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], and Daniele Schön[1, 2, 4]

[1]Institute for Language, Communication, and the Brain, Aix-Marseille Univ, Marseille, France
[2]Aix Marseille Univ, Inserm, INS, Inst Neurosci Syst, Marseille, France
[3]APHM, Hôpital de la Timone, Service de Neurophysiologie Clinique, Marseille, France
[4]Senior authorship

Abstract

To what extent do speech and music processing rely on domain-specific and domain-general neural networks? Adopting a dynamical system framework, we investigate the presence of frequency-specific and network-level selectivity and combine it with a statistical approach in which a clear distinction is made between shared, preferred, and category-selective neural responses. Using intracranial EEG recordings in 18 epilepsy patients listening to natural and continuous speech and music, 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 regional selectivity. Instead, neural selectivity is restricted to distributed and frequency-specific coherent oscillations, typical of spectral fingerprints. Our work addresses a longstanding debate and redefines an epistemological posture on how to map cognitive and brain functions.

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

PREPRINT

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

Conflict of interests: The authors declare no competing interests.

Acknowledgments: We thank all patients for their willing participation. We thank Patrick Marquis for helping with the data acquisition and all colleagues from the Institut de Neuroscience des Systèmes for useful discussions.

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., M.M., B.M., and D.S. wrote the paper; N.t.R., M.M., A.T., B.M., and D.S. edited the paper.

Funding sources: ANR-20-CE28-0007-01 (to B.M), ANR-21-CE28-0010 (to D.S), ANR-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI), ANR-17-EURE-0029 (NeuroMarseille), the French government under the Programme "Investissements d'Avenir", the Initiative d'Excellence d'Aix-Marseille Université (A*MIDEX, AMX-19-IET-004).

1 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). Such a question 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 language and other cognitive processes (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; 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 simplest element concerns the choice of the stimuli. Many studies claiming functional selectivity used rather short auditory stimuli (Boebinger et al., 2021; Norman-Haignere et al., 2015, 2022). Besides the low ecological validity of such stimuli, they further rely 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 typically lasts less 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. Another element concerns the data at stake. Most studies that compared language and music processing, examined fMRI data. However, ignoring individual variability in the precise locations of functional areas, by spatially smoothing the results, may favor lack of differences. Taking into account individual variability by examining functional responses at the level of individual brains seems an appropriate choice (Chen et al., 2021; Fedorenko et al., 2011; Nieto-Castañón & Fedorenko, 2012). Nonetheless, generalization at the population level often concluded in favor of highly selective areas or networks and a rather modular view of brain functions. Indeed, studies on the selectivity of speech and music have so far mostly focused on regional selectivity, i.e. the presence of a simple anatomo-functional spatial code (Chen et al., 2021; Fedorenko et al., 2011; Norman-Haignere et al., 2015).

Here, we would like to consider cognition as resulting from interactions among functionally specialized but widely distributed brain networks and adopt a dynamical system 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 to, rather than focusing only on high-frequency

PREPRINT

activity (HFa), a common approach in the literature involving human intracranial EEG recordings (Martin et al., 2019; Norman-Haignere et al., 2022; Oganian & Chang, 2019), investigate the entire frequency spectrum of neural activity. Indeed, while HFa is a good proxy of focal neural spiking (Le Van Quyen et al., 2001; 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 (e.g. only the raw signal or only HFa) 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, 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 HFa). Finally, we carefully distinguished between the three different categories of neural responses described above: shared, selective, and preferred. Our results reveal that the majority of neural responses are shared between natural speech and music, and they highlight an absence of regional selectivity. Instead, we found neural selectivity to be restricted to distributed and frequency-specific coherent oscillations, typical of spectral fingerprints.

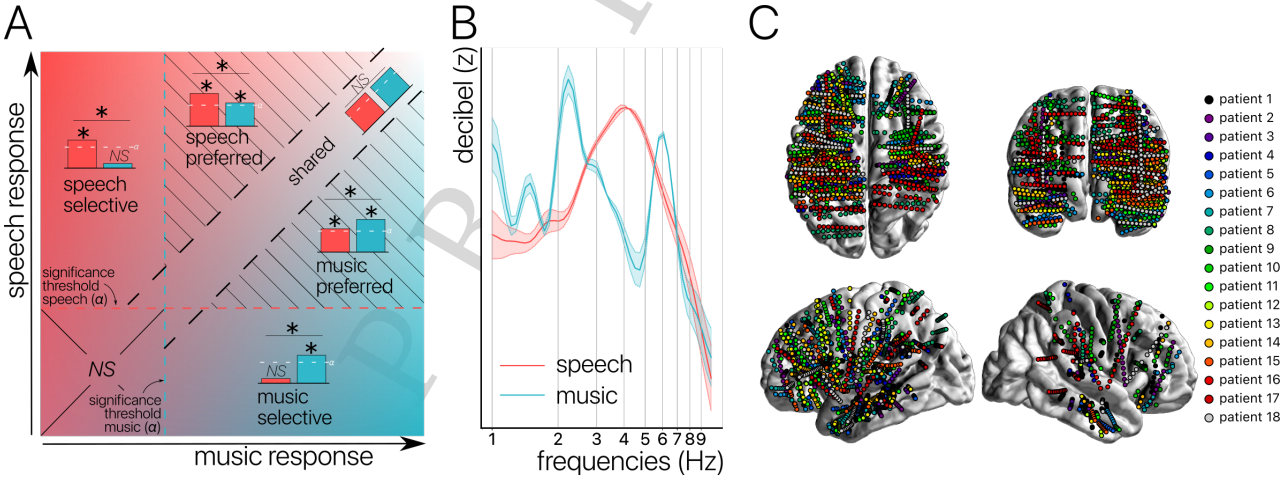


Figure 1: Concepts, Stimuli and Recordings. ***(A) Conceptual definition of selective, shared, and preferred neural processes.*** Illustration of the continua between speech and music selectivity, speech and music preference and shared resources. A "selective" neural response is a response that is significant for one domain but not the other, and where there is a significant difference between domains (for speech top-left; for music bottom-right). "Shared" neural responses occur when there is a significant neural response to at least one of the two stimuli, while there are no significant differences between domains (visible along the diagonal). And, finally, "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). 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*** ($N = 18$).

