Neural coding of spectrotemporal modulations in the auditory cortex supports speech and music categorization

AUTHOR LIST

Jérémie Ginzburg[1,2,3,4,*], Émilie Cloutier Debaque[2,3,4], Arthur Borderie[2,3,4], Benjamin Morillon[5], Laurence Martineau[4], Paule Lessard Bonaventure[4], Robert J Zatorre[1,2,6], Philippe Albouy[2,3,4,6]

AUTHOR AFFILIATIONS AND FOOTNOTES

1: Montreal Neurological Institute, McGill University, Montreal, Canada

2: Centre for Research in Brain, Language and Music (CRBLM), and International Laboratory for Brain Music and Sound Research (BRAMS), Montreal, Canada

3: CERVO Brain research center, Laval University, Québec, Canada

4: Centre Hospitalier Universitaire de Québec, Laval University, Québec, Canada

5: Aix-Marseille Université, INSERM, Institut de Neurosciences des Systèmes, Marseille, France

6: Senior authors

*: Lead and corresponding author

CONTACT INFORMATION

*Correspondance: jeremie.ginzburg@mcgill.ca

SUMMARY

Humans effortlessly distinguish speech from music, but the neural basis of this categorization remains debated. Here, we combined intracranial recordings in humans with behavioral categorization of speech or music taken from a naturalistic movie soundtrack to test whether the encoding of spectrotemporal modulation (STM) acoustic features is sufficient to categorize the two classes. We show that primarily temporal and primarily spectral modulation patterns characterize speech and music, respectively, and that non-primary auditory regions robustly track these acoustic features over time. Critically, cortical representations of STMs predicted perceptual judgments gathered in an independent sample. Finally, speech- and music-related STM representations showed stronger tracking of category-specific acoustical features in the left versus right non-primary auditory regions, respectively. These findings support a domain-general framework for auditory categorization, such that perceptual categories emerge from efficient neural coding of low-level acoustical cues, highlighting the computational relevance of STMs in shaping auditory representations that support behavior.

KEYWORDS

auditory cortex, intracranial EEG, spectrotemporal modulations, speech perception, music perception, efficient neural coding, humans

INTRODUCTION

Understanding how the brain represents perceptual categories remains a central question in cognitive neuroscience. A prominent example of this problem is provided by arguments about the neural systems underlying categorization of the two most important forms of human auditory communication: speech and music (Mehr et al., 2021). This debate specifically centers on whether speech and music categorization rely on domain-specific, modular representations in distinct neural populations that cannot be accounted for by acoustic features (Angulo-Perkins et al., 2014; Boebinger et al., 2021; Norman-Haignere et al., 2015, 2022) or whether their categorization can be accounted for by domain-general models through the hierarchical organization of lower-level neural resources capable of processing acoustical cues relevant for a wide range of auditory inputs (Albouy et al., 2020; Flinker et al., 2019; Zatorre, 2022).

The nature of these neural representations carries significant implications for our understanding of brain function, including the question of whether representations are localized or distributed across broader networks (Bizley & Cohen, 2013; Rissman & Wagner, 2012; Staeren et al., 2009). While some neuroimaging studies argue for clear functional specialization, identifying distinct bilateral brain regions dedicated to speech versus song or instrumental music (Chen et al., 2023; Friederici, 2020; Norman-Haignere et al., 2015, 2022), others propose that the representation of speech and music is based upon consistent differences in the acoustical cues that characterize these two classes of sounds, which also have different hemispheric weighting (Albouy et al., 2020; Fadiga et al., 2009; Koelsch, 2011; Robert et al., 2024; Schön et al., 2010; Te Rietmolen et al., 2024).

The domain-specific view posits that the human auditory cortex is organized to support specialized processing pathways for behaviorally meaningful sound categories (i.e. speech, music). While early stages of the auditory pathway are largely explained by tuning to basic acoustic features, this framework suggests that category-specific representations emerge just beyond these early areas, where neural processing becomes increasingly non-linear (Kell et al., 2018). These transformations are thought to give rise to representations in non-primary auditory regions that encode auditory objects in ways that are more invariant to low-level acoustic variability (McDermott & Simoncelli, 2011). Supporting this view, several studies have reported distinct neural populations in bilateral non-primary auditory areas that respond selectively and non-linearly to speech, music, or even specifically to songs (Boebinger et al., 2021; Kell et al., 2018; Norman-Haignere et al., 2015, 2022; Zuk et al., 2020), concluding that these responses cannot be explained by acoustic models (Norman-Haignere et al., 2015, 2022).

The domain-general perspective, in contrast, posits that neural populations must be capable of encoding a wide range of acoustic features while at the same time being capable of more abstract categorical representations of stimulus classes, which raises the need for a biologically plausible model that explains how activity patterns within these populations can develop representations of sound categories. Such representations must be rich enough to capture the complexity of the signal yet structured enough to allow the extraction of meaningful invariants from them. The concept of spectrotemporal receptive fields (Chi et al., 2005) offers a computationally grounded and neurophysiologically plausible framework for understanding how auditory neurons decompose incoming acoustic signals (Shamma, 2001; Singh & Theunissen, 2003; Woolley et al., 2005) without making any unnecessary assumptions about domain specificity. This model suggests that auditory neurons function as spectrotemporal modulation (STM) rate filters, a proposal supported by single-cell recordings in animals (Depireux et al., 2001; Massoudi et al., 2015; Rodríguez et al., 2010; Theunissen et al., 2000; Woolley et al., 2005), and human neuroimaging studies (Hullett et al., 2016; Santoro et al., 2014; Schönwiesner & Zatorre, 2009; Venezia et al., 2019). A recent study has shown that STM-based representations are sufficient to capture spectrotemporal features that can robustly discriminate song from speech across diverse cultures, outperforming a wide range of other acoustic descriptors (Albouy et al., 2024). Furthermore, this STM framework has been shown to account for the functional hemispheric lateralization of speech and music processing (Albouy et al., 2020; Haiduk et al., 2024), and that this lateralization, rooted in the STM encoding scheme, extends beyond speech and music (Boemio et al., 2005; Jamison et al., 2006; Robert et al., 2024; Schönwiesner et al., 2005; Zatorre & Belin, 2001) emphasizing its domain-general nature.

The present study aimed to address the domain-general versus domain-specific debate by investigating whether STM features of naturalistic sounds are sufficient to explain categorical responses to speech and music, or whether such higher-order representations cannot be accounted for by low-level acoustic features, as proposed by the domain-specific view. To address this issue, we recorded stereo-electroencephalography (sEEG) activity via implanted electrodes from patients with intractable epilepsy while they passively listened to a naturalistic 7-minute excerpt from the audio track of a movie. The excerpt contained speech and music presented both separately and simultaneously, allowing us to approach natural listening conditions with neural measures of high spatial and temporal precision. Separately, behavioral categorization data were collected from healthy participants listening to the same excerpt to identify segments of speech and/or music. These datasets allowed us to test whether the STM features of the audio track could be predicted from oscillatory magnitude recorded intracranially, and whether these features, as predicted by a domain-general account are in

turn sufficient to predict behavioral categorization in an independent group (out-of-sample prediction).

We formulated five key predictions: (i) If the behavioral categorization of the continuous naturalistic stimulus is based on STM features, we should observe distinct STM patterns in the acoustics of the waveform for segments classified as speech and music, replicating findings from studies using shorter and more controlled stimuli (Albouy et al., 2020, 2024). (ii) If the brain tracks these STM patterns in naturalistic settings, they should be robustly encoded in the auditory cortex, particularly in non-primary regions (Albouy et al., 2020; Flinker et al., 2019). (iii) If STM features alone are sufficient to support perceptual categorization, then brain-reconstructed STMs (i.e., STMs predicted from neural signals) should be able to account for behavioral categorization judgments obtained from an independent group of participants. (iv) Under the same assumption, the brain-reconstructed STMs should closely match the acoustically derived STMs that are most relevant for categorization in the healthy listeners, as determined in the previous analysis step. (v) Finally, the similarity between brain-reconstructed and acoustically derived STM patterns should exhibit hemispheric asymmetries, stronger in the left hemisphere for speech, and in the right hemisphere for music.

RESULTS

STM features predict perceptual categorization: pure spectral modulations for music, pure temporal modulations for speech

We collected behavioral categorization data from 19 healthy participants listening to the first 7 minutes and 50 seconds of the French-dubbed audio track of *Harry Potter and the Philosopher's Stone*. Participants continuously indicated whether they perceived speech and/or music by pressing corresponding keys (see Figure 1a and STAR method), resulting in binary vectors for each participant reflecting perceived categories over time. To relate these judgments to acoustic properties, we decomposed the audio signal using the STM framework (Elliott & Theunissen, 2009), extracting STM patterns at each time point with a 1.5-second sliding window (see Figure 1b and STAR method). This yielded time-resolved 2D matrices capturing spectral (0–9 cycles/kHz, sampled over 109 points) and temporal (0–15 Hz, sampled over 20 points) modulation power.

Our first goal was to investigate if STM features are sufficient to explain categorization of speech and music, and if so which specific STM patterns contribute most to each category. We applied stratified 5-fold cross-validated logistic regression models to predict behavioral categorization based on the STM decomposition of the audio track. Each model used the STM data of all audio track samples as features to predict the binary behavioral judgment (presence/absence) outcome, for each category (speech or music). Subsequently, we applied

the Haufe transformation (Haufe et al., 2014) to derive interpretable feature patterns from the feature weights, which were averaged for each judgement category across behavioral subjects (Figure 1c). For the speech category, the STM features most important for the binary classification contained pure temporal modulations. Specifically, the 90th percentile of highest values spanned from 0.8 to 15 Hz in temporal modulations, but only from 0 to 1 cyc/kHz in spectral modulations (dotted line in Fig 1c, left). Peak activity (99th percentile; solid line in Fig 1c, left) was concentrated within 5.6 to 8.8 Hz in temporal modulations but 0 to 0.4 cyc/kHz in spectral modulations. For the music category, pure spectral modulations were instead the primary distinguishing features. High activity in the 90th percentile was observed across most of the spectral modulation range (except for a narrow gap between 3.3 and 3.7 cyc/kHz), and within a range of 0 to 1.6 Hz in temporal modulations (dotted line in Fig 1c, right). Peaks (99th percentile) were localized in a wide range of spectral modulation values, around 0.3, 1.2, 5.7, 6.5, 7.4, and 8.6 cyc/kHz, and temporal modulation range of 0 to 0.8 Hz (solid lines in Fig 1c, right). These findings align with the expected dynamics of high temporal/low spectral modulation for speech signals and low temporal/high spectral modulation for musical signals (Albouy et al., 2020, 2024; Elliott & Theunissen, 2009; Flinker et al., 2019; Haiduk et al., 2024). Importantly, these results validate the use of time-resolved tracking in STM space to differentiate auditory categories. Building on this outcome, we next applied the STM tracking approach to assess which sEEG contacts in our patient sample could effectively track these STM features in the audio tracks that they listened to.

