Speech and music recruit frequency-specific distributed and overlapping cortical networks

Noémie te Rietmolen[1,2,*], Manuel Mercier[2], Agnès Trébuchon[1,2,3], Benjamin Morillon[1,2,4,*], Daniele Schön[1,2,4,*]

[1] Institute for Language, Communication, and the Brain, Aix-Marseille Université, Marseille, France

[2] Aix Marseille Université, INSERM, INS, Institut de Neurosciences des Systèmes, Marseille, France

[3] APHM, Hôpital de la Timone, Service de Neurophysiologie Clinique, Marseille, France

[4] Senior authorship

* corresponding authors: noemieter@gmail.com; daniele.schon@univ-amu.fr; bnmorillon@gmail.com

Keywords: auditory neurosciences; network dynamics; language; neurophysiology; cognition

Conflict of interests: The authors declare no competing interests.

Acknowledgments: We thank all patients for their willingful participation. We thank Patrick Marquis for helping with the data acquisition, and Anne-Catherine Tomei and all colleagues from the Institut de Neuroscience des Systèmes for useful discussions.

Funding sources: ANR-20-CE28-0007-01 (to B.M), ANR-21-CE28-0010 (to D.S), ANR-17-EURE-0029 (NeuroMarseille), and co-funded by the European Union (ERC, SPEEDY, ERC-CoG-101043344). This work, carried out within the Institute of Convergence ILCB, was also supported by grants from France 2030 (ANR-16-CONV-0002), the French government under the Programme «Investissements d'Avenir», and the Excellence Initiative of Aix-Marseille University (A*MIDEX, AMX-19-IET-004).

Author contributions: B.M. and D.S. designed research. A.T. and M.M. acquired data; N.t.R., M.M., B.M. and D.S. analyzed data; N.t.R., B.M., M.M., and D.S. wrote the paper; N.t.R., A.T., M.M., B.M., and D.S. edited the paper.

Data availability statement: The conditions of our ethics approval do not permit public archiving of anonymised study data. Readers seeking access to the data should contact Dr. Daniele Schön (daniele.schon@univ-amu.fr). Access will be granted to named individuals in accordance with ethical procedures governing the reuse of clinical data, including completion of a formal data sharing agreement.

Code availability statement: Data analyses were performed using custom scripts in Python, available on Github: github.com/noemietr/iSpeech

Abstract

To what extent does speech and music processing rely on domain-specific and domain-general neural networks? Using whole-brain intracranial EEG recordings in 18 epilepsy patients listening to natural, continuous speech or music, we investigated the presence of frequency-specific and network-level brain activity. We combined it with a statistical approach in which a clear distinction is made between *shared*, *preferred,* and domain-*selective* neural responses. We show that the majority of focal and network-level neural activity is shared between speech and music processing. Our data also reveal an absence of anatomical regional selectivity. Instead, domain-selective neural responses are restricted to distributed and frequency-specific coherent oscillations, typical of spectral fingerprints. Our work highlights the importance of considering natural stimuli and brain dynamics in their full complexity to map cognitive and brain functions.

Introduction

The advent of neuroscience continues the longstanding debate on the origins of music and language -that fascinated Rousseau and Darwin (Kivy, 1959; Rousseau, 2009)- on new biological ground: evidence for the existence of selective and/or shared neural populations involved in their processing. The question on functional selectivity versus domain-general mechanisms is closely related to the question of the nature of the neural code: Are representations *sparse* (and localized) or *distributed*? While the former allows to explicitly represent any stimulus in a small number of neurons, it would require an intractable number of neurons to represent all possible stimuli. Experimental evidence instead suggests that stimulus identification is achieved through a population code, implemented by neural coupling in a distributed dynamical system (Bizley & Cohen, 2013; Rissman & Wagner, 2012). The question of the nature of the neural code has tremendous implications: it defines an epistemological posture on how to map cognitive and brain functions. This, in turn, affects both the definition of cognitive operations - what is actually computed - as well as the way we look at the data - looking for differences or similarities.

Neuroimaging studies report mixed evidence of selectivity and resource sharing. On one hand, one can find claims for a clear distinction between brain regions dedicated to language versus other cognitive processes (Chen et al., 2023; Fedorenko et al., 2011; Fedorenko & Blank, 2020; Friederici, 2020) and for the existence of specific and separate neural populations for speech, music, and song (Boebinger et al., 2021; Norman-Haignere et al., 2022). On the other hand, other neuroimaging studies suggest that the brain regions that support language and speech also support nonlinguistic functions (Albouy et al., 2020; Fadiga et al., 2009; Koelsch, 2011; Menon et al., 2002; Robert et al., 2023; Schön et al., 2010). This point is often put forward when interpreting the positive impact music training can have on different levels of speech and language processing (Flaugnacco et al., 2015; François et al., 2013; Kraus & Chandrasekaran, 2010; Schön et al., 2004).

Several elements may account for these different findings. The very first may rely on the definition of a brain region. This can be considered as a set of functionally homogeneous but spatially distributed voxels, or, alternatively, as an anatomical landmark as those used in brain atlases (e.g. inferior frontal gyrus). However, observing functional regional selectivity in a distributed pattern is not incompatible with the observation of an absence of anatomical regional selectivity: a selective set of voxels may exist within an anatomically non-selective region. A second element concerns the choice of the stimuli. Some of the studies claiming functional selectivity used rather short auditory stimuli (Boebinger et al., 2021; Norman-Haignere et al., 2015; Norman-Haignere et al., 2022). Besides the low

ecological validity of such stimuli that may reduce the generalizability of the findings (Theunissen et al., 2000), their comparison further relies on the assumption that speech and music share similar cognitive time constants. However, speech unfolds faster than music (Ding et al., 2017), and while a linguistic phrase is typically shorter than a second (Inbar et al., 2020), a melodic phrase is an order of magnitude longer. Moreover, balancing the complexity/simplicity of linguistic and musical stimuli can be challenging, and musical stimuli are often reduced to very simple melodies played on a synthesizer. These simple melodies mainly induce pitch processing in associative auditory regions (Griffiths et al., 2010) but do not recruit the entire dual-stream auditory pathways (Zatorre et al., 2007). Overall, while short and simple stimuli may be sufficient to induce linguistic processing, they might not be cognitively relevant musical stimuli. Finally, another element concerns the data at stake. Most studies that compared language and music processing, examined functional MRI data (Chen et al., 2023; Fedorenko et al., 2011; Nieto-Castañón & Fedorenko, 2012). Here, we would like to consider cognition as resulting from interactions among functionally specialized but widely distributed brain networks and adopt an approach in which large-scale and frequency-specific neural dynamics are characterized. This approach rests on the idea that the canonical computations that underlie cognition are bound to spectral fingerprints consisting of network- and frequency-specific coherent oscillations (Siegel et al., 2012). This framework requires relying on time-resolved neurophysiological recordings (M/EEG) and —rather than focusing only on the amplitude of the high-frequency activity (HF), a common approach in the literature involving human intracranial EEG recordings (Martin et al., 2019; Norman-Haignere et al., 2022; Oganian & Chang, 2019)— to investigate the entire frequency spectrum of neural activity. Indeed, while HF amplitude is a good proxy of focal neural spiking (Le Van Quyen et al., 2010; Ray & Maunsell, 2011), large-scale neuronal interactions mainly rely on slower dynamics (Kayser et al., 2012; Kopell et al., 2000; Siegel et al., 2012).

Following the reasoning developed above, we suggest that the study of selectivity of music and language processing should carefully consider the following points: First, the use of ecologically valid stimuli, both in terms of content and duration. Second, a within-subject approach comparing both conditions. Third, aiming for high spatial sensitivity. Fourth, considering not only one type of neural activity (broadband, HF amplitude…) but the entire frequency spectrum of the neurophysiological signal. Fifth, use a broad range of complementary analyses, including connectivity, and take into account individual variability. Finally, we suggest that a clear distinction should be made between shared, selective, and preferred activity. That is, be *A* and *B* two investigated cognitive functions, “shared” would be a neural population that significantly and equally contributes to the processing of both *A* and *B*; “selective” would be a neural population that significantly contributes to the processing of *A* but not of *B*; “preferred” would be a neural population that significantly contributes to the processing of both *A* and *B*, but more prominently for *A* or *B* (Figure 1A).

In an effort to take into account all the above challenges and to precisely quantify the degree of shared, preferred, and selective responses both at the levels of the channels and anatomical regions (Figure 1C-D), we conducted an experiment on 18 pharmacoresistant epileptic patients explored with stereotactic EEG (sEEG) electrodes. Patients listened to long and ecological audio-recordings of speech and music (10-minutes each). We investigated stimulus encoding, spectral content of the neural activity, and brain connectivity over the entire frequency spectrum (from 1-120 Hz; i.e. delta band to HF). Finally, we carefully distinguished between the three different categories of neural responses described above: shared, selective, and preferred across the two investigated cognitive domains. Our results reveal that the majority of neural responses are shared between natural speech and music, and they highlight an absence of anatomical regional selectivity. Instead, we found neural selectivity to be restricted to distributed and frequency-specific coherent oscillations, typical of spectral fingerprints.

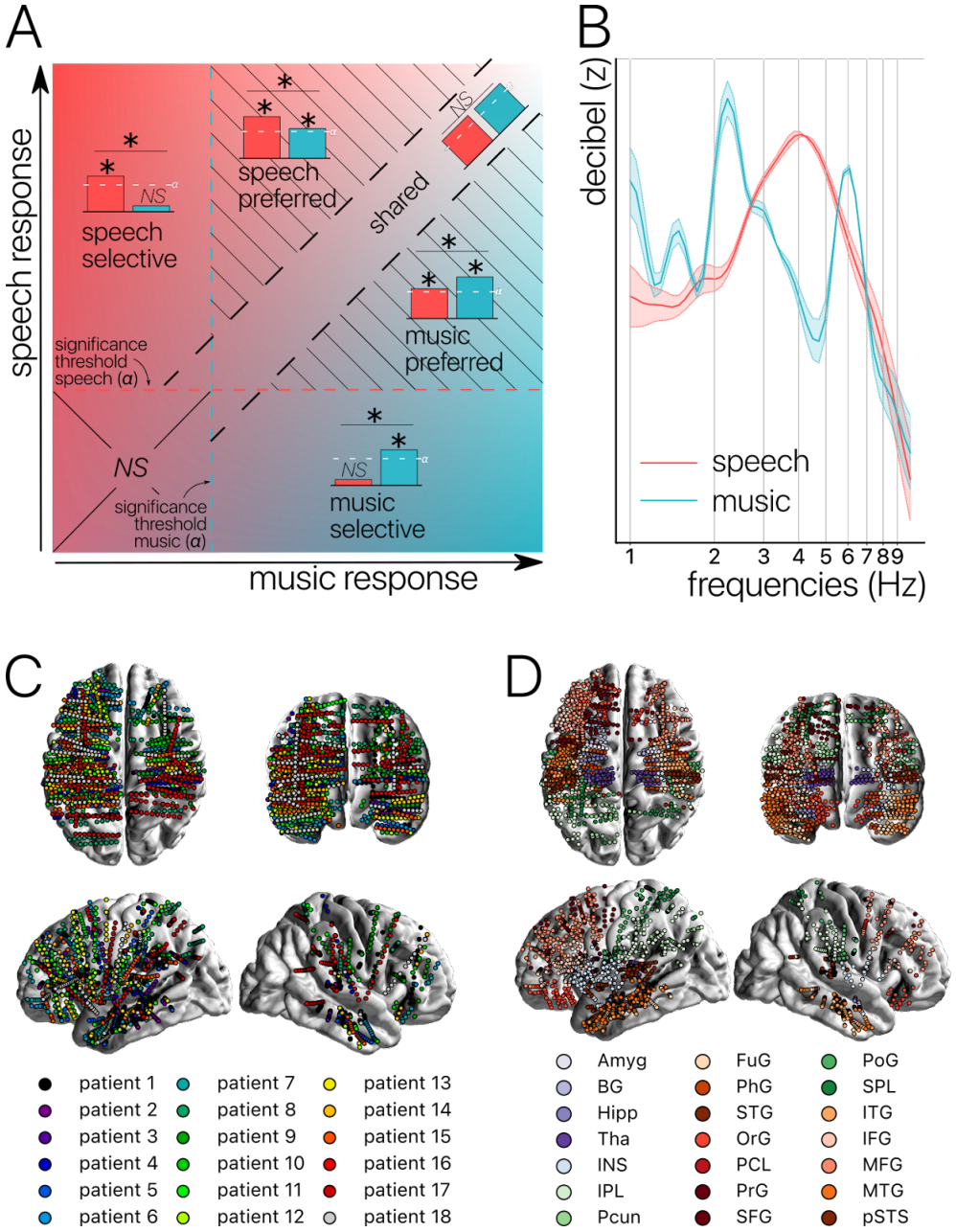

Figure 1. Concepts, Stimuli and Recordings. ***(A) Conceptual definition of selective, preferred and shared neural processes***. *Illustration of the continua between speech and music selectivity, speech and music preference and shared resources. "Selective" responses are neural responses significant for one domain but not the other, and with a significant difference between domains (for speech top-left; for music bottom-right). "Preferred" responses correspond to neural responses that occur during both speech and music processing, but with a significantly stronger response for one domain over the other (striped triangles). Finally, "shared" responses occur when there are no significant differences between domains, and there is a significant neural response to at least one of the two stimuli (visible along the diagonal). If neither domain produces a significant neural response, the difference is not assessed (lower left square).* ***(B) Stimuli.*** *Modulation spectrum of the acoustic temporal envelope of the continuous, 10-minutes long speech and music stimuli.* ***(C) Anatomical localization of the sEEG electrodes for each patient*** *(N=18).* ***(D) Anatomical localization of the sEEG electrodes for each anatomical region***. *Abbreviations according to the Brainnetome Atlas (Fan et al., 2016).*