PREPRINT

2 Results

2.1 Regional neural activity is mostly non-category selective to speech or music.

To investigate the presence of category selectivity during ecological perception of speech and music, we first analyzed the neural response to these two cognitive domains in both a spatially and spectrally resolved manner. Strikingly, in all frequency bands, we observed that the significant neural responses are predominantly shared between the two domains, accounting for 48% (theta; 5−8 Hz) to 64% (low-gamma; 30−50 Hz) of the overall responses (Figure 2). The preferred category is also substantially present, accounting for 30% in the delta band (1−4 Hz) to 35% for the high-frequency activity (80−120 Hz; HFa). The shared and preferred categories then globally dominate all bands, in particular above 15 Hz where they represent >90% of all the responses. Selective responses are much less common, and appear more present in the lower frequency bands, in particular for music listening (Figure 2A-C).

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. 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 (see Table S1). Crucially, there was no evidence of regional selectivity, i.e. we found no evidence of a simple anatomo-functional spatial code. Indeed, firstly, selective responses appear in regions always showing shared responses as well. Secondly, selective responses to music and speech often co-occur in the same regions. This result held between as well as within patients. That is, while we found some regional selectivity at the individual level, selectivity clearly remained rare, with fewer patients showing selective processing than patients showing shared processing (Figure S5). Finally, there seems to be an interaction between regions and frequency bands. For instance, while the left anterior middle temporal gyrus shows some low frequency activity specific to music (Figure 2A), it shows activity rather specific to speech in the low-gamma band (Figure 2E).

Overall, these results reveal an absence of regional selectivity to music or speech under ecological conditions. Instead, selective responses coexist in space across different frequency bands. But, while selectivity may not be striking at the level of 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) and which neural dynamics (low vs. high frequencies) best encode the dynamics of the auditory stimuli.

2.2 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 sound types, we extracted the continuous, instantaneous envelope and the discrete acoustic onset edges (henceforth 'peakRate'; see Methods) and we quantified how well these two sound features are encoded by either the low frequency (LF) or the amplitude of the high frequency activity (HFa; Figure 3A). The model for which most channels significantly encoded the acoustic structure corresponded to the model in which LF neural activity encoded the speech or music peakRates (Figure 3A). In particular, the low frequency signal encodes the acoustic signal much better than does the amplitude of the HFa (peakRate & LF vs. peakRate & HFa comparison). As a side note, these results also show that the peakRates are better encoded by neural activity than the instantaneous envelope. Thus, we replicate previous findings showing that neural responses to the speech envelope primarily reflect peakRate landmarks and not the instantaneous envelope (Doelling et al., 2014; Oganian & Chang, 2019).

Furthermore, we show that the peakRates are encoded by the LF neural activity throughout the cortex, for both speech and music (Figure 3B-C). More precisely, the regions wherein neural activity significantly