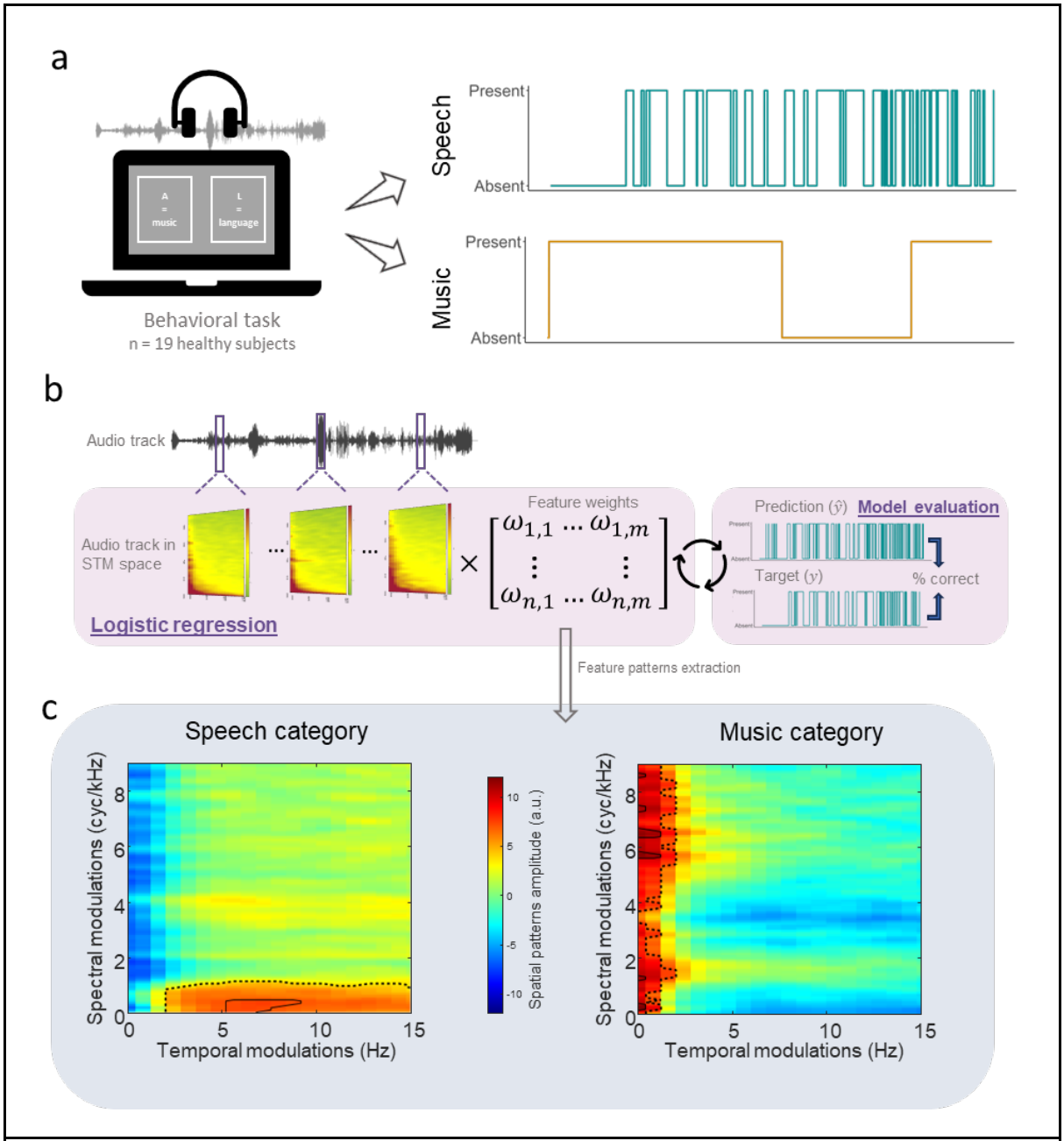

Figure 1: STM features predict perceptual categorization: pure spectral modulations for music, pure temporal modulations for speech

(a): Online behavioral task. Participants (n = 19) judged whether music and/or speech were present during the time-course of the audio track. The right panel illustrates judgement vectors obtained for speech and music categorization from one subject. Two judgement vectors were obtained for each participant.

(b): Analysis pipeline. The audio track was decomposed into the STM space using a 1.5-second sliding window, generating a STM matrix for each time sample. For each category (speech/music), a binary logistic regression model was trained on the audio track in the STM space to classify whether the corresponding category was present or absent along the time-course of the audio track. The feature weights of the model were then extracted and transformed into interpretable feature patterns using the Haufe transformation (Haufe et al., 2014).

(c): Results. Averaged feature patterns across participants in the STM space for each categorical judgement. Threshold-based contours highlight regions of high feature patterns amplitude: the 90th percentile of the distribution is outlined with a dotted line, and the 99th percentile is outlined with a solid line.

Bilateral non-primary auditory areas track the spectrotemporal modulations of a naturalistic sounds

Intracranial EEG recordings were collected from eleven patients with focal drug-resistant epilepsy as they passively listened to the audio track. After bipolar re-referencing, this resulted in a total of 846 sEEG contacts, with 51.4% located in the left hemisphere and 48.6% in the right hemisphere (Figure 2a, see supplementary Figure S1 for detail). For each sEEG contact, the magnitude of delta (0.5-3 Hz), theta (4-8 Hz), alpha (9-13 Hz), beta (14-25 Hz), low gamma (26-50 Hz), and high gamma (50-150 Hz) oscillations were extracted from the raw signal during the audio track listening, using Hilbert transform (z-score-normalized over time).

To uncover brain regions responsive to STMs, we applied 10-fold cross-validated ridge regression models that linked oscillatory magnitude across these six frequency bands, recorded from the 846 sEEG contacts, to the STM features of the audio track. For each sEEG contact, the model predicted the time course of STM amplitude for each of the 2,180 spectrotemporal combinations— corresponding to 109 spectral (0–9 cycles/kHz) × 20 temporal (0–15 Hz) modulation points extracted from the 2D STM matrices —producing an accuracy map of fisher-transformed correlation coefficients within the STM space. sEEG contacts that significantly predicted any portion of the STM space (t-tests vs 0, Bonferroni-corrected for number of STM points, sEEG contacts and cross-validation folds) were then spatially clustered in the MNI space, where clusters with overlapping contacts from at least three patients were considered significant (see Borderie et al., 2024 and STAR methods). Using this method, we identified four spatial clusters comprising a total of 39 sEEG contacts, bilaterally distributed in non-primary auditory cortical areas (see Supplementary Table S1 for MNI coordinates and atlas labels). Two clusters were located in each hemisphere. In both hemispheres, one cluster included sEEG contacts primarily positioned in the anterior portion of the superior temporal region, with 11 sEEG contacts in the left hemisphere and 13 in the right. These anterior clusters included sEEG contacts located in the medial belt, in area A5, and the anterior distal superior temporal sulcus (STS), with the right cluster additionally containing an sEEG contact in area A4 according to the HCP atlas (Glasser et al., 2016). The second cluster in the left hemisphere was located more posteriorly, encompassing sEEG contacts in the posterior distal STS and area A5. In the right hemisphere, the remaining cluster included sEEG contacts in the medial and lateral belt, and in area A5, corresponding to mid-

to-posterior portions of the non-primary auditory cortex. Each cluster included sEEG contacts spanning at least three patients, ensuring the generalizability of these findings (see Supplementary Table 1 for the MNI coordinates and atlas labels of the sEEG contacts). Overall, we observed that sEEG contacts located in non-primary auditory areas demonstrate the ability to track STM features of complex, naturalistic sounds. We then used the signal from these sEEG contacts for all subsequent analyses.

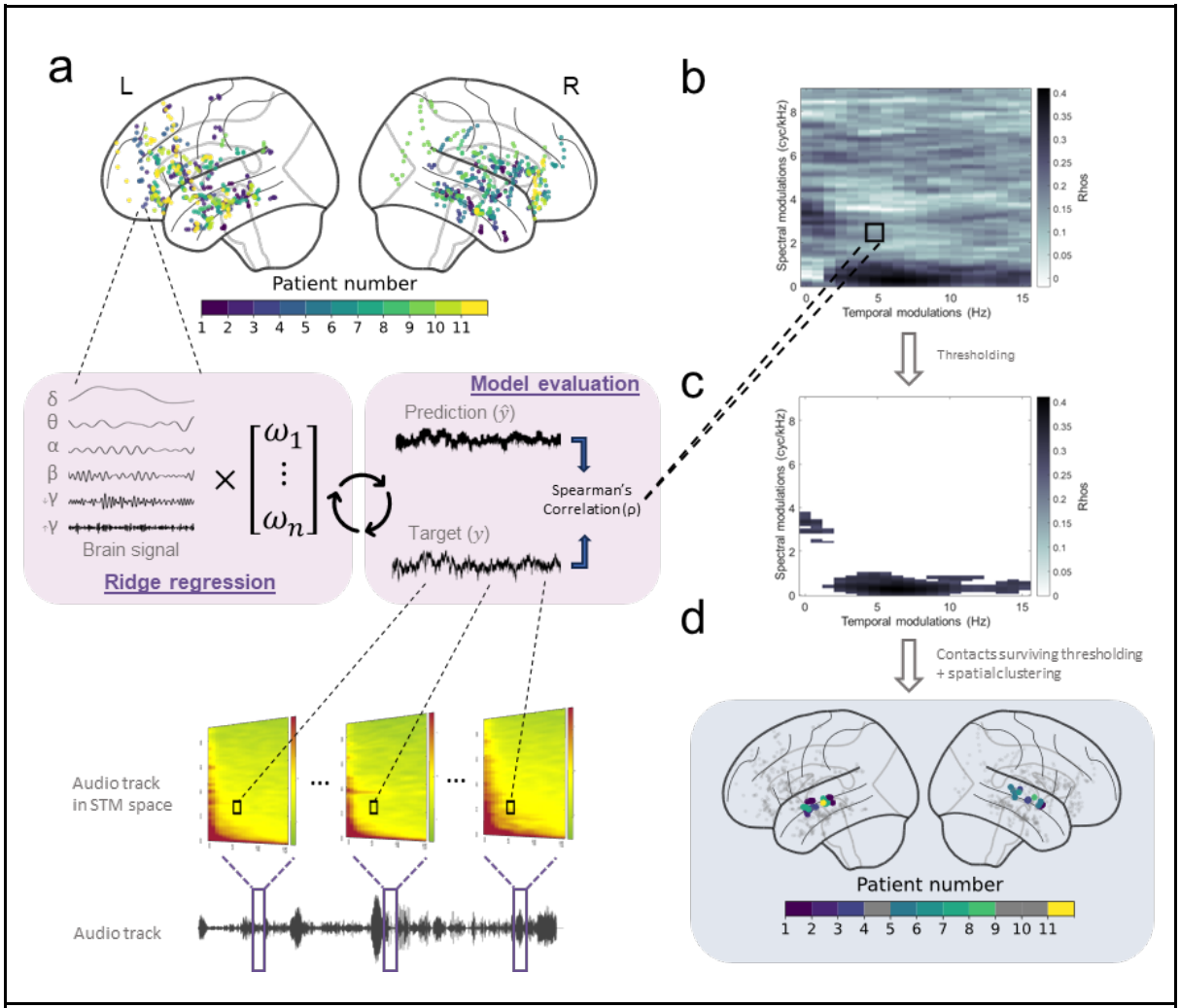

Figure 2: Bilateral non-primary auditory areas track the spectrotemporal modulations of a naturalistic sound

(a): Schematics of the regression model. Each sEEG contact (top part of the figure, showing all sEEG contacts from the eleven patients in standard MNI space on the left and right sides of a glass half-brain visualization). Colors on the scale correspond to each patient's implantation scheme. Signal was decomposed into six frequency bands and used as features in a ridge regression model (ridge regression panel: middle part of panel a). The audio track that the patients listened to (bottom part of panel a) was decomposed into the STM space as in Figure 1b. Each spectral/temporal combination time course from the matrix was then used as the target variable (y) in the regression model. Model predictions ($\hat{y}$) were evaluated (model evaluation panel) against the target (y) using Spearman's correlations.

(b): Example accuracy maps. An accuracy map was generated for each contact by running

the model (depicted in (a)) at each spectral and temporal combination/coordinate in the STM space, yielding a map of correlation coefficient (ρ).
(c): Thresholding. Each accuracy map was thresholded by performing one-tailed one-sample t-tests against zero on Fisher z-transformed correlation coefficients for each STM point (Bonferroni corrected adjusted for number of STM points, sEEG contacts and cross-validation folds, $\alpha = 1x10^{-30}$). The heatmap represents the thresholded accuracy map from (b).
(d): STM-sensitive contacts. Main panel: contacts that survived statistical thresholding and spatial clustering (see STAR Methods). Four clusters were identified in bilateral non-primary auditory areas (see supplementary Table S1 for coordinates and Figure S1 for detailed visualization).

Behavioral categorization judgments can be decoded from brain-reconstructed STMs in bilateral non-primary auditory areas.

