 to localize electrodes from a pre-implantation MRI, a post-implantation CT scan and coverage information mapping electrodes to channel numbers in the functional data. After co-registration of the CT scan to the MRI, we performed brain-shift compensation with a hull obtained using scripts from the iso2mesh toolbox (Fang & Boas, 2009; Blenkmann et al., 2017). Cortical surfaces were extracted using the Freesurfer toolbox (Dale et al., 1999). We used volume-based normalization to convert patient-space electrode coordinates into MNI coordinates for illustration purposes, and surface-based normalization using the Freesurfer's fsaverage template to automatically obtain anatomical labels from the aparc+aseg atlas. Labels were then confirmed by a neurologist (RTK).

Encoding – Data preparation. We used Spectro-Temporal Receptive Fields (STRFs) as encoding models, with the 32 frequency bins of the stimulus spectrogram as features or predictors, and the HFA of a given electrode as target to be predicted.

We log-transformed the auditory spectrogram to compress all acoustic features into the same order of magnitude (e.g., low-sound-level musical background and high-sound-level lyrics). This ensured modeling would not be dominated by high-volume musical elements.

We then computed the feature lag matrix from the song's auditory spectrogram. As HFA is elicited by the song stimulus, we aim at predicting HFA from the preceding song spectrogram. We chose a time window between 750 ms and 0 ms before HFA, to allow a sufficient temporal integration of auditory-related neural responses, while ensuring a reasonable features-to-

observations ratio to avoid overfitting. This resulted in 2,400 features (32 frequency bins by 75 time lags at fs=100Hz).

We obtained 18,898 observations per electrode, each one consisting of a set of one target HFA value and its preceding 750ms auditory spectrogram excerpt (19,072 samples of the whole song, minus 74 samples at the beginning for which there is no preceding 750ms window).

At each electrode, we rejected observations for which the HFA value exceeded 7 standard deviations (Z units), resulting in an average rejection rate of 1.83% (min 0% - max 15.02%, SD 3.2%).

Encoding – Model fitting. To obtain a fitted STRF for a given electrode, we iterated through the following steps 250 times.

We first split the dataset into training, validation and test sets (60-20-20 ratio, respectively) using a custom group-stratified-shuffle-split algorithm (based on the StratifiedShuffleSplit cross-validator in scikit-learn). We defined relatively long, 2-second groups of consecutive samples as indivisible chunks of data. This ensured that training and test sets would not contain neighbor, virtually identical samples (as both music and neural data are highly correlated over short periods of time), and was critical to prevent overfitting. We used stratification to enforce equal splitting ratios between the vocal (13 to 80s) and instrumental parts of the song. This ensured stability of model performance across all 250 iterations, by avoiding that a model could be trained on the instrumentals only and tested on the vocals. We used shuffle splitting, akin to bootstrapping with replacement between iterations, which allows us to determine test set size independently from the number of iterations (as opposed to KFold cross-validation).

We then standardized the features, by fitting a robust scaler to the training set only (estimates the median and the 2-98 quantile range; RobustScaler in sklearn package), and using it to transform all training, validation and test sets. This gives comparable importance to all features, i.e., every time lag and frequency of the auditory spectrogram.

We employed linear regression with RMSProp optimizer for efficient model convergence, Huber loss cost function for robustness to outlier samples, and early stopping to further prevent overfitting. In early stopping, a generalization error is estimated on the validation set at each training step, and model fitting ends after this error stops diminishing for 10 consecutive steps. This model was implemented in Tensorflow 1.6 and Python 3.6. The learning rate hyperparameter of the RMSProp optimizer was manually tuned to ensure fast model convergence all by avoiding exploding gradients (overshooting of the optimization minimum).

We evaluated prediction accuracy of the fitted model by computing both the correlation coefficient (Pearson's r) and the R-squared between predicted and actual test-set target (i.e., HFA at a given electrode). Along with these two performance metrics, we also saved the fitted model coefficients.

Then, we combined these 250 split-scale-fit-evaluate iterations in a bootstrap-like approach to obtain one STRF and assess its significance (i.e., whether we can linearly predict HFA, at a given electrode, from the song spectrogram). For each STRF, we z-scored each coefficient across the 250 models. For the prediction accuracy, we computed the 95% confidence interval (CI) from the 250 correlation coefficients, and deemed an electrode as significant if its 95% CI did

not contain 0. As an additional criterion, we rejected significant electrodes with an average R-squared (across the 250 models) at or below 0.