Results

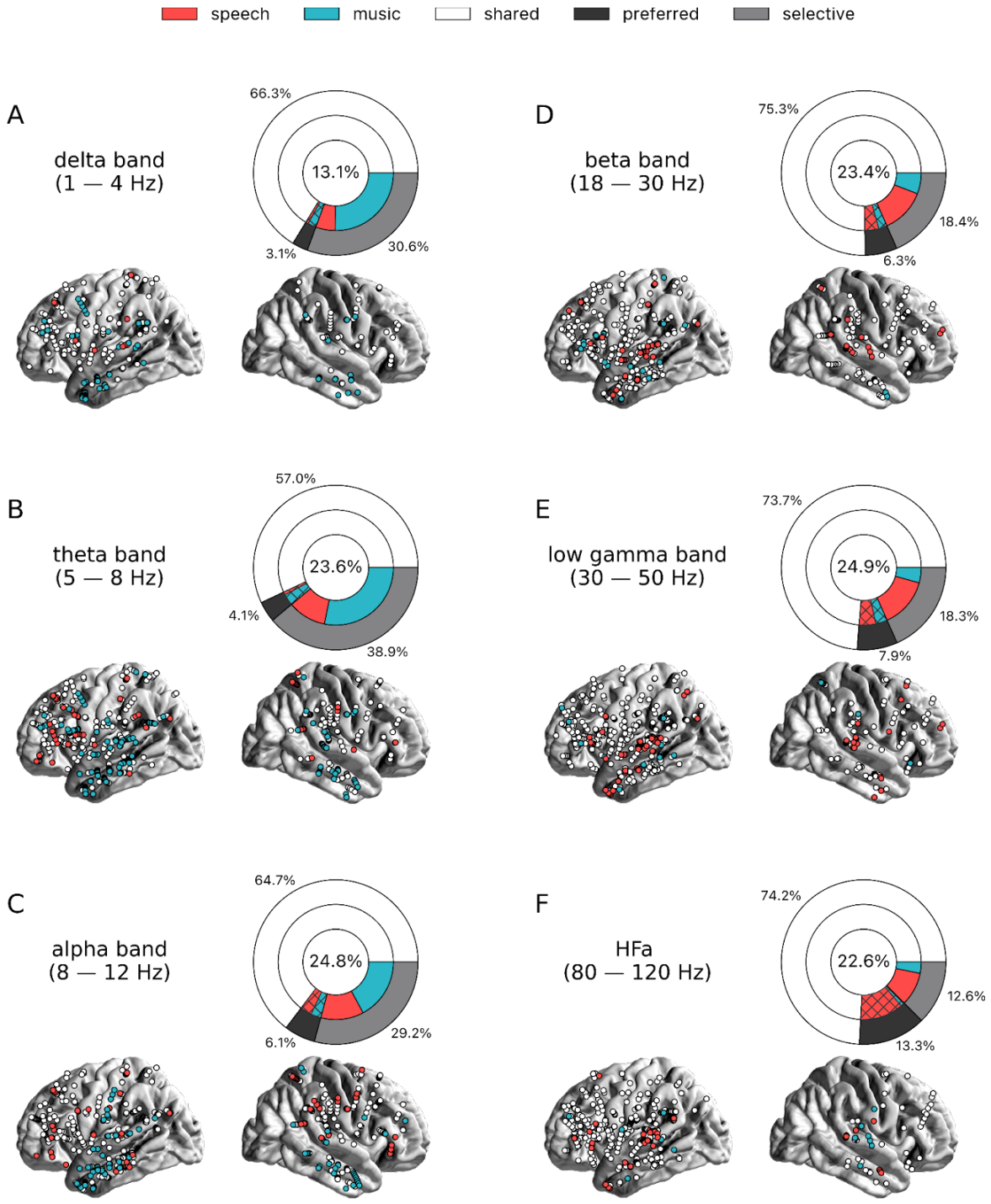

Figure 2. Power spectrum analyses of activations (speech or music > tones). ***(A)-(F)*** *Neural responses to speech and/or music for the 6 canonical frequency bands. Only significant activations compared to the baseline condition (pure tones listening) are reported (see Figure S8 for non-thresholded results).* ***Nested pie charts indicate****: (1) in the center, the percentage of channels that showed a significant response to speech and/or music. (2) The outer pie indicates the percentage of channels, relative to the center, classified as shared (white), selective (light gray), and preferred (dark gray). (3) The inner pie indicates, for the selective (plain) and preferred (pattern) categories, the proportion of channels that were (more) responsive to speech (red) or*

*music (blue). **Brain plots indicate**: Distribution of shared (white) and selective (red/blue) sEEG channels projected on the brain surface. Results are significant at α < 0.01 (N=18).*

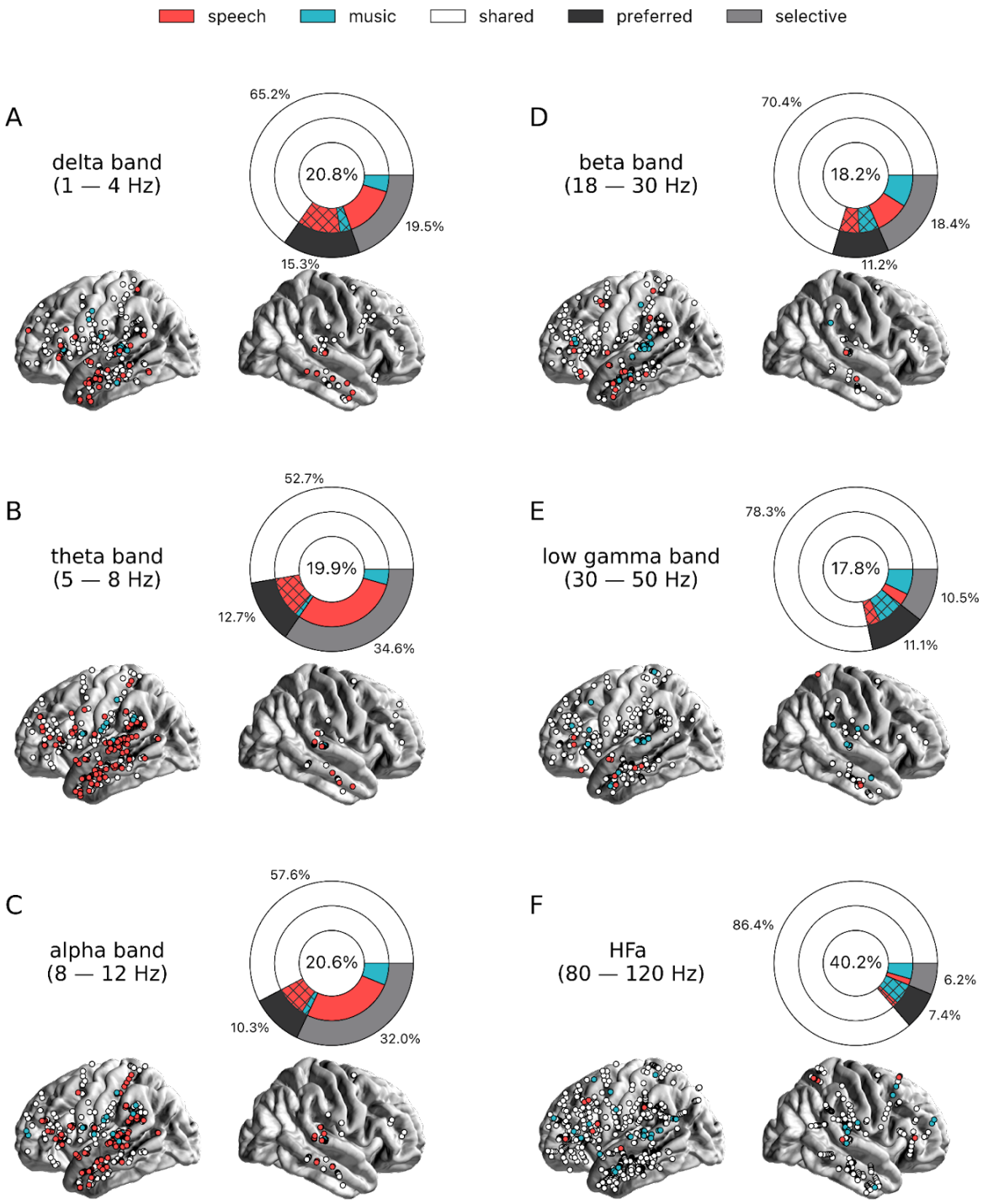

***Figure 3. Power spectrum analyses of deactivations (speech or music < tones)**. **(A)-(F)** Neural responses to speech and/or music for the 6 canonical frequency bands. Only significant deactivations compared to the baseline condition (pure tones listening) are reported. Same conventions as in Figure 2. Results are significant at α < 0.01 (N=18).*

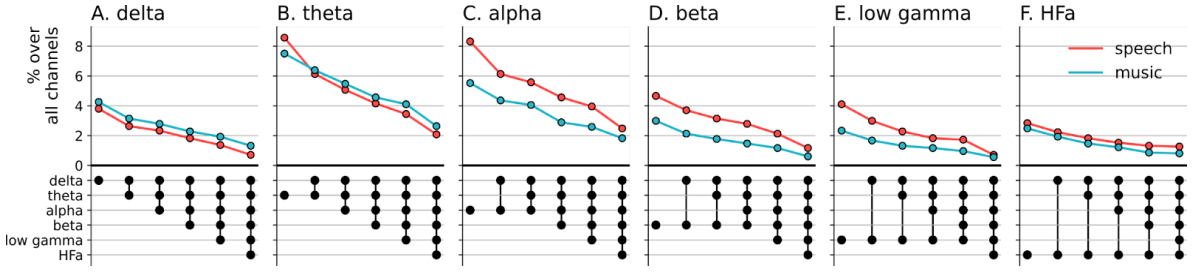

Figure 4. Cross-frequency channel selectivity for the power spectrum analyses. *Percentage of channels that selectively respond to speech (red) or music (blue) across different frequency bands. For each plot, the first (leftmost) value corresponds to the percentage (%) of channels displaying a selective response in a specific frequency band (either activation or deactivation, compared to the baseline condition of pure tones listening). In the next value, we remove the channels that are significantly responsive in the other domain for the following frequency band (e.g. in panel A: speech selective in delta; speech selective in delta XOR music responsive in theta; speech selective in delta XOR music responsive in theta XOR music responsive in alpha; and so forth). The black dots at the bottom of the graph indicate which frequency bands were successively included in the analysis. Note that channels remaining selective across frequency bands did not necessarily respond selectively in every band. They simply never showed a significant response to the other domain in the other bands.*

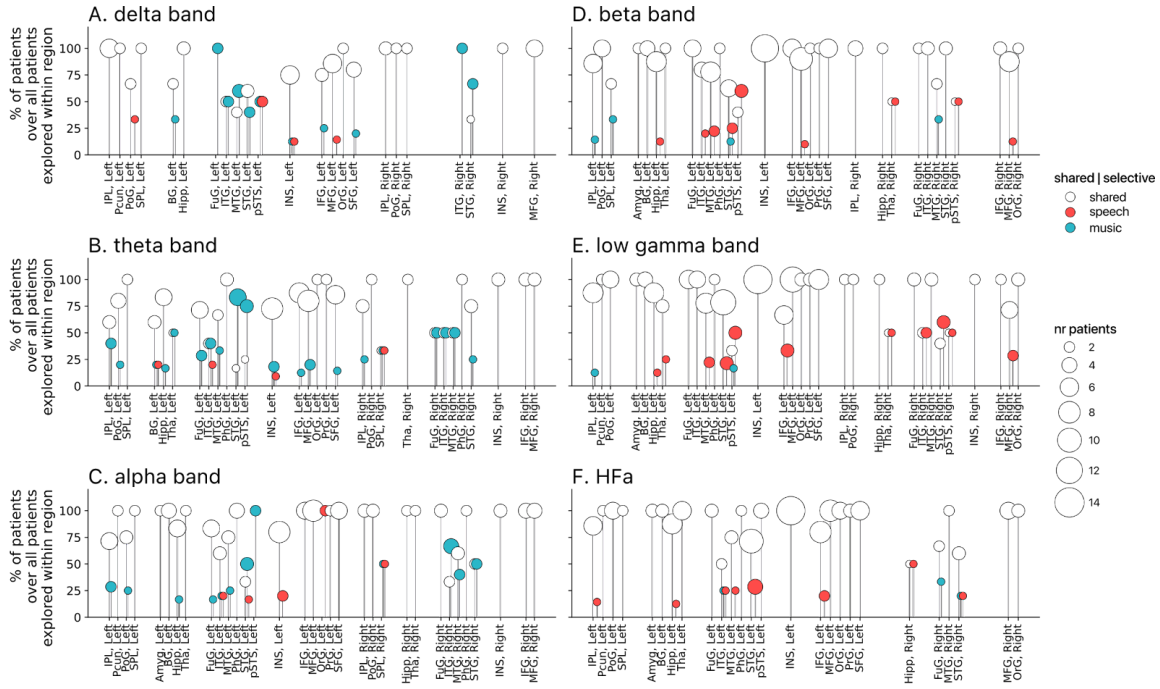

Figure 5. Population prevalence for the power spectral analyses of activations (speech or music > tones; N = 18). (A)-(F) *Population prevalence of shared or selective responses for the 6 canonical frequency bands, per anatomical region (note that preferred responses are excluded). Only significant activations compared to the baseline condition (pure tones listening) are reported. Regions (on the x-axis) were included in the analyses if they had been explored by*

minimally 2 patients with minimally 2 significant channels. Patients were considered to show regional selective processing when all their channels in a given region responded selectively to either speech (red) or music (blue). When regions contained a combination of channels with speech selective, music selective or shared responses, the patient was considered to show shared (white) processing in this region. The height of the lollipop (y-axis) indicates the percentage of patients over the total number of explored patients in that given region. The size of the lollipop indicates the number of patients. As an example, in panel F (HFa band), most lollipops are white with a height of 100%, indicating that, in these regions, all patients presented a shared response profile. However, in the left inferior parietal lobule (IPL, Left) one patient (out of the 7 explored) shows speech selective processing (filled red circle). A fully selective region would thus show a fixed-color full height across all frequency bands. Abbreviations according to the Brainnetome Atlas (Fan et al., 2016).

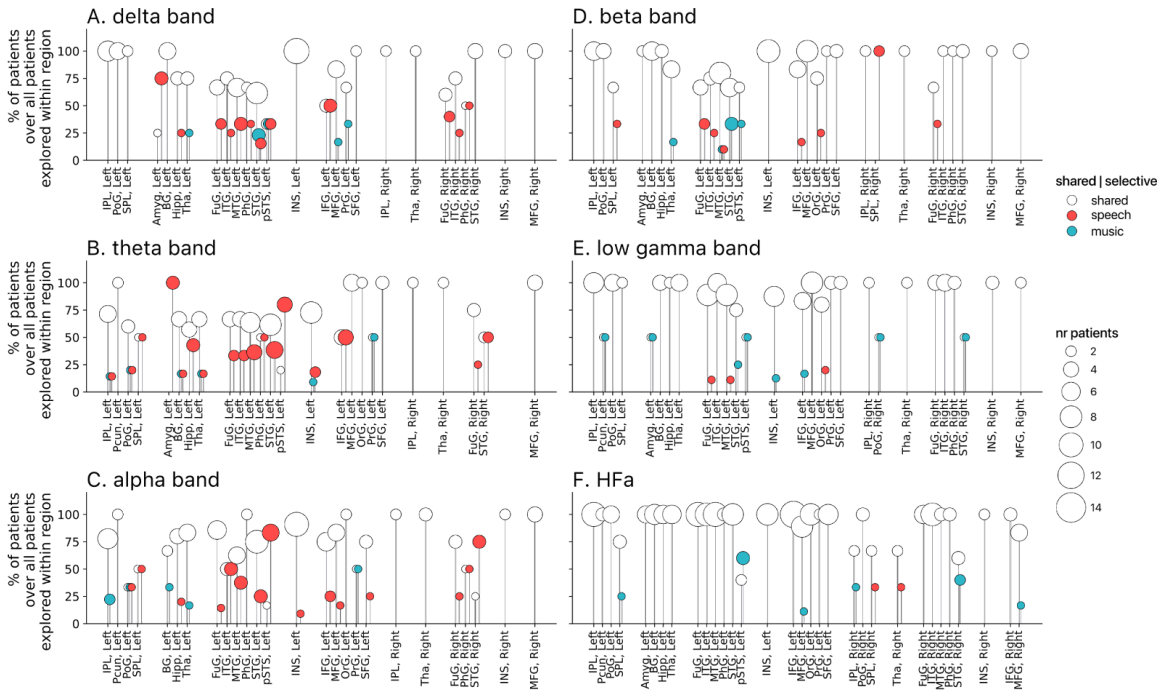

***Figure 6**. **Population prevalence for the power spectral analyses of deactivations (speech or music < tones; N = 18)**. **(A)-(F)** Population prevalence of shared or selective responses for the 6 canonical frequency bands, per anatomical region. Only significant deactivations compared to the baseline condition (pure tones listening) are reported. Same conventions as in Figure 5.*

Anatomical regional neural activity is mostly non-domain selective to speech or music.