We next address the key question of whether brain-reconstructed STMs (i.e., STMs predicted from neural signal) are sufficient to decode behavioral categorization judgments of speech and music. Our objective was to define how brain-reconstructed STMs can effectively link neural activity to behavioral judgment obtained in healthy individuals through acoustic features. At each time point, reconstructed STMs from STM-sensitive sEEG contacts (see STAR Methods) were used as features to classify participants' binary judgments of whether the track contained speech or music. Reconstructed STMs from 36 of the 39 STM-sensitive sEEG contacts significantly predicted whether the audio track contained speech (all corrected $p < .046$), and all 39 contacts significantly predicted the presence of music (all corrected $p < .001$). These findings highlight that the category (speech or music) in audio tracks can be accurately predicted from brain activity, when represented in the STM space.

To identify the specific regions of the STM space contributing to category decoding across sEEG contacts, we performed, within each contact, a statistical test against zero on feature patterns at each point in the STM space across behavioral subjects (see STAR Methods). We then counted, for each STM point, the number of contacts showing a significant feature pattern (Figure 3e). For the speech category, the STM regions with the highest number of sEEG contacts showing significant feature patterns correspond to pure temporal modulations. Specifically, the top 10% of sEEG contact counts (90th percentile) ranged from 0.6 to 15 Hz in temporal modulations and from 0 to 1 cyc/kHz in spectral modulations (dotted line in Fig 3e, left). The peak sEEG contact count (99th percentile; solid line in Fig 3e, left) was concentrated in the range of 4.8 to 8 Hz for temporal modulations and 0 to 0.25 cyc/kHz for spectral modulations. For the music category, the STM regions with the highest number of sEEG contacts showing significant feature patterns corresponded to pure spectral modulations. A high sEEG contact count in the 90th percentile was observed across most of the spectral

modulation range, except for two narrow gaps between 3.4 and 3.7 cyc/kHz (consistent with the analysis using the actual audio track) and between 4.5 and 4.8 cyc/kHz (dotted lines in Fig 3e, right). The peak sEEG contact count (99th percentile; solid lines in Fig 3e, right) was localized within spectral modulation values around 0.2, 6, 6.6, 7.7, and 8.7 cyc/kHz and temporal modulation of 0 to 0.8 Hz. These results demonstrate that brain-reconstructed STMs can accurately predict behavioral categorization even though behavioral data were derived from different samples of participants (out of sample prediction). To assess the behavioral relevance of these brain-derived features, we will next test whether the STM patterns driving categorization are similar when derived from brain data (Figure 3e) and when derived directly from the acoustics (Figure 1c).

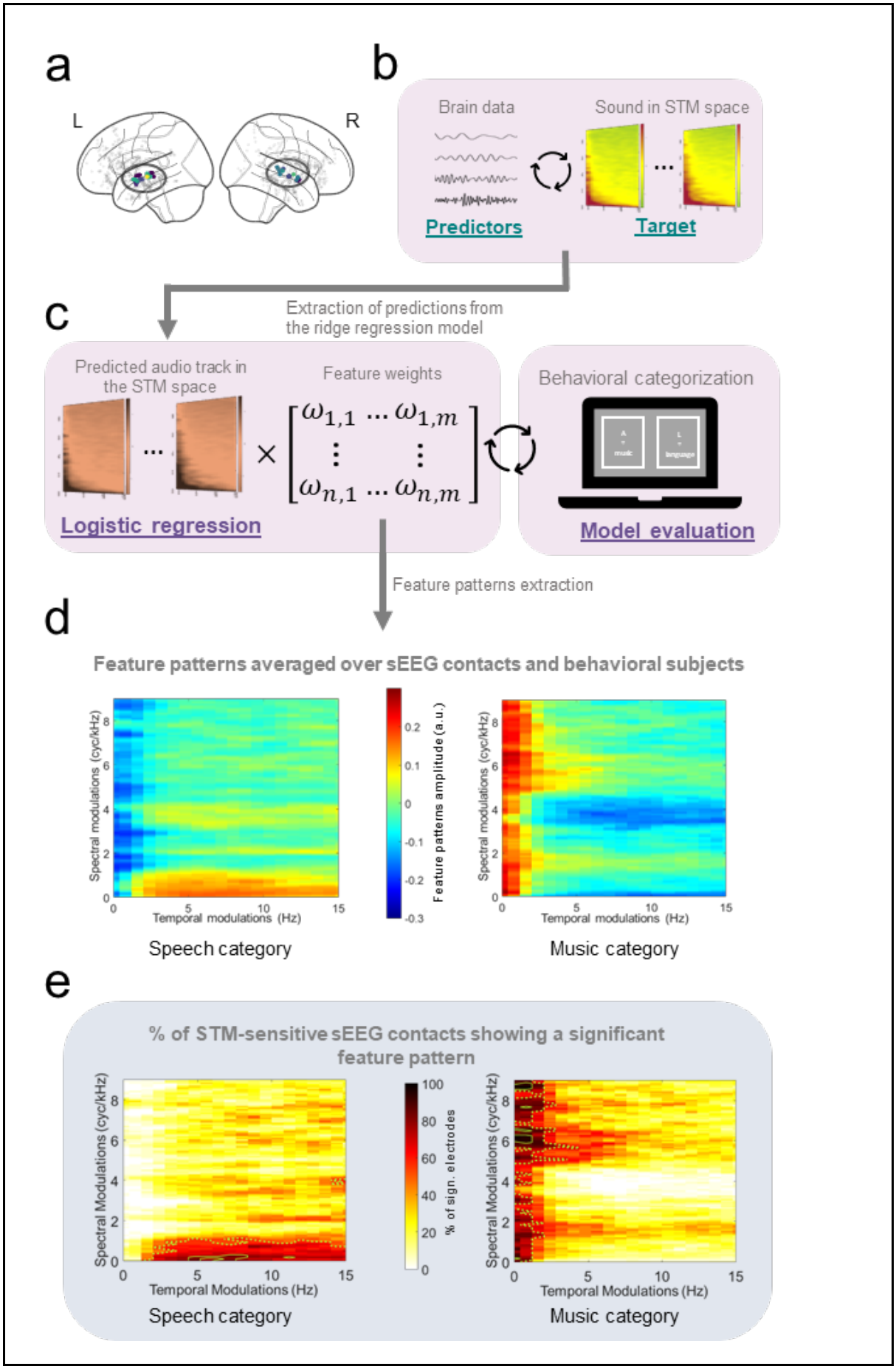
a
L
R
b
Brain data
Sound in STM space
Predictors
Target
c
Extraction of predictions from the ridge regression model
Predicted audio track in the STM space
Feature weights
Logistic regression
Behavioral categorization
Model evaluation
Feature patterns extraction
d
Feature patterns averaged over sEEG contacts and behavioral subjects
Spectral modulations (cyc/kHz)
Temporal modulations (Hz)
Feature patterns amplitude (a.u.)
Speech category
Music category
e
% of STM-sensitive sEEG contacts showing a significant feature pattern
Spectral Modulations (cyc/kHz)
Temporal Modulations (Hz)
% of sign. electrodes
Speech category
Music category

Figure 3: Behavioral categorization judgments can be decoded from brain-reconstructed STMs in bilateral non-primary auditory areas.
(a) STM-sensitive sEEG contacts (same as Figure 2d).
(b) Schematics of the regression model. The model used brain signals as predictors and the time courses of each spectral/temporal combination of the audio track as targets, as depicted in Figure 2a. The predicted time courses for each spectral/temporal combination were extracted, providing the audio track in the STM space predicted from brain signals.
(c) Schematics of the logistic regression model. Based on the predicted audio track in the STM space obtained in (b), binary logistic regression models were trained to classify whether music or speech was present or absent. The feature weights of the model were then extracted and transformed into feature patterns using the Haufe transformation (Haufe et al., 2014). Feature patterns were obtained for each STM-sensitive sEEG contact, behavioral participant and category.
(d) Feature patterns in the STM space averaged over all sEEG contacts and all behavioral subjects.
(e) Heatmaps showing, for each category, the percentage of STM-sensitive sEEG contacts that exhibited a significant feature pattern across behavioral subjects at each spectral and temporal combination/coordinate in the STM space. Threshold-based contours highlight regions of high sEEG contact count with significant feature patterns: the 90th percentile of the distribution is outlined with a dotted line, and the 99th percentile is outlined with a solid line.

Brain-reconstructed and acoustic STM representations are highly similar

To assess whether the brain-reconstructed STM representations truly reflect the acoustic features relevant for perceptual categorization, we evaluated the similarity between the feature patterns (Figure 4a) of: (1) the STM of the actual audio track (Fig 1c), and (2) the brain-reconstructed STM (Fig 3d) averaged across hemispheres. For each category and each hemisphere, nineteen comparisons were performed between STM patterns derived from brain-reconstructed STMs and from acoustic STMs, both based on the prediction of each behavioral judgment vector (n = 19 behavioral participants). STM patterns were highly similar, showing significant correlation coefficients (Spearman correlation; $p < .001$, Bonferroni-corrected adjusted for number of subjects, categories, and hemispheres) for all comparisons, confirming the strong alignment between the STM features of the audio and those derived from brain activity (Figure 4b). We confirmed this result at the group-level, by comparing the distribution of correlation coefficients across behavioral participants to shuffled data, (see STAR Methods; Figure 4b) revealing highly significant differences for each category (both $p = 2.9 \times 10^{-22}$). These key results demonstrate that the STM patterns in behaviorally relevant acoustic features are represented in a continuous and sufficiently similar manner in the brain in non-primary auditory areas. Next, we examined whether the representation of these STM

patterns exhibits hemispheric asymmetries that have been previously associated with differential STM patterns processing.

Similarity between brain-reconstructed and acoustic STM representations reveals hemispheric lateralization

Finally, we tested whether the similarity between the two representations reflected the previously described hemispheric specialization of the auditory cortices (Figure 4c), with the left hemisphere being more sensitive to temporal modulations and the right to spectral modulations (Albouy et al., 2020; Flinker et al., 2019; Zatorre & Belin, 2001). To do so, we carried out a 2x2 aligned rank transform test on the correlation coefficients used in the previous similarity analysis, with category (speech/music) and hemisphere (left/right) as factors. The main effect of category was significant ($F(1, 55.65) = 72.20$, $p < .001$), with higher correlation coefficients for the music category compared to the speech category. This result was expected, as models involving speech categorization consistently performed worse than those for music, likely due to the more fragmented and less continuous nature of speech categorization vectors, reflecting the conversational structure of the audio tracks (Figure 1a, right panel). No main effect of hemisphere was observed ($F(1, 55.65) = 0.68$, $p = 0.412$), indicating that both hemispheres similarly encode auditory categorical information. Finally, the category-by-hemisphere interaction was significant ($F(1, 55.65) = 23.64$, $p < .001$). Post-hoc analysis showed that for the speech category, correlation coefficients were significantly higher in the left hemisphere (FDR-corrected $p = .027$) compared to the right, whereas for the music category, correlation coefficients were significantly higher in the right hemisphere (FDR-corrected $p < .001$) compared to the left. Additional post-hoc analyses of music vs speech comparisons in each hemisphere revealed significantly lower correlation coefficients for speech than for music in both the left and right hemispheres (both FDR-corrected $p < .001$), consistent with the previously noted lower accuracy for models involving speech categorization. Overall, we found that the high similarity between feature patterns derived from acoustical-STM and brain-reconstructed STM showed hemispheric asymmetry with greater similarity in the left hemisphere for speech categorization and in the right hemisphere for music categorization.