Encoding – Analysis of prediction accuracy. To assess how strongly each brain region encodes the song, we performed a two-way ANOVA on the correlation coefficients of all electrodes showing a significant STRF, with laterality (left or right hemisphere) and area (STG, sensorimotor, IFG or other) as factors. We then performed a multiple comparison (post hoc) test to disentangle any differences between factor levels.

Encoding – Analysis of model coefficients. We analyzed the STRF tuning patterns using an independent component analysis (ICA), to highlight electrode populations tuned to distinct STRF features. Firstly, we ran an ICA with 10 components on the centered STRF coefficients, to identify components individually explaining more than 5% of variance. We computed explained variance by back-projecting each component and using the following formula: $pvaf_i = 100 - 100*mean(var(STRF - backproj_i))/mean(var(STRF))$; with i from 1 to 10 components, $pvaf_i$ being the percent variance accounted for by ICA component i, STRF being the centered STRF coefficients, and $backproj_i$ being the back-projection of ICA component i in electrode space. We found 3 ICA components explaining more than 5% of variance. To optimize the unmixing process, we ran a new ICA asking for 3 components. Then, we determined each component sign by setting as positive the sign of the most salient coefficient. Lastly, for each ICA component, we defined electrodes as representing the component if their ICA coefficient was positive.

To look at rhythmic tuning patterns, we computed the temporal modulations of each STRF. Indeed, due to their varying frequencies and latencies, they were not captured by the combined component analysis. We quantified temporal modulations between 1 and 16 Hz over the 32 spectral frequency bins of each STRF, and extracted the maximum modulation value across all 32 frequency bins between 6 and 7 Hz of temporal modulations, corresponding to the song rhythmicity of 16th notes at 99 bpm. We defined electrodes as representing the component if their maximum modulation value was above a manually defined threshold of .3.

Encoding – Musical elements. To link STRF components to musical elements in the song, we ran a sliding-window correlation between each component and the song spectrogram. Positive correlation values indicate specific parts of the song or musical elements (i.e., vocals, lead guitar…) that elicit an increase of HFA.

Decoding - Ablation analysis. To assess the contribution of different brain regions and STRF components in representing the song, we performed an ablation analysis. We quantified the impact of ablating sets of electrodes on the prediction accuracy of a linear decoding model computed using all 347 significant electrodes. Firstly, we constituted sets of electrodes based on anatomical or functional criteria. We defined 12 anatomical sets by combining two factors – area (whole hemisphere, STG, SMC, IFG, or other areas) and laterality (bilateral, left or right). We defined 12 functional sets by combining two factors – STRF component identified in the STRF coefficient analyses (onset, sustained, late onset, and rhythmic) and laterality (bilateral, left or right). See Fig. 5 for the exact list of electrode sets. Secondly, we computed the decoding models using the same algorithm as for the encoding models. Decoding models aim at predicting the

song spectrogram from the elicited neural activity. Here, we used HFA from a set of electrodes as input, and a given frequency bin of the song spectrogram as output. For each of the 24 ablated sets of electrodes, we obtained 32 models (one per spectrogram frequency bin), and compared each one of them to the corresponding baseline model computed using all 347 significant electrodes (repeated-measure one-way ANOVA). We then performed a multiple comparison (post hoc) test to assess differences between ablations.

We based our interpretation of ablations results on the following assumptions. Collectively, as they had significant STRFs, all 347 significant electrodes represent acoustics information on the song. If ablating a set of electrodes resulted in a significant impact on decoding accuracy, we considered that this set represented unique information. Indeed, were this information shared with another set of electrodes, a compensation-like mechanism could occur and void the impact on decoding accuracy. If ablating a set of electrodes resulted in no significant impact on decoding accuracy, we considered that this set represented redundant information, shared with other electrodes (as the STRFs were significant, we ruled out the possibility that it could be because this set did not represent any acoustic information). Also, comparing the impact of a given set and one of its subsets of electrodes provided further insights on the unique or redundant nature of the represented information.