To investigate the presence of domain-selectivity during ecological perception of speech and music, we first analyzed the neural responses to these two cognitive domains in both a spatially and spectrally resolved manner, with respect to two baseline conditions: one in which patients listened to pure tones, the other in which patients listened to isolated syllables (see Methods). Here we will report the results using pure tones data as baseline, but note that the results using syllables data as baseline were highly similar (see Figures S1-5). We classified, for each canonical frequency band, each channel into one of the categories mentioned above, i.e. shared, selective, or preferred (Figure

1A), by examining whether speech and/or music differ from baseline and whether they differ from each other. We also considered both activations and deactivations, compared to baseline, as both index a modulation of neural population activity, and have both been linked with cognitive processes (Pfurtscheller & Lopes da Silva, 1999; Proix et al., 2022). Overall, neural responses are predominantly shared between the two domains, accounting for ~70% of the channels which showed a significant response compared to baseline (Figures 2-3). The preferred category is also systematically present, accounting for 3 to 15% of significant neural responses, across frequency bands. Selective responses are more present in the lower frequency bands (~30% up to the alpha band), and quite marginal in the HFa band (6-12%).

The spatial distribution of the spectrally-resolved responses corresponds to the network that is typically involved in speech and music perception and includes both ventral and dorsal auditory pathways, well beyond the auditory cortex activity that is accounted for by the baseline conditions. This is the case for overall responses but also when only looking at shared responses. For instance, HFa shared responses represent 74-86% of the overall significant HFa responses, and are visible in the left superior and middle temporal gyri, inferior parietal lobule, and the precentral, middle and inferior frontal gyri (Figures 2F & 3F). The left hemisphere appears to be more strongly involved, but this result is biased by the inclusion of a majority of patients with a left hemisphere exploration (Figure 1C-D and Table S1). Also, when inspecting left and right hemispheres separately, the patterns of shared, selective, and preferred responses remain similar across hemispheres across frequency band (see figures S6-7 for activation and deactivation, respectively)

Next, we investigated whether the channels selectivity (to speech or music) observed in a given frequency band was robust across frequency bands (Figure 4). We estimated the cross-frequency channel selectivity, that is the percentage of channels that selectively respond to speech or music across different frequency bands. We first computed the percentage of total channels selective for speech and music (either activated or deactivated compared to baseline) in a given frequency band. We then verified whether these channels were unresponsive to the other domain in the other frequency bands. This was done by examining each frequency band in turn and deducting any channels that showed a significant neural response to the other domain. When considering the entire frequency spectrum, the percentage of total channels being selective to speech or music is ~4 times less than when considering a single frequency band. For instance, while up to 8% of the total channels are selective for speech (or music) in the theta band, this percentage always drops to ~2% when considering the cross-frequency channel selectivity.

Critically, we found no evidence of anatomical regional selectivity, i.e. of a simple anatomo-functional spatial code (see Figure 1D for the definition of anatomical regions). We estimated, for each frequency band, activation/deactivation responses, and anatomical region, the proportion of patients showing selectivity for speech or music, by means of a population prevalence analysis (Figures 5-6; see Methods). This analysis revealed that, for the majority of patients, first of all, in most regions there were channels that responded to both speech and music (indicative of shared responses at the anatomical regional level), and, second of all, for the minority of anatomical regions for which a selectivity for the same domain (speech or music) was observed across multiple patients, this selectivity does not hold when also considering other frequency bands and activation/deactivation responses. For instance, while the left anterior middle temporal gyrus shows delta activity selective to music (Figure 2A and 5A), it shows low-gamma activity selective to speech (Figure 2E and 5E). The left STG and pSTS, which show selective activations in the theta and alpha bands for music (Figure 5B-C), show selective deactivations in the same bands for speech (Figure 6B-C) and a majority of shared activations in the HFa (Figure 5F). This absence of anatomical regional selectivity is also evident when looking at the non-thresholded results (Figure S8).

Overall, these results reveal an absence of regional selectivity to speech or music under ecological conditions. Instead, selective responses coexist in space across different frequency bands. But, while selectivity may not be striking at the level of anatomical regional activity, it may still be present at the network level. To investigate this hypothesis, we explored the connectivity between the auditory cortex and the rest of the brain. And, to functionally define the auditory cortex for each patient, we first investigated the relation between the auditory signal itself and the brain response to identify which sEEG channels (spatial) best encode the dynamics of the auditory stimuli.

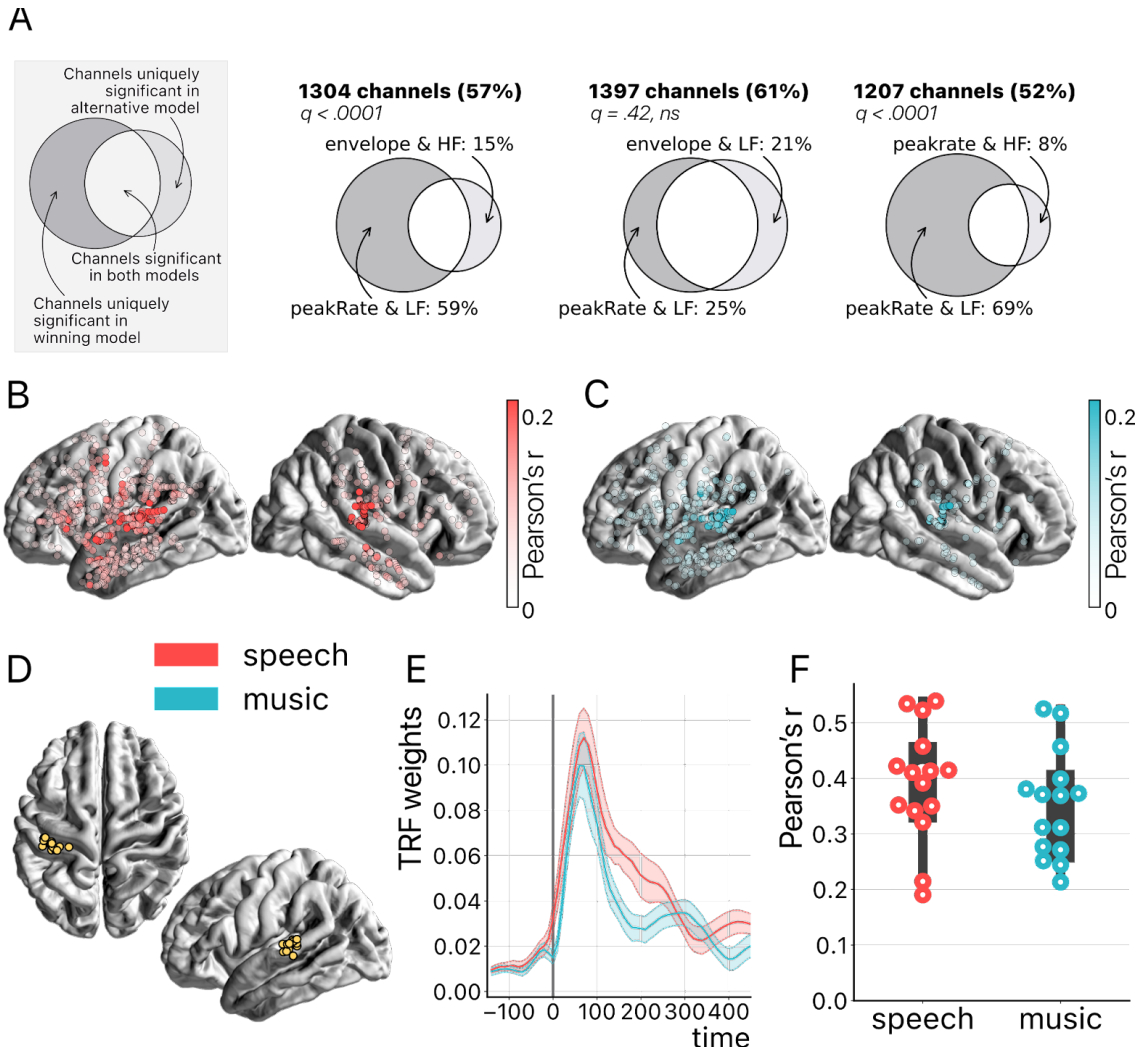

***Figure 7. Temporal Response Function (TRF) analyses**. **(A)** Model comparison. On the top left a toy model illustrates the use of Venn diagrams **comparing the winning model** (peakRate in LF) to each of the three other models for speech and music (pooled). Four TRF models were investigated to quantify the encoding of the instantaneous envelope and the discrete acoustic onset edges (peakRate) by either the low frequency band (LF) or the high frequency amplitude (HF). The 'peakRate & LF' model significantly captures the largest proportion of channels, and is, therefore, considered the winning model. The percentages on the top (in bold) indicate the percentage of total channels for which a significant encoding was observed during speech and/or music listening in either of the two compared models. In the Venn diagram, we indicate, out of all significant channels, the percentage that responded in the winning model (left) or in the alternative model (right). The middle part indicates the percentage of channels shared between the winning and the alternative model (percentage not shown). Q-values indicate pairwise model comparisons (Wilcoxon sign rank test, FDR-corrected). **(B)** and **(C) peakRate & LF model**: Spatial*

distribution of sEEG channels wherein low frequency (LF) neural activity significantly encodes the speech (red) and music (blue) peakRates. Higher prediction accuracy (Pearson's r) is indicated by increasing color saturation. All results are significant at q < 0.01 (N=18). ***(D) Anatomical localization of the best encoding channel within the left hemisphere for each patient*** *(N=15), as estimated by the 'peakRate and LF' model (averaged across speech and music). These channels are all within the auditory cortex and serve as seeds for subsequent connectivity analyses.* ***(E) TRFs averaged across the seed channels*** *(N=15), for speech and music.* ***(F) Prediction accuracy*** *(Pearson's r) of the neural activity of each seed channel, for speech and music.*

Low-frequency neural activity best encodes acoustic dynamics.

We linearly modeled the neurophysiological responses to continuous speech and music using temporal response functions (TRF). Based on previous studies (Oganian & Chang, 2019; Zion Golumbic et al., 2013; Zuk et al., 2021), we compared four TRF models. From both stimuli, we extracted the continuous, broadband temporal envelope (henceforth 'envelope') and the discrete acoustic onset edges (henceforth 'peakRate'; see Methods) and we quantified how well these two acoustic features are encoded by either the low frequency band (LF, 1-9 Hz) or the high frequency amplitude (HF, 80-120 Hz) bands. For each model, we estimated the percentage of total channels for which a significant encoding was observed during speech and/or music listening. The model for which most channels significantly encoded speech and/or music acoustic features corresponded to the model in which LF neural activity encoded the *peakRates* (Figure 7A). In general, the LF activity encodes the acoustic features in significantly more channels than the HF amplitude (*peakRate* & LF vs. *peakRate* & HF comparison: t = 13.39, q < .0001; *peakRate* & LF vs. *envelope* & HF comparison: t = 9.55, q < .0001). While the *peakRates* are encoded by numerically more channels than the instantaneous envelope, this difference was not significant (*peakRate* & LF vs. *envelope* & LF comparison: t = 1.93, q = .42).

Furthermore, we show that the *peakRates* are encoded by the LF neural activity throughout the cortex, for both speech and music (Figure 7B-C). More precisely, the regions wherein neural activity significantly encodes the acoustic structure of the stimuli go well beyond auditory regions and extend to the temporo-parietal junction, motor cortex, inferior frontal gyrus, and anterior and central sections of the superior and middle temporal gyrus. In particular, the strongest encoding values for speech are observed in the typical left-hemispheric language network, comprising the upper bank of the superior temporal gyrus, the posterior part of the inferior frontal gyrus, and the premotor cortex (Malik-Moraleda et al., 2022). Still, as expected, the best cortical tracking of the acoustic structure takes place in the auditory cortex, for both speech and music (Figure 7D). In other words, the best encoding channels are the same for speech and music and are those located closest to —or in— the primary auditory cortex. While the left hemisphere appears to be more strongly involved, this result is biased by the inclusion of a majority of patients with a left hemisphere exploration (see Figure 1C-D and Table S1). Proportionally, we found no difference in the number of significant channels between hemispheres (i.e. speech: 41% and 44% for left and right hemispheres respectively; music: 22% and 24% for left and right hemispheres respectively). Finally, the *peakRate* & LF model, i.e. the model that captures the largest proportion of significant channels during speech and/or music perception (Figure 7A), yields for both classes of stimuli a similar TRF shape (Figure 7E) as well as similar prediction accuracy scores (Pearson's r), of up to 0.55 (Figure 7F).