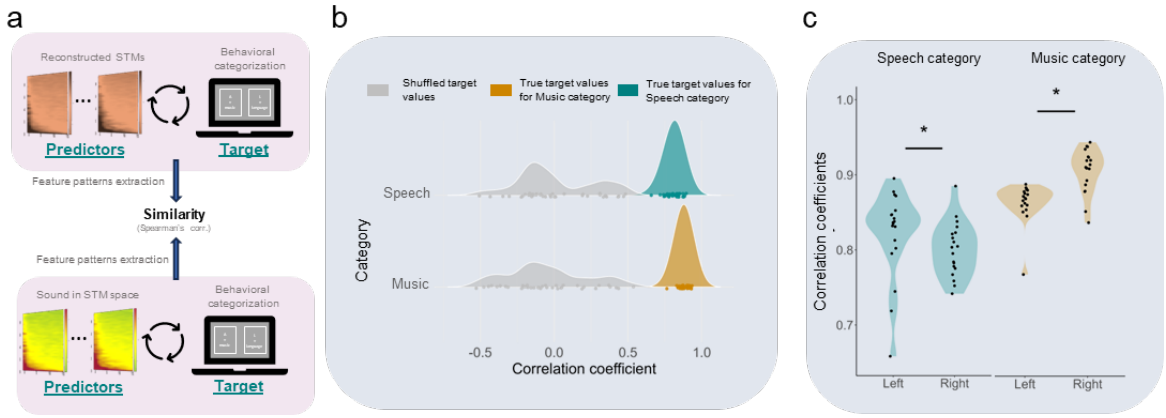

Figure 4: Similarity between brain-reconstructed and acoustic STM representations reveals hemispheric lateralization

(a) Schematic of the analysis pipeline. Feature patterns were extracted from the logistic regression models described in Figure 3c, where brain-reconstructed STMs served as predictors and behavioral categorization responses as target variables (top panel). This was done separately for each STM-sensitive sEEG contact, category, and behavioral participant, and the resulting feature patterns were then averaged across sEEG contacts within each hemisphere. In parallel, feature patterns were obtained from the logistic regression models described in Figure 1b, where the actual acoustic STMs were used as predictors of behavioral categorization (bottom panel), again for each category and behavioral participant. To evaluate similarity between brain-reconstructed and acoustic feature representations, Spearman correlations were computed between the two sets of feature patterns. For each category and each hemisphere, this resulted in 19 comparisons (one per behavioral participant) between the hemisphere-averaged brain-reconstructed STM patterns and the corresponding acoustic STM patterns.

(b) Feature pattern similarity distribution. Density curves represent the distribution of correlation coefficients between feature patterns derived from the two models depicted in (a). The top (aqua) distribution corresponds to the speech category, and the bottom (orange) distribution corresponds to the music category. For comparison, a separate distribution (gray) shows the correlation coefficients obtained when feature patterns from the first model (top panel of (a)) were compared with those from the second model after shuffling the behavioral categorization labels. All Spearman's correlation coefficients calculated from the two original models were significant at $p < .001$ after Bonferroni correction adjusted for number of subjects, categories, and hemispheres.

(c) Similarity lateralization. Correlation coefficients are displayed for each behavioral participant in the speech category (left plot) and the music category (right plot), with data from sEEG contacts implanted in the left and right hemisphere (x-axis). *$p < .01$, with pairwise Wilcoxon signed-rank tests, FDR-corrected.

DISCUSSION

In the present study, we investigated how the human brain encodes and categorizes naturalistic speech and music, integrating continuous behavioral categorization judgments

from healthy participants with intracranial recordings from the auditory cortex of epileptic patients. We specifically tested whether perceptual categories depend on the encoding of acoustical features, as predicted by a domain-general model, or whether such features are insufficient for categorization, as predicted by a domain-specific model. Our results provide converging evidence favoring the domain-general hypothesis, demonstrating that spectro-temporal modulation (STM) features of naturalistic sounds are sufficient for speech and music categorization to emerge from non-primary auditory cortical activity. Specifically, we found that (i) distinct STM patterns are associated with the perception of naturalistic speech and music categories in healthy individuals (Figure 1c), validating our analysis approach to naturalistic auditory input; (ii) STM features are robustly tracked in non-primary auditory areas (Figure 2d); (iii) STMs reconstructed from brain signals in these regions are sufficient to predict behavioral categorization judgments from a separate group of participants (Figure 3e), showing that brain-derived representations align closely with perceptual decisions; (iv) STM patterns from the brain-reconstructed STMs are highly similar to those obtained from the actual audio track (Figure 4b), and (v), this effect is lateralized (Figure 4c) with stronger acoustic/brain similarity in the right hemisphere for music and in the left hemisphere for speech, consistent with well-known hemispheric asymmetries in auditory processing. We next discuss each of these points in turn.

Distinct STM patterns are associated with the perception of naturalistic speech and music categories

Using a time-resolved tracking approach, we found that human judgments of perceiving speech and/or music at any given moment within a soundtrack can be well differentiated based on their STM content, with speech associated with high temporal (peak 5.6–8.8 Hz) and low spectral modulations (0–1 cyc/kHz), and music with high spectral (peaks between 5.7 and 8.6 cyc/kHz) and low temporal modulations (0–1 Hz , Figure 1c). This finding is consistent with prior literature demonstrating that speech and music rely on distinct STM features, which not only support their perceptual differentiation but also provide an optimal representational framework for capturing their complex structures. Specifically, temporal modulations in the range of 4–7 Hz are known to be critical for speech intelligibility, as this range aligns with the syllabic rate of spoken language (Albouy et al., 2024; Ding et al., 2017; Elliott & Theunissen, 2009; Giroud et al., 2023, 2024; R. V. Shannon et al., 1995; Varnet et al., 2017). In the spectral domain, speech recognition has been shown to rely primarily on low spectral modulations (Albouy et al., 2024), with critical modulations for speech comprehension typically falling below 4 cyc/kHz (Elliott & Theunissen, 2009). In contrast, music perception operates on a different timescale, with key temporal modulations typically falling below 2 Hz (Chang et al., 2024), reflecting the slower amplitude modulations of musical phrases across genres (Ding et al.,

2017) and cultures (Albouy et al., 2024). Regarding spectral modulations, previous studies have shown that for songs that contain only vocal elements, important spectral modulations range from approximately 3 to 8 cyc/kHz (Albouy et al., 2020, 2024). Our present results and previous literature thus confirm that STM representations can effectively capture the statistical properties of speech and music that are most relevant for perceptual categorization. These observations raise the question of whether such behaviorally relevant acoustic representations are also reflected in the tuning properties of the human auditory cortex.

The auditory cortex robustly tracks STMs

Our second finding demonstrated that the auditory cortex is tuned to track the STM regularities present in natural speech and music. We identified sEEG contacts that significantly tracked STM patterns, using a robust ridge regression approach combined with the temporal response function method (Figure 2a). We demonstrated that STM patterns are dynamically tracked in non-primary auditory regions (Figure 2d). Specifically, STM tracking was observed in the left hemisphere within the medial belt, in area A5, and in both anterior and posterior portions of the distal superior temporal sulcus (STS); in the right hemisphere, tracking was found in the medial and lateral belt regions, area A5, and the anterior distal STS. No such tracking was observed in sEEG contacts outside these auditory regions. Importantly, none of our participants had sEEG contact coverage over the primary auditory cortex, a common limitation of intracranial studies, so we cannot exclude the possibility that STM tracking may also occur in primary or adjacent auditory areas.

Our findings align with a broad body of evidence from both animal and human studies showing that STM sensitivity is a fundamental property of the auditory system. In animal models, single-unit recordings across species such as zebra finches, ferrets, cats, and monkeys have consistently revealed tuning to STM throughout the auditory pathway. Tuning to STM has been reported from the inferior colliculus and medial geniculate body to primary and secondary auditory cortex (primarily using methods that estimate spectrotemporal receptive fields (STRFs) of individual neurons) in response to controlled synthetic ripple stimuli as well as more ecologically valid sounds like birdsong and vocalizations (Calhoun & Schreiner, 1998; Depireux et al., 2001; Eggermont, 2002; Escabí et al., 2003; Escabí & Schreiner, 2002; Hsu et al., 2004; Kowalski et al., 1996; Massoudi et al., 2015; Miller et al., 2002; Rodríguez et al., 2010; Theunissen et al., 2000; Woolley et al., 2005). Similarly, human neuroimaging research has confirmed STM sensitivity across primary, secondary, and even higher-order auditory regions. This has been demonstrated through multiple approaches: spectrotemporal modulation transfer functions measured with synthetic ripple stimuli (Schönwiesner & Zatorre, 2009), STRFs estimated with naturalistic sounds (Venezia et al., 2019, 2021), classic activation-based fMRI mapping using dynamic ripple stimuli (Langers et al., 2003), and

encoding/decoding models that explicitly linked neural responses to the STM content of speech, music, vocalizations, and environmental sounds (Santoro et al., 2014).

Further evidence for STM tuning has been shown recently in three human studies conducted within the STM framework. These studies demonstrated that degrading the STM of speech and tonal sounds leads to reduced cortical discriminability, with effects localized in the same regions as our STM-sensitive contacts. Albouy et al. (2020) showed that STM degradation of sentences and melodies impaired category classification in left A4 and medial belt for temporal modulations, and right medial belt for spectral ones. Robert et al. (2024) used synthetic sounds mimicking simple actions (containing mainly temporal modulations) and object properties (containing mainly spectral modulations) and found that STM degradation disrupted classification in left A4/A5 and IFG for actions and right parabelt, A4/A5 and pSTS for objects. Flinker et al. (2019), using electrocorticography and degraded speech, reported cortical tracking of STM features in left A5 (temporal) and right medial belt (spectral) for speech- and vocal timbre-related features respectively. Across these studies, despite differences in stimuli and methods, sensitivity to STM structure consistently emerged in areas overlapping with the STM-sensitive contacts identified in the present study.

In studies adopting a domain-specific perspective, Norman-Haignere et al. (2015, 2022) proposed distinct neural components for speech and music categories and distinct from other components sensitive to acoustic features based on fMRI voxel decomposition method. Yet, the topographies they report largely overlap with the STM-sensitive regions identified in the present and previous studies. In their fMRI work (Norman-Haignere et al., 2015), independent components tuned to STMs were located near the borders of primary auditory cortex, both anteriorly and posteriorly, while speech- and music-selective category components emerged in lateral STG, planum polare, and planum temporale. Similarly, in their electrocorticography (ECOG) study (Norman-Haignere et al., 2022), they reported STM sensitivity in both primary auditory cortex and more anterior non-primary auditory areas and revealed speech- and music-selective components in middle and anterior STG. Thus, the findings across studies converge regarding both the existence and the location of higher-order representations for speech and music, a point we shall return to in the final discussion.

Brain-reconstructed STMs are sufficient to decode behavioral categorization judgments

Having established that STMs provide an efficient shared acoustic representation for distinguishing speech and music, and that the auditory cortex is finely tuned to these features, we reconstructed the STM features of the audio track from intracranial signals recorded at STM-sensitive sEEG contacts (Figure 3a-c) and found that these predicted STMs were highly

effective in decoding continuous categorization judgments. This crucial result shows that behavioral categorical decisions can be predicted from brain-reconstructed acoustic representations. In addition, a closer look at the feature patterns of the brain-reconstructed STMs that were most predictive of behavioral categorization (Figure 3d) revealed a match with those identified using the STM patterns derived directly from the real audio track (Figure 3e). This observation critically suggests a functional convergence between acoustically and neurally derived features. The next important step was therefore to formally test the similarity between these two sets of STM patterns to determine whether the same representational structure underlies both acoustically and neurally derived data.

High similarity between brain-reconstructed and acoustically-derived STMs

A similarity analysis (Figure 4a) revealed a strong correspondence between the acoustically and neurally derived STM patterns that are critical for successful behavioral categorization (Figure 4b). Notably, because our neural and behavioral data were collected from separate participant groups, the observed neural–perceptual mapping reflects a robust out-of-sample prediction, a rigorous validation approach in cognitive neuroscience (Gabrieli et al., 2015; Scheinost et al., 2019; Walters et al., 2022), indicating that the underlying coding principles generalize across individuals rather than overfitting to subject-specific variability or noise. These findings offer our most compelling demonstration that brain-reconstructed STM representations alone are sufficient to accurately predict behavioral categorization judgments of speech and music, establishing a direct link between neural encoding of STM features in the auditory cortex and high-order perceptual outcomes. The final analytical step focused on whether this acoustic-neural coupling reflects meaningful organizational principles of auditory processing, by testing for hemispheric specialization.