PREPRINT

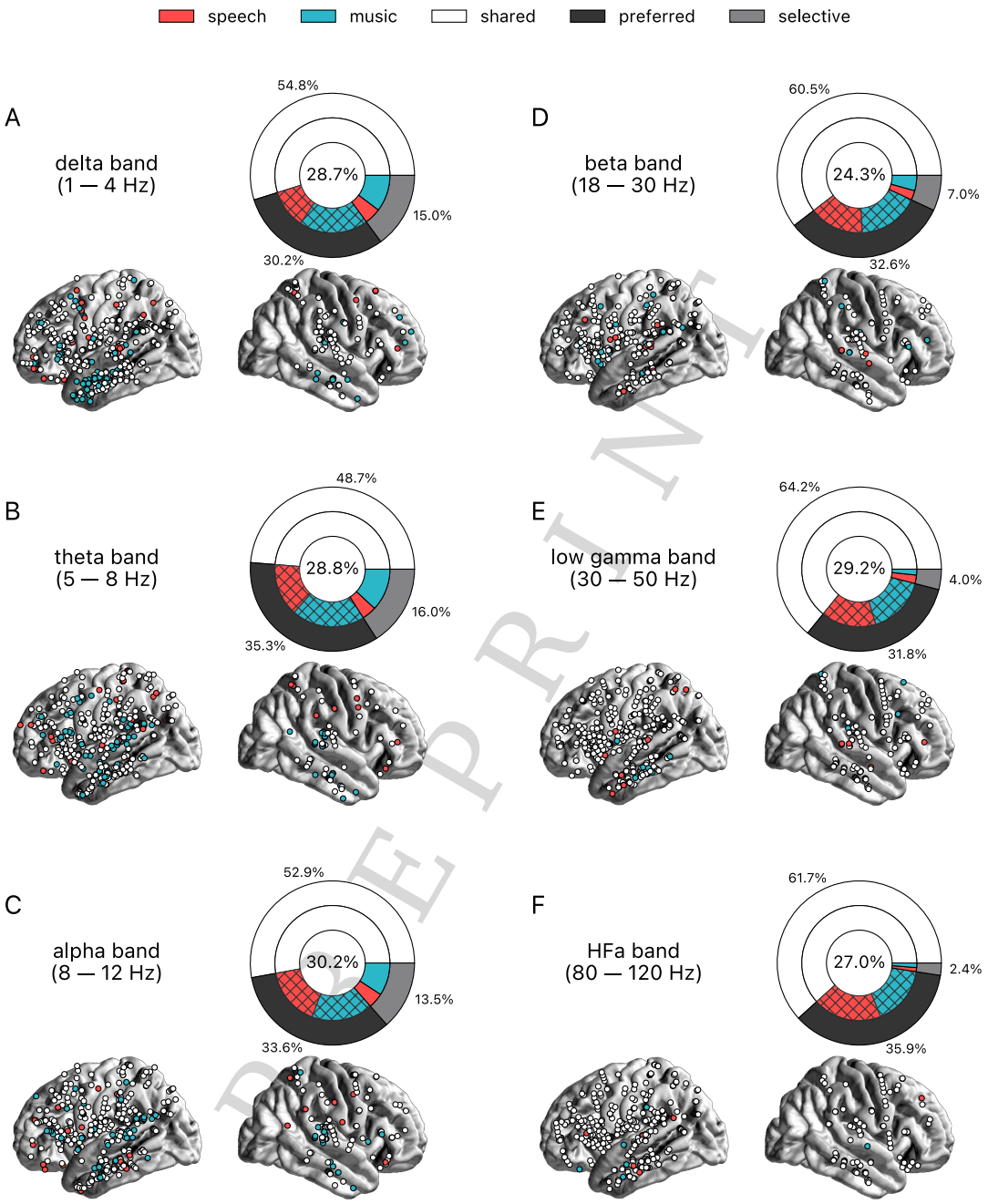


Figure 2: Power spectrum analyses. A-F. Regional neural responses to speech and/or music for the 6 canonical frequency bands. ***Nested pie charts indicate***: (1) in the center, the percentage of channels that showed a strong response to speech and/or music. A response was considered strong if it exceeded 1 standard deviation above the mean neural response across channels. (2) The outer pie indicates the percentage, relative to the center, of shared (white), selective (light gray) and preferred (dark gray) channels. (3) The inner pie indicates, for the selective (plain) and preferred (pattern) categories, the proportion of channels that were 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 (note that, here, the preferred category is not represented). Results are significant at $q < 0.01$ ($N = 18$).

PREPRINT

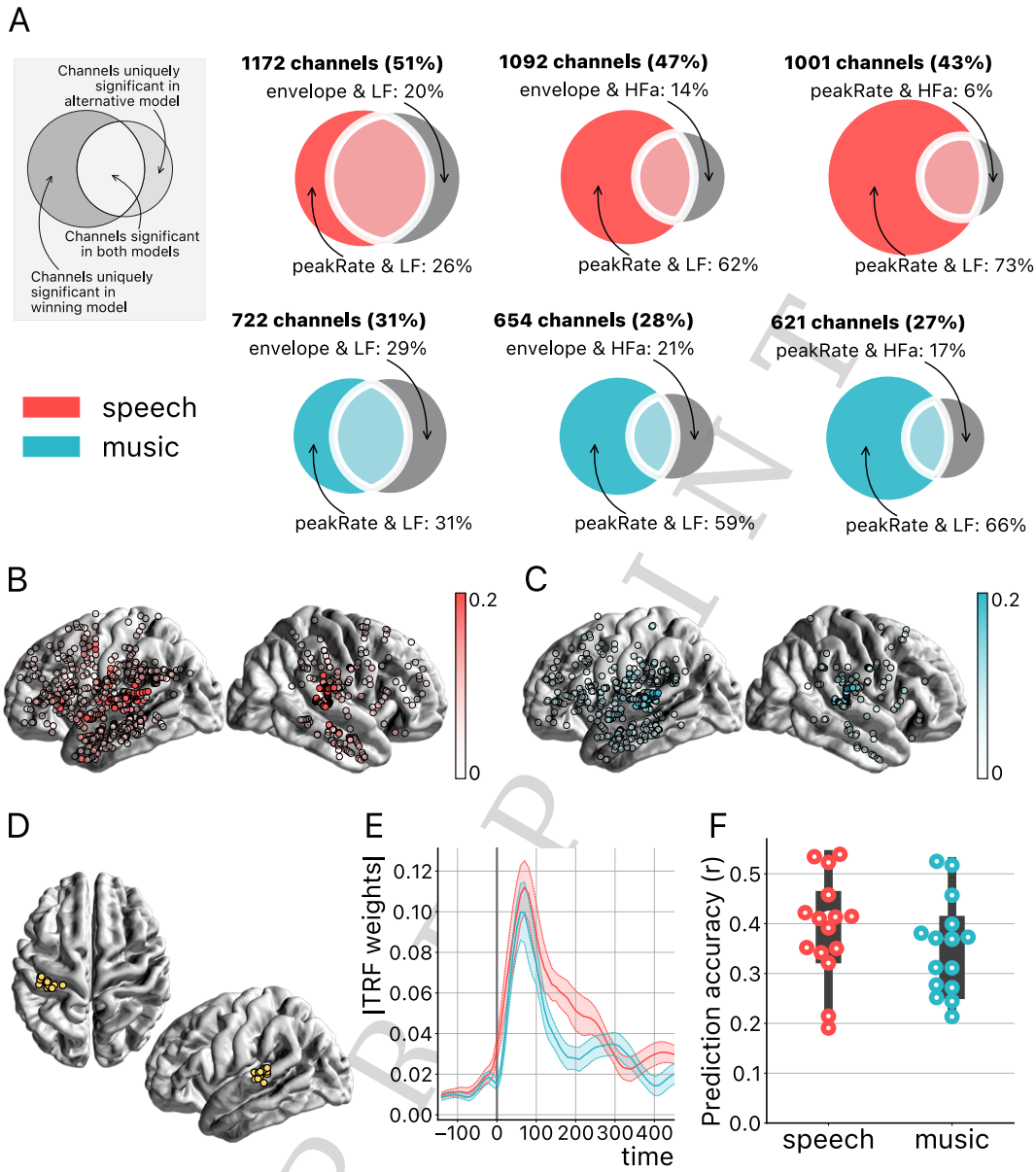


Figure 3: Temporal Response Function (TRF) analyses. ***(A) Venn diagrams comparing the winning model*** (peakRate in LF) to each of the three other models (see top left toy model) for speech (top, red) and music (bottom, blue). Four TRF models were investigated to quantify the encoding of the instantaneous envelope and the discrete acoustic onset edges (peakRate) by either the low frequency (LF) or the amplitude of the high frequency activity (HFa). Percentages indicate the ratio of significant channels for each model comparison. The peakRate & LF model captures the largest proportion of channels. ***(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). 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.

PREPRINT

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 3D). 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. Finally, the peakRate & LF model, i.e. the model that captures the largest proportion of significant channels for both speech and music perception (Figure 3A), is associated with similar TRF for speech and music (Figure 3E) and prediction accuracy scores (Pearson's r) of up to 0.55 (Figure 3F).

2.3 Domain selectivity occurs at the level of frequency-specific distributed networks.

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 40% of the channels significantly connected at delta rate, and only 10% at HFa (Figure 4). Across frequencies, most of the significant connections are shared between the two cognitive domains, followed by preferred connections and with selective connections in the minority. Selectivity is nonetheless present across frequencies, with a decreasing gradient of selectivity from low to high frequencies. In other words, selectivity is mostly observed for connections in the delta range (26% of the significant delta connections) and only rarely in HFa (5% of the significant HFa connections).