Decoding – Parametric analyses. We quantified the influence of different methodological factors (number of electrodes, dataset duration, and model type) on the prediction accuracy of decoding models. In a bootstrapping approach, we randomly constituted subsets of 5, 10, 20, 40, 80, 160 and 320 electrodes (sampling without replacement) to be used as inputs of linear decoding models. We processed 100 bootstrap resamples (i.e., 100 sets of 5 electrodes, 100 sets of 10 electrodes…), and normalized for each of the 32 frequency bins the resulting correlation coefficients by the correlation coefficients of the full, 347-electrode decoding model. For each resample, we averaged the correlation coefficients from all 32 models (1 per frequency bin of the song spectrogram). This yielded 100 prediction accuracy estimates per number of electrodes. We then fitted a two-term power series model to these estimates, to quantify the apparent power-law behavior of the obtained bootstrap curve. We adopted the same approach for dataset duration, with excerpts of 15, 30, 60, 90, 120, 150 and 180 consecutive seconds. For model type, we trained linear and non-linear decoding models to predict each of the 128 frequency bins of the song spectrogram (full spectral resolution). We used the multilayer perceptron (MLP)—a simple, fully connected neural network, as non-linear model. We tuned model hyperparameters (number of hidden layers, number of units per layer, and the L2-regularization) through grid search using a validation set, and evaluated model performance on a 20-second, fixed held-out test set.

Decoding - Song reconstruction. To explore the extent to which we could reconstruct the song from neural activity, we collected the 128 predicted targets for both linear and MLP decoding models. As both magnitude and phase information are required to properly reconstruct a waveform from a spectrogram, we used an iterative phase-estimation algorithm to transform the magnitude-only decoded auditory spectrogram into the song waveform (Chi et al., 2005).

Acknowledgements

This work was funded by the Fondation Pour l'Audition (FPA RD-2015-2, LB), the National Institutes of Health NIH/NIBIB (R01-EB026439, P41-EB018783, PB), NIH/NINDS (U24-NS109103, U01-NS108916, R13-NS118932, PB; R01-NS21135, RTK), NIH/NIDCD (1R21-DC018374, BNP), and the National Science Foundation (NSF 1850687, BNP). We thank Stephanie Martin, Christopher R. Holdgraf, and Christian Mikutta for their helpful comments on data analysis.

Author contributions

Study design and data acquisition (GS, PB), data preprocessing and analysis (LB, BNP), writing (LB, AL, DM), editing (RTK, BP).

Competing interests

The authors confirm that there are no relevant financial or non-financial competing interests to report.

References

Abrams, D. A., Bhatara, A., Ryali, S., Balaban, E., Levitin, D. J., & Menon, V. (2011). Decoding Temporal Structure in Music and Speech Relies on Shared Brain Resources but Elicits Different Fine-Scale Spatial Patterns. *Cerebral Cortex (New York, NY)*, *21*(7), 1507–1518. https://doi.org/10.1093/cercor/bhq198

Aertsen, A., & Johannesma, P. I. M. (1980). Spectro-temporal receptive fields of auditory neurons in the grassfrog—I. Characterization of tonal and natural stimuli. *Biological Cybernetics*, *38*, 223–234. https://doi.org/10.1007/BF00337015

Aertsen, A. M. H. J., Johannesma, P. I. M., & Hermes, D. J. (1980). Spectro-temporal receptive fields of auditory neurons in the grassfrog—II. Analysis of the Stimulus-Event Relation for Tonal Stimuli. *Biological Cybernetics*, *38*(4), 235–248. https://doi.org/10.1007/BF00337016

Aertsen, A., Olders, J. H. J., & Johannesma, P. I. M. (1981). Spectro-temporal receptive fields of auditory neurons in the grassfrog—III. Analysis of the stimulus-event relation for natural stimuli. *Biological Cybernetics*, *39*, 195–209. https://doi.org/10.1007/BF00342772

Akbari, H., Khalighinejad, B., Herrero, J. L., Mehta, A. D., & Mesgarani, N. (2019). Towards reconstructing intelligible speech from the human auditory cortex. *Scientific Reports*, *9*(1), 874. https://doi.org/10.1038/s41598-018-37359-z