Similarity between brain-reconstructed and acoustic STM representations reveals hemispheric lateralization

Hemispheric asymmetries (favoring the left auditory cortex for speech and the right for music) have been robustly documented over decades of research, including lesion studies, electrophysiological recordings, and neuroimaging investigations in both neurotypical individuals and those with neurodevelopmental disorders such as amusia (Albouy et al., 2013; Johnsrude et al., 2000; Loui et al., 2009; Milner, 1962; Okamoto & Kakigi, 2015; Samson & Zatorre, 1988; Sihvonen et al., 2019; Zatorre, 2022). Our data revealed a hemispheric lateralization in the similarity between brain-reconstructed STM patterns (extracted from STM-sensitive sEEG contacts) and the STM patterns of the actual audio stimuli, when both were used to predict perceptual categorization outcomes (Figure 4c). Specifically, speech categorization showed higher similarity in sEEG contacts implanted in the left-hemisphere,

while music categorization showed higher similarity in sEEG contacts implanted in the right-hemisphere. In other words, the observed hemispheric lateralization of the similarity patterns between brain-reconstructed and stimulus-derived STMs reveals that left auditory regions are more sensitive to the pure temporal modulations characteristic of speech, while right auditory regions are more sensitive to the pure spectral modulations typical of music. Our results thus show that the brain leverages the statistical structure of STM in a category- and hemisphere-specific manner to support auditory categorization.

In line with our results, hemispheric specialization for STM processing has been well-documented: the left auditory cortex operates at faster temporal integration windows than the right (Boemio et al., 2005; Giroud et al., 2020; Luo & Poeppel, 2012; Teng et al., 2016, 2017), making it better suited for tracking rapid temporal modulations, such as those prominent in speech. Conversely, the right auditory cortex exhibits sharper spectral tuning and heightened sensitivity to pitch-related information (Cha et al., 2016; Coffey et al., 2016; Hyde et al., 2008; Jamison et al., 2006; Zatorre, 2022; Zatorre & Belin, 2001), aligning with its preferential processing of fine spectral cues relevant for music. More recently, this hypothesis has been directly tested using the STM framework across fMRI, MEG, and sEEG/ECOG, showing that degradation of speech, songs, or environmental sounds in the temporal or spectral modulation domains selectively impact brain activity in the left STG for temporal degradation and in the right STG for spectral degradation (Albouy et al., 2020; Flinker et al., 2019; Robert et al., 2024). The present findings and recent literature reinforce the view that auditory perception relies on a shared STM encoding framework, with category-related hemispheric biases emerging as a consequence of distinct sensitivity in the processing of temporal and spectral modulations rather than as a function of intrinsic domain specificity.

Efficient neural coding in the auditory system

Our results can be understood within the broader efficient neural coding framework which posits that perceptual systems have evolved to efficiently encode behaviorally relevant environmental stimuli for any given species (Attneave, 1954; Barlow, 1961). Computationally, efficient coding can be achieved if neural representations match the statistical regularities of the signals they represent (C. E. Shannon, 1948). While this principle has been extensively documented in the visual domain (Graham & Field, 2007; Simoncelli & Olshausen, 2001), it has also been increasingly supported in the auditory modality (Gervain & Geffen, 2019; Guevara Erra & Gervain, 2016; Lewicki, 2002; Ming & Holt, 2009; Rieke et al., 1995; Smith & Lewicki, 2006). Among various ways to represent natural sounds, STMs are particularly well-suited for capturing the statistical structure of complex sounds (Figure 1c), including vocalizations, as they reflect joint dependencies between temporal and spectral modulations that are characteristic of natural stimuli, dependencies that are not captured by the raw

spectrum or spectrogram alone (Chi et al., 2005; Shamma, 2001; Singh & Theunissen, 2003; Theunissen & Elie, 2014).

At the neural level, STMs have been shown to capture response properties of auditory neurons that are not accounted for by traditional frequency tuning curves across multiple species and auditory regions, including the midbrain, thalamus, and both primary and secondary auditory cortices in birds and non-human mammals (Depireux et al., 2001; Hsu et al., 2004; Massoudi et al., 2015; Theunissen et al., 2000; Woolley et al., 2005). In humans, such specific tuning to STMs in the auditory cortex, including non-primary auditory regions, has been demonstrated in neuroimaging studies using synthetic stimuli like dynamic ripples (Langers et al., 2003; Schönwiesner & Zatorre, 2009) and natural stimuli such as speech (Santoro et al., 2014; Venezia et al., 2019, 2021), as well as in an intracranial recording study using speech stimuli (Hullett et al., 2016). In addition, speech signals can be reconstructed from patterns of activity in the STG using STM-based STRF models (Pasley et al., 2012). Accordingly, under the assumption of efficient neural coding and given the known tuning of central auditory neurons to STMs, we propose that category-specific neural representations in the auditory cortex, such as those distinguishing speech from music that we have studied here, can be largely explained by differential sensitivity to STM features.

While we argue that sensitivity to STM features provides a domain-general and low-dimensional framework for explaining auditory cortical representations of complex sounds, we do not dispute the likely existence of subsequent non-linear transformations that may support higher-level categorical processing. Indeed, abstract categorical representations must exist to account for humans' ability to effortlessly categorize complex sounds. A critical consideration in this context is the hierarchical nature of auditory cortical organization (cf. Kell et al., 2018), as supported by extensive evidence in both human and non-human primates (Kaas & Hackett, 1999; Okada et al., 2010; Patterson et al., 2002; Rauschecker & Scott, 2009), paralleling the functional architecture observed in the visual system (Goodale & Milner, 1992; Maunsell & Newsome, 1987; Vogels & Orban, 1996). However, our findings, consistent with a large body of literature demonstrating STM tuning throughout the auditory cortex, show in a direct manner that neural representations of continuous, naturalistic sounds in auditory areas are fundamentally structured by sensitivity to STM features. These representations are sufficient to support category-level distinctions, such as between speech and music, without the need to invoke categorical abstraction at this processing stage.

This interpretation does not preclude additional, potentially non-linear transformations at later stages of the auditory object recognition hierarchy. Rather, it highlights that the input to such stages already carries rich, domain-relevant information structured by the statistical properties of natural sounds. From this perspective, category-selective responses observed in non-

primary auditory cortex through non-linear voxel decomposition methods (e.g., Norman-Haignere et al., 2015, 2022) may reflect higher-order transformations that build on, rather than replace, STM-based representations. Crucially, the brain regions identified in those studies as selectively responsive to speech or music largely overlap with the STM-sensitive regions revealed in our present work and in previous reports, with the additional property that they are lateralized. This overlap suggests that the emergence of categorical selectivity in these areas may be a consequence of their differential sensitivity to the statistical structure of specific sound categories, as captured by STMs. In this view, the observed selectivity is not at odds with a domain-general encoding scheme but rather emerges naturally from it.

Conclusions and future perspectives

Speech and music can arguably be considered the two most important and species-characteristic human auditory communication signals (Mehr et al., 2021). Our findings that neural STM representations efficiently capture the statistical properties of naturalistic speech and music thus can be viewed in the context of functional adaptation of the human brain to these important classes of sounds. Importantly, our study bridges a longstanding theoretical debate by showing that domain-general mechanisms, grounded in the efficient neural encoding of STMs, can provide a biologically plausible framework that fully supports category-level distinctions, both at the neural and behavioral levels. While our findings do not exclude domain-specific mechanisms at later stages of processing in higher-order associative brain regions, we argue that the foundation for these distinctions lies in the acoustic domain.

Finally, it is important to recall that the auditory processing hierarchy does not start at the cortical level. Subcortical structures such as the inferior colliculus and the medial geniculate body exhibit tuning to spectrotemporal features in animal models (e.g., Woolley et al., 2005, Escabi & Schreiner, 2002). Investigating these subcortical contributions in humans could provide valuable insights into the early stages of hierarchical auditory encoding and help refine our understanding of how STM representations emerge and propagate along the auditory pathway.

RESOURCE AVAILABILITY

Lead contact

For further information or resource requests, please contact the lead investigator, Jérémie Ginzburg (jeremie.ginzburg@mcgill.ca).

Materials availability

This study did not generate new materials.

Data and code availability

De-identified and pre-processed patients and behavioral data have been deposited at https://doi.org/10.17605/OSF.IO/FT289. They are publicly available as of the date of publication. Note that raw SEEG and neuroimaging (T1-MPRAGE) data are protected and cannot be shared (2022-5890).

All original code has been deposited on OSF at https://doi.org/10.17605/OSF.IO/FT289 and is publicly available as of the date of publication.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGEMENTS

This work was supported via a project grant (486895) from the Canadian Institutes of Health Research to R.J.Z. and P.A., by the Fonds de Recherche du Québec via funding to the Center for Research in Brain, Language and Music (RSMA-340954), and by an NSERC Discovery grant to P.A. (RGPIN-2019-06162). R.J.Z. is funded via the Canada Research Chair program, and by the Scientific Grand Prize (FPA RD-2021-6) from the Fondation pour l'Audition (Paris, France). P.A. is supported by FRQS Junior 2 (329968) grant, and a Brain Canada Future leaders Grant (https://braincanada.ca/). B.M. is supported by the European Union (ERC, SPEEDY, ERC-CoG-101043344), Fondation Pour l'Audition (FPA RD-2022-09), the ANR (ANR-20-CE28-0007; ANR-16-CONV-0002) and the Excellence Initiative of Aix-Marseille University (A*MIDEX).

We thank the patients for their participation, and Romain Quentin and Oscar Bedford for their valuable discussions and insights.

AUTHOR CONTRIBUTIONS

JG: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. ECD: Data curation, Formal analysis, Software; AB: Data curation, Software; BM: Writing – review & editing. LM, PLB: Data curation; RJZ: Conceptualization, Funding acquisition, Project administration, Resources, Investigation, Methodology, Supervision, Writing – review & editing, PA: Conceptualization, Data curation, Funding acquisition, Project administration, Resources, Formal analysis, Investigation, Methodology, Validation, Supervision, Writing – review & editing. The manuscript's contents have been approved by all the co-authors, and they have provided consent for its publication.

DECLARATION OF INTERESTS

The authors declare no competing interests.

SUPPLEMENTAL INFORMATION

Document S1: Table S1, Figure S1.

STAR METHODS

Experimental model and subject details

Patients

sEEG data from 11 patients suffering from drug-resistant focal epilepsy were collected (self-reported biological sex, 4 females, mean age = 29.6 years, SD = 3.2 years) following the Research Opportunities in Humans Consortium guidelines (Feinsinger et al., 2022). All patients were native French speakers from Canada and they all completed the task as instructed. The study protocol was explained verbally by both the research and clinical teams, and participation was discussed with each patient prior to hospitalization. All patients provided verbal assent and signed written informed consent forms. This research complies with all relevant ethical regulations; the experimental procedures were approved by the Ethics Review Board of the CHU de Québec - Université Laval (2022-5890, Québec, QC, Canada).

Healthy Participants

20 healthy adults participated in the behavioral experiment online. No statistical method was used to predetermine sample size. All participants were native French speakers from France or Canada (self-reported biological sex, 11 females, mean age = 36.4 years, SD = 14.7 years). Participants reported normal hearing and no history of neurological or psychiatric disease. Data from one participant was excluded due to a problem with data collection on the online server. All participants provided written informed consent, and the experimental procedures were approved by the Ethics Review Board of the CIUSSS de la Capitale Nationale (2022–2476). The study has been conducted according to the principles expressed in the Declaration of Helsinki.

Method details

Stimuli and procedure

Stimuli

The stimuli used in both the sEEG and behavioral experiments were identical and consisted of the original audio track from the first 7 minutes and 50 seconds of the French-dubbed version of *Harry Potter and the Philosopher's Stone* (titled *Harry Potter à l'école des sorciers* in French). This audio track included the complete soundtrack from the movie, including speech, music, and environmental sounds as they appear in the cinematic experience.

Procedure: sEEG experiment

sEEG patients, who had electrodes implanted based on clinical requirements, were individually tested in a dedicated sEEG monitoring room at the *Hôpital de l'Enfant Jésus*, CHU de Québec (QC, Canada). sEEG data were recorded continuously throughout the session. Stimuli were presented using Presentation software (Neurobehavioral Systems, Albany, CA, USA). Patients were given the option to use their own earphones or those provided by the research team. After ensuring a comfortable sound level with a separate audio file, patients were instructed to listen attentively to the entire audio track.

