Cortical tracking of naturalistic music and speech across frequency bands and brain regions: functional mapping and temporal dynamics

Sergio Osorio[1,2,*] & M. Florencia Assaneo[3,*]

1. Department of Neurology, Harvard Medical School, Massachusetts General Hospital, Boston, MA
2. Martinos Center for Biomedical Imaging, Massachusetts General Hospital, Boston, MA
3. Laboratorio de Percepción y Producción del Habla, Instituto de Neurobiología, Universidad Autónoma de México, Santiago de Querétaro, MX.

* Corresponding authors: *sosoriogaleano@mgh.harvard.edu*, *fassaneo@inb.unam.mx*

Abstract

Music and speech encode hierarchically organized structural complexity at the service of human expressiveness and communication. Carefully controlled experiments have suggested that populations of neurons in cortical auditory regions track temporal modulations within rhythmic acoustic signals, physiologically supporting perception of both music and speech. However, whether cortical tracking of music and speech extends to less controlled (i.e., naturalistic) signals remains contentious. Here, we investigated whether cortical tracking can be observed under more natural perceptual scenarios, and how stimulus type, frequency band or anatomical localization modulates this effect. We analyzed intracranial recordings from 30 subjects while they passively watched a movie where visual scenes were accompanied by either music or speech stimuli. Cross-correlation between brain and acoustic signals, along with density-based clustering analyses and linear mixed effect modeling, revealed both anatomically overlapping and functionally distinct mapping of the tracking effect as a function of stimulus type and frequency band. We observed widespread tracking of music and speech signals in the Slow Frequency Band (1-8Hz), with near zero temporal lags and high mixed-selectivity. In contrast, High Frequency Band (70-120Hz) tracking was higher during speech perception, was more densely concentrated in classical language processing areas, and showed a clear frontal-to-temporal gradient in lag values that was not observed during perception of musical stimuli. Our results highlight the recruitment of domain-general and domain-specific mechanisms during perception of naturalistic music and speech signals, as well as a complex interaction between cortical region and frequency band that shapes temporal dynamics during processing of hierarchically organized temporal structures in speech.

Introduction

Music and speech are acoustic signals that encode hierarchically organized structural complexity and support emotional and semantic expressiveness in human communication (Aboitiz, 2017; Asano, 2022; Fujita & Fujita, 2022; Lerdahl & Jackendoff, 1983; McFee et al., 2017; Miyagawa et al., 2013; Schön & Morillon, 2019). For this, small structural units (e.g., notes or syllables) are linearly combined to build up longer units (e.g., phrases or sentences) in order to establish non-linear relationships between high-level abstract representations. Hence, understanding how the brain physiologically supports perception of music and speech is of great importance for cognitive and auditory neuroscience.

Because both signals deploy in time, hierarchical structure is physically embedded in quasi-periodic modulations of acoustic features, which can be characterized by corresponding spectral peaks observable in the signal via power density and Fourier analyses. Across multiple languages and different musical genres, temporal modulations are dominated by the mean syllabic rate, which converges around 5 Hz (N. Ding et al., 2017; Pellegrino et al., 2011), and the mean note rate, which concentrates around 2 Hz (N. Ding et al., 2017; van Noorden & Moelants, 1999), correspondingly. Despite subtle differences in their temporal modulations, however, perception of both music and speech imposes essentially the same neurocognitive demands, namely, multiscale tracking of structural units and their dependencies. This has led some to propose that the neural architecture supporting perception of both phenomena must derive from similar evolutionary affordances and must be supported by common domain-general mechanisms (Asano, 2022; Fiveash et al., 2021; Tillmann, 2012).

Remarkably, during perception of music and speech, neuronal populations synchronize their activity to the incoming acoustic signal. Magnetoencephalography (MEG) and electroencephalography (EEG) research has shown synchronization of brain signals in the Slow Frequency Band (SFB, 1-8 Hz) with musical features, such as beat, tempo and meter in temporal and prefrontal electrodes (Doelling & Poeppel, 2015; Harding et al., 2019; Kaneshiro et al., 2020; Nozaradan, 2014; Nozaradan et al., 2012). This effect is modulated by note rate (Nozaradan et al., 2012), musical expertise (Doelling & Poeppel, 2015), stimulus familiarity (Kumagai et al., 2017) and stimulus complexity (Wollman et al., 2020). Similarly, temporal fluctuations of acoustic features in the speech signal, such as syllable and word onsets, are tracked by brain activity within the SFB (Assaneo & Poeppel, 2018; Chalas et al., 2023; N. Ding et al., 2015), an effect that is in turn modulated by syllabic rate (Assaneo & Poeppel, 2018; N. Ding et al., 2015), speech intelligibility (N. Ding et al., 2015; Doelling et al., 2014; Peelle et al., 2013) and attention (Vanthornhout et al., 2019; Viswanathan et al., 2019).

Neural tracking of music and speech has also been robustly observed within the range of the High Frequency gamma Band (HFB, >70 Hz), particularly through the modulation of HFB amplitude (Y. Ding et al., 2019; Herff et al., 2020; Kubanek et al., 2013). Electrocorticography (EcoG) research has shown that passive perception of musical stimuli is associated with both time-locked and sustained HFB neural responses in the temporal lobe, supramarginal gyrus and inferior frontal gyrus (Berezutskaya et al., 2017; Y. Ding et al., 2019). Cortical tracking of the envelope of the music signal within the 70-170Hz frequency range has been additionally observed in right prefrontal regions (Y. Ding et al., 2019; Herff et al., 2020). For speech stimuli, synchronization of brain activity within

75-115 Hz band to the signal envelope has been previously observed, with an increasing gradient from perisylvian and prefrontal regions towards primary auditory areas (Kubanek et al., 2013). Despite limited spatial resolution and signal-to-noise ratios compared to electrocorticography, MEG and EEG studies further confirm tracking of the speech envelope in temporal electrodes and source-reconstructed auditory areas in the HFB (Harding et al., 2019; Kulasingham et al., 2020; Synigal et al., 2020).

Physiologically, SFB and HFB activity are thought to reflect distinct, yet related neural mechanisms of information processing and transmission. SFB activity is presumed to reflect fluctuating windows of temporal coordination between large neuronal populations that are crucial for information transfer between globally distributed cortical regions (Fries, 2015; Fujisawa & Buzsáki, 2011; Sirota & Buzsáki, 2005). In turn, activity within the HFB is thought to reflect sparse encoding of information via transient excitation of local cell assemblies during sensory perception and cognition (Fries et al., 2007; Ray, Crone, et al., 2008; Rieder et al., 2011). What is more, a potential interplay between both SFB and HFB activity has been proposed, where the power of gamma-band neuronal activity depends on the phase of its slower counterpart (Buzśaki & Wang, 2012; Canolty et al., 2006).

In spite of the evidence that cortical tracking of acoustic signals within the SFB and HFB supports perception of music and speech, a couple of questions deserve closer attention. First, previous studies have made use of carefully controlled stimuli of relative short duration (e.g., single notes, syllables, words, short tonal melodies, or sentences) and enhanced periodicity (Assaneo & Poeppel, 2018; N. Ding et al., 2015; Nozaradan et al., 2012). Such studies have been important in helping us understand how the brain processes music and speech, but under naturalistic scenarios, music and speech signals are far less

periodic, are embedded in noise, and compete for attentional resources with input from other sensory modalities. For this reason, there have been recent calls and attempts to investigate brain responses to speech and music using less controlled materials and in the context of sensory multimodality (Broderick et al., 2018; Hamilton & Huth, 2020; He et al., 2015; Kern et al., 2022; Osorio et al., 2024; Poikonen et al., 2018; Sturm et al., 2015; Willems et al., 2020). Whether cortical tracking extends to more natural conditions therefore requires further inquiry. Only a handful of EcoG studies have addressed this question, with evidence converging on preliminary findings that tracking of natural signals is more distributed than previously thought and recruits overlapping anatomical regions with little or no anatomical regional selectivity (Berezutskaya et al., 2017; Te Rietmolen et al., 2024).

Another question that remains underexplored is whether cortical tracking of natural music and speech signals is differentially expressed in the SFB and HFB. Indeed, evidence for a role of SFB and HFB activity in tracking of hierarchical structures in music and speech usually comes from studies separately focusing on either frequency range depending on the acquisition method that was used. One previous EcoG study, however, directly compared how the brain responds to music and speech in these two frequency bands. Results showed that while the temporal envelope of the acoustic signal is tracked by both SFB and HFB activity, acoustic edges, which correspond to the peak rate of temporal modulations in the signal, are best predicted by activity in the SFB (Te Rietmolen et al., 2024). However, whether more complex interactions between stimulus type, frequency band and anatomical region exist remains unknown.

To address these questions, we analyzed an open-access database of electrocortical recordings acquired during naturalistic, multimodal perception of music and speech stimuli (Berezutskaya et al., 2022). Prior to surgical procedures, patients passively watched 30-second videos containing either music (multi-instrument clips) or speech (single voice and dialogues) audios, while brain activity was recorded. Data from a subsample of 30 subjects with electrode grids located in left prefrontal, central and/or temporal regions was analyzed, thus offering a coverage of most cortical areas classically involved in music and speech perception. We were interested in understanding 1) whether cortical tracking of music and speech can be observed under more natural perceptual scenarios in response to less controlled acoustic signals, and 2) whether the spatial and temporal dynamics of cortical tracking are modulated by frequency band or stimulus type.

To quantify cortical tracking of music and speech in a straightforward way, we conducted cross-correlation analyses between the cochlear envelope of acoustic signals and the bandpass-filtered brain signals (SFB, the brain activity filtered between 1-8 Hz; and HFB, the amplitude modulation of the 70-120 Hz filtered brain signal). To investigate whether frequency-band-dependent tracking effects can be mapped to similar or different cortical areas, we conducted density-based clustering analyses. Finally, to investigate the effect of condition and cortical region in the strength of the tracking effect and its temporal dynamics, we conducted mixed effect modeling analyzes. Cross-correlation coefficients and temporal lags were set as predicted variables, whereas frequency band (SFB, HFB), stimulus type (music, speech), and anatomical location of electrodes (cortical region) were set as regressors of interest. For all models, subject and anatomical region were set as

potential random effects to control for the within-subject nature of the data and differences in EcoG grid locations across individuals.

Materials and methods

Data was obtained from a publicly available dataset (Berezutskaya et al., 2022) collected at the University Medical Center Utrecht from 63 patients that underwent intracortical electrode implantation as part of diagnosis procedures for resection surgery due to a clinical history of drug-resistant epileptic seizures. For this study, we used data from a subsample of 30 subjects (mean age = 27.33, SD=15.28, 19 females, see supplementary table s1), selected on the basis of electrode grid coverage (i.e., only patients with grids in left prefrontal, central and/or temporal regions). Language dominance for all subjects was localized to the left hemisphere (table s1). This resulted in a total number of 1,858 electrodes across all subjects (mean = 61.93, SD = 19.09, supplementary table s2) distributed across left-hemisphere prefrontal, central and temporal regions (figure s1).