Albouy, P., Benjamin, L., Morillon, B., & Zatorre, R. J. (2020). Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science*, *367*(6481), 1043–1047. https://doi.org/10.1126/science.aaz3468

Albouy, P., Mattout, J., Bouet, R., Maby, E., Sanchez, G., Aguera, P.-E., Daligault, S., Delpuech, C., Bertrand, O., Caclin, A., & Tillmann, B. (2013). Impaired pitch perception and memory in congenital amusia: The deficit starts in the auditory cortex. *Brain: A Journal of Neurology*, *136*(Pt 5), 1639–1661. https://doi.org/10.1093/brain/awt082

Alluri, V., Toiviainen, P., Jääskeläinen, I. P., Glerean, E., Sams, M., & Brattico, E. (2012). Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage*, *59*(4), 3677–3689. https://doi.org/10.1016/j.neuroimage.2011.11.019

Anumanchipalli, G. K., Chartier, J., & Chang, E. F. (2019). Speech synthesis from neural decoding of spoken sentences. *Nature*, *568*(7753), 493. https://doi.org/10.1038/s41586-019-1119-1

Baltzell, L. S., Srinivasan, R., & Richards, V. (2019). Hierarchical organization of melodic sequences is encoded by cortical entrainment. *NeuroImage*, *200*, 490–500. https://doi.org/10.1016/j.neuroimage.2019.06.054

Blenkmann, A. O., Phillips, H. N., Princich, J. P., Rowe, J. B., Bekinschtein, T. A., Muravchik, C. H., & Kochen, S. (2017). iElectrodes: A Comprehensive Open-Source Toolbox for Depth and Subdural Grid Electrode Localization. *Frontiers in Neuroinformatics*, *11*, 14. https://doi.org/10.3389/fninf.2017.00014

Brattico, E., Tervaniemi, M., Näätänen, R., & Peretz, I. (2006). Musical scale properties are automatically processed in the human auditory cortex. *Brain Research*, *1117*(1), 162–174. https://doi.org/10.1016/j.brainres.2006.08.023

Bruns, A. (2004). Fourier-, Hilbert- and wavelet-based signal analysis: Are they really different approaches? *Journal of Neuroscience Methods*, *137*(2), 321–332. https://doi.org/10.1016/j.jneumeth.2004.03.002

Caclin, A., Brattico, E., Tervaniemi, M., Näätänen, R., Morlet, D., Giard, M.-H., & McAdams, S. (2006). Separate neural processing of timbre dimensions in auditory sensory memory. *Journal of Cognitive Neuroscience*, *18*(12), 1959–1972. https://doi.org/10.1162/jocn.2006.18.12.1959

Caclin, A., Giard, M.-H., Smith, B. K., & McAdams, S. (2007). Interactive processing of timbre dimensions: A Garner interference study. *Brain Research*, *1138*, 159–170. https://doi.org/10.1016/j.brainres.2006.12.065

Casey, M. A. (2017). Music of the 7Ts: Predicting and Decoding Multivoxel fMRI Responses with Acoustic, Schematic, and Categorical Music Features. *Frontiers in Psychology*, *8*, 1179. https://doi.org/10.3389/fpsyg.2017.01179

Chi, T., Ru, P., & Shamma, S. A. (2005). Multiresolution spectrotemporal analysis of complex sounds. *The Journal of the Acoustical Society of America*, *118*(2), 887–906.

Choi, I., Rajaram, S., Varghese, L. A., & Shinn-Cunningham, B. G. (2013). Quantifying attentional modulation of auditory-evoked cortical responses from single-trial electroencephalography. *Frontiers in Human Neuroscience*, *7*, 115. https://doi.org/10.3389/fnhum.2013.00115

Conner, C. R., Ellmore, T. M., Pieters, T. A., DiSano, M. A., & Tandon, N. (2011). Variability of the Relationship between Electrophysiology and BOLD-fMRI across Cortical Regions in Humans. *Journal of Neuroscience*, *31*(36), 12855–12865. https://doi.org/10.1523/JNEUROSCI.1457-11.2011

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

David, S. V., Mesgarani, N., & Shamma, S. A. (2007). Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network (Bristol, England)*, *18*(3), 191–212. https://doi.org/10.1080/09548980701609235