Procedure: Behavioral experiment

Participants completed a judgment task to determine whether music or speech was present in an audio track. The experiment was designed using PsychoPy (Peirce et al., 2019) and hosted on Pavlovia for online data collection (available here https://pavlovia.org/palbouy/hp_categories_task_2). Upon accessing the experiment, participants provided demographic information, including age and biological sex. They were instructed to ensure they would be undisturbed for at least 15 minutes, wear personal headphones, and adjust the volume to a comfortable level using an audio excerpt. The task instructions explained that participants were to press the "Q" key whenever they heard music and the "L" key whenever they heard speech, holding the key as long as the respective category was present and releasing it when it was absent. A visual feedback screen displayed two rectangles labeled "Q = Music" and "L = Language" (see Figure 1a) which turned red when the corresponding key was pressed and reverted to white when released. Prior to starting, participants were asked to try pressing both keys to familiarize themselves with the feedback. They then completed a 2-minute training session using an unrelated audio track with both music and conversation before proceeding to the main task, which involved the 7-minute-and-50-second experimental audio stimuli described earlier.

Extraction of spectro-temporal modulations

The acoustic signal was analyzed over its time course using the STM framework (see Albouy et al., 2020, 2024; Elliott & Theunissen, 2009 for similar procedures). STM patterns (i.e., modulograms) were extracted to obtain 512 STM per second, a rate corresponding to the sEEG sampling frequency using the ModFilter algorithm (Elliott & Theunissen, 2009) with a 1.5-second window starting from the current time sample (see Figure 1b). For each window, this method applies a 2D fast Fourier transform to the autocorrelation matrix of the sound stimulus in its spectrographic representation. This process generates, at every time sample, a 2D matrix capturing the power of the audio signal across spectral and temporal combinations. The STM pattern extraction was performed separately for each channel of the

stereo audio track, and the resulting patterns were averaged together in the STM space to obtain a unified representation of the audio stimulus. Because speech and music STM patterns are typically symmetric between positive and negative temporal modulations (Elliott & Theunissen, 2009) and to reduce data size for faster computation, the negative temporal modulation values were averaged with the corresponding positive values. This resulted in STM patterns at each time sample, with temporal modulations ranging from 0 to 15 Hz (20 points) and spectral modulations spanning 0 to 9 cycles/kHz (109 points), corresponding to a total of 2180 spectrotemporal combinations.

Behavioral data

For each participant and each category (speech/music), a binary judgement vector was created, with 1 indicating that the participant perceived the category in the audio track, and 0 indicating its absence.

sEEG recording and preprocessing

sEEG contacts localisation

Patients were stereotactically implanted with 16 to 18 semi-rigid multi-lead electrodes (AdTech; diameter: 0.8 mm). Brain activity was recorded from sEEG contacts, which were 2.0 mm wide and spaced 3, 4, 5, or 6 mm apart, targeting clinically defined brain structures. A 3D T1-weighted anatomical MRI (MPRAGE, 3T Siemens Trio, Siemens AG, Erlangen, Germany) was acquired pre-implantation, consisting of 160 sagittal slices with 1 mm^3 isotropic voxels covering the entire brain. Cortical surfaces were reconstructed from the T1-weighted MRI, and the precise locations of the sEEG contact were identified using a post-implantation CT scan co-registered to the pre-implantation MRI (ImaGIN toolbox, https://f-tract.eu/software/imagin/). The MNI coordinates of the sEEG contacts were determined using the SPM toolbox (http://www.fil.ion.ucl.ac.uk/spm/). Finally, atlas labels were extracted from the extended Human Connectome Project (HCPex) multimodal parcellation atlas (Huang et al., 2022), a modified and expanded version of the HCP-MMP v1.0 atlas (Glasser et al., 2016).

Recording

Intracranial brain activity was recorded simultaneously from sEEG contacts (sampling rate: 512 Hz) using a Natus video-sEEG monitoring system. Online referencing was performed using a contact located in white matter. During acquisition, signals were bandpass filtered between 0.1 and 200 Hz. sEEG contacts contaminated by pathological epileptic activity or environmental artifacts were rejected in collaboration with the clinical team

Preprocessing

The Brainstorm software with default parameters (Tadel et al., 2011) was used to preprocess and Hilbert transform sEEG data.

A notch filter was applied to reduce powerline interference (60 Hz and its harmonics, 120 and 160 Hz). The sEEG signal was segmented from -10 seconds to 490 seconds relative to the audio track onset, (-10 to 0 seconds baseline period). Each sEEG signal was subsequently re-referenced to its immediate neighbor using bipolar derivations. This bipolar montage enhances local specificity by removing artifacts common to adjacent contacts, such as powerline noise (e.g., 60 Hz) and distant physiological artifacts, while also reducing the effects of distant sources propagating through volume conduction. The resulting spatial resolution of each sEEG bipolar contact is estimated at 3–4 mm (Jerbi et al., 2009). Brain activity was recorded from a total of 846 bipolar contacts (ranging from 47 to 88 across patients).

Hilbert transform

To investigate the relationship between oscillatory rhythm fluctuations and STM in the audio track, we applied the Hilbert transform to extract the amplitude envelope (i.e., magnitude) of oscillatory rhythms across the entire neurophysiological range. This analysis was performed for each sEEG contact during the audio track listening period. Oscillatory rhythms were segmented into the following frequency bands: delta (0.5–3 Hz), theta (4–8 Hz), alpha (9–13 Hz), beta (14–25 Hz), low gamma (26–50 Hz), and high gamma (50–150 Hz).

Quantification and statistical analysis

The logistic regression analyses were conducted with the *MNE-Python* (Gramfort et al., 2013) and *Scikit-learn* (Pedregosa et al., 2011) libraries.

The ridge regression analyses were conducted with the *spyeeg* library (Guilleminot et al., 2021; latest version: https://github.com/phg17/sPyEEG).

Statistical analyses were conducted with MATLAB Version: 9.12.0 (R2022a).

Spatial clustering was conducted with custom-made MATLAB scripts (https://doi.org/10.17605/OSF.IO/FT289).

STM features used in perceptual categorisation

For each behavioral subject and each category (speech/music), a logistic regression model was applied, where the STM of all samples of the audio track served as features, and the binary behavioral judgment of the category acted as the target variable. Both the features and target variables were temporally downsampled to 50 Hz to optimize processing efficiency. The data were shuffled before splitting into folds to ensure representative and unbiased partitions, and the model was evaluated using stratified 5-fold cross-validation. This combined approach

ensured robust generalization and mitigated potential biases from the data's original ordering. To verify that all models were successfully trained, we computed the classifier's performance as the mean accuracy across the 5 stratified cross-validation folds. Model accuracies for each category were significantly above chance across subjects (all adjusted $p < .001$), as determined by a Wilcoxon signed-rank test against chance level (0.5, Bonferroni-corrected to account for the number of subjects with an alpha level of 0.05). A non-parametric test was used because it is better suited to handle the non-parametric nature of accuracy values. Then, feature patterns were derived by applying the Haufe transformation to the feature weights (Haufe et al., 2014). The Haufe transformation makes feature weights interpretable by converting them into feature patterns that directly reflect the contributions of features' spatial sources to the model's predictions. While raw feature weights often mix statistical relevance with noise and are influenced by the properties of the data representation, the transformation projects these weights back to the data space, disentangling the predictive components from the data structure. A feature pattern of zero indicates no contribution from the corresponding neural source, while positive patterns indicate that increased activity in those sources enhances the predicted outcome. Conversely, negative patterns suggest that increased activity in those sources suppresses or inversely relates to the predicted outcome, providing interpretable results. In other words, this process provided, for each behavioral subject and category, interpretable feature importance within the STM space, highlighting the specific regions of the modulation space used to decode each category. For each category, feature patterns were then averaged across behavioral subjects (Figure 1c).

Identification of STM-sensitive sEEG contacts

Ridge regressions

The goal of the analysis was to identify sEEG contacts that exhibited significant responses to the spectrotemporal modulations of the audio track. To achieve this, a ridge regression model was implemented, using the brain signals recorded from each sEEG contact (n=846)—decomposed into six frequency bands (as described above)—as features, and the time course of each spectrotemporal combination of the STM-decomposed audio track (n=2180) as the target variable. To enhance computational efficiency, both the feature and target variables were temporally downsampled to 50 Hz. To account for potential delays between brain responses and audio features, the ridge regression model was applied using a temporal response function (TRF), incorporating time lags ranging from 0 to 1.5 seconds. A backward-modeling approach was used, mapping brain signals onto audio features, as opposed to the more conventional forward-modeling approach. Data were shuffled prior to splitting into folds to ensure robust generalization, and the model was evaluated using 10-fold cross-validation. The optimal regularization parameter (α) was selected by testing values sampled

logarithmically across 20 points, spanning from 10^{-10} to 10^{10}. Model performance for each fold was assessed using Spearman's correlation between the predicted ($\hat{y}$) and actual (y) time courses of the spectrotemporal combinations. The final accuracy was calculated as the average correlation across all folds.

For subsequent analyses involving model weight extraction, only the models trained with the optimal regularization parameter—determined based on the highest average accuracy across folds—were used. The extracted weights were then averaged across all time lags tested with the TRF, providing a summary measure of feature importance that integrated responses over the entire lag period.

Accuracy map and thresholding

The analysis described earlier resulted in, for each sEEG contact, a correlation coefficient (rho) from a ridge regression model predicting the time course of STM amplitude based on brain signal for each spectrotemporal combination. In other words, an accuracy map of rhos in the STM space was obtained for each sEEG contact (see Figure 2b for an example). After applying a Fisher z-transformation to the rhos, statistical significance was evaluated using Bonferroni-corrected one-tailed t-tests for correlation coefficients on each rho of each accuracy map at an alpha level of 10^{-30} and adjusting for the number of combinations in the STM space, sEEG contacts, and folds (as used in a similar procedure in Berezutskaya et al., 2022). Only sEEG contacts showing at least six adjacent significant rhos within the 2D STM space were retained.

Spatial clustering

To identify brain areas responsive to STM, the MNI coordinates of STM-sensitive sEEG contacts were extracted. For each significant sEEG contact, a 5-mm radius 3D sphere was projected onto an MNI-standardized MRI volume using the MarsBar toolbox (https://marsbar-toolbox.github.io/). The radius was chosen based on the estimated spatial resolution of the bipolar montage (Jerbi et al., 2009). Only clusters consisting of at least two overlapping sEEG contacts from at least three different patients (to ensure generalizability) were considered significant, while isolated contacts were excluded to minimize Type I error rate.

Using brain-reconstructed STMs to decode behavioral categorization

In this analysis, we extracted from STM-sensitive sEEG contacts the prediction outputs from the previously described ridge regression model. For each sEEG contact, a predicted time-course of the fluctuation of amplitude of each spectrotemporal combination was generated from the corresponding ridge regression model, thus obtaining a predicted audio track in the STM space (i.e., brain-reconstructed STMs). Then, we used the brain-reconstructed STMs as

features in a logistic regression model and the binary behavioral judgment of the category (speech/music) as the target variable. The data were shuffled before splitting into folds and the model was evaluated using stratified 5-fold cross-validation. For each category, only sEEG contacts showing significant model accuracy across behavioral subjects were retained, as determined by a Wilcoxon signed-rank test against chance level (0.5) for each category and sEEG contact across behavioral subjects (Bonferroni-corrected to account for the number of sEEG contacts with an alpha level of 0.05). Feature patterns were then derived from significant models by applying the Haufe transformation to the feature weights (Haufe et al., 2014). This resulted in, for each sEEG contact, behavioral subject and category, interpretable feature importance within the STM space. To precisely identify the specific regions of the STM space used to decode each category, one-tailed t-tests (right-tailed) against 0 on z-scored feature patterns were conducted for each combination within the STM space and each sEEG contact across behavioral subjects (Bonferroni-corrected adjusting for the number of sEEG contacts, and combinations, with an alpha level of 0.05). The one-tailed approach was used because only positive feature patterns indicate that increased activity in the feature space enhances prediction outcomes (Haufe et al., 2014). This analysis resulted in a representation that indicated, for each point in the STM space, the number of sEEG contacts exhibiting a significant feature pattern (Figure 3e).

Similarity between feature patterns using brain-reconstructed vs. actual audio track for behavioral categorization.