Participants were passively presented 30-second interleaved excerpts from the movie “Pipi on the run” (Pårymmen med Pippi Långstrump, 1970) dubbed to the Dutch language, where visual scenes were accompanied by music (7 blocks) or speech (6 blocks) stimuli. The task was originally designed as part of a language mapping procedure prior to surgical intervention. All participants and legal guardians, when applicable, provided their informed consent to participate in the study and to make their de-identified data publicly available (Berezutskaya et al., 2022). Procedures were revised and approved by the Medical Ethical Committee of the University Medical Center Utrecht in accordance with the Declaration of Helsinki (2013).

Data analysis

Acoustic signals

Acoustic data was analyzed using MATLAB Version 9.6.0 (2019a). The acoustic signal was extracted from the audiovisual stimulus and segmented according to stimulus type (speech or music). The last musical segment was excluded from analyses to have a balanced number of segments across the two conditions. The segmented signals were then down-sampled to 16,000 Hz. Next, each segment's auditory representation in the human cochlea was obtained by detrending, resampling (200hz), and filtering each signal in 128 logarithmically spaced frequencies within the range of human audition (180 to 7,246 Hz). The cochlear spectrogram was obtained using the NSL toolbox (Chi et al., 2005). Finally, signals were averaged across the 128 frequencies to obtain the cochlear envelope of each acoustic segment. For each envelope, the Power Spectrum Density was obtained using Welch's method.

EcoG and neuroanatomical data preprocessing

Electrophysiological and neuroanatomical data was preprocessed using Brainstorm (Tadel et al., 2011). Electrodes were first re-referenced to the local average of each individual EcoG grid. Next, data was bandpass filtered between 0.5 and 120Hz using an FIR Keyser filter (attenuation level = 60 dB, Order = 3714). A second-order IIR notch filter was applied to remove line noise (60Hz) and harmonics (120Hz). Then, electrodes reported as bad in the original dataset and additional noisy channels visually identified on the basis of their power spectrum properties via Welch's PSD analysis were rejected from further analysis (mean = 3.5, SD = 4.6).

To inspect EcoG grid location, electrodes were plotted against each subject's cortical surfaces, which are available in the original dataset. After replicating individual electrode locations for each participant, we then projected all electrodes in a the MNI-ICBM152 cortical template (figure s1). In a few cases, this procedure resulted in spatial distortion of some electrodes due to conversion between coordinate systems. In such cases, electrode location was corrected by visually comparing the electrode's position in the subject's anatomy and relocating it if necessary to the corresponding anatomical area in the common cortical space. For more details on structural MRI preprocessing and post-surgical grid corregistration procedures in the original data, see Berezutskaya et al., 2022.

Electrophysiological signals were then bandpass-filtered and epoched using Fieldtrip (Oostenveld et al., 2011). A third order Butterworth IIR filter was used to obtain continuous brain signals within the Slow Frequency Band (SFB, 1-8Hz) and High-Frequency Band (HFB, 70-120Hz). For the latter frequency band, the envelope was additionally obtained by extracting the absolute value of the signal's Hilbert transform. Finally, data was epoched into twelve 30-second segments, corresponding to the six music and six speech clips.

Cross-correlation analysis

Cross-correlation analyses were conducted using MATLAB Version 9.6.0 (2019a) and 9.14.0 (2023a). The cross-correlation function was estimated for each electrode between the z-normalized brain signals and the z-normalized cochlear envelopes of the music and speech segments. This was done per subject, for the two frequency bands of interest (SFB and HFB). Previous studies have used small temporal windows of approximately 500ms, which excludes the slowest temporal modulations in the acoustic

signals. Here, we used a maximum lag of ± 400 samples, corresponding to a four-second window between -2000 and 2000ms. This is in line with findings that larger temporal windows are associated with better model performance in decoding analyses of brain signals using Artificial Neural Networks (Berezutskaya et al., 2020).

The maximum correlation coefficient and its corresponding temporal lag were extracted from the correlation function. The mean coefficient estimates and corresponding temporal lags were obtained by averaging across the values for the six segments within each condition (figure 1b, for data distributions see supplementary figure s2). Statistical significance was estimated for each individual by obtaining a null distribution of cross-correlation coefficients via permutation (figure 1b). For each subject, the cross-correlation function was estimated 1,000 times between the z-normalized brain signals and the z-normalized cochlear envelopes of randomly generated white-noise signals. P-values for the observed data were subsequently obtained by estimating the probability of observing the same or more extreme cross-correlation coefficients given the empirically-obtained null distributions. The threshold for statistical significance was set at an alpha value of 0.001.

Density-based cluster analysis

All statistically significant electrodes across our 30 subjects were then plotted in a normalized cortical space (MNI-ICBM152). Electrodes that significantly correlated with the acoustic envelope were identified performing multiple comparisons (i.e., 1858 electrodes were studied). Accordingly, a high number of false positives are expected, even if a more conservative alpha value was chosen a priori. Because the permutation procedure is agnostic of the anatomical location of electrodes, and in order to identify cortical regions where the group-level effect was localized, we additionally conducted a density-based

cluster classification analysis on the collapsed data across all subjects using the DBSCAN algorithm (Ester et al., 1996) as implemented by MATLAB. The assumption was that the overall effect of cortical tracking should be more pronounced in task-relevant cortical regions. DBSCAN should therefore spatially constrain any statistical effect in a data-driven manner to functionally relevant cortical areas, while taking care of isolated electrodes which would otherwise be considered outliers.

DBSCAN requires only two initial parameters: the minimum number of data points for a cluster to be identified and the minimum radius within which those data points should be contained (i.e., the epsilon parameter). This means that no starting data point or number of clusters need to be defined a priori, unlike other clustering methods. To derive the optimal parameters for the DBSCAN algorithm, we conducted a procedure to obtain the parameter combination that optimized clustering within each condition. A matrix of parameter optimization indices (*OI*) was obtained by the following formula.

$$OI = \left(\frac{E}{\epsilon}\right) * (PC)$$

Where ***E*** is the matrix containing the percentage of electrodes obtained for each combination of initial parameters, **Є** is a vector containing the list of possible minimum radiuses, ***P*** is a vector containing a range of possible minimum data points for a cluster to be identified, and ***C*** is the number of clusters obtained by each combination of parameters. Finally, we normalized ***OI*** so that values range between 0 and 1, with 0 indicating poor algorithm performance and 1 indicating optimal performance.

Supplementary figure s3 shows the normalized OI values as a function of different parameter combination for the SFB and HFB analyses. The ideal parameters are those

which maximize OI_{norm} with the lowest possible cluster radius and the highest possible number of electrodes within such cluster. The final parameters used for DBSCAN clustering analyses for each frequency band and stimulus type are provided in table 1.

	Frequency band	**Є**	**Min points**	**OI_{norm}**
Speech	HFB	0.014	12	0.94
	SFB	0.012	12	0.92
Music	HFB	0.006	12	1.00
	SFB	0.006	16	1.00

Table 1. Parameters for density-based clustering analysis. The epsilon (Є) parameter represents the radius of search, and the minimum points stands for the minimum number of electrodes within the search area for cluster identification.

Electrode anatomical labeling

DBSCAN returns a numbered list of identified clusters under the specified parameter combination. We therefore grouped all clusters resulting from DBSCAN analyses as a single group and anatomically labeled each electrode using the Desikan-Killiany anatomical atlas (Desikan et al., 2006). We took this approach because clusters where the result of an algorithmic operation performed in space that is agnostic of brain anatomy, which in practice means that two small clusters could be identified within the same anatomical region. Electrodes were therefore labeled according to the cortical area in which they were located by plotting each electrode on the MNI-ICBM152 template against overlaid parcellations from the Desikan-Killiany atlas. This resulted in nine cortical regions

of interest (table 2), which were used as categorical predictors in our linear mixed effect modeling analyses.

Linear mixed effect modeling

Linear Mixed Effect (LME) modeling was conducted using the LME4 package (Bates et al., 2015) in R (R Core Team, 2023) using R Studio version 2023.6.1 (R Studio Corte Team, 2023). Data was modeled for the two response variables of interest: cross-correlation values and temporal lags. For each LME analysis, the minimum number of datapoints within each cortical region for inclusion in the model was set to nine. For music, electrodes within the STG, MTG, PFC and Supramarginal Gyrus (SG) were included in the analysis. The IFG, premotor cortex, precentral and postcentral gyri were excluded because not enough electrodes were found within these regions across frequency bands. For speech, electrodes within the STG, MTG, precentral and postcentral gyri, IFG, and SG were included in the analysis. Not enough electrodes were found across frequency bands in the premotor or prefrontal cortices, so these regions were excluded. We additionally performed a joint LME analysis for music and speech including the STG, MTG and SG as cortical regions of interest.

Region	**Music**		**Speech**	
	SFB	*HFB*	*SFB*	*HFB*
STG	37	16	159	123
MTG	38	33	198	43
PFC	15	28	36	2
IFG	4	4	118	53
IP	0	12	0	0
Premotor	0	5	24	0
Precentral	5	1	137	46
Postcentral	15	4	95	23
SG	9	28	90	30

Table 2. Number of statistically significant electrodes per anatomical region. *STG*, (Superior Temporal Gyrus), *MTG* (Middle Temporal Gyrus), PFC (Prefrontal Cortex) *IFG* (Inferior Frontal Gyrus), *IP* (Inferior Parietal), SG (Supramarginal Gyrus)

A forward stepwise model selection procedure using Maximum Likelihood Estimation was conducted to identify the models that best fit our data (table 3 and supplementary tables s3, s4 and s5). Single predictor models for frequency band and cortical region including a random intercept for subject were tested against a null (random intercept-only) model. If any single-predictor model was significantly better than the null model, the effect of adding additional terms and interactions while keeping the random intercept unchanged was tested. Once the optimal combination of fixed effect terms was determined, the effect of adding other random effects was tested. This procedure was conducted for music and speech separately. However, we additionally conducted a joint LME analysis where condition (music and speech) was considered an additional fixed effect. Table 3 shows the best models predicting cross-correlation coefficients and temporal lags per condition with their corresponding Akaike Information Criterion (AIC) and Bayes Information Criterion (BIC). For music, no model outperformed a null model in predicting temporal lags. For more information on model selection procedures, see supplementary materials.

	LME model	**AIC**	**BIC**
Music	r ~ frequency + (1|subject)	-1299,136	-1285,854
	lag ~ (1+subject)	277.78	287.73
Speech	r ~ frequency*region + (1|subject) + (1|subject:region)	6351,467	-6276,218
	lag ~ frequency*region + (1|subject) + (1|subject:region)	13426,74	13501,99

Both	r ~ frequency*region*condition + 1(subject) + (1|subject:region)	-4379.39	-4309.04
	lag ~ frequency*region*condition + 1(subject)	9624.23	9689.89

Table 3. Best fit models for each condition and for the collapsed model and for each response variable of interest. For music, no model was significantly better in predicting temporal lags than a null model.