Deike, S., Gaschler-Markefski, B., Brechmann, A., & Scheich, H. (2004). Auditory stream segregation relying on timbre involves left auditory cortex. *Neuroreport*, *15*(9), 1511–1514. https://doi.org/10.1097/01.wnr.0000132919.12990.34

Ding, Y., Zhang, Y., Zhou, W., Ling, Z., Huang, J., Hong, B., & Wang, X. (2019). Neural Correlates of Music Listening and Recall in the Human Brain. *Journal of Neuroscience*, *39*(41), 8112–8123. https://doi.org/10.1523/JNEUROSCI.1468-18.2019

Dvorak, D., & Fenton, A. A. (2014). Toward a proper estimation of phase–amplitude coupling in neural oscillations. *Journal of Neuroscience Methods*, *225*, 42–56. https://doi.org/10.1016/j.jneumeth.2014.01.002

Fang, Q., & Boas, D. A. (2009). Tetrahedral mesh generation from volumetric binary and grayscale images. *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 1142–1145. https://doi.org/10.1109/ISBI.2009.5193259

Foldal, M. D., Blenkmann, A. O., Llorens, A., Knight, R. T., Solbakk, A.-K., & Endestad, T. (2020). The brain tracks auditory rhythm predictability independent of selective attention. *Scientific Reports*, *10*, 7975. https://doi.org/10.1038/s41598-020-64758-y

Geiser, E., Ziegler, E., Jancke, L., & Meyer, M. (2009). Early electrophysiological correlates of meter and rhythm processing in music perception. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *45*(1), 93–102. https://doi.org/10.1016/j.cortex.2007.09.010

Giraud, A.-L., Kleinschmidt, A., Poeppel, D., Lund, T. E., Frackowiak, R. S. J., & Laufs, H. (2007). Endogenous cortical rhythms determine cerebral specialization for speech perception and production. *Neuron*, *56*(6), 1127–1134. https://doi.org/10.1016/j.neuron.2007.09.038

Gordon, C. L., Cobb, P. R., & Balasubramaniam, R. (2018). Recruitment of the motor system during music listening: An ALE meta-analysis of fMRI data. *PLOS ONE*, *13*(11), e0207213. https://doi.org/10.1371/journal.pone.0207213

Goydke, K. N., Altenmüller, E., Möller, J., & Münte, T. F. (2004). Changes in emotional tone and instrumental timbre are reflected by the mismatch negativity. *Brain Research. Cognitive Brain Research*, *21*(3), 351–359. https://doi.org/10.1016/j.cogbrainres.2004.06.009

Grahn, J. A., & McAuley, J. D. (2009). Neural bases of individual differences in beat perception. *NeuroImage*, *47*(4), 1894–1903. https://doi.org/10.1016/j.neuroimage.2009.04.039

Grahn, J. A., & Rowe, J. B. (2009). Feeling the beat: Premotor and striatal interactions in musicians and nonmusicians during beat perception. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *29*(23), 7540–7548. https://doi.org/10.1523/JNEUROSCI.2018-08.2009

Gupta, D., Hill, N., Brunner, P., Gunduz, A., & Schalk, G. (2014). Simultaneous Real-Time Monitoring of Multiple Cortical Systems. *Journal of Neural Engineering*, *11*, 056001. https://doi.org/10.1088/1741-2560/11/5/056001

Hall, E. L., Robson, S. E., Morris, P. G., & Brookes, M. J. (2014). The relationship between MEG and fMRI. *NeuroImage*, *102*, 80–91. https://doi.org/10.1016/j.neuroimage.2013.11.005

Hamilton, L. S., Edwards, E., & Chang, E. F. (2018). A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Current Biology*, *28*(12), 1860-1871.e4. https://doi.org/10.1016/j.cub.2018.04.033

Harding, E. E., Sammler, D., Henry, M. J., Large, E. W., & Kotz, S. A. (2019). Cortical tracking of rhythm in music and speech. *NeuroImage*, *185*, 96–101. https://doi.org/10.1016/j.neuroimage.2018.10.037

Herff, S. A., Herff, C., Milne, A. J., Johnson, G. D., Shih, J. J., & Krusienski, D. J. (2020). Prefrontal High Gamma in ECoG Tags Periodicity of Musical Rhythms in Perception and Imagination. *ENeuro*, *7*(4). https://doi.org/10.1523/ENEURO.0413-19.2020