For this analysis, we used two sets of feature patterns. First, we used the feature patterns from the previous logistic regression, where the STMs of all audio track samples served as the features, and the binary behavioral judgment of the category acted as the target variable for each category and participant. Second, we used the feature patterns from the logistic regression where the brain-reconstructed audio track in the STM space served as the features, and the behavioral categorization served as the target variable for each STM-sensitive sEEG contact (for which model accuracy was significantly greater than chance, as described in the previous section), category, and behavioral participant. For the second set, feature patterns from STM-sensitive sEEG contacts were averaged for each hemisphere of the brain. We used spearman correlations to quantify similarity between feature patterns, and their significance was tested using Bonferroni correction to account for the number of participants, categories, and hemispheres, with an alpha level of 0.05. This process resulted in a correlation coefficient for each behavioral participant, category, and hemisphere. In addition to the main analysis described above, we conducted a control analysis to assess the statistical significance of the observed similarity patterns. Specifically, for the first model, in which a logistic regression was applied to binary behavioral judgments using the STM representation of the actual audio track,

we disrupted the relationship between predictors and data by shuffling the binary judgments independently for each model. We then performed the same similarity analysis on these shuffled datasets, allowing us to obtain a random distribution of correlation coefficients. To evaluate whether the observed correlation coefficients were significantly different from those obtained with shuffled predictors, we conducted a Wilcoxon signed-rank test for each category, comparing the true distribution to the shuffled distribution of correlation coefficients. Finally, to assess hemisphere preference for each category, we conducted a 2x2 non-parametric factorial ANOVA using the aligned rank transform (ART) method as implemented in the ARTool package in R (Elkin et al., 2021) on correlation coefficients, which served as similarity scores. The design included category (speech/music) as a within-subject factor and hemisphere (left/right) as a between-subject factor. Post-hoc comparisons were performed using pairwise Wilcoxon signed-rank tests, with false discovery rate (FDR) correction applied to control for multiple comparisons.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work the lead author used ChatGPT 4 for purposes of language and grammar clean-up exclusively. After using this tool/service, the lead author reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

REFRENCES

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*, *136*(5), 1639–1661. https://doi.org/10.1093/brain/awt082

Albouy, P., Mehr, S. A., Hoyer, R. S., Ginzburg, J., Du, Y., & Zatorre, R. J. (2024). Spectro-temporal acoustical markers differentiate speech from song across cultures. *Nature Communications*, *15*(1), 4835. https://doi.org/10.1038/s41467-024-49040-3

Angulo-Perkins, A., Aubé, W., Peretz, I., Barrios, F. A., Armony, J. L., & Concha, L. (2014). Music listening engages specific cortical regions within the temporal lobes:

Differences between musicians and non-musicians. *Cortex*, *59*, 126–137. https://doi.org/10.1016/j.cortex.2014.07.013

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review*, *61*(3), 183.

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication*, *1*(01), 217–233.

Berezutskaya, J., Vansteensel, M. J., Aarnoutse, E. J., Freudenburg, Z. V., Piantoni, G., Branco, M. P., & Ramsey, N. F. (2022). Open multimodal iEEG-fMRI dataset from naturalistic stimulation with a short audiovisual film. *Scientific Data*, *9*(1), 91. https://doi.org/10.1038/s41597-022-01173-0

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, *14*(10), 693–707. https://doi.org/10.1038/nrn3565

Boebinger, D., Norman-Haignere, S. V., McDermott, J. H., & Kanwisher, N. (2021). Music-selective neural populations arise without musical training. *Journal of Neurophysiology*, *125*(6), 2237–2263. https://doi.org/10.1152/jn.00588.2020

Boemio, A., Fromm, S., Braun, A., & Poeppel, D. (2005). Hierarchical and asymmetric temporal sensitivity in human auditory cortices. *Nature Neuroscience*, *8*(3), 389–395. https://doi.org/10.1038/nn1409

Borderie, A., Caclin, A., Lachaux, J.-P., Perrone-Bertollotti, M., Hoyer, R. S., Kahane, P., Catenoix, H., Tillmann, B., & Albouy, P. (2024). Cross-frequency coupling in cortico-hippocampal networks supports the maintenance of sequential auditory information in short-term memory. *PLOS Biology*, *22*(3), e3002512. https://doi.org/10.1371/journal.pbio.3002512

Calhoun, B. M., & Schreiner, C. E. (1998). Spectral envelope coding in cat primary auditory cortex: Linear and non-linear effects of stimulus characteristics. *European Journal of Neuroscience*, *10*(3), 926–940. https://doi.org/10.1046/j.1460-9568.1998.00102.x

Cha, K., Zatorre, R. J., & Schönwiesner, M. (2016). Frequency Selectivity of Voxel-by-Voxel Functional Connectivity in Human Auditory Cortex. *Cerebral Cortex*, *26*(1), 211–224. https://doi.org/10.1093/cercor/bhu193

Chang, A., Teng, X., Assaneo, M. F., & Poeppel, D. (2024). The human auditory system uses amplitude modulation to distinguish music from speech. *PLOS Biology*, *22*(5), e3002631. https://doi.org/10.1371/journal.pbio.3002631

Chen, X., Affourtit, J., Ryskin, R., Regev, T. I., Norman-Haignere, S., Jouravlev, O., Malik-Moraleda, S., Kean, H., Varley, R., & Fedorenko, E. (2023). The human language system, including its inferior frontal component in "Broca's area," does not support music perception. *Cerebral Cortex*, *33*(12), 7904–7929. https://doi.org/10.1093/cercor/bhad087

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. https://doi.org/10.1121/1.1945807

Coffey, E. B. J., Herholz, S. C., Chepesiuk, A. M. P., Baillet, S., & Zatorre, R. J. (2016). Cortical contributions to the auditory frequency-following response revealed by MEG. *Nature Communications*, *7*(1), 11070. https://doi.org/10.1038/ncomms11070

Depireux, D. A., Simon, J. Z., Klein, D. J., & Shamma, S. A. (2001). Spectro-Temporal Response Field Characterization With Dynamic Ripples in Ferret Primary Auditory Cortex. *Journal of Neurophysiology*, *85*(3), 1220–1234. https://doi.org/10.1152/jn.2001.85.3.1220

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. https://doi.org/10.1016/j.neubiorev.2017.02.011

Eggermont, J. J. (2002). Temporal modulation transfer functions in cat primary auditory cortex: Separating stimulus effects from neural mechanisms. *Journal of Neurophysiology*, *87*(1), 305–321.

Elkin, L. A., Kay, M., Higgins, J. J., & Wobbrock, J. O. (2021). An Aligned Rank Transform Procedure for Multifactor Contrast Tests. *The 34th Annual ACM Symposium on User Interface Software and Technology*, 754–768. https://doi.org/10.1145/3472749.3474784

Elliott, T. M., & Theunissen, F. E. (2009). The Modulation Transfer Function for Speech Intelligibility. *PLoS Computational Biology*, *5*(3), e1000302. https://doi.org/10.1371/journal.pcbi.1000302

Escabí, M. A., Miller, L. M., Read, H. L., & Schreiner, C. E. (2003). Naturalistic auditory contrast improves spectrotemporal coding in the cat inferior colliculus. *Journal of Neuroscience*, *23*(37), 11489–11504. https://doi.org/10.1523/JNEUROSCI.23-37-11489.2003

Escabí, M. A., & Schreiner, C. E. (2002). Nonlinear Spectrotemporal Sound Analysis by Neurons in the Auditory Midbrain. *The Journal of Neuroscience*, *22*(10), 4114–4131. https://doi.org/10.1523/JNEUROSCI.22-10-04114.2002

Fadiga, L., Craighero, L., & D'Ausilio, A. (2009). Broca's Area in Language, Action, and Music. *Annals of the New York Academy of Sciences*, *1169*(1). https://doi.org/10.1111/j.1749-6632.2009.04582.x

Flinker, A., Doyle, W. K., Mehta, A. D., Devinsky, O., & Poeppel, D. (2019). Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. *Nature Human Behaviour*, *3*(4), 393–405. https://doi.org/10.1038/s41562-019-0548-z

Friederici, A. D. (2020). Hierarchy processing in human neurobiology: How specific is it? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1789), 20180391. https://doi.org/10.1098/rstb.2018.0391

Gabrieli, J. D. E., Ghosh, S. S., & Whitfield-Gabrieli, S. (2015). Prediction as a Humanitarian and Pragmatic Contribution from Human Cognitive Neuroscience. *Neuron*, *85*(1), 11–26. https://doi.org/10.1016/j.neuron.2014.10.047

Gervain, J., & Geffen, M. N. (2019). Efficient Neural Coding in Auditory and Speech Perception. *Trends in Neurosciences*, *42*(1), 56–65. https://doi.org/10.1016/j.tins.2018.09.004

Giroud, J., Lerousseau, J. P., Pellegrino, F., & Morillon, B. (2023). The channel capacity of multilevel linguistic features constrains speech comprehension. *Cognition*, *232*, 105345. https://doi.org/10.1016/j.cognition.2022.105345

Giroud, J., Trébuchon, A., Mercier, M., Davis, M. H., & Morillon, B. (2024). The human auditory cortex concurrently tracks syllabic and phonemic timescales via acoustic spectral flux. *Science Advances*, *10*(51), eado8915. https://doi.org/10.1126/sciadv.ado8915

Giroud, J., Trébuchon, A., Schön, D., Marquis, P., Liegeois-Chauvel, C., Poeppel, D., & Morillon, B. (2020). Asymmetric sampling in human auditory cortex reveals spectral processing hierarchy. *PLOS Biology*, *18*(3), e3000207. https://doi.org/10.1371/journal.pbio.3000207

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., Ugurbil, K., Andersson, J., Beckmann, C. F., Jenkinson, M., Smith, S. M., & Van Essen, D. C. (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178. https://doi.org/10.1038/nature18933

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, *15*(1), 20–25. https://doi.org/10.1016/0166-2236(92)90344-8

Graham, D. J., & Field, D. J. (2007). Statistical regularities of art images and natural scenes: Spectra, sparseness and nonlinearities. *Spatial Vision*, *21*.

Guevara Erra, R., & Gervain, J. (2016). The Efficient Coding of Speech: Cross-Linguistic Differences. *PLOS ONE*, *11*(2), e0148861. https://doi.org/10.1371/journal.pone.0148861

Haiduk, F., Zatorre, R. J., Benjamin, L., Morillon, B., & Albouy, P. (2024). Spectrotemporal cues and attention jointly modulate fMRI network topology for sentence and melody

perception. *Scientific Reports*, *14*(1), 5501. https://doi.org/10.1038/s41598-024-56139-6

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

Hsu, A., Woolley, S. M. N., Fremouw, T. E., & Theunissen, F. E. (2004). Modulation Power and Phase Spectrum of Natural Sounds Enhance Neural Encoding Performed by Single Auditory Neurons. *The Journal of Neuroscience*, *24*(41), 9201–9211. https://doi.org/10.1523/jneurosci.2449-04.2004

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*, *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

Jamison, H. L., Watkins, K. E., Bishop, D. V. M., & Matthews, P. M. (2006). Hemispheric Specialization for Processing Auditory Nonspeech Stimuli. *Cerebral Cortex*, *16*(9), 1266–1275. https://doi.org/10.1093/cercor/bhj068

Jerbi, K., Ossandón, T., Hamamé, C. M., Senova, S., Dalal, S. S., Jung, J., Minotti, L., Bertrand, O., Berthoz, A., Kahane, P., & Lachaux, J. (2009). Task-related gamma-band dynamics from an intracerebral perspective: Review and implications for surface EEG and MEG. *Human Brain Mapping*, *30*(6), 1758–1771. https://doi.org/10.1002/hbm.20750

Johnsrude, I. S., Penhune, V. B., & Zatorre, R. J. (2000). Functional specificity in the right human auditory cortex for perceiving pitch direction. *Brain*, *123*(1), 155–163. https://doi.org/10.1093/brain/123.1.155

Kaas, J. H., & Hackett, T. A. (1999). “What” and “where” processing in auditory cortex. *Nature Neuroscience*, *2*(12), 1045–1047. https://doi.org/10.1038/15967

Kell, A. J. E., Yamins, D. L. K., Shook, E. N., Norman-Haignere, S. V., & McDermott, J. H. (2018). A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron*, *98*(3), 630-644.e16. https://doi.org/10.1016/j.neuron.2018.03.044