Comparing regional activity (Figure 2) and network-level connectivity (Figure 4), we observe that the latter reveals more selectivity (e.g. 15% vs. 26% of the significant contacts at delta rate). However, here again no regional selectivity is observed, i.e. not a single cortical region is solely selective to speech or music. Rather, in every cortical region showing selective channels, there is a mix of shared, speech selective, music selective, or preferred channels, both between as well as within patients (see population prevalence in Figure S6). Finally, selectivity is again frequency-specific. Hence, selectivity, in addition to being the exception rather than the rule, is only visible at the level of frequency-specific distributed networks.

3 Discussion

In this study, we investigated the existence of category 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 regional selectivity when patients listened to a tale or to instrumental music. With this approach, we addressed the fundamental question of the extent to which natural speech and music processing is performed by shared neural populations and by category-selective neural populations. By combining classic sEEG investigations of high-frequency activity (HFa) with the analyses of other frequency bands, 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 strikingly show that speech and music mostly rely on shared resources. Further, while selective responses seem absent at the level of cortical regions, selectivity can be observed at the level of frequency-specific distributed networks. Previous work has reported that written or spoken language selectively activated diverse left-lateralized functional cortical regions (Chen et al., 2021; Fedorenko et al., 2011; Fedorenko & Blank, 2020; Malik-Moraleda et al., 2022). In particular, using functional MRI, it was found that these strong and selective cortical responses were not visible during the presentation of musical excerpts (Chen et al., 2021; Fedorenko et al., 2011). A direct interpretation of these results would state that language and music are processed by distinct anatomical substrates. The link between structure and

PREPRINT

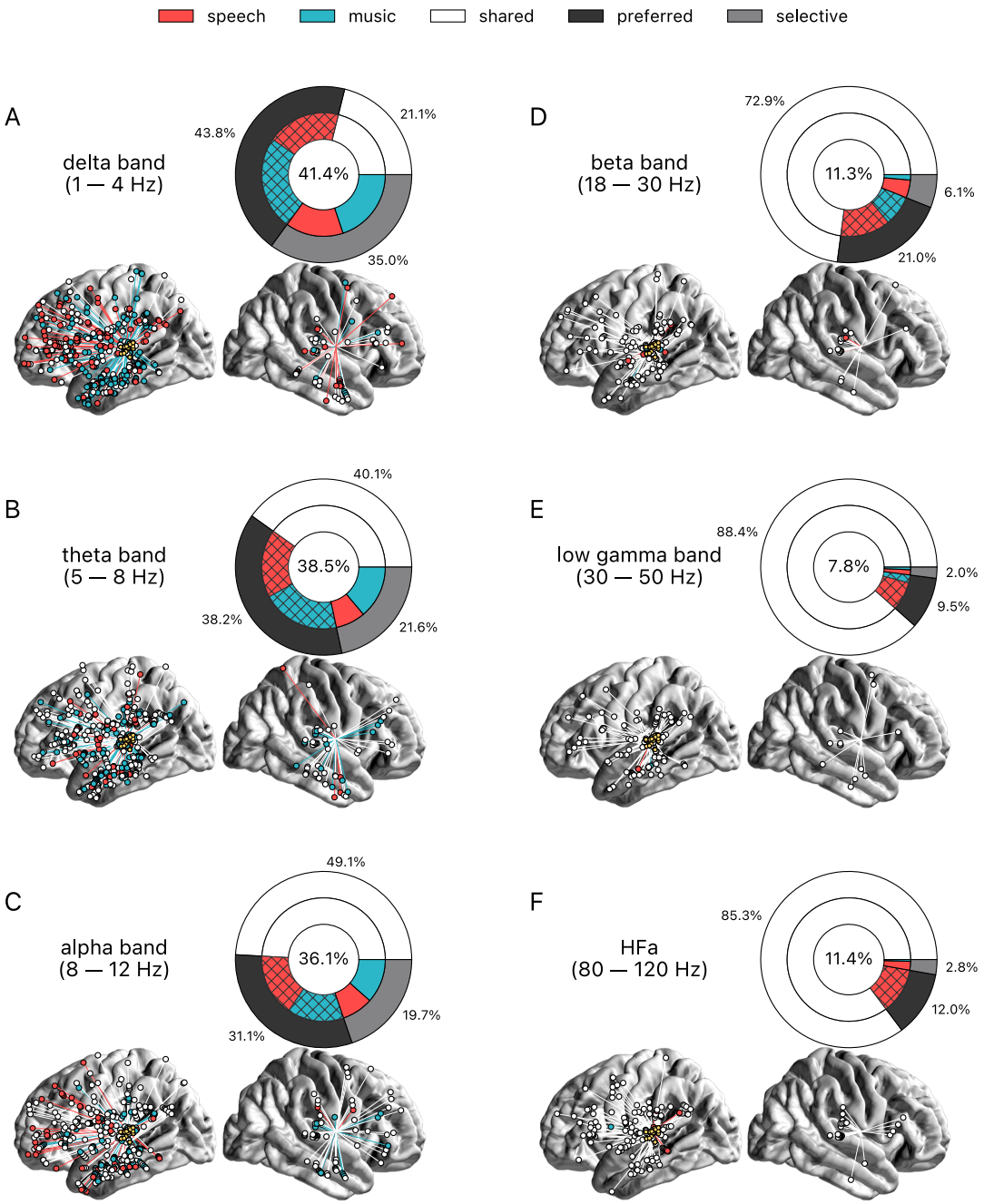


Figure 4: Seed-based functional connectivity analyses. ***(A)-(E). Coherence*** and ***(F) Amplitude correlation*** analyses, during speech and music perception. The seed was located in the left auditory cortex (see Figure 3D). Same conventions as in Figure 2, except for the center percentage in the nested pie charts which, here, reflects the percentage of channels significantly connected to the seed. Results are significant at $q < 0.01$ ($N = 15$).