Results

We first studied the properties of our acoustic stimuli. Top panels in figure 1a show the soundwave of two representative music and speech acoustic segments (see also figure s4 for all other segments). A cochlear filter was applied to obtain the approximate spectrotemporal representation of each music and speech segment in the human cochlea (figure 1a, middle panels, see methods). Next, the cochlear envelope of each segment was estimated by averaging across all the cochlear frequencies (figure 1a, bottom panels). Figure 1b shows the normalized spectral power for the six music (left) and speech (right) segments correspondingly, obtained via Welch´s Power Spectrum Density (PSD) method. PSD analyses show high variability in spectral peaks across the different music and speech segments, consistent with the uncontrolled nature of the stimuli. The mean spectrum across music cochlear envelopes, however, shows peaks at around 0.2, 2 and 5 Hz, suggesting that across segments, music signals are somewhat more rhythmic than speech signals. Analyses confirmed statistically significant weak-to-moderate positive correlations between the six music segments ($\alpha = 0.05$, two-tailed, table 4, top panel). For speech, no clear dominant peak was observed when the cochlear envelopes were averaged, reflecting marked differences in the temporal structure across the six different speech segments. This is

consistent with the nature of the speech stimuli, which included both monologues and conversations between multiple individuals. Correlation analysis showed weak negative correlations for the acoustic envelopes of speech (table 4, bottom panel).

Segments	1	2	3	4	5	6	
1							*Music*
2	0.397***						
3	0.232***	0.103***					
4	0.442***	0.560***	0.299***				
5	0.290***	0.253***	0.012 n.s.	0.225***			
6	0.029*	-0.131***	0.147***	-0.037***	-0.006 n.s.		
1							*Speech*
2	0.036 n.s.						
3	0.030*	0.006 n.s.					
4	-0.022 n.s.	0.030*	-0.050***				
5	0.049***	-0.102***	-0.042***	-0.098***			
6	0.043***	0.164***	-0.059***	0.002 n.s.	-0.006 n.s.		

**** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$*

Table 4. Correlation across the music and speech segments.

We then studied neural tracking of the music and speech envelopes. For this, we estimated the cross-correlation function between the brain and acoustic signals. Figure 1c shows a schematic representation of the cross-correlation analysis pipeline. For each subject (n = 30), we obtained the cross-correlation function between the cochlear envelopes of the music (n = 6) and speech (n = 6) segments, and the brain signal recorded at each electrode (n = 1858, mean = 61.93, SD = 19.09) in the Slow Frequency Band (SFB, 1 to

8Hz bandpass filtered signal) and the High Frequency Band (HFB, envelope of the 70 to 120Hz bandpass filtered signal). For each cross-correlation, we then extracted the maximum brain-to-stimulus coefficient and its corresponding temporal lag. We finally averaged the coefficients across the six segments within each stimulus type. Statistical significance was estimated using a null distribution of cross-correlation coefficients obtained via permutation ($n_{permutations} = 1000$, $\alpha = 0.001$, figure 1c). Finally, we additionally implemented a density-based clustering classification analysis using the DBSCAN algorithm to remove spatial outliers and identify cortical regions where significant electrodes were most densely grouped.

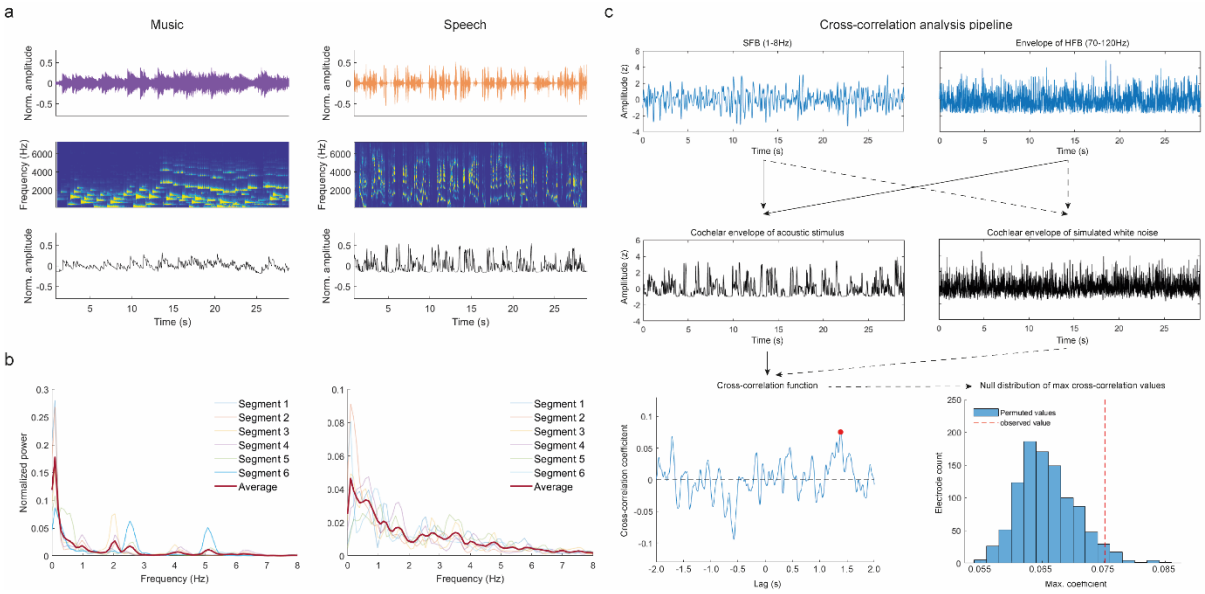

Figure 1. Analysis of acoustic signals. **a.** Sound waves of representative acoustic segments for music (top-left panel) and speech (top-right panel), their cochlear spectrograms (middle panels) and cochlear envelope (bottom panels). **b.** PSD for all music (left) and speech (right) segments. Red line shows average power across segments. **c.** Schematic representation of the cross-correlation analysis. For brain signals (SFB, 1-8Hz, top left and envelope of HFB, 70-120Hz, top right), cross-correlation was estimated against the

cochlear envelope of stimuli signals (middle left) to obtain the maximum correlation coefficient and its corresponding lag (bottom left). A permutation procedure was conducted by estimating the cross-correlation function between SFB and HFB brain signals and the cochlear envelope of simulated white-noise (middle right), to obtain a null distribution of maximum correlation coefficients which was used to estimate significance thresholds. Solid black arrows represent the pipeline for real data. Dotted black lines represent the permutation procedure.

In both frequency bands, permutation and density-based clustering analyses showed a higher number of statistically significant electrodes tracking speech compared to music (figure 2). Electrodes tracking the music envelope in the SFB concentrated in perisylvian, temporal and prefrontal regions, including the middle and posterior portions of the STG, anterior, middle, and posterior portions of the MTG and ITG, the lowermost portions of the precentral and postcentral gyri, and the most anterior portion of the dorsolateral Prefrontal Cortex (dlPFC, figure 2a, left). For speech, cortical tracking in the SFB was abundantly found in perisylvian, ventral prefrontal, pre and postcentral, and temporal regions, including the anterior, middle, and posterior portions of both the STG and MTG, the lowermost portions of the SG, middle and lower portions of the precentral and postcentral gyri, premotor areas, the IFG, and the dlPFC (figure 2a, middle). In the SFB, the number of statistically significant electrodes tracking the speech signals (n = 823, percent from total = 44.22%) was 6.5 times higher than the number of electrodes tracking the music signals (n = 125, percent from total = 6.72%, figure 2b). Interestingly, within the SFB, 84% of the

electrodes tracking the music signals (n = 105, percent from total = 5.65%) also tracked the envelope of the speech signals (figure 2a, right and figure 2b).

Electrodes tracking music stimuli in the HFB concentrated around PFC and premotor areas, as well as in SG, IP, STG and MTG regions (figure 2c, left). In turn, electrodes tracking speech in the HFB range were more densely concentrated in perisylvian regions, including the middle and posterior STG, lowermost portions of the precentral and postcentral gyri, and the middle and posterior portions of the IFG (figure 2c, middle). There was an approximate two-fold increase in the number of electrodes showing statistically significant tracking of speech in the HFB (n = 320, percent from total = 17.22%) compared to music (n = 136, percent from total = 7.32%, figure 2d). Only 18.32% of electrodes tracking the music envelope also tracked the speech envelope in the HFB. These electrodes were restricted to posterior portions of the STG (n = 25, percent from total = 1.34%, figure 2c, right).

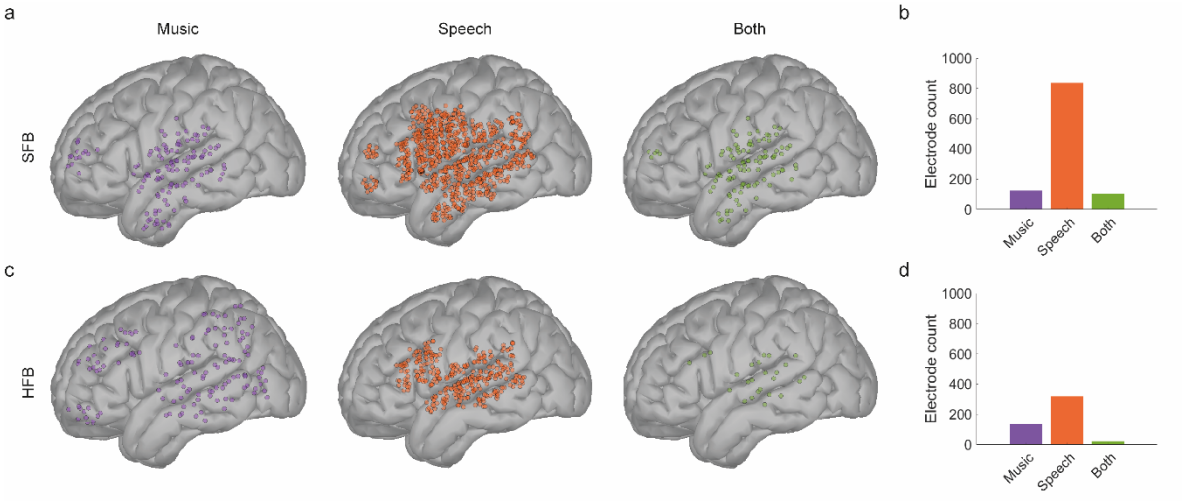

Figure 2. Statistically significant electrodes and their anatomical localization in the MNI-ICBM152 cortical template. **a.** and **c.** Spatial distribution of significant electrodes in the

SFB (**a**) and HFB (**c**) after density-based clustering analyses for music (orange) and speech (purple). Electrodes that show mixed-selectivity (i.e., respond to both stimuli) are presented in green. **b.** and **d.** Number of electrodes that survive permutation statistics and density-based clustering classification in the SFB (**b**) and HFB (**d**).

Next, we investigated the organization of cross-correlation values and temporal lags (figure 3). For tracking of music signals in the SFB, cross-correlation coefficients (mean = 0.09, SD = 0.006) showed a gradient of increasing values from association areas, including anterior portions of the ITG and MTG, towards primary auditory areas and posterior portions of the STG (figure 3a, top-left). Temporal lags were maximally concentrated around time zero (mean = 12ms, median = 50ms, SD = 470ms, figure 3a, top-right), following a similar gradient with more positive lags found in anterior and middle portions of the STG (figure 3c, left). In the HFB, coefficients (mean = 0.96, SD = 0.013) showed a more spatially-distributed gradient, with increasing values towards association areas, including superior parietal, middle temporal, and prefrontal regions (figure 3a, bottom-left). Similar to the SFB, temporal lags for the HFB concentrated around time zero (mean = -14ms, median = 57ms, SD = 480ms, figure 3a, bottom-right), with maximum lags in middle and posterior portions of the STG (figure 3c, right).