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews. Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

Hoefle, S., Engel, A., Basilio, R., Alluri, V., Toiviainen, P., Cagy, M., & Moll, J. (2018). Identifying musical pieces from fMRI data using encoding and decoding models. *Scientific Reports*, *8*(1), 2266. https://doi.org/10.1038/s41598-018-20732-3

Holdgraf, C. R., Heer, W. de, Pasley, B., Rieger, J., Crone, N., Lin, J. J., Knight, R. T., & Theunissen, F. E. (2016). Rapid tuning shifts in human auditory cortex enhance speech intelligibility. *Nature Communications*, *7*, 13654. https://doi.org/10.1038/ncomms13654

Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017). Encoding and Decoding Models in Cognitive Electrophysiology. *Frontiers in Systems Neuroscience*, *11*. https://doi.org/10.3389/fnsys.2017.00061

Hullett, P. W., Hamilton, L. S., Mesgarani, N., Schreiner, C. E., & Chang, E. F. (2016). Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *36*(6), 2014–2026. https://doi.org/10.1523/JNEUROSCI.1779-15.2016

Hyde, K. L., Peretz, I., & Zatorre, R. J. (2008). Evidence for the role of the right auditory cortex in fine pitch resolution. *Neuropsychologia*, *46*(2), 632–639. https://doi.org/10.1016/j.neuropsychologia.2007.09.004

Janata, P. (2015). Chapter 11—Neural basis of music perception. In M. J. Aminoff, F. Boller, & D. F. Swaab (Eds.), *Handbook of Clinical Neurology* (Vol. 129, pp. 187–205). Elsevier. https://doi.org/10.1016/B978-0-444-62630-1.00011-1

Janata, P., Birk, J. L., Van Horn, J. D., Leman, M., Tillmann, B., & Bharucha, J. J. (2002). The cortical topography of tonal structures underlying Western music. *Science (New York, N.Y.)*, *298*(5601), 2167–2170. https://doi.org/10.1126/science.1076262

Jongsma, M. L. A., Desain, P., & Honing, H. (2004). Rhythmic context influences the auditory evoked potentials of musicians and non-musicians. *Biological Psychology*, *66*(2), 129–152. https://doi.org/10.1016/j.biopsycho.2003.10.002

Koelsch, S. (2011). Toward a Neural Basis of Music Perception – A Review and Updated Model. *Frontiers in Psychology*, *2*. https://doi.org/10.3389/fpsyg.2011.00110

Kohoutová, L., Heo, J., Cha, S., Lee, S., Moon, T., Wager, T. D., & Woo, C.-W. (2020). Toward a unified framework for interpreting machine-learning models in neuroimaging. *Nature Protocols*, *15*(4), 1399–1435. https://doi.org/10.1038/s41596-019-0289-5

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., & Schalk, G. (2013). The Tracking of Speech Envelope in the Human Cortex. *PLOS ONE*, *8*(1), e53398. https://doi.org/10.1371/journal.pone.0053398

Kumar, S., Sedley, W., Nourski, K. V., Kawasaki, H., Oya, H., Patterson, R. D., Howard, M. A., Friston, K. J., & Griffiths, T. D. (2011). Predictive coding and pitch processing in the auditory cortex. *Journal of Cognitive Neuroscience*, *23*(10), 3084–3094. https://doi.org/10.1162/jocn_a_00021

Kumar, S., Stephan, K. E., Warren, J. D., Friston, K. J., & Griffiths, T. D. (2007). Hierarchical Processing of Auditory Objects in Humans. *PLOS Computational Biology*, *3*(6), e100. https://doi.org/10.1371/journal.pcbi.0030100

Lachaux, J. P., Rudrauf, D., & Kahane, P. (2003). Intracranial EEG and human brain mapping. *Journal of Physiology, Paris*, *97*(4–6), 613–628. https://doi.org/10.1016/j.jphysparis.2004.01.018

Leszczyński, M., Barczak, A., Kajikawa, Y., Ulbert, I., Falchier, A. Y., Tal, I., Haegens, S., Melloni, L., Knight, R. T., & Schroeder, C. E. (2020). Dissociation of broadband high-frequency activity and neuronal firing in the neocortex. *Science Advances*, *6*(33), eabb0977. https://doi.org/10.1126/sciadv.abb0977