function can also be envisioned through a more subtle articulation of brain networks with shared, preferred, or selective neural responses. Within this framework, the localisasionsim view applies to highly specialized processes (i.e. functional niches), while general domains are mostly 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), spectral modulations for melodies, etc (Albouy et al., 2020; Arnal et al., 2015, 2019; Flinker et al., 2019). Which features —which neural mechanisms— are necessary and sufficient to route communicative sounds towards selective neural networks remains a promising field of investigation to explore. An exciting avenue would be to anchor our approach in an ontogenetic and phylogenetic perspective (Tinbergen, 1963) which is most appropriate to establish hypotheses about the emergence of specialized networks. In the current study, we did not observe a single cortical region for which speech-selectivity was present, in any of our analyses. On the contrary, we report patterns of shared, preferred, or selective neural responses and connectivity fingerprints. Several non-exclusive explanations may account for this ground-breaking finding: First, our results rest on a conceptually and statistically rigorous definition of the concept of selectivity (Figure 1). Importantly, we part with the simple selective versus shared dichotomy and adopt a more biologically valid and continuous framework (Buzsàki, 2019; Zatorre & Gandour, 2008) by adding a new category that is often neglected in the literature: preferred responses. Indeed, responses in this category are usually reported as shared or selective and most often the statistical approach does not allow to have a more nuanced view (*cf.* Chen et al., 2021). 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 redefines an epistemological posture 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 HFa (Bellier et al., 2022; Norman-Haignere et al., 2020; Oganian & Chang, 2019). However, while HFa 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). Finally, most studies only investigated local neural activity and did not consider the brain as a distributed dynamical 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. 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),

PREPRINT

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. 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). We further confirmed previous findings showing that neural responses to the speech envelope primarily reflect acoustic onset edge (peakRate) landmarks and not the instantaneous envelope (Doelling et al., 2014; Oganian & Chang, 2019). 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 signal clearly better encodes the acoustic signal than the HFa (Figure 3A).

In conclusion, we emphasize that adopting an a priori stance towards regional selectivity could prevent us from understanding the true nature of neural computations, because, if we eventually consider that there is language-selectivity while there is not, we are doomed to fail our understanding of the computations because we will only probe the system with linguistic stimuli. While we adopted 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 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.

4 Methods

4.1 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).

4.2 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. Patients were recorded either in an insulated Faraday cage ($N = 6$) or in the bedroom. In the Faraday cage, data were recorded using a 256-channels amplifier (Brain Products), sampled at 1 kHz 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.

PREPRINT

4.3 Experimental design.

Patients passively listened to 10 minutes of storytelling (576,7 secs, La sorcière de la rue Mouffetard, Gripari, 2004) and 10 minutes of music (580.36 secs, Reflejos del Sur, Oneness, 2006) separated by 3 minutes of rest. The order of conditions was counterbalanced across patients (see Table S1). 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).

4.4 General preprocessing.

To increase spatial sensitivity and reduce passive volume conduction from neighboring regions (Mercier et al., 2017), the signal was offline re-referenced using bipolar montage. That is, for a pair of adjacent electrodes, the rereferencing led to a virtual channel located at the mid-point of locations the original electrodes. To precisely localize the channels, a procedure similar to the one used in the iELVis toolbox was applied (Groppe et al., 2017). 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 the MNI template (MNI 152 Linear), 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 Destrieux atlas (Destrieux et al., 2010). 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.

4.5 Extraction of high-frequency activity (HFa).

A bandpass-Hilbert approach was used to extract the high frequency activity (HFa, $80-120$ Hz), with four 10 Hz wide sub-bands spanning from 80 to 120 Hz. For each sub-band, the signal was bandpass-filtered and the Hilbert transform was computed to obtain the envelope. Each sub-band envelope was standardized by dividing it by its mean and, finally, all sub-bands were averaged together to give the HFa estimate (Ossandón et al., 2012; Vidal et al., 2012).

4.6 Artifact rejection.

To define artifacted channel we used both the broadband (raw) and HFa signals, channels with a variance greater than $2 * IQR$ (interquartile range, i.e. a non-parametric estimate of the standard deviation) were tagged as artifacted channels (on average 18% of the channels). Then the data (again both the broadband and the HFa) 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 4.8). Statistical definition of selective, preferred, and shared categories.

PREPRINT

We defined these categories by capitalizing on both the main effects of —and contrast between— the neural responses to speech and music stimuli (see Figure 1A). A "selective" neural response is a response significant for one domain (speech or music) but not the other, and with a significant difference (one main effect and difference effect). "Preferred" responses correspond to neural responses that occur during both speech and music processing, but with a significantly stronger response for one of the two domains (two main effects and difference effect). Finally, "shared" neural responses occur when there are no significant differences between domains, and that there is a significant neural response to at least one of the two stimuli (one or two main effects and no difference effect). If none of the two domains produces a significant neural response, the difference is not investigated (case "neither" main effect).

4.7 Spectral decomposition.

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 HFa ($80-120$ Hz, see section 4.5). For the five lower frequency bands, the oscillatory power of each sEEG channel was calculated using the Continuous Morlet Wavelet Transform (CWT; Morlet 1983). More specifically, the continuous (unepoched) broadband data was zero-padded with 5 seconds segments to prevent edge artifacts and then decomposed in 100 frequencies that logarithmically ranged from 1 to 150 Hz. We used the Morlet wavelet transform implemented in the MNE-python function time_frequency.tfr_array_morlet using 7 cycles (Gramfort et al., 2014). Afterwards, the 5 second zero-pads and the previously tagged bad epochs (see section 4.6) were removed and remaining data were averaged in non-overlapping epochs of 5 seconds. Finally, data were averaged across the relevant frequencies to obtain the 5 canonical bands.

In order to categorize each channel in one of the categories mentioned above (shared, selective, and preferred; Figure 1A), for each frequency band, we examined (i) the main effects (i.e. where in the brain do we observe a strong response to speech and/or music) and (ii) the difference effect (i.e. where do we observe a significantly greater response to one of the sounds over the other). (i) Those channels where the neural response was strongest, relative to the power at the other channels, were considered significant, as determined by a 'significance threshold' (1 standard deviation above the mean) from the distribution of the power (averaged between the speech and music data) across non-excluded channels. While examining these distributions, we noticed a few remaining artifacted outlier channels with possibly non-physiological activity values. Therefore, we first dismissed these channels by applying an exclusion criterion to all channels where the average power exceeded 2 standard deviations above the mean, prior to determining the significance threshold (as described above). These criteria led to about 30% of the channels to be considered as significant in each frequency band. As a sanity check, we verified that neither the additional outlier rejection procedure, nor the exact 'significance threshold' (be it 1, 2 or 3 standard deviations above the mean), changed the observed pattern in the results. That is, the proportions of channels belonging to the different categories (shared, selective, preferred) and their spatial distribution (spatial overlap between categories) was robust with regard to the threshold used to define significance. (ii) For the difference effect, we used a standard Student's t-test (FDR corrected) comparing power values between conditions over epochs, for each channel and each frequency band.