The cortical tracking effect for the speech signal in the SFB (mean = 0.10, SD = 0.001) was particularly distributed in cortical space, with higher coefficients found along the anterior-to-posterior axis of the STG and MTG, but also in portions of the IFG, precentral gyrus and SG (figure 3b, top-left). Similar to music, temporal lags for speech in the SFB band converged around zero (mean = -6ms, median = -29ms, SD = 470ms, figure

3b, top-right, figure 3d, left). In turn, cortical tracking of the speech signal in the HFB (mean = 0.10, SD = 0.023) showed a densely-grouped distribution and sharp increase in cross-correlation coefficient values from prefrontal to posterior temporal regions, with a concentration of the highest cross-correlation coefficients in the middle and posterior STG (figure 3b, bottom-left). Interestingly, unlike all other conditions, mean temporal lags in the HFB band were biased towards positive values (mean = 440ms, median = 490ms, SD = 590ms, figure 3b, bottom-right). Temporal lags in the HFB followed a similar gradient as cross-correlation coefficients, with more positive lags (i.e. higher delays) concentrating around the middle portion of the STG (figure 3d, right).

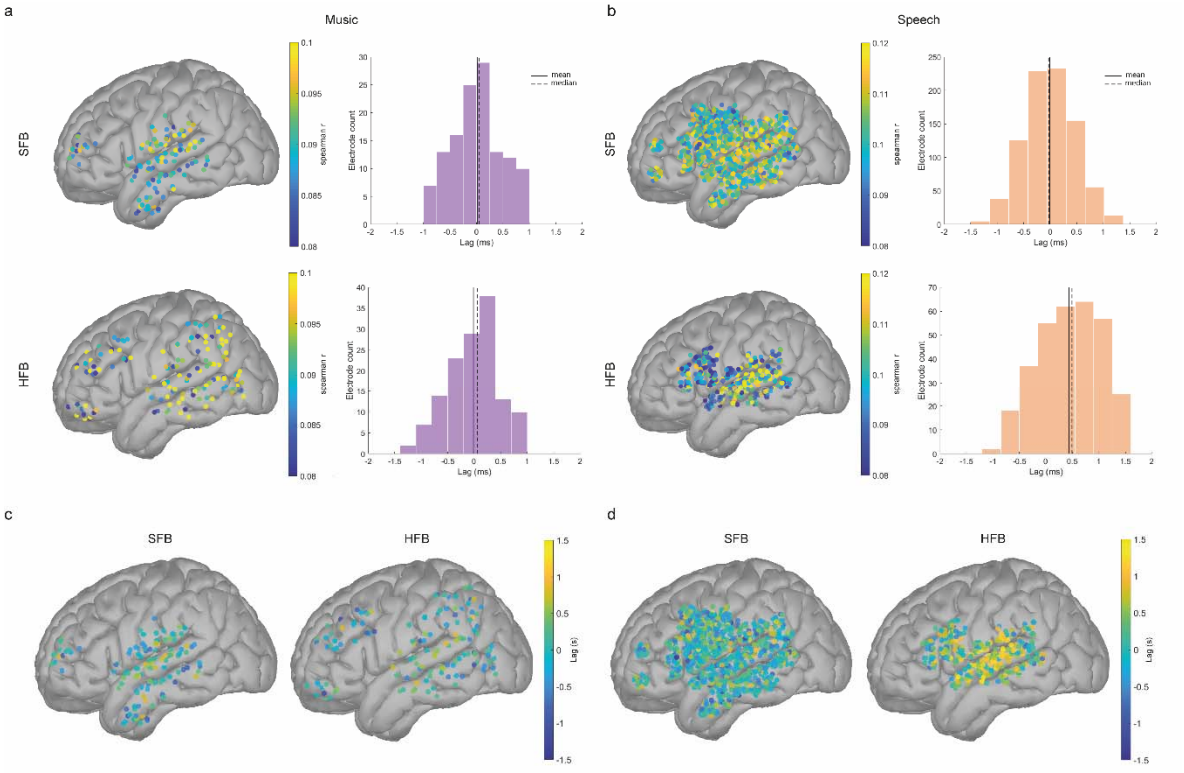

Figure 3. Cortical tracking of music and speech envelopes. **a.** Distribution of mean correlation values in the MNI-ICBM152 cortical template (left column panels) and histogram of their corresponding temporal lags (right column panels) in the SFB and HFB

frequencies for music. **b**. Spatial distribution of mean correlation values (left column panels) and histogram of their corresponding temporal lags (right column panels) for speech. Positive lags indicate that the acoustic signal precedes brain signals, whereas negative lags indicate that the brain signals precede the acoustic signal. All electrodes shown are statistically significant ($p < 0.001$, uncorrected) and survive clustering analyses. **c.** and **d.** Spatial organization of temporal lags during cortical tracking of music (c) and speech (d) in both the SFB and HFB.

To quantify the effect of cortical tracking more accurately across anatomical regions, electrodes were next labeled according to their anatomical location (see methods). Electrodes were classified within one of nine left-hemisphere cortical areas (figure 4a and table 2). Cortical tracking of music was more prominent within the MTG and STG regions in the SFB, and within the MTG, SG and PFC in the HFB (figure 4b). For speech stimuli, cortical tracking was higher within MTG, STG and postcentral regions for the SFB, and within STG and IFG for the HFB (figure 4c). Next, we resorted to mixed effect modeling to investigate how anatomical location and frequency band influence cortical tracking. The two response variables of interest were cross-correlation coefficients and temporal lags. The fixed effects of interest were frequency band and cortical region. The random effects of interest were subject (to account for the within subject nature of the task) and the interaction between subject and cortical region (to account for the different electrode grid locations across subjects). We performed a forward stepwise model selection procedure to select the best model explaining the data (see supplementary materials).

For music, a model including frequency band as fixed effect and subject as random effect best fit the observed cross-correlation coefficients (AIC = -1299.13, BIC = -1285.85, $R^2_{Conditional}$ = 0.12, $R^2_{Marginal}$ = 0.04, table s3). This model showed a statistically significant main effect of frequency (F (1, 184.87) = 9.00, p = 0.003, figure 4d). Estimated Marginal Means (emmeans) suggested that cortical tracking of the acoustic signals was significantly higher in the HFB (emmean = 0.094, C.I. = [0.092, 0.097]) than in the SFB (emmean = 0.089, C.I. = [0.086, 0.092]). However, no model was significantly better than a null model in predicting temporal lags during cortical tracking of music.

For speech, a model considering the interaction between frequency band and cortical region as fixed effect, and both subject and the subject-by-region interaction as random effects best explained the cross-correlation coefficients (AIC = -6351.47, BIC = -6276.22, $R^2_{Conditional}$ = 0.19, $R^2_{Marginal}$ = 0.08, table s4). This model showed a significant main effect of frequency (F (1, 1066.49) = 25.73, p = 4.64e-07) and cortical region (F (5, 162.87) = 8.81, p = 2.01e-07), and a significant interaction between frequency and cortical region (F (5, 1058.98) = 7.47, p = 6.71e-07). Post-hoc pairwise comparisons showed that cortical tracking of the speech envelope was significantly higher in the SFB compared to the HFB within the IFG (emmean $_{SFB}$ = 0.105, CI = [0.102, 0.108]; emmean $_{HFB}$ = 0.095, CI =[0.091, 0.099], p < 0.0001), MTG (emmean $_{SFB}$ = 0.10, CI = [0.101, 0.107]; emmean $_{HFB}$ = 0.99, CI = [0.091, 0.010], p = 0.022) and precentral gyrus (emmean $_{SFB}$ = 0.104, CI = [0.101, 0108]; emmean $_{HFB}$ = 0.0942, CI = [0.0897, 0.099], p < 0.0001, figure 4e), and in the HFB compared to the SFB within the STG (emmean $_{SFB}$ = 0.107, CI = [0.105, 0.110]; emmean $_{HFB}$ = 0.102, CI = [0.108, 0.114], p = 0.016, figure 4e). In the HFB, cortical tracking was higher within STG (emmean = 0.109, CI = [0.107, 0.112]) compared to the

IFG (emmean = 0.100, CI = [0.097, 0.103], $p < 0.0001$), MTG (emmean = 0.102, CI = [0.099, 0.105], $p = 0.0002$), precentral gyrus (emmean = 0.993, CI = [0.096, 0.102], $p < 0.0001$), postcentral gyrus (emmean = 0.993, CI = [0.095, 0.103], $p = 0.0002$) and SG (emmean = 0.104, CI = [0.100, 0.108], $p = 0.040$, figure 4e). No statistically significant differences were observed in cross-correlation coefficients across the different anatomical regions for the SFB.

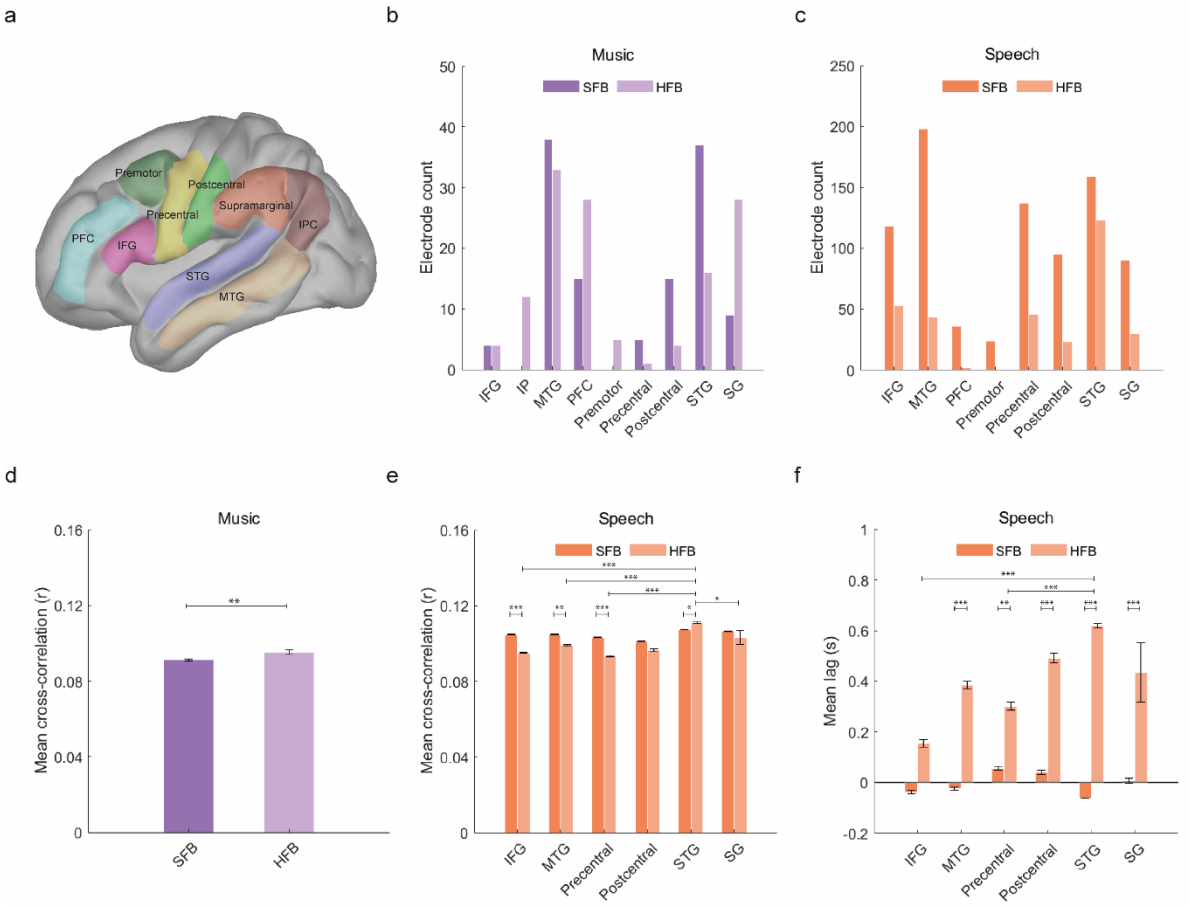

Figure 4. Anatomical regions of interest and significant effects as per mixed effect modeling analyses. **a.** Anatomical location of statistically significant electrodes after density-based clustering analyses. **b.** and **c.** Number of electrodes per anatomical parcellation for music (**a**) and speech (**c**). **d.** Main effect of frequency band for mean cross-

correlation values during cortical tracking of music. **e.** Main and interaction effects for cross-correlation values during cortical tracking of speech. **f.** Main and interaction effects for temporal lags during cortical tracking of speech. Whiskers represent the Standard Error of the Mean (SEM). For all panels, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