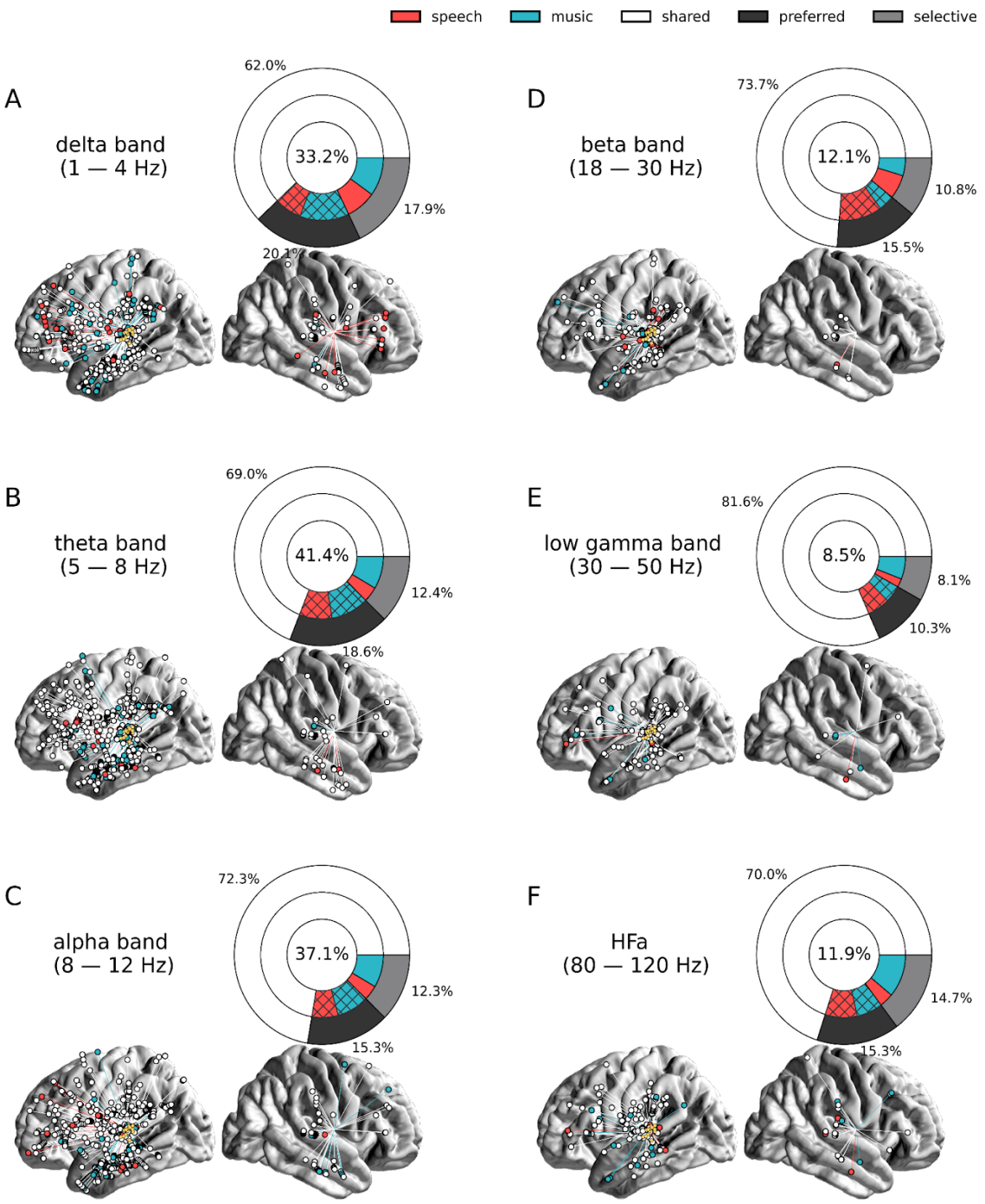

Figure 8. Seed-based functional connectivity analyses. (A)-(F) Significant coherence *responses to speech and/or music for the 6 canonical frequency bands (see Figure S8 for non-thresholded results). The seed was located in the left auditory cortex (see Figure 7D). Same conventions as in Figure 2, except for the center percentage in the nested pie charts which, here, reflects the percentage of channels significantly connected to the seed. Results are significant at α < 0.01 (N=15).*

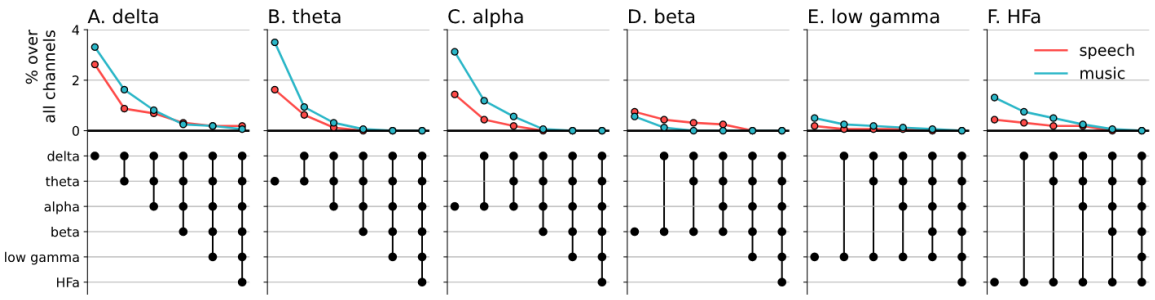

Figure 9. Cross-frequency channel selectivity for the connectivity analyses. *Percentage of channels that showed selective coherence with the primary auditory cortex in speech (red) or music (blue) across different frequency bands. Same conventions as in Figure 4.*

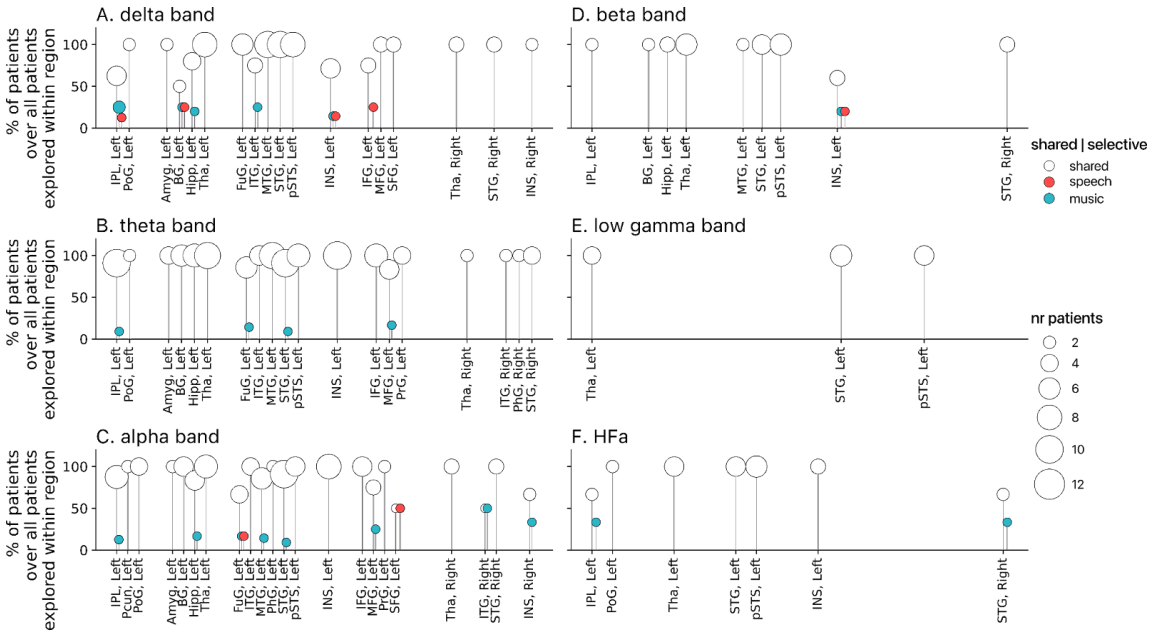

Figure 10. Population prevalence for the connectivity analyses (N=15). *Same conventions as in Figure 5 .*

Connections of the auditory cortex are also mostly non-domain selective to speech or music.

Seed-based connectivity analyses first revealed that, during speech or music perception, the auditory cortex is mostly connected to the rest of the brain through slow neural dynamics, with ~33% of the channels showing coherence values higher than the surrogate distribution at delta rate, and only ~12% at HFa (Figure 8, see also Figure S9 for non-thresholded results). Across frequencies, most of the significant connections are shared between the two cognitive domains (~70%), followed by preferred (~15%) and selective connections (~12%). Selectivity is nonetheless homogeneously present in all frequency bands (Figure 8). Importantly, selectivity is again frequency-specific (Figure 9). Estimating the cross-frequency channel selectivity, the percentage of total connections being selective to speech or music is at zero for all frequency bands except for the delta range (speech =

0.19%; music = 0.06%). Hence, selectivity is only visible at the level of frequency-specific distributed networks. Finally, here again no anatomical regional selectivity is observed, i.e. not a single cortical region is solely selective to speech or music. Rather, in every cortical region, the majority of patients show shared responses at the regional level, as estimated by the population prevalence analysis (Figure 10).

Discussion

In this study, we investigated the existence of domain-selectivity for speech and music under ecological conditions. We capitalized on the high spatiotemporal sensitivity of human stereotactic recordings (sEEG) to thoroughly evaluate the presence of selective neural responses —estimated both at the level of individual sEEG channels and anatomical cortical regions— when patients listened to a story or to instrumental music. More precisely, we quantified the extent to which natural speech and music processing is performed by shared, preferred, or domain-selective neural populations. By combining sEEG investigations of high-frequency activity (HFa) with the analyses of other frequency bands (from delta to low-gamma), the neural encoding of acoustic dynamics and spectrally-resolved connectivity analyses, we obtained a thorough characterization of the neural dynamics at play during natural and continuous speech and music perception. Our results show that speech and music mostly rely on shared neural resources. Further, while selective responses seem absent at the level of atlas-based cortical regions, selectivity can be observed at the level of frequency-specific distributed networks in both power and connectivity analyses.

Previous work has reported that written or spoken language selectively activates a left-lateralized functional cortical network (Chen et al., 2023; Fedorenko et al., 2011; Fedorenko & Blank, 2020; Malik-Moraleda et al., 2022). In particular, in previous functional MRI studies, these strong and selective cortical responses were not visible during the presentation of short musical excerpts, and are hypothesized to index linguistic processes (Chen et al., 2023; Fedorenko et al., 2011). Moreover, in the superior temporal gyrus, specific and separate neural populations for speech, music, and song are visible (Boebinger et al., 2021; Norman-Haignere et al., 2022). These selective responses, not visible in primary cortical regions, seem independent of both low-level acoustic features and higher-order linguistic meaning (Norman-Haignere et al., 2015), and could subtend intermediate representations (Giordano et al., 2023). Within this framework, the localizationism view applies to highly specialized processes (i.e. functional niches), while general cognitive domains are mostly spatially distributed. Recent studies have shown that some communicative signals (e.g. alarm, emotional, linguistic) can exploit distinct acoustic niches to target specific neural networks and trigger reactions adapted to the intent of the emitter (Albouy et al., 2020; Arnal et al., 2019). Using neurally relevant spectro-temporal representations (MPS), these studies show that different subspaces encode distinct information types: slow temporal modulations for meaning (speech), fast temporal modulations for alarms (screams), and spectral modulations for melodies (Albouy et al., 2020; Arnal et al., 2015, 2019; Flinker et al., 2019). Which acoustic features –and which neural mechanisms– are necessary and sufficient to route communicative sounds towards selective neural networks remains a promising field of investigation to explore.

In this context, in the current study we did not observe a single anatomical region for which speech-selectivity was present, in any of our analyses. On the contrary, we report spatially distributed and frequency-specific patterns of shared, preferred, or selective neural responses and connectivity fingerprints. This indicates that domain-selective brain regions should be considered as a set of functionally homogeneous but spatially distributed voxels, instead of anatomical landmarks. Several non-exclusive explanations may account for this finding. First, our results part with the simple selective versus shared dichotomy and adopt a more biologically valid and continuous framework (Buzsáki, 2019; Zatorre & Gandour, 2008) by adding a new category that is often neglected in the literature: *preferred* responses (Figure 1A). Indeed, responses in this category are usually reported as shared or selective and most often the statistical approach does not allow a more nuanced view (*cf* (Chen et al., 2023)). However preferred responses, namely responses that are stronger to a given class of stimuli but that are also present with other stimuli, are relevant and should not be collapsed into either the selective or shared categories. Introducing this intermediate category refines the epistemological and statistical approach on how to map cognitive and brain functions. It points

toward the presence of gradients of neural activity across cognitive domains, instead of all-or-none responses. This framework is more compatible with the notion of distributed representations wherein specific regions are more-or-less recruited depending on their relative implication in a distributed manifold (Elman, 1991; Rissman & Wagner, 2012).

Second, most of the studies that reported regional-selectivity are grounded on functional MRI data that lack a precise temporal resolution. Furthermore, the few studies assessing selectivity with intracranial EEG recordings analyzed only the HF amplitude (Bellier et al., 2022; Norman-Haignere et al., 2020; Oganian & Chang, 2019). However, while this latter reflects local (Kopell et al., 2000) and possibly feedforward activity (Bastos et al., 2015; Fontolan et al., 2014; Fries, 2015) other frequency bands are also constitutive of the cortical dynamics and involved in cognition. For instance, alpha/beta rhythms play a role in predicting upcoming stimuli and modulating sensory processing and associated spiking (Arnal & Giraud, 2012; Bastos et al., 2020; Morillon & Baillet, 2017; Saleh et al., 2010; van Kerkoerle et al., 2014). Also slower dynamics in the delta/theta range have been described to play a major role in cognitive processes and in particular for speech perception, contributing to speech tracking, segmentation and decoding (Ding et al., 2017; Doelling et al., 2014; Giraud & Poeppel, 2012; Gross et al., 2013; Keitel et al., 2017). Importantly, we here addressed both activations and deactivations that can co-occur in the same spatial location across different frequency bands (Pfurtscheller & Lopes da Silva, 1999) (Proix et al., 2022) and indeed observed that domain-selectivity is frequency-specific, meaning that domain-selectivity is marginal when considering the entire spectrum of activity of a given sEEG channel. Finally, most studies only investigated local neural activity and did not consider the brain as a distributed system, analyzed through the lens of functional connectivity analyses. While topological approaches are more complex, they also provide more nuanced and robust characterization of brain functions. Critically, our approach reveals the limitation of adopting a reductionist approach – either by considering the brain as a set of independent regions instead of distributed networks, or by overlooking the spectral complexity of the neural signal.

Third, the ecological auditory stimuli we used are longer and more complex than stimuli used in previous studies and hence more prone to elicit distributed and dynamical neural responses (Hasson et al., 2010; Sonkusare et al., 2019; Theunissen et al., 2000) and they require, in the case of music, for instance, more complex representations of melody and rhythm motifs contributing to stronger representations of meter, tonality, and groove (Boebinger et al., 2021). While natural speech and music directly stimulate cognitively relevant neural processes, they do not allow to directly isolate specific brain operations. Computational models -which can be as diverse as acoustic (Chi et al., 2005), cognitive (Giordano et al., 2021), information-theoretic (Di Liberto et al., 2020; Donhauser & Baillet, 2019), or neural network (Millet et al., 2022) models- are hence necessary to further our understanding of the type of computations performed by our reported frequency-specific distributed networks. Moreover, incorporating such insights can help us avoid misattributing differences between speech and music driven by unmatched sensitivity factors (e.g., arousal, emotion, or attention) as inherent speech or music selectivity (Mas-Herrero et al., 2013; Nantais & Schellenberg, 1999).

Our modeling approach was targeting the temporal dynamics of the speech and music stimuli. Beyond confirming that acoustic dynamics are strongly tracked by auditory neural dynamics, it revealed, investigating the entire cortex, that such neural tracking also occurs well outside of auditory regions – up to motor and inferior frontal areas (Figure 3B; see also (Chalas et al., 2022; Zion Golumbic et al., 2013). Of note, this spatial map of speech dynamics encoding is very similar to former reports of the brain regions belonging to the language system (Diachek et al., 2020). But, here again, adopting an approach that investigates both low and high frequencies of the neural signal – an approach that is not enough embraced in intracranial EEG studies (Proix et al., 2022) – reveals that the low frequency activity clearly better encodes acoustic features than the HF amplitude (Figure 3A).