4.8 Temporal Response Function (TRF) analyses.

We used the Temporal Response Function (TRF) technique to estimate the encoding of the acoustic stimuli by neural activity. 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 stimulus acoustic features and neural activity. 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 tested four

PREPRINT

models wherein we examined how well two different acoustic features —envelope and peakRate (see below; Oganian & Chang, 2019)— map onto two types of neural activity —low frequency activity (LF; $1-9$ Hz) and amplitude of the HFa ($80-120$ Hz; see section 4.5). 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.

Envelope and peakRate acoustic features. To estimate the temporal envelope of the two stimuli (speech, music), the auditory 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. Next, and following previous findings that showed the relevance of acoustic onset edges in speech perception (Doelling et al., 2014; Oganian & Chang, 2019), we also estimated them from the speech and music envelopes (henceforth 'peakRate'). These were defined as peaks in the rate of change (first derivative) of the envelope. Finally, both the envelopes and peakRates were downsampled to 100 Hz and z-scored to be mapped to the neural data in the TRF analyses.

4.9 Connectivity analysis.

We used a seed-based approach to examine the frequency-specific functional connectivity maps in response to speech and music between the entire brain and the auditory cortex (excluding the channels immediately neighboring the seed-channels). All analyses were done using the MNE-python library (Gramfort et al., 2014). As seeds we selected those channels that were revealed to best encode the speech and music stimuli by our TRF analyses (see Figure 3). The same canonical frequency bands as in the spectral analysis were investigated: delta ($1-4$ Hz), theta ($5-8$ Hz), alpha ($8-12$ Hz), beta ($18-30$ Hz), low-gamma ($30-50$ Hz), and HFa ($80-120$ Hz). For the five lower frequency bands, we used spectral coherence as measure of connectivity, while for HFa, wherein the phase is unlikely to be synchronized, we focused on envelope correlations. For each frequency band, we categorized each channel in one of the categories mentioned above (see Figure 1A) by again examining both the main effects (i.e. which channels show coherent signal with the auditory cortex in response to speech and/or music) and the difference effects (i.e. which channels reveal significantly greater connectivity in response to one sound over the other).

The approach we followed to compute coherence was slightly different between our main effects and difference effects. That is, for the main effect, for which we used surrogate data, we used a discrete Fourier decomposition to lighten computation times, while for the difference effect, we computed the spectral coherence using continuous Morlet wavelets. These two spectral decomposition methods are generally considered as equivalent and lead to similar results (Bruns, 2004; Le Van Quyen et al., 2001). For the main effects, for the coherence measures (i.e. when examining the connectivity patterns in the 5 lower frequency bands), the epoched data was decomposed in the frequency bands of interest using the discrete Fourier transform. To prevent edge artifacts, prior to extracting the bands, epochs were zero-padded with 3.5-seconds segments which were later removed. For the envelope correlations, pairwise non-orthogonal envelope correlations were assessed for the epoched HFa between the seed and the rest of the channels and then Fisher-z-scored. Statistical significance of the connectivity values was assessed using surrogate data with 100 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 100 connectivity values for each channel, which were used as null-distribution to calculate the probability threshold associated with genuine connectivity. For the difference effects, we computed the spectral coherence using Morlet wavelets to time-frequency decompose the zero-padded, epoched data in 100 frequencies that logarithmically ranged from 1 to 120 Hz. Connectivity maps corresponding to the

PREPRINT

delta, theta, alpha, beta, and low-gamma bands were reconstructed by averaging the coherence values calculated for each individual frequency within the relevant range. The spectral coherence was computed for each time-point across epochs, so that we obtained 2500 measures (5 sec $*$ 500 samples per epoch). Then we computed a sliding non-overlapping average (smoothing window = 25 measures) to obtain 100 robust estimates of connectivity for each channel (i.e. this procedure reduces the risk of observing spurious high connectivity estimates). Finally, significance was assessed using a standard permutation test, i.e. we generated a null-distribution with 1000 iterations wherein, at each iteration, we switched condition-labels of a random number of epochs prior to subtracting. We then used this null-distribution to calculate the probability threshold associated with a true difference effect.

PREPRINT

References

Albouy, P., Benjamin, L., Morillon, B., & Zatorre, R. J. 2020. Distinct sensitivity to spectrotemporal modulation supports brain asymmetry for speech and melody. *Science* 367.1043–1047.

Arnal, L. H., Flinker, A., Kleinschmidt, A., Giraud, A.-L., & Poeppel, D. 2015. Human screams occupy a privileged niche in the communication soundscape. *Curr. Biol.* 25.2051–2056.

Arnal, L. H. & Giraud, A.-L. 2012. Cortical oscillations and sensory predictions. *Trends Cogn. Sci.* 16.390–398.

Arnal, L. H., Kleinschmidt, A., Spinelli, L., Giraud, A.-L., & Mégevand, P. 2019. The rough sound of salience enhances aversion through neural synchronisation. *Nat. Commun.* 10.3671.

Bastos, A. M., Lundqvist, M., Waite, A. S., Kopell, N., & Miller, E. K. 2020. Layer and rhythm specificity for predictive routing. *Proc. Natl. Acad. Sci. U. S. A.* 117.31459–31469.

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* 85.390–401.

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.

Bizley, J. K. & Cohen, Y. E. 2013. The what, where and how of auditory-object perception. *Nat. Rev. Neurosci.* 14.693–707.

Boebinger, D., Norman-Haignere, S. V., McDermott, J. H., & Kanwisher, N. 2021. Music-selective neural populations arise without musical training. *J. Neurophysiol.* 125.2237–2263.

Bruns, A. 2004. Fourier-, Hilbert- and wavelet-based signal analysis: are they really different approaches? *J. Neurosci. Methods* 137.321–332.

Buzsàki, 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. 2021. The human language system does not support music processing.

Chi, T., Ru, P., & Shamma, S. A. 2005. Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* 118.887–906.