For temporal lags during tracking of speech (figure 4f), the best model also included frequency, region and their interaction as fixed effects, and subject as well as the subject-by-region interaction as random effects (AIC = 1611.49, BIC = 1686.74, $R^2_{Conditional} = 0.25$, $R^2_{Marginal} = 0.17$, table s4). This model showed significant main effects of frequency ($F(1, 1066.70) = 117.27$, $p < 2.20e-16$), cortical region ($F(5, 150.79) = 3.54$, $p = 0.0047$) and the interaction between frequency and region ($F(5, 1060.53) = 7.11$, $p = 1.48e-06$). Post-hoc pairwise comparisons showed more positive lags for HFB compared to SFB within the IFG (emmean $_{HFB}$ = 114ms, CI = [-33ms, 262ms]; emmean $_{SFB}$ = -42ms, CI = [-152ms, 67ms], $p = 0.050$), precentral gyrus (emmean $_{HFB}$ = 260ms, CI = [99ms, 420ms]; emmean $_{SFB}$ = 18ms, CI = [9ms, 128ms], $p = 0.004$), MTG (emmean $_{HFB}$ = 380ms, CI = [220ms, 541ms]; emmean $_{SFB}$ = -23ms, CI = [-118ms, 72ms], $p < 0.0001$), postcentral gyrus (emmean $_{HFB}$ = 431ms, CI = [219ms, 643ms]; emmean $_{SFB}$ = 2ms, CI = [-122ms, 126ms], $p = 0.0002$), SG (emmean $_{HFB}$ = 455ms, CI = [266ms, 643ms]; emmean $_{SFB}$ = 32ms, CI [-9ms, 154ms], $p < 0.0001$) and STG (emmean $_{HFB}$ = 618ms, CI = [510ms, 725ms]; emmean $_{SFB}$ = -59ms, CI = [-158ms, 40ms], $p < 0.0001$), suggesting that delays between the acoustic envelope of speech and brain signals are higher in the HFB across all cortical regions of interest. In the HFB, tracking within the STG (emmean = 618ms) occurred 503ms later compared to the IFG (emmean = 114ms, $p < 0.0001$) and 358ms later compared to the postcentral gyrus

(emmean = 260ms, CI = [100ms, 420ms], $p = 0.002$). No differences in temporal lags across cortical regions were observed within the SFB range (figure 4f).

Finally, we investigated the overall effect of cortical tracking and temporal lags by designing a joint model for music and speech. For this, we used a subsample of electrodes corresponding to the cortical regions where a significant effect of cortical tracking was observed for both music and speech. This included the STG, MTG and the SG (figure 5a). For mean cross-correlation values, the model that best fit the data was one including the interaction between condition, frequency band, and cortical region as fixed effects, and subject and the interaction between subject and cortical region as random effects (AIC = -4502.26, BIC = -4431.91, $R^2_{Conditional} = 0.26$, $R^2_{Marginal} = 0.17$, table s5). Results of this model suggest a main effect of condition ($F(1, 799.99) = 76.49$, $p = 2.0e-16$, figure 5b), where cortical tracking of speech (emmean = 0.105, C.I.= [0.103, 0.107]) is significantly higher than cortical tracking of music (emmean = 0.092, C.I. = [0.089, 0.095]). No effects were found for frequency band or cortical region. For mean temporal lags, the model included the interaction between condition, frequency band and region as fixed effects, and a random intercept for subject (AIC = 1191.96, BIC = 1257.61, $R^2_{Conditional} = 0.22$, $R^2_{Marginal} = 0.20$, table s5). Results of this model (figure 5c) show statistically significant main effects of frequency band ($F(1, 787.56) = 14.16$, $p = 0.0002$), condition ($F(1, 776.45) = 13.27$, $p = 0.0003$), and the interaction between frequency and condition ($F(1, 749.70) = 35.98$, $p = 3.096e-09$). Post-hoc pairwise comparisons showed that cortical tracking of speech in the HFB (emmean = 484ms, CI = [394ms, 574ms]) occurs, on average, 510ms later than tracking of speech in the SFB (emmean = -25ms, CI = [-85ms, 34ms], $p < 0.0001$), 504ms later than tracking of music in the HFB (emmean = -20ms, CI = [-147ms,

107ms], $p < 0.0001$), and 385ms later than tracking of music in the SFB (emmean = 99ms, CI = [-39ms, 237ms], $p < 0.0001$). No differences in temporal lags were observed for cortical tracking of music across frequency bands.

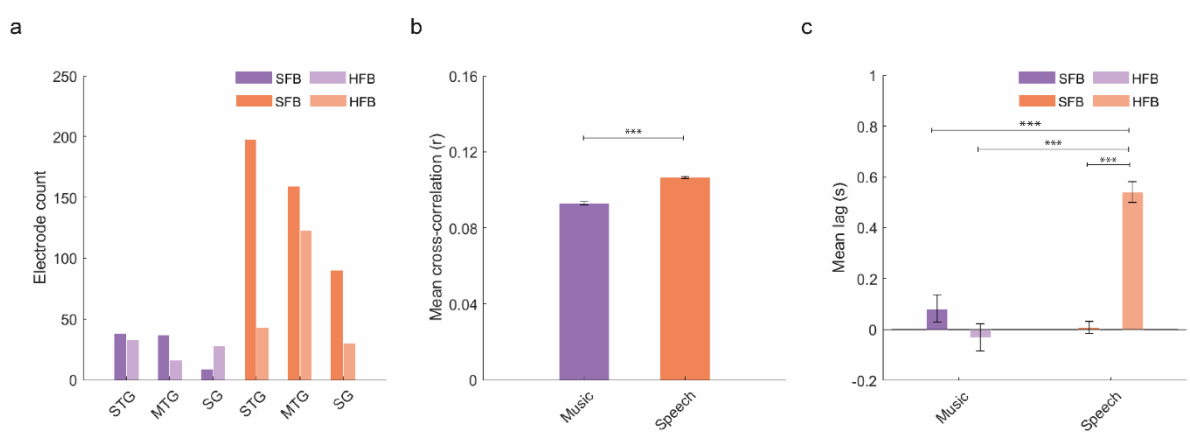

Figure 5. Number of electrodes per anatomical location and statistical effects of joint mixed effect model. **a.** Number of electrodes in the three anatomical location where statistically significant electrodes were found for both conditions. **b.** Main effect of condition for cross-correlation coefficients. **c.** Main effect and interaction for temporal lags. Whiskers represent the SEM. For all panels, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Discussion

In this study, we investigated how neural signals track naturalistic music and speech stimuli using a publicly available EcoG dataset. Cortical tracking of music and speech has been observed in the range of slow (1-8Hz, SFB) and fast (>70Hz, HFB) oscillatory activity under carefully controlled conditions and in response to stimuli of enhanced periodicity, using multiple non-invasive neuroimaging methods (N. Ding et al., 2015; Nozaradan, 2014; Nozaradan et al., 2012; Poeppel & Assaneo, 2020). Here, we asked whether this effect

could also be observed under more ecologically valid scenarios, by leveraging the increased SNR afforded by intracranial recordings. We also asked whether complex interactions between anatomical region, frequency band and stimulus type exist that have not been previously observed due to methodological or analytical limitations.

Cortical organization of the tracking effect in the slow and fast frequency bands

In line with similar EcoG studies (Berezutskaya et al., 2017; Te Rietmolen et al., 2024), we found distributed cortical tracking in left hemisphere regions classically involved in language and music perception. Density-based clustering analyses showed overlapping tracking of music and speech in the SFB and HFB within the MTG, STG, SG, precentral and postcentral gyri. These findings are in line with previous research using source-localized MEG signals (Ahissar et al., 2001; Assaneo & Poeppel, 2018; Chalas et al., 2023; Park et al., 2020) and intracortical techniques (Berezutskaya et al., 2017; Y. Ding et al., 2019; Harding et al., 2019; Herff et al., 2020; Kubanek et al., 2013; Te Rietmolen et al., 2024). Overlap of cortical tracking around perisylvian and temporal areas for both stimuli suggests that domain-general neuroanatomical mechanisms are involved in the processing of temporal modulations in acoustic signals (Asano, 2022; Fiveash et al., 2021; Tillmann, 2012). This was particularly true for the SFB, where 84% of electrodes that showed a statistically significant effect for music also showed a significant effect for speech, thus suggesting not only high anatomical overlap, but also high mixed selectivity within the range of slow oscillations. Mixed selectivity has been proposed to optimize computational power during complex cognitive operations, thus supporting high-dimensional neural representations (Fusi et al., 2016; Rigotti et al., 2013; Tye et al., 2024).

However, we also found frequency-band-dependent and stimulus-dependent differences. Tracking of music in both frequency bands, and of speech in the SFB, extended beyond classic auditory regions to prefrontal, middle temporal, and parietal areas. In contrast, tracking of speech in the HFB was densely localized to perisylvian and ventral prefrontal regions. One previous EcoG study found a similar spatial organization pattern of neural activations during speech perception, with SFB activity showing a global distribution in cortical space, while HFB activity was mainly localized to the STG (Zion Golumbic et al., 2013). This frequency-dependent pattern of global/local spatial organization is reminiscent of the hypotheses that slow oscillatory activity reflects communication between globally distributed neuronal ensembles, and high frequency activity reflects local activation of small neural populations in response to specific features in the sensory scene (Buzsaki, 2006; Buzśaki & Wang, 2012; Canolty et al., 2006; Fries, 2015; Fries et al., 2007; Fujisawa & Buzsáki, 2011; Ray, Niebur, et al., 2008; Rieder et al., 2011; Sirota & Buzsáki, 2005). Notably, the importance of cross-frequency interactions between slow and fast gamma oscillations for speech perception and language processing has also been previously documented (Hyafil et al., 2015; Lizarazu et al., 2019, 2023; Luo & Poeppel, 2012), suggesting that an interplay between slow and fast oscillatory activity probably underlies multiscale extraction of both acoustic features and hierarchical structures during speech perception and language comprehension.