In conclusion, our results point to a massive amount of shared neural response to speech and music, well beyond the auditory cortex. They also show the interest of considering shared, preferred, and selective responses when investigating domain selectivity. Importantly these three classes of responses should be considered in respect to 1) activation or deactivation patterns compared to a baseline, 2) different frequency bands and 3) power spectrum (activity) and connectivity approaches. Combining all these points of view, gives a richer although possibly more complex view of brain functions. While our data point to an absence of anatomical regional selectivity for speech and music, such a selectivity still exists at the level of a spatially distributed and frequency-specific network. Thus, the inconsistency with previous findings may be limited to the idea that some anatomical regions are selective to speech or music processing. However, the two points of view can be reconciled when considering a fine-grained network approach allowing selectivity to coexist for speech and music within the same anatomical region. Finally, in adopting here a comparative approach of speech and music – the two main auditory domains of human cognition – we only investigated one type of speech and of music. Future work is needed to investigate for instance whether different sentences or melodies activate the same selective frequency-specific distributed networks and whether environmental or alarm sounds induce different patterns of neural selectivity.

Methods

Participants.

18 participants (10 females, mean age 30 y, range 8 - 54 y) with pharmacoresistant epilepsy participated in the study. All patients were French native speakers. Neuropsychological assessments carried out before stereotactic EEG (sEEG) recordings indicated that all participants had intact language functions and met the criteria for normal hearing. In none of them were the auditory areas part of their epileptogenic zone as identified by experienced epileptologists. Recordings took place at the Hôpital de La Timone (Marseille, France). Patients provided informed consent prior to the experimental session, and the experimental protocol was approved by the Institutional Review board of the French Institute of Health (IRB00003888).

Data acquisition.

The sEEG signal was recorded using depth electrodes shafts of 0.8 mm diameter containing 10 to 15 electrode contacts (Dixi Medical or Alcis, Besançon, France). The contacts were 2 mm long and were spaced from each other by 1.5 mm. The locations of the electrode implantations were determined solely on clinical grounds. Patients were included in the study if their implantation map covered at least partially the Heschl's gyrus (left or right). The cohort consists of 13 unilateral implantations (10 left, 3 right) and 5 bilateral implantations, yielding a total of 271 electrodes and 3371 contacts (see Figure 1C-D for electrodes localization).

Patients were recorded either in an insulated Faraday cage or in the bedroom. In the Faraday cage, participants laid comfortably in a chair, the room was sound attenuated and data were recorded using a 256-channels amplifier (Brain Products), sampled at 1kHz and high-pass filtered at 0.016 Hz. In the bedroom, data were recorded using a 256-channels Natus amplifier (Deltamed system), sampled at 512 Hz and high-pass filtered at 0.16 Hz.

Experimental design.

Patients completed three separate sessions. In one session they completed the main experimental paradigm and the two additional sessions served as baseline for the spectral analysis (see below).

In the main experimental session, patients passively listened to ~10 minutes of storytelling (Gripari, 2004); 577 secs, *La sorcière de la rue Mouffetard, (Gripari, 2004)* and ~10 minutes of music (580 secs, *Reflejos del Sur, (Oneness, 2006)* separated by 3 minutes of rest. The order of conditions was counterbalanced across patients (see Table S1). This session was conducted in the Faraday cage (N=6) or in the bedroom (N=12).

In the two baseline sessions, patients passively listened to two types of auditory stimuli: 1) 30-ms–long pure tones, presented binaurally at 500 Hz or 1 kHz (with a linear rise and fall time of 0.3 ms) 110 times each, with an ISI of 1,030 (±200) ms; and 2) /ba/ or /pa/ syllables, pronounced by a French female speaker and presented binaurally 250 times each, with an ISI of 1,030 (±200) ms. These stimuli were designed for a clinical purpose in order to functionally map the auditory cortex. These two recording sessions (lasting ~ 2 and 4 minutes) were performed in the Faraday cage.

In the Faraday cage, a sound Blaster X-Fi Xtreme Audio, an amplifier Yamaha P2040 and Yamaha loudspeakers (NS 10M) were used for sound presentation. In the bedroom, stimuli were presented using a Sennheiser HD 25 headphone set. Sound stimuli were presented at 44.1 kHz sample rate and 16 bits resolution. Speech and music excerpts were presented at ~75 dBA (see Figure 1B).

General preprocessing related to electrodes localisation.

To increase spatial sensitivity and reduce passive volume conduction from neighboring regions (Mercier et al., 2017, 2022), the signal was offline re-referenced using bipolar montage. That is, for a pair of adjacent electrode contacts, the referencing led to a virtual channel located at the midpoint locations of the original contacts. To precisely localize the channels, a procedure similar to the one used in the iELVis toolbox and in the fieldtrip toolbox was applied (Groppe et al., 2017; Stolk et al., 2018). First, we manually identified the location of each channel centroid on the post-implant CT scan using the Gardel software (Medina Villalon et al., 2018). Second, we performed volumetric segmentation and cortical reconstruction on the pre-implant MRI with the Freesurfer image analysis suite (documented and freely available for download online http://surfer.nmr.mgh.harvard.edu/). Third, the post-implant CT scan was coregistered to the pre-implant MRI via a rigid affine transformation and the pre-implant MRI was registered to MNI152 space, via a linear and a non-linear transformation from SPM12 methods (Penny et al., 2011), through the FieldTrip toolbox (Oostenveld et al., 2011). Fourth, applying the corresponding transformations, we mapped channel locations to the pre-implant MRI brain that was labeled using the volume-based Human Brainnetome Atlas (Fan et al., 2016).

Based on the brain segmentation performed using SPM12 methods through the Fieldtrip toolbox, bipolar channels located outside of the brain were removed from the data (3%). The remaining data (Figure 1C) was then bandpass filtered between 0.1 Hz and 250 Hz, and, following a visual inspection of the power spectral density (PSD) profile of the data, when necessary, we additionally applied a notch filter at 50 Hz and harmonics up to 200 Hz to remove power line artifacts (N=12). Finally, the data were downsampled to 500 Hz.

Artifact rejection.

To define artifacted channel we used both the broadband signal and the amplitude of the high frequency activity. This latter was obtained by computing —with the Hilbert transform— the analytic amplitude of four 10-Hz-wide sub-bands spanning from 80 to 120 Hz. Each sub-band was standardized by dividing it by its mean and, finally, all sub-bands were averaged together (Ossandón et al., 2012; Vidal et al., 2012). Channels with a variance greater than 2*IQR (interquartile range, i.e. a non-parametric estimate of the standard deviation) —on either the broadband or high frequency signals— were tagged as artifacted channels (on average 18% of the channels). Then the data were epoched in non-overlapping segments of 5 seconds (2500 samples). To exclude artifacted epochs, epochs wherein the maximum amplitude (over time) summed across non-excluded channels was greater than 2*IQR were tagged as artifacted epochs. Overall, 6% of the speech epochs and 7% of the music epochs were rejected. Channels and epochs defined as artifacted were excluded from subsequent analyses, except if specified otherwise (see TRF section).

Spectral analysis.

Six canonical frequency bands were investigated: delta (1-4 Hz), theta (5-8 Hz), alpha (8-12 Hz), beta (18-30 Hz), low-gamma (30-50 Hz), and high-frequency activity (HFa; 80-120 Hz). To prevent edge artifacts, prior to extracting the power spectrum, epochs were zero-padded on both sides with 3.5-seconds segments which were later removed. For each patient, channel, epoch, and frequency band, the power of the neural signal was calculated using the Welch approach on Discrete Fourier Transform from the scipy-python library (Virtanen et al., 2020) and then averaged across the relevant frequencies to obtain these six canonical bands.

For each canonical band and each channel, we classified the time-averaged neural response as being selective, preferred, or shared across the two investigated cognitive domains (speech, music). We defined these categories by capitalizing on both the simple effects of —and contrast between— the neural responses to speech and music stimuli compared to a baseline condition (see Figure 1A).

“Selective” responses are neural responses that are significantly different compared to the baseline for one domain (speech or music) but not the other, and with a significant difference between domains (i.e. speech or music is different from baseline + difference effect between the domains). “Preferred” responses correspond to neural responses that occur during both speech and music processing, but with a significantly stronger response for one domain over the other (i.e. both speech and music are significantly different from baseline + difference effect between the domains). Finally, “shared” responses occur when there are no significant differences between domains, and there is a significant neural response to at least one of the two stimuli (one or two simple effects + no difference). If none of the two domains produces a significant neural response, the difference is not assessed (case “neither” simple effect). As neural populations can synchronize or desynchronize in response to sensory stimulation, we estimated these categories separately for significant activations and significant deactivations compared to baseline.

For each frequency band and channel, the statistical difference between each condition and the baseline —as well as between conditions— was estimated with paired sample permutation tests based on the t-statistic from the mne-python library (Gramfort et al., 2014) with 1000 permutations and the *tmax* method to control the family-wise error rate (Nichols & Holmes, 2002).

Temporal Response Function (TRF) analysis.

We used the Temporal Response Function (TRF) to estimate the encoding of acoustic features by neural activity. Two acoustic features were extracted from our stimuli (speech, music): the envelope and the peakRate. To estimate the temporal envelope of the two stimuli, the acoustic waveforms were decomposed into 32 narrow frequency bands using a cochlear model, and the absolute value of the Hilbert transform was computed for each of these narrowband signals. The broadband temporal envelope (henceforth ‘envelope’) resulted from the summation of these absolute values. The acoustic onset edges (henceforth ‘peakRate’) were defined as peaks in the rate of change (first derivative) of the envelope (Doelling et al., 2014; Oganian & Chang, 2019). Finally, both the envelopes and peakRates were downsampled to 100 Hz and z-scored to be mapped to the neural data.

All computations of the TRF used the pymTRF library (Steinkamp, 2019), a python adaption of the mTRF toolbox (Crosse et al., 2016). A TRF is a model that, via linear convolution, serves as a filter to quantify the relationship between two continuous signals, here stimulus features and neural activity. Hence, for this analysis, the entire duration of the recordings were preserved, i.e., no artifacted epochs were excluded. When applied in a forward manner, the TRF approach describes the mapping of stimulus features onto the neural response (henceforth 'encoding'; (Crosse et al., 2016)). Using ridge regression to avoid overfitting, we examined how well the two different acoustic features —envelope and peakRate— map onto low frequency activity (LF; 1-9 Hz) or the amplitude of the high frequency activity (HF; 80-120 Hz) (Ding et al., 2016; Zion Golumbic et al., 2013). Hence four encoding models were estimated: envelope/peakRate acoustic features * LF/HF neural activity. For each model and patient, the optimal ridge regularization parameter (λ) was estimated using cross-validation on the sEEG channels situated in the auditory cortex. We considered time lags from -150 to 1000 ms for the TRF estimations. 80% of the data was used to derive the TRFs and the remaining 20% was used as a validation set. The quality of the predicted neural response was assessed by computing Pearson’s product moment correlations (Fisher-z-scored) between the predicted and actual neural data for each channel and model using the scipy-python library (p-values FDR-corrected).

Models were finally compared in terms of the percentage of channels that significantly encoded the acoustic structure of speech and/or music. This percentage was estimated at the single-subject level and combined with non-parametric Wilcoxon sign rank tests at the group level

to define the winning model. In other words, the winning model is the model for which the percentage of channels significantly encoding speech and/or music acoustic features is the largest. Multiple comparison across pairs of models was controlled for with a FDR correction.

Connectivity analysis.

We examined the frequency-specific functional connectivity maps in response to speech and music, between the entire brain and the auditory cortex using a seed-based approach (we excluded the channels immediately neighboring the seed-channel). As seed, we selected, per patient, the channel that best encoded the speech and music acoustic features (see TRF analysis; Figure 7). We used spectral coherence as a connectivity measure for all canonical bands (see above) and all analyses were performed using the mne-python library (Gramfort et al., 2014).

For each frequency band, we classified each channel into selective, preferred, or shared categories (see Figure 1A) by examining both the simple effects (i.e. which channels display a significantly coherent signal with the seed during speech and/or music processing) and the difference effects (i.e. is coherence significantly stronger for one domain over the other).

Statistical significance was assessed for each frequency band and channel using surrogate data with 1000 iterations, which were generated by modifying the temporal structure of the sEEG-signal recorded at the seeds (i.e. shuffling the epochs) prior to computing connectivity. This process led to a total of 1000 connectivity values, which were used as null-distribution to calculate the probability threshold associated with genuine connectivity.

Population prevalence.

For both the spectral and the connectivity analyses, in order to make sure that the results are not driven by the heterogeneity of electrode locations across participants, we examined, for each region, the proportions of patients showing only shared or selective responses. That is, for both the spectral and connectivity results, we examined results representativeness as follows: for each anatomical region wherein at least two patients have at least two significantly responsive channels, we computed the percentage of patients that showed a pattern of selective (i.e. all channels selective to speech or music) or a shared (i.e. a mixture of channels responding to speech and/or music) responses. This approach is inspired by the population prevalence, where an equivalent metric is introduced (i.e., the Maximum A Posterior estimate see (Ince et al., 2021).