Limb, C. J. (2006). Structural and functional neural correlates of music perception. *The Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology*, *288A*(4), 435–446. https://doi.org/10.1002/ar.a.20316

Liu, Y., Coon, W. G., de Pesters, A., Brunner, P., & Schalk, G. (2015). The effects of spatial filtering and artifacts on electrocorticographic signals. *Journal of Neural Engineering*, *12*(5), 056008. https://doi.org/10.1088/1741-2560/12/5/056008

Martin, S., Mikutta, C., Leonard, M. K., Hungate, D., Koelsch, S., Shamma, S., Chang, E. F., Millán, J. D. R., Knight, R. T., & Pasley, B. N. (2017). Neural Encoding of Auditory Features during Music Perception and Imagery. *Cerebral Cortex (New York, N.Y.: 1991)*, 1–12. https://doi.org/10.1093/cercor/bhx277

Merriam, A. P., & Merriam, V. (1964). *The Anthropology of Music*. Northwestern University Press.

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2009). Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *Journal of Neurophysiology*, *102*(6), 3329–3339. https://doi.org/10.1152/jn.91128.2008

Meyes, R., Lu, M., Puiseau, C. W. D., & Meisen, T. (2019). Ablation Studies in Artificial Neural Networks. *ArXiv*.

Moses, D. A., Metzger, S. L., Liu, J. R., Anumanchipalli, G. K., Makin, J. G., Sun, P. F., Chartier, J., Dougherty, M. E., Liu, P. M., Abrams, G. M., Tu-Chan, A., Ganguly, K., & Chang, E. F. (2021). Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria. *New England Journal of Medicine*, *385*(3), 217–227. https://doi.org/10.1056/NEJMoa2027540

Nakai, T., Koide-Majima, N., & Nishimoto, S. (2021). Correspondence of categorical and feature-based representations of music in the human brain. *Brain and Behavior*, *11*(1), e01936. https://doi.org/10.1002/brb3.1936

Nan, Y., & Friederici, A. D. (2013). Differential roles of right temporal cortex and Broca's area in pitch processing: Evidence from music and Mandarin. *Human Brain Mapping*, *34*(9), 2045–2054. https://doi.org/10.1002/hbm.22046

Norman-Haignere, S. V., Feather, J., Boebinger, D., Brunner, P., Ritaccio, A., McDermott, J. H., Schalk, G., & Kanwisher, N. (2020). *Intracranial recordings from human auditory cortex reveal a neural population selective for song* (p. 696161). https://doi.org/10.1101/696161

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. https://doi.org/10.1155/2011/156869

Pasley, B. N., David, S. V., Mesgarani, N., Flinker, A., Shamma, S. A., Crone, N. E., Knight, R. T., & Chang, E. F. (2012). Reconstructing speech from human auditory cortex. *PLoS Biology*, *10*(1), e1001251. https://doi.org/10.1371/journal.pbio.1001251

Peretz, I. (2006). The nature of music from a biological perspective. *Cognition*, *100*(1), 1–32. https://doi.org/10.1016/j.cognition.2005.11.004

Peretz, I., & Zatorre, R. J. (2005). Brain organization for music processing. *Annual Review of Psychology*, *56*, 89–114. https://doi.org/10.1146/annurev.psych.56.091103.070225

Plack, C. J., Barker, D., & Hall, D. A. (2014). Pitch coding and pitch processing in the human brain. *Hearing Research*, *307*, 53–64. https://doi.org/10.1016/j.heares.2013.07.020

Potes, C., Brunner, P., Gunduz, A., Knight, R. T., & Schalk, G. (2014). Spatial and temporal relationships of electrocorticographic alpha and gamma activity during auditory processing. *NeuroImage*, *97*, 188–195. https://doi.org/10.1016/j.neuroimage.2014.04.045

Potes, C., Gunduz, A., Brunner, P., & Schalk, G. (2012). Dynamics of electrocorticographic (ECoG) activity in human temporal and frontal cortical areas during music listening. *NeuroImage*, *61*(4), 841–848. https://doi.org/10.1016/j.neuroimage.2012.04.022