Functional anatomical selectivity during tracking of natural music and speech signals

In the SFB, electrodes tracking the envelope of speech were observed within the IFG, which were notably absent during cortical tracking of music stimuli. In contrast, in the HFB, electrodes tracking the envelope of music stimuli were observed within the dlPFC

and IP regions, which were absent during tracking of speech. Both of these findings are in line with extensive fMRI literature on the role of the dlPFC and parietal cortex during music perception, and of the IFG during speech perception (Chan & Han, 2022; Enge et al., 2020; Green & Swets, 1966; Lankinen et al., 2023; Ohnishi et al., 2001; Tremblay & Small, 2011). In the EcoG literature, selective dlPFC recruitment in response to music has been previously reported during perception of musical stimuli, which has been attributed to attentional and emotional processing, as well as recall of complex melodic features in music stimuli (Y. Ding et al., 2019; Herff et al., 2020; Sturm et al., 2014). The involvement of the IFG during tracking of speech stimuli in both SFB and HFB has also been extensively reported in previous studies using intracortical methods (Berezutskaya et al., 2020; Sturm et al., 2014; Zion Golumbic et al., 2013). This suggests that in spite of partial anatomical overlap in perisylvian regions, cortical tracking of music and speech signals can also be mapped to distinct parietal and prefrontal brain regions in a stimulus-dependent manner.

Multiple roles have been previously attributed to the IFG during speech perception, including syntactic processing, semantic processing and phonological working memory (Bulut, 2023; Enge et al., 2020; Friederici, 2011; Perrachione et al., 2017). Admittedly, activation of the IFG has also been reported during music perception, however, recent metanalyses and clinical studies suggest this effect is right lateralized (Asano et al., 2022; Chen et al., 2023). Finally, increased BOLD activation of IP regions has been observed in response to music compared to speech (Merrill et al., 2012), and increased musical training is associated with increased IP activity among musicians (Liu et al., 2018).

Not many EcoG studies have directly investigated the extent to which tracking of music and speech can be mapped to distinct cortical areas and, more importantly, whether there is any interaction between cortical regions and oscillatory range. However, a recent study found no substantial differences in anatomical regional selectivity in how brain signals track music and speech in different frequency bands (Te Rietmolen et al., 2024). Our results are at odds with these findings, showing that functional anatomical differences are mainly observable in the range of HFB activity, and showing functional activations similar to those observed in fMRI research. One possible explanation for this discrepancy in results could be the use of density-based clustering analyses, which topographically restricted the cortical tracking effect to areas with a higher concentration of significant electrodes, thus revealing subtle differences in the cortical organization of the tracking effect that may have been missed if the spatial distribution of electrodes had not been controlled for. However, an alternative explanation is the use of unimodal auditory stimuli in that study, compared to the use of audiovisual stimuli in this study. Multimodal stimuli may have recruited association cortices or could have introduced attentional competition from both modalities resulting in both stimulus types being integrated differently.

Increased cortical tracking of speech across frequency bands

The net effect of cortical tracking was higher for speech than for music. This was reflected in a higher proportion of electrodes tracking the speech envelope in both frequency bands, as well as in the results of our joint linear mixed effect models, which show a main effect of stimulus type in predicting mean cross-correlation coefficient values. Such findings are aligned with the most recent literature using naturalistic stimuli, which have also shown that brain signals recover the envelope of speech signals more faithfully

than that of music signals (Berezutskaya et al., 2017; Te Rietmolen et al., 2024). For this particular sample, this could reflect the participant's everyday experience and familiarity with processing speech stimuli, as opposed to a lack of expertise or formal training in music. In support of this view, both speech intelligibility and musical training have been shown to increase cortical tracking of the corresponding signals (Di Liberto et al., 2020; N. Ding et al., 2015; Doelling & Poeppel, 2015; Harding et al., 2019; Peelle et al., 2013), which implies that previous experience processing a particular stimulus type modulates the magnitude of the cortical tracking effect. However, we cannot rule out the possibility that, regardless of musical expertise, subjects simply allocated increased attention to the speech stimuli because, unlike music, this information was crucial for understanding the story. Indeed, attentional allocation has been associated with increased cortical tracking in previous studies (Vanthornhout et al., 2019; Viswanathan et al., 2019). The lack of behavioral data makes it difficult to empirically back up one of these competing explanations, although if attention had mediated the effect, we should have observed increased HFB clustering within attentional prefrontal areas during speech perception, which was not the case.

Cortical gradient in lag values during HFB tracking of speech signals

Our analyses also revealed interesting temporal dynamics during cortical tracking of speech stimuli in the range of HFB activity. Positive lags are indicative of brain signals following acoustic modulations in auditory stimuli. Physiologically, this could reflect event-related activations or entrainment of brain signals in response to temporal modulations and acoustic features (Doelling et al., 2019; Zou et al., 2021). In turn, negative lags indicate that brain signals anticipate fluctuations in the acoustic input. Physiologically,

this could reflect predictive processes that presumably optimize sensory encoding (Patel & Morgan, 2017; Teng et al., 2020; Tillmann, 2012). For music and speech in the SFB, mean and median lag values were close to zero across all anatomical regions of interest, thus suggesting that neural activity in the slow oscillatory range tracks fast temporal modulations in the incoming acoustic signal in a relatively time-locked manner. Similarly, median temporal lags during HFB tracking followed the music signal by approximately 60ms on average. This could additionally reflect fast neural responses to specific acoustic features present in the music signal. In line with this, previous EcoG research has demonstrated that fluctuations in HFB amplitude within prefrontal, pre/postcentral, and temporal regions encode features such as loudness, harmonization, and the presence or absence of lyrics (Sturm et al., 2014). In the same study, onset of lyrics was characterized by additional increases in HFB power, localized to posterior portions of the STG.

For speech, estimated marginal means show that HFB tracking always occurred at higher lag values compared to the SFB. Our results showed an increasing gradient of lag values from prefrontal, parietal and middle temporal towards superior temporal regions, suggesting a highly localized effect of late-latency HFB tracking of speech within middle and posterior portions of the STG. The joint LME model further confirmed that HFB tracking of speech occurred significantly later compared to tracking of speech in the SFB, and to tracking of music in both the SFB and HFB in the three anatomical regions included in the analysis (STG, MTG and SG). This HFB gradient, which is observed for speech but not for music, could reflect tracking of structural units in the speech signal at different timescales. If this was the case, a prediction would be that, among trained musicians, a similar gradient in lag values across functionally relevant anatomical regions should be

observed, reflecting their ability to extract multiscale structural units embedded in music signals.

Limitations

Our study has some important limitations. First off, we lacked any behavioral measure that facilitates the interpretation of observed differences in cortical tracking. This precludes us from ruling out potential confounding factors, such as attentional allocation (Simon et al., 2023; Vanthornhout et al., 2019; Viswanathan et al., 2019). Another potential confounding factor in our study is the presentation of a concurrent visual stream during passive auditory perception of music and speech. Previous literature has shown that during audiovisual perception, temporal modulations in visual stimuli can be crossmodally represented in auditory areas (Besle et al., 2008; Mégevand et al., 2020; Park et al., 2018). Therefore, our effect could be partially driven by the concomitant representation of visual scenes which may have been integrated differently with music and speech stimuli.

Third, we used a very broad temporal window for our cross-correlation analyses (4000ms). This might have reduced our statistical power and increased the chance of spurious results. However, our motivation for using this broad time window was twofold: on the one hand, previous research has focused on temporal modulations such as beat, notes, syllables, and words, using windows of analysis of about 500ms. This ignores temporal structures such as musical and speech phrases occurring at much slower rates. Second, a study using the same dataset showed that a data-driven Artificial Neural Network (ANN) best performs in predicting brain signals and generalizes better to new stimuli when longer segments (~6 seconds) are used for model training (Berezutskaya et al., 2020). We

therefore considered that using a broader window could unveil interesting temporal dynamics that may have been overlooked in the past.

Finally, we derive our interpretations from marginal mean estimates obtained from LME models. While this allows us to draw some general conclusions at the group level, we acknowledge that this ignores high intra-individual variability in lag estimates, showing a range of both negative and positive lags in all regions and conditions. Positive and negative lags might reflect predictive and evoked responses to the stimuli. Morphosyntactic information has been shown to facilitate predictive processing during speech comprehension (N. Ding et al., 2015; Patel & Morgan, 2017; Willems et al., 2016), whereas neural mechanisms devoted for rhythm perception are involved in the self-generation of temporal predictions during processing of music stimuli (Merchant et al., 2015; Quiroga-Martinez et al., 2020; Rohrmeier & Koelsch, 2012; Tillmann, 2012). On this note, phase presession during ongoing auditory processing of temporal structures has been proposed as a potential neural mechanism facilitating predictive processing during speech perception (Teng et al., 2020, 2021). However, pinpointing in what way and to what extent negative and positive lag values reflect prediction, evoked responses, or neural entrainment is beyond the scope of this study. Relatedly, we cannot account here for the exact structures that are being tracked within each temporal bin and anatomical region observed in the HFB during speech perception. However, methods already exist to answer these questions, so the open-access nature of this dataset will hopefully motivate other researchers to examine whether the range of lag estimates can be associated with different structural units in acoustic signals.

Conclusion

In summary, our results reveal widespread cortical tracking, as well as both functional overlap and anatomical specialization in the spatial and temporal dynamics associated with the tracking of naturalistic music and speech stimuli. Among our non-specialized sample, passive perception of music and speech was associated with distributed tracking of acoustic signals in the SFB with near zero delays in mostly overlapping temporal and perisylvian areas. While some overlap was also observed for both type of signals in the HFB, marked anatomical-functional selectivity was also found in this frequency range, with dlPFC and IP regions preferentially tracking music and IFG regions preferentially tracking speech. The overall magnitude of the tracking effect was higher for speech across frequency bands, and a gradient in lag values was observed during HFB tracking of speech from association areas in prefrontal, middle temporal, and parietal regions towards the superior temporal areas, which was nonetheless absent during HFB tracking of music.

Our study provides evidence that cortical tracking of music and speech extends beyond carefully controlled experimental settings to complex naturalistic signals that are far less controlled, less rhythmic, and that are presented under conditions of sensory multimodality. Our results also highlight the fact that the human brain exploits both domain-general mechanisms and frequency-dependent anatomical specialization during cortical tracking of natural acoustic signals. Finally, we unveil a complex interaction between stimulus type, region, and frequency band that might be modulated by an individual’s previous experience, which may underly the ability to extract hierarchical structures from temporal modulations in natural acoustic signals.

Conflict of interest

The authors declare no conflict of interest.

Data availability statement

Data for this study is open and can be accessed at https://openneuro.org/datasets/ds003688. Code used for data analysis can be made available upon request.