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. *Front. Hum. Neurosci.* 10.604.

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. 2010. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage* 53.1–15.

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.

PREPRINT

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. *J. Neurosci.* 40.4536–4550.

Ding, N., Melloni, L., Yang, A., Wang, Y., Zhang, W., & Poeppel, D. 2017. Characterizing Neural Entrainment to Hierarchical Linguistic Units using Electroencephalography (EEG). *Front. Hum. Neurosci.* 11.481.

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 2.761–768.

Donhauser, P. W. & Baillet, S. 2019. Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron*.

Elman, J. L. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Mach. Learn.* 7.195–225.

Fadiga, L., Craighero, L., & D'Ausilio, A. 2009. Broca's area in language, action, and music. *Ann. N. Y. Acad. Sci.* 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. *Cereb. Cortex* 26.3508–3526.

Fedorenko, E., Behr, M. K., & Kanwisher, N. 2011. Functional specificity for high-level linguistic processing in the human brain. *Proc. Natl. Acad. Sci. U. S. A.* 108.16428–16433.

Fedorenko, E. & Blank, I. A. 2020. Broca's Area Is Not a Natural Kind. *Trends Cogn. Sci.* 24.270–284.

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.e0138715.

Flinker, A., Doyle, W. K., Mehta, A. D., Devinsky, O., & Poeppel, D. 2019. Spectrotemporal modulation provides a unifying framework for auditory cortical asymmetries. *Nat Hum Behav* 3.393–405.

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. *Nat. Commun.* 5.4694.

François, C., Chobert, J., Besson, M., & Schön, D. 2013. Music training for the development of speech segmentation. *Cereb. Cortex* 23.2038–2043.

Friederici, A. D. 2020. Hierarchy processing in human neurobiology: how specific is it? *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 375.20180391.

Fries, P. 2015. Rhythms for cognition: communication through coherence. *Neuron*.

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. *Nat Hum Behav* 5.1203–1213.

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

PREPRINT

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. *Curr. Biol.* 20.1128–1132.

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. *J. Neurosci. 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 Biol.* 11.e1001752.

Inbar, M., Grossman, E., & Landau, A. N. 2020. Sequences of Intonation Units form a 1 Hz rhythm. *Sci. Rep.* 10.15846.

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 Comput. Biol.* 8.e1002717.

Keitel, A., Gross, J., & Kayser, C. 2017. Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features.

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. *Proc. Natl. Acad. Sci. U. S. A.* 111.14332–14341.

Kivy, P. 1959. Charles Darwin on Music. *Journal of the American Musicological Society* 12.42–48.

Koelsch, S. 2011. Toward a neural basis of music perception - a review and updated model. *Front. Psychol.* 2.110.

Kopell, N., Ermentrout, G. B., Whittington, M. A., & Traub, R. D. 2000. Gamma rhythms and beta rhythms have different synchronization properties. *Proc. Natl. Acad. Sci. U. S. A.* 97.1867–1872.

Kraus, N. & Chandrasekaran, B. 2010. Music training for the development of auditory skills. *Nat. Rev. Neurosci.* 11.599–605.

Le Van Quyen, M., Foucher, J., Lachaux, J., Rodriguez, E., Lutz, A., Martinerie, J., & Varela, F. J. 2001. Comparison of Hilbert transform and wavelet methods for the analysis of neuronal synchrony. *J. Neurosci. Methods* 111.83–98.

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. *Nat. Neurosci.* 25.1014–1019.

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 Lang.* 193.73–83.

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. *J. Neurosci. Methods* 303.7–15.

PREPRINT

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* 17.1742–1754.

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.

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.

Morillon, B. & Baillet, S. 2017. Motor origin of temporal predictions in auditory attention. *Proc. Natl. Acad. Sci. U. S. A.* 114.E8913–E8921.

Nieto-Castañón, A. & Fedorenko, E. 2012. Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. *Neuroimage* 63.1646–1669.

Norman-Haignere, S., Kanwisher, N. G., & McDermott, J. H. 2015. Distinct Cortical Pathways for Music and Speech Revealed by Hypothesis-Free Voxel Decomposition. *Neuron* 88.1281–1296.

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.

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. *Curr. Biol.*

Oganian, Y. & Chang, E. F. 2019. A speech envelope landmark for syllable encoding in human superior temporal gyrus.

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. *Comput. Intell. Neurosci.* 2011.156869.

Ossandón, T., Vidal, J. R., Ciumas, C., Jerbi, K., & others. 2012. Efficient “pop-out” visual search elicits sustained broadband gamma activity in the dorsal attention network. *Journal of*.

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.

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. *Nat. Commun.* 13.48.

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

Rissman, J. & Wagner, A. D. 2012. Distributed representations in memory: insights from functional brain imaging. *Annu. Rev. Psychol.* 63.101–128.

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* 65.461–471.

PREPRINT

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* 51.450–461.

Schön, D., Magne, C., & Besson, M. 2004. The music of speech: music training facilitates pitch processing in both music and language. *Psychophysiology* 41.341–349.

Siegel, M., Donner, T. H., & Engel, A. K. 2012. Spectral fingerprints of large-scale neuronal interactions. *Nat. Rev. Neurosci.* 13.121–134.

Steinkamp, S. R. 2019. pymtrf: Translation of the mtrf-Toolbox for Matlab.

Tinbergen, N. 1963. On aims and methods of Ethology. *Z. Tierpsychol.* 20.410–433.

Vidal, J. R., Freyermuth, S., Jerbi, K., & others. 2012. Long-distance amplitude correlations in the high gamma band reveal segregation and integration within the reading network. *Journal of*.

Zatorre, R. J., Chen, J. L., & Penhune, V. B. 2007. When the brain plays music: auditory–motor interactions in music perception and production. *Nat. Rev. Neurosci.* 8.547–558.