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

Kowalski, N., Depireux, D. A., & Shamma, S. A. (1996). Analysis of dynamic spectra in ferret primary auditory cortex. I. Characteristics of single-unit responses to moving ripple spectra. *Journal of Neurophysiology*, *76*(5), 3503–3523. https://doi.org/10.1152/jn.1996.76.5.3503

Langers, D. R., Backes, W. H., & Dijk, P. V. (2003). Spectrotemporal features of the auditory cortex: The activation in response to dynamic ripples. *NeuroImage*, *20*(1), 265–275. https://doi.org/10.1016/S1053-8119(03)00258-1

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience*, *5*(4), 356–363. https://doi.org/10.1038/nn831

Loui, P., Alsop, D., & Schlaug, G. (2009). Tone Deafness: A New Disconnection Syndrome? *The Journal of Neuroscience*, *29*(33), 10215–10220. https://doi.org/10.1523/JNEUROSCI.1701-09.2009

Luo, H., & Poeppel, D. (2012). Cortical Oscillations in Auditory Perception and Speech: Evidence for Two Temporal Windows in Human Auditory Cortex. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00170

Massoudi, R., Van Wanrooij, M. M., Versnel, H., & Van Opstal, A. J. (2015). Spectrotemporal response properties of core auditory cortex neurons in awake monkey. *PLoS One*, *10*(2), e0116118. https://doi.org/doi.org/10.1371/journal.pone.0116118

Maunsell, J. H. R., & Newsome, W. T. (1987). Visual Processing in Monkey Extrastriate Cortex. *Annual Review of Neuroscience*, *10*(1), 363–401. https://doi.org/10.1146/annurev.ne.10.030187.002051

McDermott, J. H., & Simoncelli, E. P. (2011). Sound Texture Perception via Statistics of the Auditory Periphery: Evidence from Sound Synthesis. *Neuron*, *71*(5), 926–940. https://doi.org/10.1016/j.neuron.2011.06.032

Mehr, S. A., Krasnow, M. M., Bryant, G. A., & Hagen, E. H. (2021). Origins of music in credible signaling. *Behavioral and Brain Sciences*, *44*, e60. https://doi.org/10.1017/S0140525X20000345

Miller, L. M., Escabí, M. A., Read, H. L., & Schreiner, C. E. (2002). Spectrotemporal Receptive Fields in the Lemniscal Auditory Thalamus and Cortex. *Journal of Neurophysiology*, *87*(1), 516–527. https://doi.org/10.1152/jn.00395.2001

Milner, B. (1962). Laterality effects in audition. *Interhemispheric Relations and Cerebral Dominance*, 177–195.

Ming, V. L., & Holt, L. L. (2009). Efficient coding in human auditory perception. *The Journal of the Acoustical Society of America*, *126*(3), 1312–1320. https://doi.org/10.1121/1.3158939

Norman-Haignere, S., Feather, J., Boebinger, D., Brunner, P., Ritaccio, A., McDermott, J. H., Schalk, G., & Kanwisher, N. (2022). A neural population selective for song in human auditory cortex. *Current Biology*, *32*(7), 1470-1484.e12. https://doi.org/10.1016/j.cub.2022.01.069

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. (2015). Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron*, *88*(6), 1281–1296. https://doi.org/10.1016/j.neuron.2015.11.035

Okada, K., Rong, F., Venezia, J., Matchin, W., Hsieh, I.-H., Saberi, K., Serences, J. T., & Hickok, G. (2010). Hierarchical Organization of Human Auditory Cortex: Evidence from Acoustic Invariance in the Response to Intelligible Speech. *Cerebral Cortex*, *20*(10), 2486–2495. https://doi.org/10.1093/cercor/bhp318

Okamoto, H., & Kakigi, R. (2015). Hemispheric Asymmetry of Auditory Mismatch Negativity Elicited by Spectral and Temporal Deviants: A Magnetoencephalographic Study. *Brain Topography*, *28*(3), 471–478. https://doi.org/10.1007/s10548-013-0347-1

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

Patterson, R. D., Uppenkamp, S., Johnsrude, I. S., & Griffiths, T. D. (2002). The Processing of Temporal Pitch and Melody Information in Auditory Cortex. *Neuron*, *36*(4), 767–776. https://doi.org/10.1016/S0896-6273(02)01060-7

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

Rauschecker, J. P., & Scott, S. K. (2009). Maps and streams in the auditory cortex: Nonhuman primates illuminate human speech processing. *Nature Neuroscience*, *12*(6), 718–724. https://doi.org/10.1038/nn.2331

Rieke, F., Bodnar, D., & Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *262*(1365), 259–265. https://doi.org/10.1098/rspb.1995.0204

Rissman, J., & Wagner, A. D. (2012). Distributed Representations in Memory: Insights from Functional Brain Imaging. *Annual Review of Psychology*, *63*(1), 101–128. https://doi.org/10.1146/annurev-psych-120710-100344

Robert, P., Zatorre, R. J., Gupta, A., Sein, J., Anton, J.-L., Belin, P., Thoret, E., & Morillon, B. (2024). Auditory hemispheric asymmetry for actions and objects. *Cerebral Cortex*, *34*(7), bhae292. https://doi.org/10.1093/cercor/bhae292

Rodríguez, F. A., Read, H. L., & Escabí, M. A. (2010). Spectral and temporal modulation tradeoff in the inferior colliculus. *Journal of Neurophysiology*, *103*(2), 887–903. https://doi.org/10.1152/jn.00813.2009

Samson, S., & Zatorre, R. J. (1988). Melodic and harmonic discrimination following unilateral cerebral excision. *Brain and Cognition*, *7*(3), 348–360. https://doi.org/10.1016/0278-2626(88)90008-5

Santoro, R., Moerel, M., De Martino, F., Goebel, R., Ugurbil, K., Yacoub, E., & Formisano, E. (2014). Encoding of Natural Sounds at Multiple Spectral and Temporal Resolutions in the Human Auditory Cortex. *PLoS Computational Biology*, *10*(1), e1003412. https://doi.org/10.1371/journal.pcbi.1003412

Scheinost, D., Noble, S., Horien, C., Greene, A. S., Lake, E. Mr., Salehi, M., Gao, S., Shen, X., O'Connor, D., Barron, D. S., Yip, S. W., Rosenberg, M. D., & Constable, R. T. (2019). Ten simple rules for predictive modeling of individual differences in neuroimaging. *NeuroImage*, *193*, 35–45. https://doi.org/10.1016/j.neuroimage.2019.02.057

Schön, D., Gordon, R., Campagne, A., Magne, C., Astésano, C., Anton, J.-L., & Besson, M. (2010). Similar cerebral networks in language, music and song perception. *NeuroImage*, *51*(1), 450–461. https://doi.org/10.1016/j.neuroimage.2010.02.023

Schönwiesner, M., Rübsamen, R., & Von Cramon, D. Y. (2005). Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. *European Journal of Neuroscience*, *22*(6), 1521–1528. https://doi.org/10.1111/j.1460-9568.2005.04315.x

Schönwiesner, M., & Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences*, *106*(34), 14611–14616. https://doi.org/10.1073/pnas.0907682106

Shamma, S. (2001). On the role of space and time in auditory processing. *Trends in Cognitive Sciences*, *5*(8), 340–348. https://doi.org/10.1016/S1364-6613(00)01704-6

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379–423.

Shannon, R. V., Zeng, F.-G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science*, *270*(5234), 303–304. https://doi.org/10.1126/science.270.5234.303

Sihvonen, A. J., Särkämö, T., Rodríguez-Fornells, A., Ripollés, P., Münte, T. F., & Soinila, S. (2019). Neural architectures of music – Insights from acquired amusia. *Neuroscience & Biobehavioral Reviews*, *107*, 104–114. https://doi.org/10.1016/j.neubiorev.2019.08.023

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural Image Statistics and Neural Representation. *Annual Review of Neuroscience*, *24*(1), 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193

Singh, N. C., & Theunissen, F. E. (2003). Modulation spectra of natural sounds and ethological theories of auditory processing. *The Journal of the Acoustical Society of America*, *114*(6), 3394–3411. https://doi.org/10.1121/1.1624067

Smith, E. C., & Lewicki, M. S. (2006). Efficient auditory coding. *Nature*, *439*(7079), 978–982. https://doi.org/10.1038/nature04485

Staeren, N., Renvall, H., De Martino, F., Goebel, R., & Formisano, E. (2009). Sound Categories Are Represented as Distributed Patterns in the Human Auditory Cortex. *Current Biology*, *19*(6), 498–502. https://doi.org/10.1016/j.cub.2009.01.066

Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D., & Leahy, R. M. (2011). Brainstorm: A User-Friendly Application for MEG/EEG Analysis. *Computational Intelligence and Neuroscience*, *2011*, 1–13. https://doi.org/10.1155/2011/879716

Te Rietmolen, N., Mercier, M., Trébuchon, A., Morillon, B., & Schön, D. (2024). *Speech and music recruit frequency-specific distributed and overlapping cortical networks*. https://doi.org/10.7554/eLife.94509.2

Teng, X., Tian, X., & Poeppel, D. (2016). Testing multi-scale processing in the auditory system. *Scientific Reports*, *6*(1), 34390. https://doi.org/10.1038/srep34390

Teng, X., Tian, X., Rowland, J., & Poeppel, D. (2017). Concurrent temporal channels for auditory processing: Oscillatory neural entrainment reveals segregation of function at different scales. *PLOS Biology*, *15*(11), e2000812. https://doi.org/10.1371/journal.pbio.2000812

Theunissen, F. E., & Elie, J. E. (2014). Neural processing of natural sounds. *Nature Reviews Neuroscience*, *15*(6), 355–366. https://doi.org/10.1038/nrn3731

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*, *20*(6), 2315–2331. https://doi.org/10.1523/JNEUROSCI.20-06-02315.2000

Varnet, L., Ortiz-Barajas, M. C., Erra, R. G., Gervain, J., & Lorenzi, C. (2017). A cross-linguistic study of speech modulation spectra. *The Journal of the Acoustical Society of America*, *142*(4), 1976–1989. https://doi.org/10.1121/1.5006179

Venezia, J. H., Richards, V. M., & Hickok, G. (2021). Speech-Driven Spectrotemporal Receptive Fields Beyond the Auditory Cortex. *Hearing Research*, *408*, 108307. https://doi.org/10.1016/j.heares.2021.108307

Venezia, J. H., Thurman, S. M., Richards, V. M., & Hickok, G. (2019). Hierarchy of speech-driven spectrotemporal receptive fields in human auditory cortex. *NeuroImage*, *186*, 647–666. https://doi.org/10.1016/j.neuroimage.2018.11.049

Vogels, R., & Orban, G. A. (1996). Chapter 14 Coding of stimulus invariances by inferior temporal neurons. In *Progress in Brain Research* (Vol. 112, pp. 195–211). Elsevier. https://doi.org/10.1016/S0079-6123(08)63330-0

Walters, J., King, M., Bissett, P. G., Ivry, R. B., Diedrichsen, J., & Poldrack, R. A. (2022). Predicting brain activation maps for arbitrary tasks with cognitive encoding models. *NeuroImage*, *263*, 119610. https://doi.org/10.1016/j.neuroimage.2022.119610

Woolley, S. M. N., Fremouw, T. E., Hsu, A., & Theunissen, F. E. (2005). Tuning for spectro-temporal modulations as a mechanism for auditory discrimination of natural sounds. *Nature Neuroscience*, *8*(10), 1371–1379. https://doi.org/10.1038/nn1536

Zatorre, R. J. (2022). Hemispheric asymmetries for music and speech: Spectrotemporal modulations and top-down influences. *Frontiers in Neuroscience*, *16*, 1075511. https://doi.org/10.3389/fnins.2022.1075511

Zatorre, R. J., & Belin, P. (2001). Spectral and temporal processing in human auditory cortex. *Cerebral Cortex*, *11*(10), 946–953. https://doi.org/10.1093/cercor/11.10.946

Zuk, N. J., Teoh, E. S., & Lalor, E. C. (2020). EEG-based classification of natural sounds reveals specialized responses to speech and music. *NeuroImage*, *210*, 116558. https://doi.org/10.1016/j.neuroimage.2020.116558