Supplementary Figures

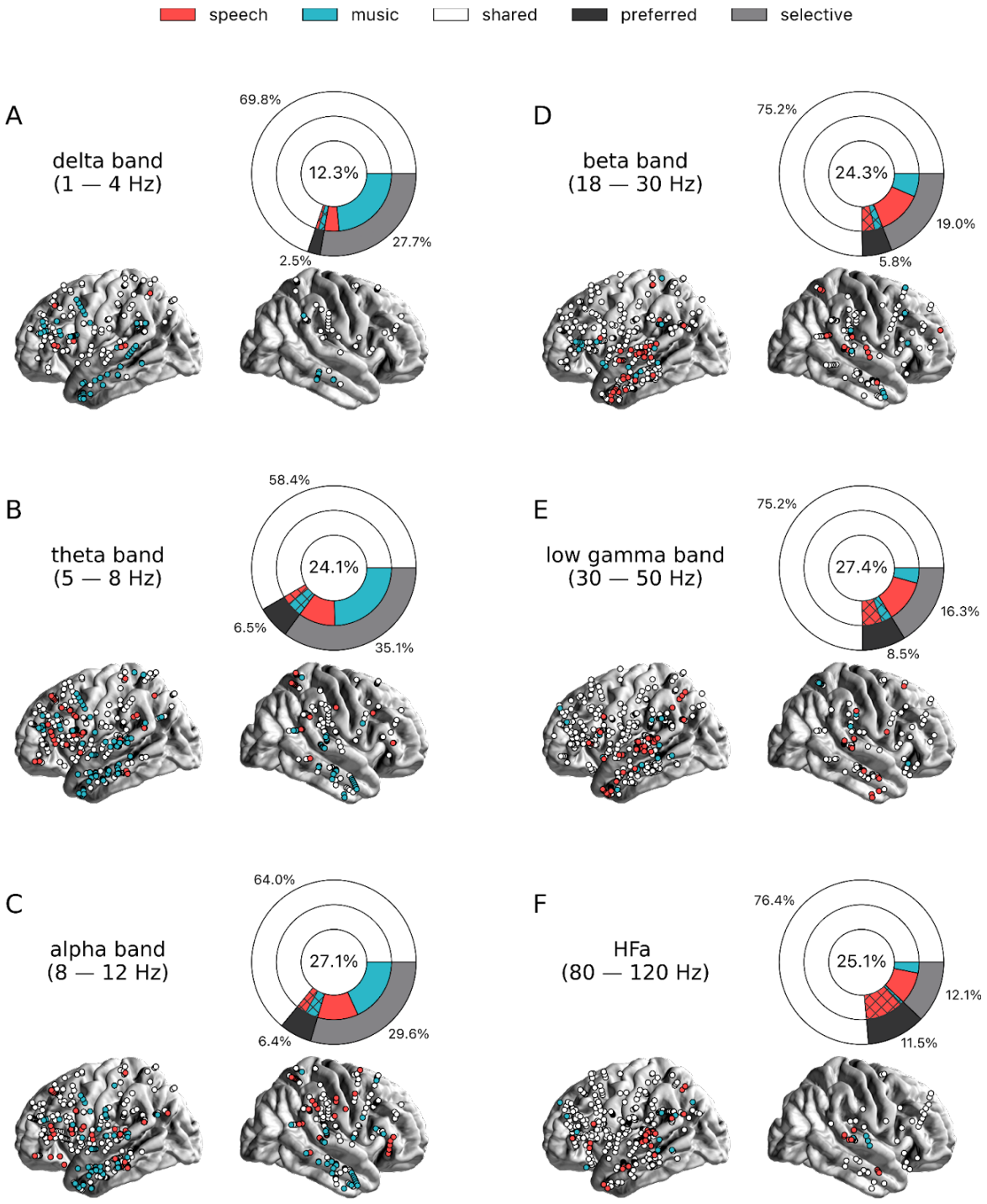

Figure S1. Power spectrum analyses of activations (speech or music > syllables). ***(A)-(F)*** *Neural responses to speech and/or music for the 6 canonical frequency bands. Only significant activations compared to the baseline condition (syllables listening) are reported. Same conventions as in Figure 2. Results are significant at α < 0.01 (N=18).*

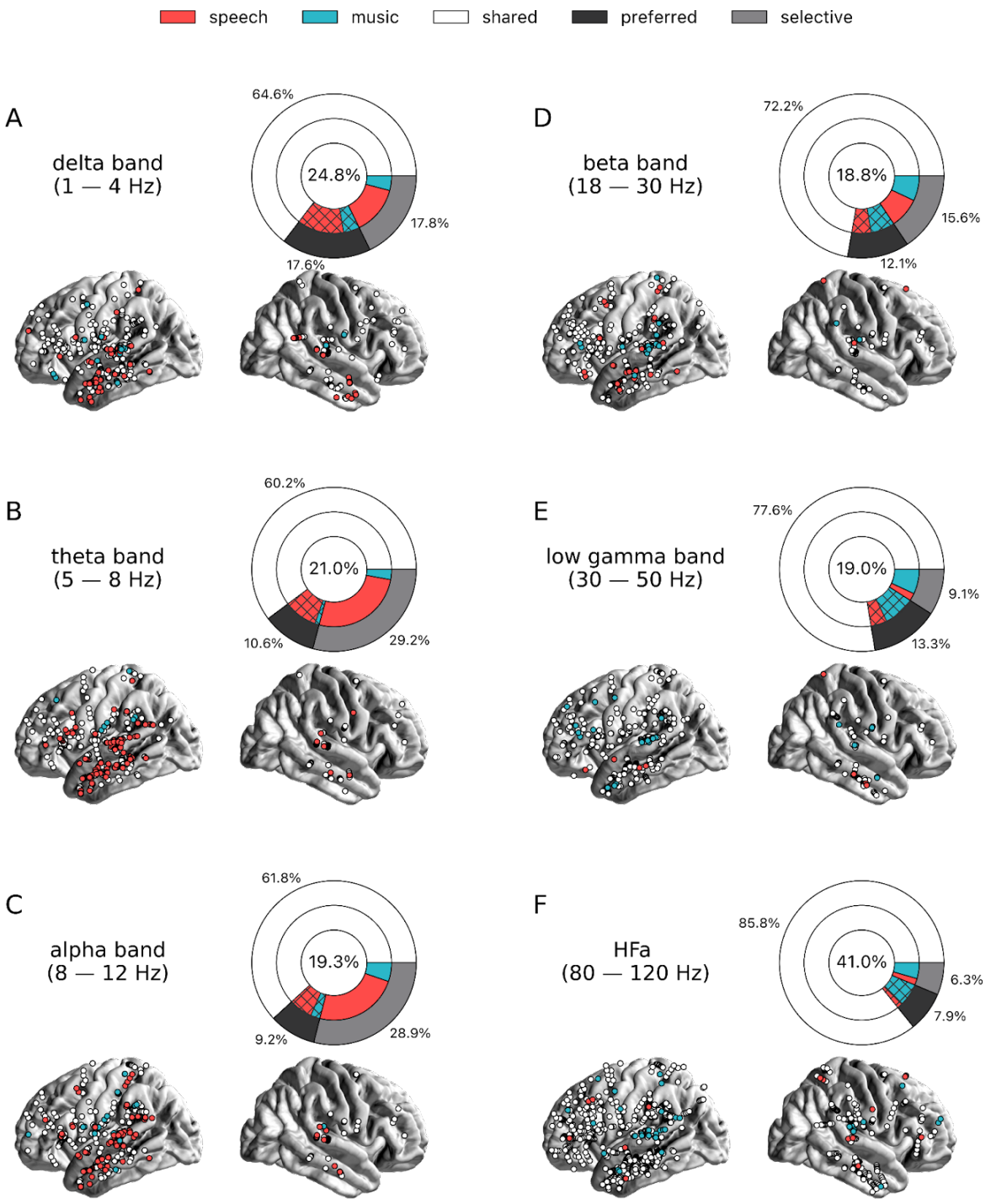

Figure S2. Power spectrum analyses of deactivations (speech or music < syllables). (A)-(F) *Neural responses to speech and/or music for the 6 canonical frequency bands. Only significant deactivations compared to the baseline condition (syllables listening) are reported. Same conventions as in Figure 2. Results are significant at α < 0.01 (N=18).*

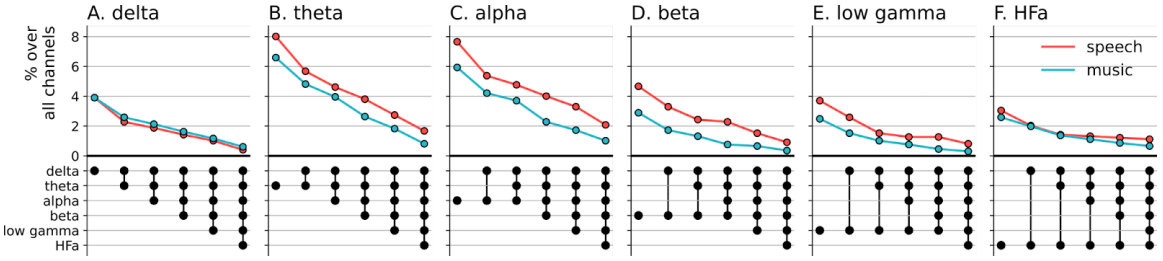

Figure S3. ***Cross-frequency channel selectivity for the power spectrum analyses (syllables baseline).*** *Percentage of channels that selectively respond to speech (red) or music (blue) across different frequency bands, compared to the baseline condition (syllable listening). Same conventions as in Figure 4.*

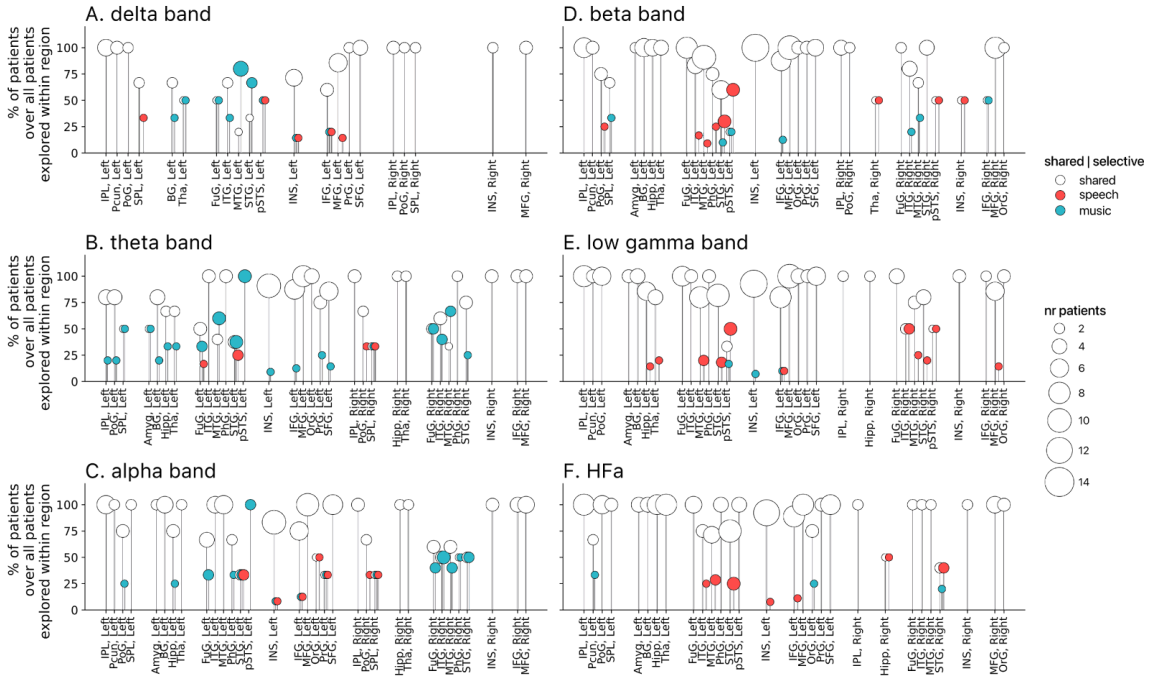

Figure S4. Population prevalence for the power spectral analyses of activations (speech or music > syllables; N = 18). (A)-(F) *Population prevalence of shared or selective responses for the 6 canonical frequency bands, per anatomical region. Only significant activations compared to the baseline condition (syllables listening) are reported. Same conventions as in Figure 5.*

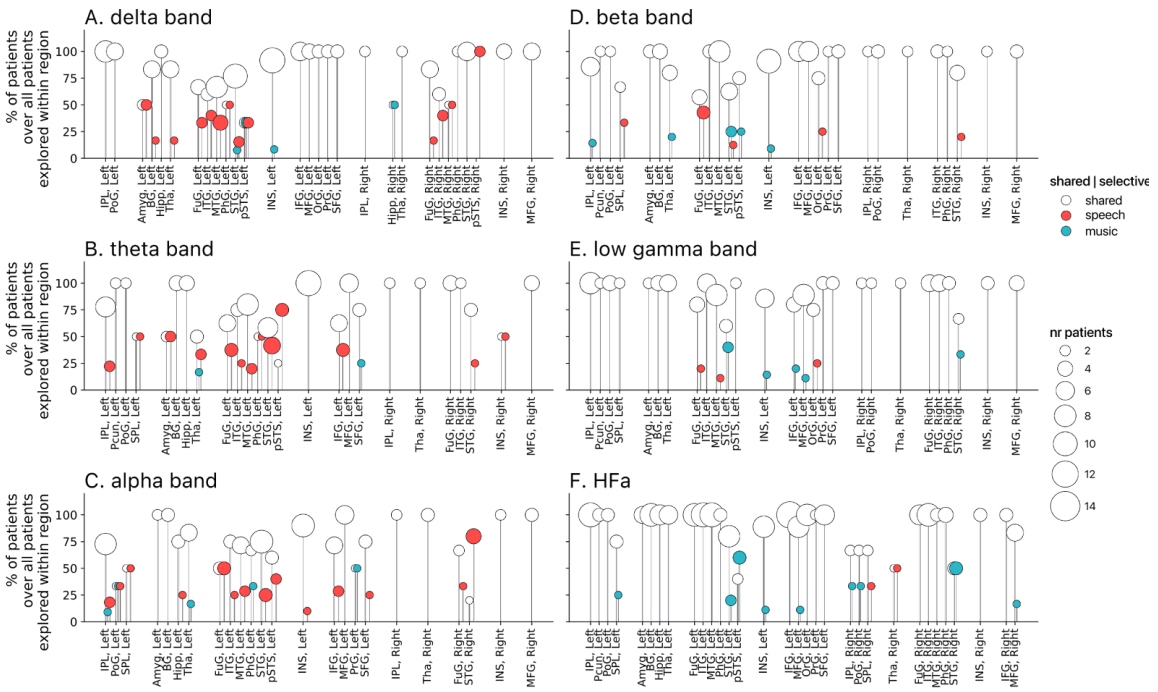

***Figure S5**. **Population prevalence for the power spectral analyses of deactivations (speech or music < syllables; N = 18)**. **(A)-(F)** Population prevalence of shared or selective responses for the 6 canonical frequency bands, per anatomical region. Only significant deactivations compared to the baseline condition (syllables listening) are reported. Same conventions as in Figure 5.*