References

Aboitiz, F. (2017). *A brain for speech*. Springer Nature. https://doi.org/10.1057/978-1-137-54060-7

Ahissar, E., Nagarajan, S., Ahissar, M., Protopapas, A., Mahncke, H., & Merzenich, M. M. (2001). Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(23), 13367–13372. https://doi.org/10.1073/pnas.201400998

Asano, R. (2022). The evolution of hierarchical structure building capacity for language and music: a bottom-up perspective. *Primates*, *63*(5), 417–428. https://doi.org/10.1007/s10329-021-00905-x

Asano, R., Lo, V., & Brown, S. (2022). The Neural Basis of Tonal Processing in Music: An ALE Meta-Analysis. *Music and Science*, *5*, 1–15. https://doi.org/10.1177/20592043221109958

Assaneo, M. F., & Poeppel, D. (2018). The coupling between auditory and motor cortices is rate-restricted: Evidence for an intrinsic speech-motor rhythm. *Science Advances*, *4*(2), 1–10. https://doi.org/10.1126/sciadv.aao3842

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*. https://doi.org/10.18637/jss.v067.i01

Berezutskaya, J., Freudenburg, Z. V., Güçlü, U., van Gerven, M. A. J., & Ramsey, N. F. (2017). Neural tuning to low-level features of speech throughout the perisylvian cortex. *Journal of Neuroscience*, *37*(33), 7906–7920. https://doi.org/10.1523/JNEUROSCI.0238-17.2017

Berezutskaya, J., Freudenburg, Z. V., Güçlü, U., van Gerven, M. A. J., & Ramsey, N. F. (2020). Brain-optimized extraction of complex sound features that drive continuous auditory perception. In *PLoS Computational Biology* (Vol. 16, Issue 7). https://doi.org/10.1371/journal.pcbi.1007992

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), 1–13. https://doi.org/10.1038/s41597-022-01173-0

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O., & Giard, M. H. (2008). Visual activation and audiovisual interactions in the auditory cortex during speech perception: Intracranial recordings in humans. *Journal of Neuroscience*, *28*(52), 14301–14310. https://doi.org/10.1523/JNEUROSCI.2875-08.2008

Broderick, M. P., Anderson, A. J., Di Liberto, G. M., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, *28*(5), 803-809.e3. https://doi.org/10.1016/j.cub.2018.01.080

Bulut, T. (2023). Domain-general and domain-specific functional networks of Broca's area underlying language processing. *Brain and Behavior*, *13*(7), 1–20. https://doi.org/10.1002/brb3.3046

Buzsaki, G. (2006). *Rhythms of the brain*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195301069.001.0001

Buzśaki, G., & Wang, X. J. (2012). Mechanisms of gamma oscillations. *Annual Review of Neuroscience*, *35*, 203–225. https://doi.org/10.1146/annurev-neuro-062111-150444

Canolty, R. T., Edwards, E., Dalal, S. S., Soltani, M., Nagarajan, S. S., Kirsch, H. E., Berger, M. S., Barbare, N. M., & Knight, R. T. (2006). High gamma power is phase-locked to theta oscillations in human neocortex. *Science*, *313*(5793), 1626–1628. https://doi.org/10.1126/science.1128115

Chalas, N., Daube, C., Kluger, D. S., Abbasi, O., Nitsch, R., & Gross, J. (2023). Speech onsets and sustained speech contribute differentially to delta and theta speech tracking in auditory cortex. *Cerebral Cortex*, *33*(10), 6273–6281. https://doi.org/10.1093/cercor/bhac502

Chan, M. M. Y., & Han, Y. M. Y. (2022). The Functional Brain Networks Activated by Music Listening:A Neuroimaging Meta-Analysis and Implications for Treatment. *Neuropsychology*, *36*(1), 4–22. https://doi.org/10.1037/neu0000777

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*(887).

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Di Liberto, G. M., Pelofi, C., Shamma, S., & de Cheveigné, A. (2020). Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening. *Acoustical Science and Technology*, *41*(1), 361–364. https://doi.org/10.1250/ast.41.361

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2015). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164. https://doi.org/10.1038/nn.4186

Ding, N., Patel, A. D., Chen, L., Butler, H., Luo, C., & Poeppel, D. (2017). Temporal modulations in speech and music. *Neuroscience and Biobehavioral Reviews*, *81*, 181–187. https://doi.org/10.1016/j.neubiorev.2017.02.011

Ding, Y., Zhang, Y., Zhou, W., Ling, Z., Huang, J., Hong, X. B., & Wang, X. X. (2019). Neural Correlates ofMusic Listening and Recall in the Human Brain. *The Journal of Neuroscience*, *39*(41), 8112–8123.

Doelling, K. B., Arnal, L. H., Ghitza, O., & Poeppel, D. (2014). Acoustic landmarks drive delta-theta oscillations to enable speech comprehension by facilitating perceptual parsing. *NeuroImage*, *85*, 761–768. https://doi.org/10.1016/j.neuroimage.2013.06.035

Doelling, K. B., Florencia Assaneo, M., Bevilacqua, D., Pesaran, B., & Poeppel, D. (2019). An oscillator model better predicts cortical entrainment to music. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(20), 10113–10121. https://doi.org/10.1073/pnas.1816414116

Doelling, K. B., & Poeppel, D. (2015). Cortical entrainment to music and its modulation by expertise. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(45), E6233–E6242. https://doi.org/10.1073/pnas.1508431112

Enge, A., Friederici, A. D., & Skeide, M. A. (2020). A meta-analysis of fMRI studies of language comprehension in children. *NeuroImage*, *215*(April), 116858. https://doi.org/10.1016/j.neuroimage.2020.116858

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise Martin. *KDD-96 Proceedings*, *2*, 565–580. https://doi.org/10.1016/B978-0-444-64165-6.03005-6

Fiveash, A., Bedoin, N., Gordon, R. L., & Tillmann, B. (2021). Processing Rhythm In Speech And Music: Shared Mechanisms And Implications For Developmental Speech And Language Disorders. *Neuropsychology*, *35*(8), 771–791.

https://doi.org/10.1037/neu0000766

Friederici, A. D. (2011). The brain basis of language processing: From structure to function. *Physiological Reviews*, *91*(4), 1357–1392. https://doi.org/10.1152/physrev.00006.2011

Fries, P. (2015). Rhythms for Cognition: Communication through Coherence. *Neuron*, *88*(1), 220–235. https://doi.org/10.1016/j.neuron.2015.09.034

Fries, P., Nikolić, D., & Singer, W. (2007). The gamma cycle. *Trends in Neurosciences*, *30*(7), 309–316. https://doi.org/10.1016/j.tins.2007.05.005

Fujisawa, S., & Buzsáki, G. (2011). A 4 Hz Oscillation Adaptively Synchronizes Prefrontal, VTA, and Hippocampal Activities. *Neuron*, *72*(1), 153–165. https://doi.org/10.1016/j.neuron.2011.08.018

Fujita, H., & Fujita, K. (2022). Human language evolution: a view from theoretical linguistics on how syntax and the lexicon first came into being. *Primates*, *63*(5), 403–415. https://doi.org/10.1007/s10329-021-00891-0

Fusi, S., Miller, E. K., & Rigotti, M. (2016). Why neurons mix: High dimensionality for higher cognition. *Current Opinion in Neurobiology*, *37*, 66–74. https://doi.org/10.1016/j.conb.2016.01.010

Green, D. M., & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley & Sons, Ltd.

Hamilton, L. S., & Huth, A. G. (2020). The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *35*(5), 573–582. https://doi.org/10.1080/23273798.2018.1499946

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

He, Y., Gebhardt, H., Steines, M., Sammer, G., Kircher, T., Nagels, A., & Straube, B. (2015). The EEG and fMRI signatures of neural integration: An investigation of meaningful gestures and corresponding speech. *Neuropsychologia*, *72*, 27–42. https://doi.org/10.1016/j.neuropsychologia.2015.04.018

Herff, S. A., Herff, C., Milne, A. J., Johnson, G. D., Shih, J. J., & Krusienski, D. J. (2020). Prefrontal high gamma in ecog tags periodicity of musical rhythms in perception and imagination. *ENeuro*, *7*(4), 1–11. https://doi.org/10.1523/ENEURO.0413-19.2020

Hyafil, A., Fontolan, L., Kabdebon, C., Gutkin, B., & Giraud, A. L. (2015). Speech encoding by coupled cortical theta and gamma oscillations. *ELife*, *4*(MAY), 1–45. https://doi.org/10.7554/eLife.06213

Kaneshiro, B., Nguyen, D. T., Norcia, A. M., Dmochowski, J. P., & Berger, J. (2020). Natural music evokes correlated EEG responses reflecting temporal structure and beat. *NeuroImage*, *214*(December 2019), 116559. https://doi.org/10.1016/j.neuroimage.2020.116559

Kern, P., Heilbron, M., de Lange, F. P., & Spaak, E. (2022). Cortical activity during naturalistic music listening reflects short-range predictions based on long-term experience. *ELife*, *11*, 1–31. https://doi.org/10.7554/ELIFE.80935

Kubanek, J., Brunner, P., Gunduz, A., Poeppel, D., & Schalk, G. (2013). The Tracking of Speech Envelope in the Human Cortex. *PLoS ONE*, *8*(1).

https://doi.org/10.1371/journal.pone.0053398

Kulasingham, J. P., Brodbeck, C., Presacco, A., Kuchinsky, S. E., Anderson, S., & Simon, J. Z. (2020). High gamma cortical processing of continuous speech in younger and older listeners. *NeuroImage*, *222*(August), 117291. https://doi.org/10.1016/j.neuroimage.2020.117291

Kumagai, Y., Arvaneh, M., & Tanaka, T. (2017). Familiarity affects entrainment of EEG in music listening. *Frontiers in Human Neuroscience*, *11*(July), 1–8. https://doi.org/10.3389/fnhum.2017.00384

Lankinen, K., Ahveninen, J., Uluç, I., Daneshzand, M., Mareyam, A., Kirsch, J. E., Polimeni, J. R., Healy, B. C., Tian, Q., Khan, S., Nummenmaa, A., Wang, Q. M., Green, J. R., Kimberley, T. J., & Li, S. (2023). Role of articulatory motor networks in perceptual categorization of speech signals: a 7T fMRI study. *Cerebral Cortex*, *33*(24), 11517–11525. https://doi.org/10.1093/cercor/bhad384

Lerdahl, F., & Jackendoff, R. (1983). An overview of hierarchical structure in music. *Music Perception: An Interdisciplinary Journal*, *1*(2), 229–252.