Rich, E. L., & Wallis, J. D. (2017). Spatiotemporal dynamics of information encoding revealed in orbitofrontal high-gamma. *Nature Communications*, *8*(1), 1139. https://doi.org/10.1038/s41467-017-01253-5

Saari, P., Burunat, I., Brattico, E., & Toiviainen, P. (2018). Decoding Musical Training from Dynamic Processing of Musical Features in the Brain. *Scientific Reports*, *8*(1), 708. https://doi.org/10.1038/s41598-018-19177-5

Schaefer, R. S., Vlek, R. J., & Desain, P. (2011). Decomposing rhythm processing: Electroencephalography of perceived and self-imposed rhythmic patterns. *Psychological Research*, *75*(2), 95–106. https://doi.org/10.1007/s00426-010-0293-4

Schalk, G., McFarland, D. J., Hinterberger, T., Birbaumer, N., & Wolpaw, J. R. (2004). BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Bio-Medical Engineering*, *51*(6), 1034–1043. https://doi.org/10.1109/TBME.2004.827072

Snyder, J. S., & Large, E. W. (2005). Gamma-band activity reflects the metric structure of rhythmic tone sequences. *Brain Research. Cognitive Brain Research*, *24*(1), 117–126. https://doi.org/10.1016/j.cogbrainres.2004.12.014

Sturm, I., Blankertz, B., Potes, C., Schalk, G., & Curio, G. (2014). ECoG high gamma activity reveals distinct cortical representations of lyrics passages, harmonic and timbre-related changes in a rock song. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00798

Talebi, V., & Baker, C. L. (2012). Natural versus synthetic stimuli for estimating receptive field models: A comparison of predictive robustness. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(5), 1560–1576. https://doi.org/10.1523/JNEUROSCI.4661-12.2012

Theunissen, F. E., David, S. V., Singh, N. C., Hsu, A., Vinje, W. E., & Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network (Bristol, England)*, *12*(3), 289–316.

Theunissen, F. E., Sen, K., & Doupe, A. J. (2000). Spectral-temporal receptive fields of nonlinear auditory neurons obtained using natural sounds. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *20*(6), 2315–2331.

Toiviainen, P., Alluri, V., Brattico, E., Wallentin, M., & Vuust, P. (2014). Capturing the musical brain with Lasso: Dynamic decoding of musical features from fMRI data. *NeuroImage*, *88*, 170–180. https://doi.org/10.1016/j.neuroimage.2013.11.017

Trainor, L. J., McDonald, K. L., & Alain, C. (2002). Automatic and controlled processing of melodic contour and interval information measured by electrical brain activity. *Journal of Cognitive Neuroscience*, *14*(3), 430–442. https://doi.org/10.1162/089892902317361949

Treder, M. S., Purwins, H., Miklody, D., Sturm, I., & Blankertz, B. (2014). Decoding auditory attention to instruments in polyphonic music using single-trial EEG classification. *Journal of Neural Engineering*, *11*(2), 026009. https://doi.org/10.1088/1741-2560/11/2/026009

Vikene, K., Skeie, G. O., & Specht, K. (2019). Compensatory task-specific hypersensitivity in bilateral planum temporale and right superior temporal gyrus during auditory rhythm and omission processing in Parkinson's disease. *Scientific Reports*, *9*(1), 1–9. https://doi.org/10.1038/s41598-019-48791-0

Warren, J. (2008). How does the brain process music? *Clinical Medicine*, *8*(1), 32–36. https://doi.org/10.7861/clinmedicine.8-1-32

Zatorre, R. J., Chen, J. L., & Penhune, V. B. (2007). When the brain plays music: Auditory–motor interactions in music perception and production. *Nature Reviews Neuroscience*, *8*(7), 547–558. https://doi.org/10.1038/nrn2152

Zatorre, R. J., & Gandour, J. T. (2008). Neural specializations for speech and pitch: Moving beyond the dichotomies. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *363*(1493), 1087–1104. https://doi.org/10.1098/rstb.2007.2161

Zatorre, R. J., & Salimpoor, V. N. (2013). From perception to pleasure: Music and its neural substrates. *Proceedings of the National Academy of Sciences of the United States of America*, *110 Suppl 2*, 10430–10437. https://doi.org/10.1073/pnas.1301228110