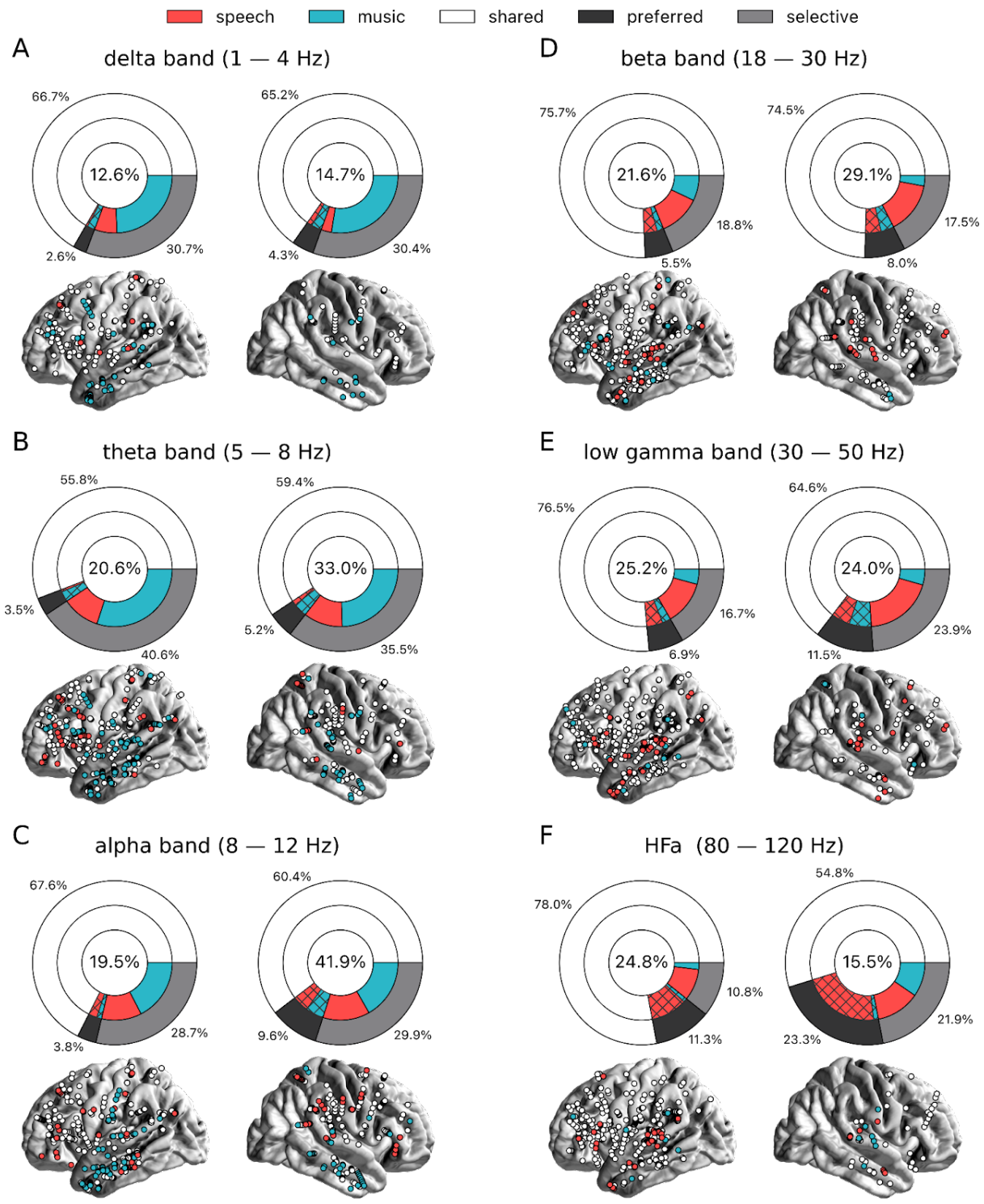

Figure S6. Power spectrum analyses of activations (speech or music > tones) for each hemisphere separately. *Same conventions as in Figure 2. Results are significant at α < 0.01 (N=18).*

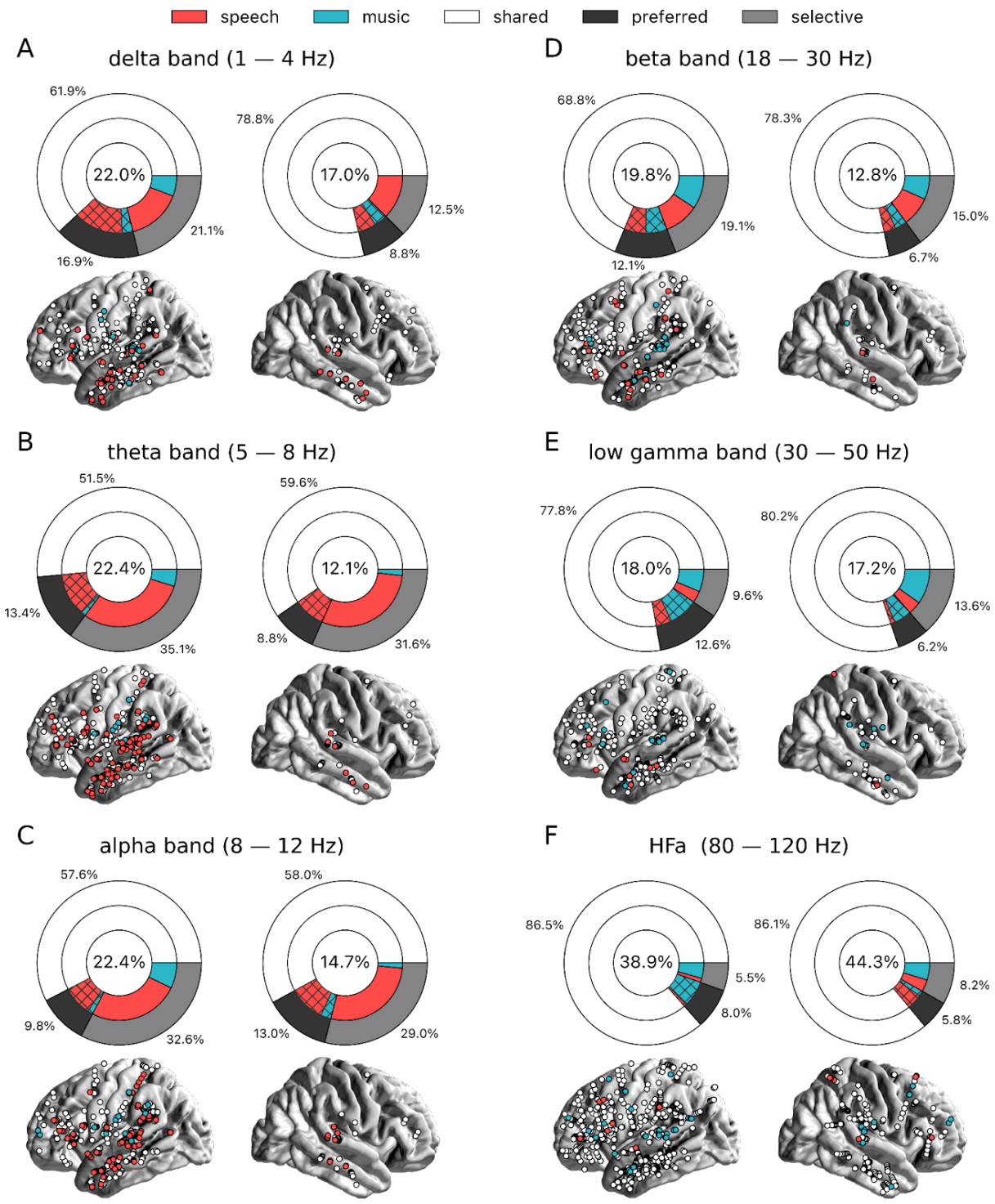

Figure S7. Power spectrum analyses of deactivations (speech or music < tones) for each hemisphere separately. *Same conventions as in Figure 2. Results are significant at α < 0.01 (N=18).*

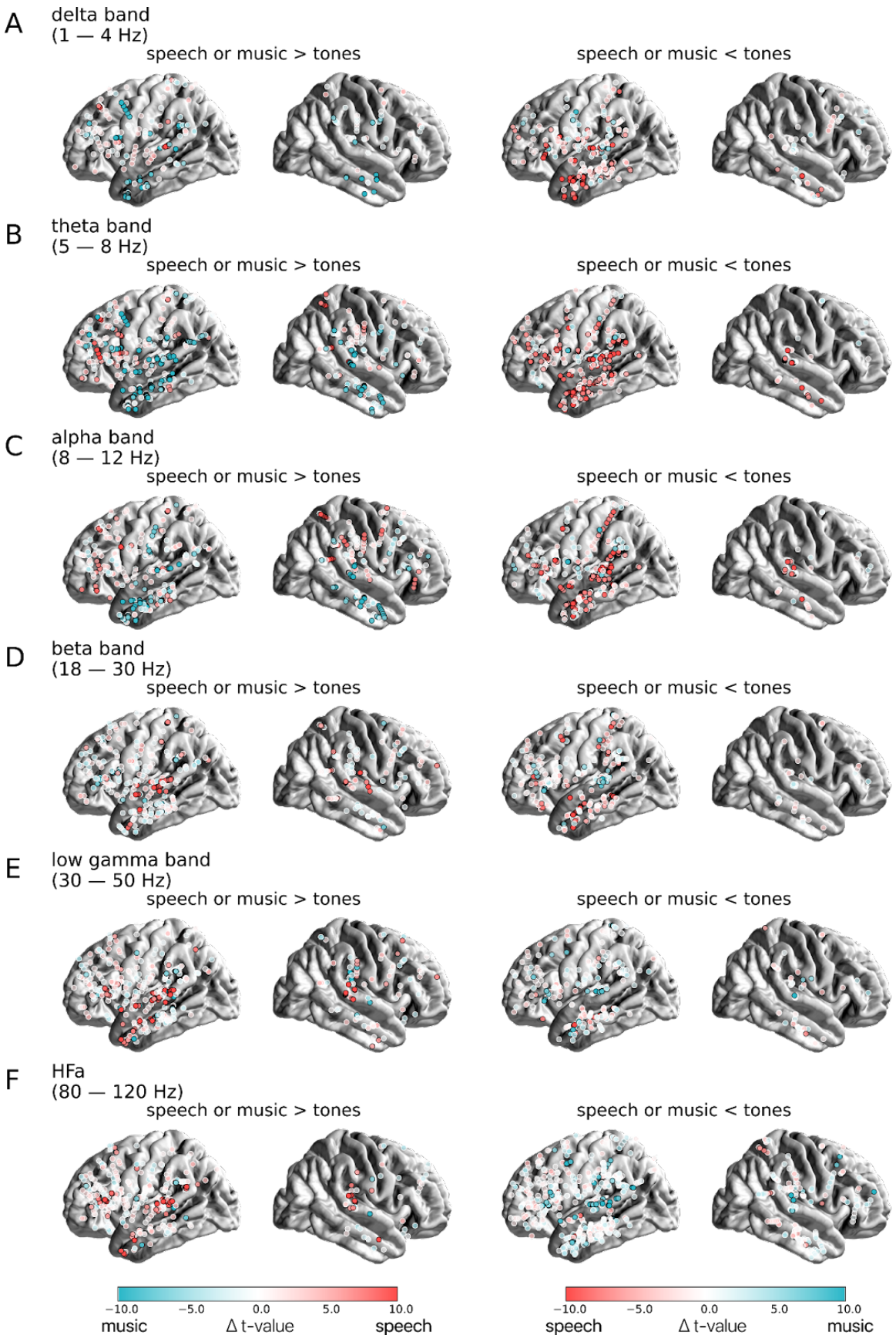

Figure S8. Contrast between the neural responses to speech and music, for the 6 canonical frequency bands *(tones baseline)*. Distribution of non-thresholded t-values from the speech vs. music

permutation test, projected on the brain surface. Only sEEG channels for which speech and/or music is significantly stronger (left column) or weaker (right column) than the baseline condition (pure tones listening) are reported. Channels for which the speech vs. music contrast is significant are circled in black. Channels for which the speech vs. music contrast is not significant are circled in white. Results are significant at α < 0.01 (N=18).

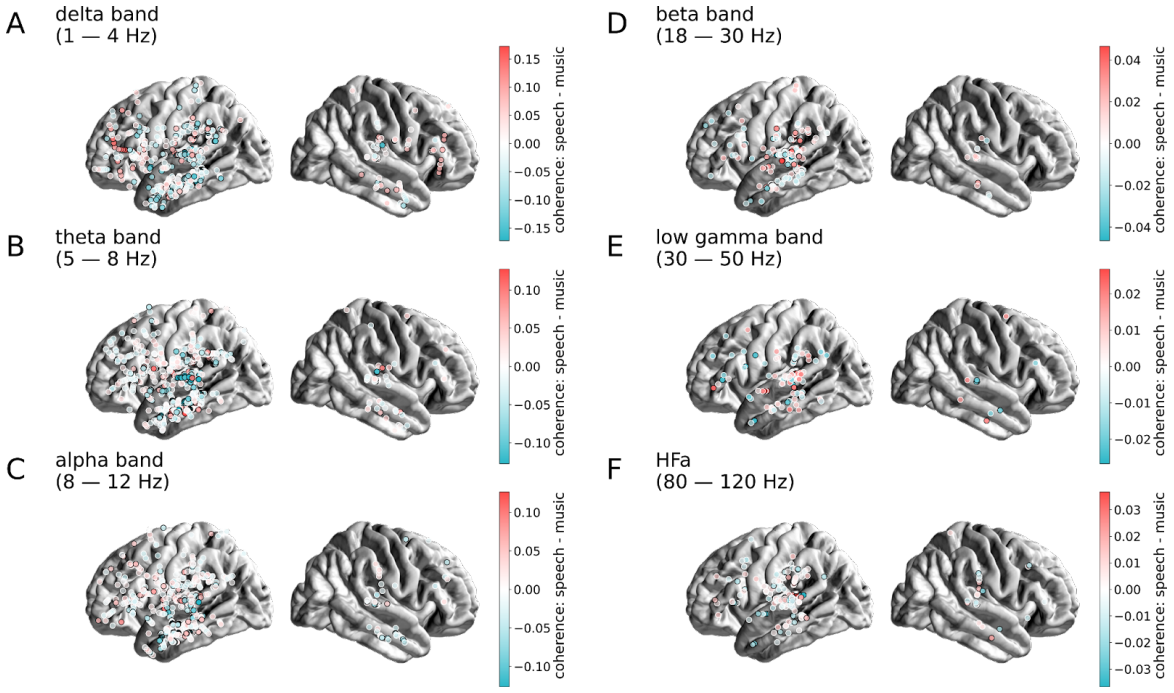

Figure S9. Contrast between the coherence responses to speech and music, for the 6 canonical frequency bands. *Distribution of non-thresholded differences in coherence values between speech and music, projected on the brain surface. Same conventions as in Figure S8. Results are significant at α < 0.01 (N=15).*

patient	gender	age	order of presentation	recording site	hemispheric dominance (typical/atypical)	H: left/right	depth electrodes
P1	f	35	speech-music	room	Typical	Bilateral	11 left; 3 right
P2	m	29	speech-music	room	AtypicalB	Left	12 left; 2 right
P3	m	39	speech-music	lab	Typical	Left	14 left; 2 right
P4	m	8	speech-music	lab	Typical	Bilateral	10 left; 2 right
P5	f	36	speech-music	room	Typical	Left	13 left; 3 right
P6	m	44	music-speech	room	Typical	Left	15 left; 2 right
P7	f	37	speech-music	lab	AtypicalB	Bilateral	7 left; 9 right
P8	f	17	speech-music	lab	Typical	Bilateral	11 left; 4 right
P9	f	54	music-speech	room	Typical	Left	15 left; 5 right
P10	f	20	music-speech	lab	Typical	Right	1 left; 11 right
P11	f	37	music-speech	room	Typical	Left	11 left; 2 right
P12	f	37	music-speech	room	Typical	Left	14 left; 2 right
P13	m	24	music-speech	room	AtypicalB	Left	13 left; 2 right
P14	f	30	music-speech	room	Typical	Left	12 left; 1 right
P15	f	22	music-speech	room	Typical	Right	1 left; 11 right
P16	m	22	music-speech	room	Typical	Bilateral	12 left; 3 right
P17	m	32	music-speech	room	Typical	Left	10 left; 2 right
P18	m	17	speech-music	lab	Typical	Right	14 left; 1 right

Table S1: Patients description.

References

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.

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A.-L., & Poeppel, D. (2015). Human screams occupy a privileged niche in the communication soundscape. *Current Biology: CB*, *25*(15), 2051–2056.