Liu, Y., Liu, G., Wei, D., Li, Q., Yuan, G., Wu, S., Wang, G., & Zhao, X. (2018). Effects of musical tempo on musicians' and non-musicians' emotional experience when listening to music. *Frontiers in Psychology*, *9*(NOV), 1–11. https://doi.org/10.3389/fpsyg.2018.02118

Lizarazu, M., Carreiras, M., & Molinaro, N. (2023). Theta-gamma phase-amplitude coupling in auditory cortex is modulated by language proficiency. *Human Brain Mapping*, *44*(7), 2862–2872. https://doi.org/10.1002/hbm.26250

Lizarazu, M., Lallier, M., & Molinaro, N. (2019). Phase−amplitude coupling between theta and gamma oscillations adapts to speech rate. *Annals of the New York Academy of Sciences*, *1453*(1), 140–152. https://doi.org/10.1111/nyas.14099

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*(MAY), 1–10. https://doi.org/10.3389/fpsyg.2012.00170

McFee, B., Nieto, O., Farbood, M. M., & Bello, J. P. (2017). Evaluating hierarchical structure in music annotations. *Frontiers in Psychology*, *8*(AUG), 1–17. https://doi.org/10.3389/fpsyg.2017.01337

Mégevand, P., Mercier, M. R., Groppe, D. M., Golumbic, E. Z., Mesgarani, N., Beauchamp, M. S., Schroeder, C. E., & Mehta, A. D. (2020). Crossmodal phase reset and evoked responses provide complementary mechanisms for the influence of visual speech in auditory cortex. *Journal of Neuroscience*, *40*(44), 8530–8542. https://doi.org/10.1523/JNEUROSCI.0555-20.2020

Merchant, H., Grahn, J., Trainor, L., Rohrmeier, M., & Fitch, W. T. (2015). Finding the beat: A neural perspective across humans and non-human primates. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1664). https://doi.org/10.1098/rstb.2014.0093

Merrill, J., Sammler, D., Bangert, M., Goldhahn, D., Lohmann, G., Turner, R., & Friederici, A. D. (2012). Perception of words and pitch patterns in song and speech. *Frontiers in Psychology*, *3*(MAR), 1–13. https://doi.org/10.3389/fpsyg.2012.00076

Miyagawa, S., Berwick, R. C., & Okanoya, K. (2013). The emergence of hierarchical

structure in human language. *Frontiers in Psychology*, *4*(February), 1–6. https://doi.org/10.3389/fpsyg.2013.00071

Nozaradan, S. (2014). Exploring how musical rhythm entrains brain activity with electroencephalogram frequency-tagging. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1658). https://doi.org/10.1098/rstb.2013.0393

Nozaradan, S., Peretz, I., & Mouraux, A. (2012). Selective neuronal entrainment to the beat and meter embedded in a musical rhythm. *Journal of Neuroscience*, *32*(49), 17572–17581. https://doi.org/10.1523/JNEUROSCI.3203-12.2012

Ohnishi, T., Matsuda, H., Asada, T., Aruga, M., Hirakata, M., Nishikawa, M., Katoh, A., & Imabayashi, E. (2001). Functional anatomy of musical perception in musicians. *Cerebral Cortex*, *11*(8), 754–760. https://doi.org/10.1093/cercor/11.8.754

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 156869. https://doi.org/10.1155/2011/156869

Osorio, S., Straube, B., Meyer, L., & He, Y. (2024). The role of co-speech gestures in retrieval and prediction during naturalistic multimodal narrative processing. *Language, Cognition and Neuroscience*, *39*(3), 367–382. https://doi.org/10.1080/23273798.2023.2295499

Park, H., Ince, R. A. A., Schyns, P. G., Thut, G., & Gross, J. (2018). Representational interactions during audiovisual speech entrainment: Redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. *PLoS Biology*, *16*(8), 1–26.

https://doi.org/10.1371/journal.pbio.2006558

Park, H., Thut, G., & Gross, J. (2020). Predictive entrainment of natural speech through two fronto-motor top-down channels. *Language, Cognition and Neuroscience*, *35*(6), 739–751. https://doi.org/10.1080/23273798.2018.1506589

Patel, A. D., & Morgan, E. (2017). Exploring cognitive relations between prediction in language and music. *Cognitive Science*, *41*, 303–320. https://doi.org/10.1111/cogs.12411

Peelle, J. E., Gross, J., & Davis, M. H. (2013). Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral Cortex*, *23*(6), 1378–1387. https://doi.org/10.1093/cercor/bhs118

Pellegrino, F., Coupe, C., & Marisco, E. (2011). A cross-language perspective on speech information rate. *Language*, *87*(3), 539–558. https://doi.org/10.2307/23011654

Perrachione, T. K., Ghosh, S. S., Ostrovskaya, I., Gabrieli, J. D. E., & Kovelman, I. (2017). Phonological working memory for words and nonwords in cerebral cortex. *Journal of Speech, Language, and Hearing Research*, *60*(7), 1959–1979. https://doi.org/10.1044/2017_JSLHR-L-15-0446

Poeppel, D., & Assaneo, M. F. (2020). Speech rhythms and their neural foundations. *Nature Reviews Neuroscience*, *21*(6), 322–334. https://doi.org/10.1038/s41583-020-0304-4

Poikonen, H., Toiviainen, P., & Tervaniemi, M. (2018). Naturalistic music and dance: Cortical phase synchrony in musicians and dancers. *PLoS ONE*, *13*(4), 1–18. https://doi.org/10.1371/journal.pone.0196065

Quiroga-Martinez, D. R., C. Hansen, N., Højlund, A., Pearce, M., Brattico, E., & Vuust, P. (2020). Musical prediction error responses similarly reduced by predictive uncertainty in musicians and non-musicians. *European Journal of Neuroscience*, *51*(11), 2250–2269. https://doi.org/10.1111/ejn.14667

Ray, S., Crone, N. E., Niebur, E., Franaszczuk, P. J., & Hsiao, S. S. (2008). Neural correlates of high-gamma oscillations (60-200 Hz) in macaque local field potentials and their potential implications in electrocorticography. *Journal of Neuroscience*, *28*(45), 11526–11536. https://doi.org/10.1523/JNEUROSCI.2848-08.2008

Ray, S., Niebur, E., Hsiao, S. S., Sinai, A., & Crone, N. E. (2008). High-frequency gamma activity (80-150 Hz) is increased in human cortex during selective attention. *Clinical Neurophysiology*, *119*(1), 116–133. https://doi.org/10.1016/j.clinph.2007.09.136

Rieder, M. K., Rahm, B., Williams, J. D., & Kaiser, J. (2011). Human gamma-band activity and behavior. *International Journal of Psychophysiology*, *79*(1). https://doi.org/10.1016/j.ijpsycho.2010.08.010

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. https://doi.org/10.1038/nature12160

Rohrmeier, M. A., & Koelsch, S. (2012). Predictive information processing in music cognition. A critical review. *International Journal of Psychophysiology*, *83*(2), 164–175. https://doi.org/10.1016/j.ijpsycho.2011.12.010

Schön, D., & Morillon, B. (2019). Music and language. In *The Oxford handbook of music and the brain* (pp. 391–416).

Simon, A., Loquet, G., Ostergaard, J., & Bech, S. (2023). Cortical auditory attention decoding during music and speech listening. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *31*, 2903–2911. https://doi.org/10.1109/TNSRE.2023.3291239

Sirota, A., & Buzsáki, G. (2005). Interaction between neocortical and hippocampal networks via slow oscillations. *Halamus & Related Systems*, *3*(4), 245–259.

Sturm, I., Biankertz, B., Potes, C., Schaik, G., & Curio, G. (2014). ECoG high gamma activity reveals distinct cortical representations of lyrics passages, Harmonic and timbre-related changes in a rock song. *Frontiers in Human Neuroscience*, *8*(OCT), 1–14. https://doi.org/10.3389/fnhum.2014.00798

Sturm, I., Dähne, S., Blankertz, B., & Curio, G. (2015). Multi-variate EEG analysis as a novel tool to examine brain responses to naturalistic music stimuli. *PLoS ONE*, *10*(10), 1–30. https://doi.org/10.1371/journal.pone.0141281

Synigal, S. R., Teoh, E. S., & Lalor, E. C. (2020). Including measures of high gamma power can improve the decoding of natural speech From EEG. *Frontiers in Human Neuroscience*, *14*(April), 1–12. https://doi.org/10.3389/fnhum.2020.00130

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*. https://doi.org/10.1155/2011/879716

Te Rietmolen, N., Mercier, M., Trebuchon, A., Morillon, B., & Schon, D. (2024). Speech and music recruit frequency-specific distributed and overlapping cortical networks. *ELife*, 2022.10.08.511398.

https://www.biorxiv.org/content/10.1101/2022.10.08.511398v1%0Ahttps://www.biorxiv.org/content/10.1101/2022.10.08.511398v1.abstract

Teng, X., Larrouy-Maestri, P., & Poeppel, D. (2021). Segmenting and predicting musical phrase structure exploits neural gain modulation and phase precession. *BioRxiv*, 2021.07.15.452556. https://doi.org/10.1101/2021.07.15.452556

Teng, X., Ma, M., Yang, J., Blohm, S., Cai, Q., & Tian, X. (2020). Constrained structure of ancient chinese poetry facilitates speech content grouping. *Current Biology*, *30*(7), 1299-1305.e7. https://doi.org/10.1016/j.cub.2020.01.059

Tillmann, B. (2012). Music and language perception: Expectations, structural integration, and cognitive sequencing. *Topics in Cognitive Science*, *4*(4), 568–584. https://doi.org/10.1111/j.1756-8765.2012.01209.x

Tremblay, P., & Small, S. L. (2011). On the context-dependent nature of the contribution of the ventral premotor cortex to speech perception. *NeuroImage*, *57*(4), 1561–1571. https://doi.org/10.1016/j.neuroimage.2011.05.067

Tye, K. M., Miller, E. K., Taschbach, F. H., Benna, M. K., Rigotti, M., & Fusi, S. (2024). Mixed selectivity: Cellular computations for complexity. *Neuron*, 1–15. https://doi.org/10.1016/j.neuron.2024.04.017

van Noorden, L., & Moelants, D. (1999). Resonance in the perception of musical pulse. *International Journal of Phytoremediation*, *21*(1), 43–66. https://doi.org/10.1076/jnmr.28.1.43.3122

Vanthornhout, J., Decruy, L., & Francart, T. (2019). Effect of task and attention on neural tracking of speech. *Frontiers in Neuroscience*, *13*(September), 1–11.

https://doi.org/10.3389/fnins.2019.00977

Viswanathan, V., Bharadwaj, H. M., & Shinn-Cunningham, B. G. (2019). Electroencephalographic signatures of the neural representation of speech during selective attention. *ENeuro*, *6*(5), 1–14. https://doi.org/10.1523/ENEURO.0057-19.2019

Willems, R. M., Frank, S. L., Nijhof, A. D., Hagoort, P., & Van Den Bosch, A. (2016). Prediction during natural language comprehension. *Cerebral Cortex*, *26*(6), 2506–2516. https://doi.org/10.1093/cercor/bhv075

Willems, R. M., Nastase, S. A., & Milivojevic, B. (2020). Narratives for neuroscience. *Trends in Neurosciences*, *43*(5), 271–273. https://doi.org/10.1016/j.tins.2020.03.003

Wollman, I., Arias, P., Aucouturier, J. J., & Morillon, B. (2020). Neural entrainment to music is sensitive to melodic spectral complexity. *Journal of Neurophysiology*, *123*(3), 1063–1071. https://doi.org/10.1152/JN.00758.2018

Zion Golumbic, E. M., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., Goodman, R. R., Emerson, R., Mehta, A. D., Simon, J. Z., Poeppel, D., & Schroeder, C. E. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party." *Neuron*, *77*(5), 980–991. https://doi.org/10.1016/j.neuron.2012.12.037

Zou, J., Xu, C., Luo, C., Jin, P., Gao, J., Li, J., Gao, J., Ding, N., & Luo, B. (2021). Θ-Band cortical tracking of the speech envelope shows the linear phase property. *ENeuro*, *8*(4), 1–7. https://doi.org/10.1523/ENEURO.0058-21.2021