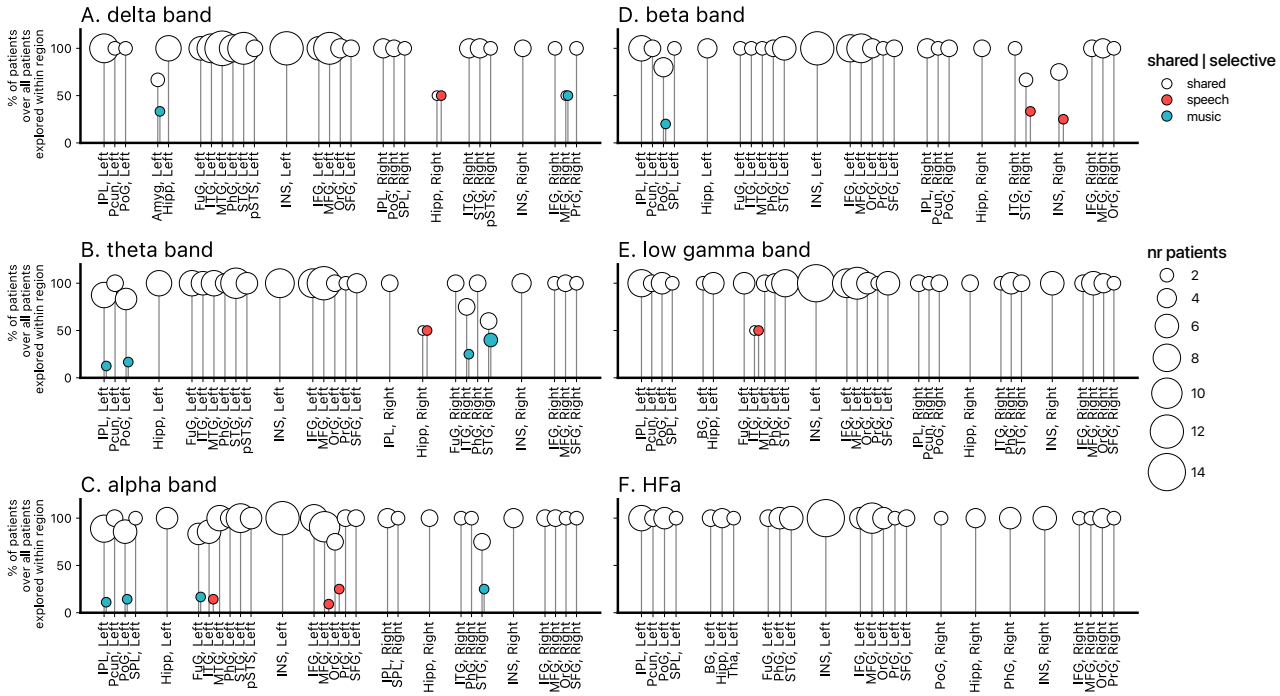
Zatorre, R. J. & Gandour, J. T. 2008. Neural specializations for speech and pitch: moving beyond the dichotomies. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 363.1087–1104.

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* 77.980–991.

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 Comput. Biol.* 17.e1009358.

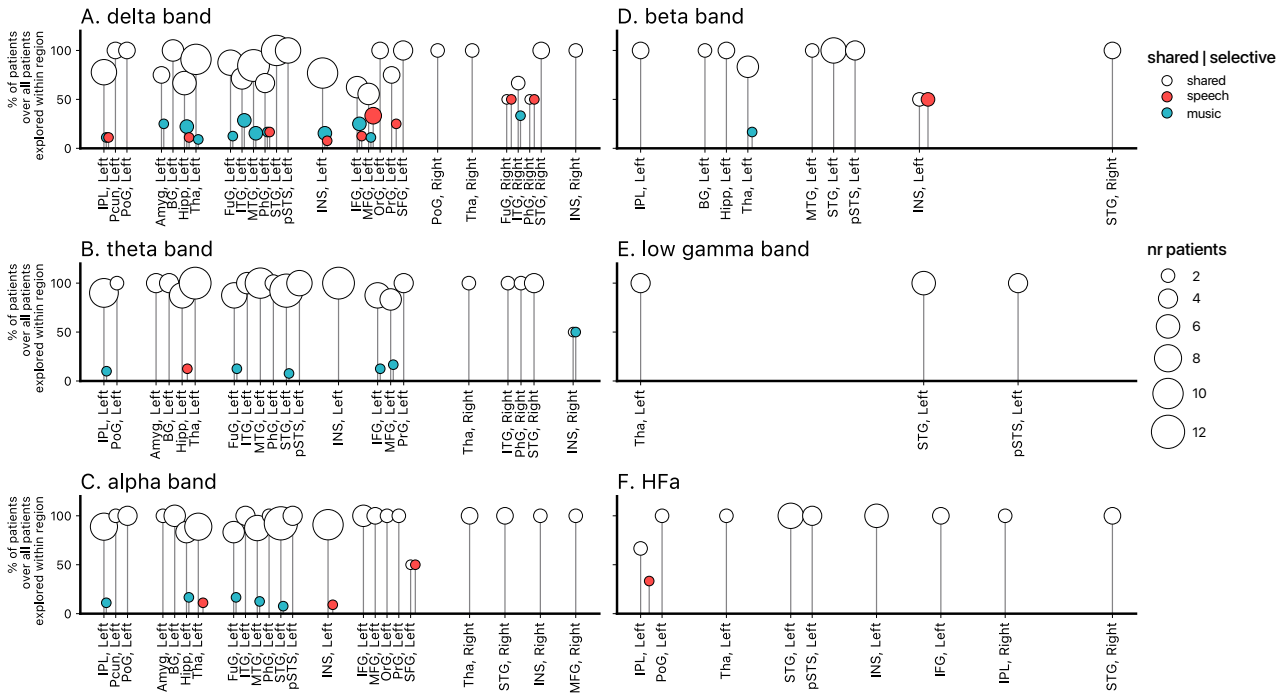
PREPRINT

S5 Supplementary



Supplemental Figure S5: Population prevalence for the power spectral analyses (N = 18). ***(A)-(F) Lollipop figure indicating population prevalence*** of shared or selective responses for the 6 canonical frequency bands per region (note that preferred responses were again excluded). 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 to either speech (red) or music (blue). When channels showed a combination of speech selective, music selective or shared responses, the patient was considered to show shared (white) processing. 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 B (theta 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 nine explored shows music selective processing (filled blue circle). Abbreviations according to the Brainnetome Atlas (Fan et al., 2016).

PREPRINT



Supplemental Figure S6: Population prevalence for the seed-based functional connectivity analyses. (N = 15). (A)-(F) Same conventions as in Figure S5.

PREPR

Supplemental Table S1: Patients description.

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