Arnal, L. H., & Giraud, A.-L. (2012). Cortical oscillations and sensory predictions. *Trends in Cognitive Sciences*, *16*(7), 390–398.

Arnal, L. H., Kleinschmidt, A., Spinelli, L., Giraud, A.-L., & Mégevand, P. (2019). The rough sound of salience enhances aversion through neural synchronisation. *Nature Communications*, *10*(1), 3671.

Bastos, A. M., Lundqvist, M., Waite, A. S., Kopell, N., & Miller, E. K. (2020). Layer and rhythm specificity for predictive routing. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(49), 31459–31469.

Bastos, A. M., Vezoli, J., Bosman, C. A., Schoffelen, J.-M., Oostenveld, R., Dowdall, J. R., De Weerd, P., Kennedy, H., & Fries, P. (2015). Visual areas exert feedforward and feedback influences through distinct frequency channels. *Neuron*, *85*(2), 390–401.

Bellier, L., Llorens, A., Marciano, D., Schalk, G., Brunner, P., Knight, R. T., & Pasley, B. N. (2022). Encoding and decoding analysis of music perception using intracranial EEG. In *bioRxiv* (p. 2022.01.27.478085). https://doi.org/10.1101/2022.01.27.478085

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

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.

Buzsáki, G. (2019). *The Brain from Inside Out*. Oxford University Press, USA.

Chalas, N., Daube, C., Kluger, D. S., Abbasi, O., Nitsch, R., & Gross, J. (2022). Multivariate analysis of speech envelope tracking reveals coupling beyond auditory cortex. *NeuroImage*, *258*, 119395.

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* ,

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.

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, *10*, 604.

Diachek, E., Blank, I., Siegelman, M., Affourtit, J., & Fedorenko, E. (2020). The Domain-General Multiple Demand (MD) Network Does Not Support Core Aspects of Language Comprehension: A Large-Scale fMRI Investigation. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *40*(23), 4536–4550.

Di Liberto, G. M., Pelofi, C., Bianco, R., Patel, P., Mehta, A. D., Herrero, J. L., de Cheveigné, A., Shamma, S., & Mesgarani, N. (2020). Cortical encoding of melodic expectations in human temporal cortex. *eLife*, *9*. https://doi.org/10.7554/eLife.51784

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. (2017). Characterizing Neural Entrainment to Hierarchical Linguistic Units using Electroencephalography (EEG). *Frontiers in Human Neuroscience*, *11*(481), 481.

Ding, N., Melloni, L., Zhang, H., Tian, X., & Poeppel, D. (2016). Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience*, *19*(1), 158–164.

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 Pt 2*(P2), 761–768.

Donhauser, P. W., & Baillet, S. (2019). Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron*. https://doi.org/10.1016/j.neuron.2019.10.019

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*(2), 195–225.

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*, 448–458.

Fan, L., Li, H., Zhuo, J., Zhang, Y., Wang, J., Chen, L., Yang, Z., Chu, C., Xie, S., Laird, A. R., Fox, P. T., Eickhoff, S. B., Yu, C., & Jiang, T. (2016). The Human Brainnetome Atlas: A New Brain Atlas Based

on Connectional Architecture. *Cerebral Cortex* , *26*(8), 3508–3526.

Fedorenko, E., Behr, M. K., & Kanwisher, N. (2011). Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(39), 16428–16433.

Fedorenko, E., & Blank, I. A. (2020). Broca's Area Is Not a Natural Kind. *Trends in Cognitive Sciences*, *24*(4), 270–284.

Flaugnacco, E., Lopez, L., Terribili, C., Montico, M., Zoia, S., & Schön, D. (2015). Music Training Increases Phonological Awareness and Reading Skills in Developmental Dyslexia: A Randomized Control Trial. *PloS One*, *10*(9), e0138715.

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.

Fontolan, L., Morillon, B., Liegeois-Chauvel, C., & Giraud, A.-L. (2014). The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nature Communications*, *5*, 4694.

François, C., Chobert, J., Besson, M., & Schön, D. (2013). Music training for the development of speech segmentation. *Cerebral Cortex* , *23*(9), 2038–2043.

Friederici, A. D. (2020). Hierarchy processing in human neurobiology: how specific is it? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *375*(1789), 20180391.

Fries, P. (2015). Rhythms for cognition: communication through coherence. *Neuron*. https://www.sciencedirect.com/science/article/pii/S0896627315008235

Giordano, B. L., Esposito, M., Valente, G., & Formisano, E. (2023). Intermediate acoustic-to-semantic representations link behavioral and neural responses to natural sounds. *Nature Neuroscience*, *26*(4), 664–672.

Giordano, B. L., Whiting, C., Kriegeskorte, N., Kotz, S. A., Gross, J., & Belin, P. (2021). The representational dynamics of perceived voice emotions evolve from categories to dimensions. *Nature Human Behaviour*, *5*(9), 1203–1213.

Giraud, A.-L., & Poeppel, D. (2012). Speech Perception from a Neurophysiological Perspective. In *The Human Auditory Cortex* (pp. 225–260). Springer, New York, NY.

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Parkkonen, L., & Hämäläinen, M. S. (2014). MNE software for processing MEG and EEG data. *NeuroImage*, *86*, 446–460.

Griffiths, T. D., Kumar, S., Sedley, W., Nourski, K. V., Kawasaki, H., Oya, H., Patterson, R. D., Brugge, J. F., & Howard, M. A. (2010). Direct recordings of pitch responses from human auditory cortex. *Current Biology: CB*, *20*(12), 1128–1132.

Gripari, P. (2004). *La sorcière de la rue Mouffetard*.

Groppe, D. M., Bickel, S., Dykstra, A. R., Wang, X., Mégevand, P., Mercier, M. R., Lado, F. A., Mehta, A. D., & Honey, C. J. (2017). iELVis: An open source MATLAB toolbox for localizing and visualizing human intracranial electrode data. *Journal of Neuroscience Methods*, *281*, 40–48.

Gross, J., Hoogenboom, N., Thut, G., Schyns, P., Panzeri, S., Belin, P., & Garrod, S. (2013). Speech rhythms and multiplexed oscillatory sensory coding in the human brain. *PLoS Biology*, *11*(12), e1001752.

Hasson, U., Malach, R., & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences*, *14*(1), 40–48.

Inbar, M., Grossman, E., & Landau, A. N. (2020). Sequences of Intonation Units form a ~ 1 Hz rhythm. *Scientific Reports*, *10*(1), 15846.

Ince, R. A., Paton, A. T., Kay, J. W., & Schyns, P. G. (2021). Bayesian inference of population prevalence. *eLife*, *10*. https://doi.org/10.7554/eLife.62461

Kayser, C., Ince, R. A. A., & Panzeri, S. (2012). Analysis of slow (theta) oscillations as a potential temporal reference frame for information coding in sensory cortices. *PLoS Computational Biology*, *8*(10), e1002717.

Keitel, A., Gross, J., & Kayser, C. (2017). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. In *PLoS Biol.* (Issue 3, p. e2004473). https://doi.org/10.1371/journal.pbio.2004473

Kivy, P. (1959). Charles Darwin on Music. *Journal of the American Musicological Society*, *12*(1), 42–48.

Koelsch, S. (2011). Toward a neural basis of music perception - a review and updated model. *Frontiers in Psychology*, *2*, 110.

Kopell, N., Ermentrout, G. B., Whittington, M. A., & Traub, R. D. (2000). Gamma rhythms and beta

rhythms have different synchronization properties. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(4), 1867–1872.

Kraus, N., & Chandrasekaran, B. (2010). Music training for the development of auditory skills. *Nature Reviews. Neuroscience*, *11*(8), 599–605.

Le Van Quyen, M., Staba, R., Bragin, A., Dickson, C., Valderrama, M., Fried, I., & Engel, J. (2010). Large-scale microelectrode recordings of high-frequency gamma oscillations in human cortex during sleep. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *30*(23), 7770–7782.

Malik-Moraleda, S., Ayyash, D., Gallée, J., Affourtit, J., Hoffmann, M., Mineroff, Z., Jouravlev, O., & Fedorenko, E. (2022). An investigation across 45 languages and 12 language families reveals a universal language network. *Nature Neuroscience*, *25*(8), 1014–1019.

Martin, S., Millán, J. D. R., Knight, R. T., & Pasley, B. N. (2019). The use of intracranial recordings to decode human language: Challenges and opportunities. *Brain and Language*, *193*, 73–83.

Mas-Herrero, E., Marco-Pallares, J., Lorenzo-Seva, U., Zatorre, R. J., & Rodriguez-Fornells, A. (2013). Individual differences in music reward experiences. *Music Perception*, *31*(2), 118–138.

Medina Villalon, S., Paz, R., Roehri, N., Lagarde, S., Pizzo, F., Colombet, B., Bartolomei, F., Carron, R., & Bénar, C.-G. (2018). EpiTools, A software suite for presurgical brain mapping in epilepsy: Intracerebral EEG. *Journal of Neuroscience Methods*, *303*, 7–15.

Menon, V., Levitin, D. J., Smith, B. K., Lembke, A., Krasnow, B. D., Glazer, D., Glover, G. H., & McAdams, S. (2002). Neural correlates of timbre change in harmonic sounds. *NeuroImage*, *17*(4), 1742–1754.

Mercier, M. R., Bickel, S., Megevand, P., Groppe, D. M., Schroeder, C. E., Mehta, A. D., & Lado, F. A. (2017). Evaluation of cortical local field potential diffusion in stereotactic electro-encephalography recordings: A glimpse on white matter signal. *NeuroImage*, *147*, 219–232.

Mercier, M. R., Dubarry, A.-S., Tadel, F., Avanzini, P., Axmacher, N., Cellier, D., Vecchio, M. D., Hamilton, L. S., Hermes, D., Kahana, M. J., Knight, R. T., Llorens, A., Megevand, P., Melloni, L., Miller, K. J., Piai, V., Puce, A., Ramsey, N. F., Schwiedrzik, C. M., … Oostenveld, R. (2022). Advances in human intracranial electroencephalography research, guidelines and good practices. *NeuroImage*, *260*, 119438.

Millet, J., Caucheteux, C., Orhan, P., Boubenec, Y., Gramfort, A., Dunbar, E., Pallier, C., & King, J.-R. (2022). Toward a realistic model of speech processing in the brain with self-supervised learning. In *arXiv [q-bio.NC]*. arXiv. http://arxiv.org/abs/2206.01685

Morillon, B., & Baillet, S. (2017). Motor origin of temporal predictions in auditory attention. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(42), E8913–E8921.

Nantais, K. M., & Schellenberg, E. G. (1999). The Mozart Effect: An Artifact of Preference. *Psychological Science*, *10*(4), 370–373.

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Human Brain Mapping*, *15*(1), 1–25.

Nieto-Castañón, A., & Fedorenko, E. (2012). Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *NeuroImage*, *63*(3), 1646–1669.

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.

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

Norman-Haignere, S. V., 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: CB*. https://doi.org/10.1016/j.cub.2022.01.069

Oganian, Y., & Chang, E. F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. In *Science Advances* (Issue 11, p. eaay6279). https://doi.org/10.1126/sciadv.aay6279

Oneness. (2006). *Reflejos del Sur*.

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.

Ossandón, T., Vidal, J. R., Ciumas, C., Jerbi, K., Hamamé, C. M., Dalal, S. S., Bertrand, O., Minotti, L., Kahane, P., & Lachaux, J.-P. (2012). Efficient “Pop-Out” Visual Search Elicits Sustained Broadband

Gamma Activity in the Dorsal Attention Network. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(10), 3414–3421.

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (2011). *Statistical Parametric Mapping: The Analysis of Functional Brain Images*. Elsevier.

Pfurtscheller, G., & Lopes da Silva, F. H. (1999). Event-related EEG/MEG synchronization and desynchronization: basic principles. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, *110*(11), 1842–1857.

Proix, T., Delgado Saa, J., Christen, A., Martin, S., Pasley, B. N., Knight, R. T., Tian, X., Poeppel, D., Doyle, W. K., Devinsky, O., Arnal, L. H., Mégevand, P., & Giraud, A.-L. (2022). Imagined speech can be decoded from low- and cross-frequency intracranial EEG features. *Nature Communications*, *13*(1), 48.

Ray, S., & Maunsell, J. H. R. (2011). Different origins of gamma rhythm and high-gamma activity in macaque visual cortex. *PLoS Biology*, *9*(4), e1000610.

Rissman, J., & Wagner, A. D. (2012). Distributed representations in memory: insights from functional brain imaging. *Annual Review of Psychology*, *63*, 101–128.

Robert, P., Zatorre, R., Gupta, A., Sein, J., Anton, J.-L., Belin, P., Thoret, E., & Morillon, B. (2023). Auditory hemispheric asymmetry as a specialization for actions and objects. In *bioRxiv* (p. 2023.04.19.537361). https://doi.org/10.1101/2023.04.19.537361

Rousseau, J.-J. (2009). *Essay on the Origin of Languages and Writings Related to Music*. UPNE.

Saleh, M., Reimer, J., Penn, R., Ojakangas, C. L., & Hatsopoulos, N. G. (2010). Fast and slow oscillations in human primary motor cortex predict oncoming behaviorally relevant cues. *Neuron*, *65*(4), 461–471.

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.

Schön, D., Magne, C., & Besson, M. (2004). The music of speech: music training facilitates pitch processing in both music and language. *Psychophysiology*, *41*(3), 341–349.

Siegel, M., Donner, T. H., & Engel, A. K. (2012). Spectral fingerprints of large-scale neuronal interactions. *Nature Reviews. Neuroscience*, *13*(2), 121–134.

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic Stimuli in Neuroscience: Critically

Acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714.

Steinkamp, S. R. (2019). pymtrf: Translation of the mtrf-Toolbox for Matlab. In *GitHub repository*. https://github.com/SRSteinkamp/pymtrf

Stolk, A., Griffin, S., van der Meij, R., Dewar, C., Saez, I., Lin, J. J., Piantoni, G., Schoffelen, J.-M., Knight, R. T., & Oostenveld, R. (2018). Integrated analysis of anatomical and electrophysiological human intracranial data. *Nature Protocols*, *13*(7), 1699–1723.

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

van Kerkoerle, T., Self, M. W., Dagnino, B., Gariel-Mathis, M.-A., Poort, J., van der Togt, C., & Roelfsema, P. R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(40), 14332–14341.

Vidal, J. R., Freyermuth, S., Jerbi, K., Hamamé, C. M., Ossandon, T., Bertrand, O., Minotti, L., Kahane, P., Berthoz, A., & Lachaux, J.-P. (2012). Long-Distance Amplitude Correlations in the High Gamma Band Reveal Segregation and Integration within the Reading Network. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *32*(19), 6421–6434.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272.

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

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

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.

Zuk, N. J., Murphy, J. W., Reilly, R. B., & Lalor, E. C. (2021). Envelope reconstruction of speech and music highlights stronger tracking of speech at low frequencies. *PLoS Computational Biology*, *17*(9), e1009358.