Auditory features modelling demonstrates sound envelope representation in striate cortex

Author list: Alice Martinelli*[1], Giacomo Handjaras*[1], Monica Betta[1], Andrea Leo[2], Luca Cecchetti[1], Pietro Pietrini[1], Emiliano Ricciardi[1], Davide Bottari[1]

Affiliations:

1 *Molecular Mind Lab, IMT School for Advanced Studies Lucca, Italy*

2 *Department of translational research and advanced technologies in medicine and surgery, University of Pisa, Italy*

*Shared

Corresponding author

Davide Bottari

davide.bottari@imtlucca.it

Competing financial interest

The authors declare no competing financial interests.

Summary

Primary visual cortex is no longer considered exclusively visual in its function. Proofs that its activity plays a role in multisensory processes have accrued. Here we provide evidence that, in absence of retinal input, V1 maps sound envelope information. We modeled amplitude changes occurring at typical speech envelope time-scales of four hierarchically-organized categories of natural (or synthetically derived) sounds. Using functional magnetic resonance, we assessed whether sound amplitude variations were represented in striate cortex and, as a control, in the temporal cortex. Sound amplitude mapping in V1 occurred regardless of the semantic content, was dissociated from the spectral properties of sounds and was not restricted to speech material. As in the temporal cortex, a spatially organized representation of amplitude modulation frequencies emerged in V1. Our results demonstrate that human striate cortex is a locus of representation of sound attributes.

Keywords

Sound Envelope, Striate Cortex, Temporal Cortex, Cross-modal, Speech, MVPA, fMRI

Introduction

Perception has been considered as highly segregated for more than a century. The existence of specialized detectors for different forms of environmental energies has prompted the enduring dominant paradigm which postulates that information from different sensory modalities is anatomo-functionally segregated in the human brain (Pascual-Leone and Hamilton, 2001; Scholvinck et al., 2010). This assumption received strong supporting evidence from seminal lesion studies demonstrating that unimodal deficits were associated to lesions in primary sensory cortices (Massopust et al., 1965; Winans, 1967). The visual system in particular has been thought to be functionally independent of cross-modal influences, and, if anything, to dominate the other senses (McGurk, 1976; Pick H. L., 1969). Despite more than a century of neurological studies (Riddoch, 1917), to what extent the primary visual cortex can be considered as exclusively unisensory remains debated.

With the advent of multisensory research (Stein and Meredith, 1993), the last decades have witnessed the challenging of such traditional view. Indeed, a burgeoning of animal and human studies proposed that cortical areas, previously believed to only process unisensory information, would be multisensory in nature instead (Ghazanfar and Schroeder, 2006; Schroeder and Foxe, 2005).

Multisensory audio-visual stimulations were found to subtly modulate the striate cortex activity in rodents (Ibrahim et al., 2016) and the awake monkey (Wang et al., 2008). The anatomical scaffolding for early audio-visual interactions has been provided by tracing studies. Sparse monosynaptic (i.e. direct) anatomical projections originating from auditory areas and terminating in primary visual cortex have been consistently observed (Falchier et al., 2002; Rockland and Ojima, 2003; Clavagnier et al., 2004; Charbonneau et al., 2012).

In humans, the corroboration that, at least to some degree, multisensory interplay occurs already at the level of V1 has been found in studies adopting multiple methodologies (for a review see Murray et al., 2016), such as hemodynamic imaging (Eckert et al., 2008; Martuzzi et al., 2007; Rohe and Noppeney, 2016; Zangenehpour and Zatorre, 2010), electrophysiological measures (Giard and Peronnet, 1999; Mercier et al., 2013; Molholm et al., 2004), as well as magnetic stimulations (Romei et al., 2009). To challenge even more the modality-segregated view, scattered evidence of pure cross-modal responses in V1, evoked by unimodal inputs of other sensory modalities, exist. Local GABAergic inhibition of V1 neurons have been linked to auditory stimulations in mice (Iurilli et al., 2012). In the cat, a shift of the preferred orientation tuning curves of striate cortex neurons was associated to prolonged exposition to noise-like auditory stimuli (Chanauria et al., 2019). Auditory-driven modulation of the ongoing oscillatory activity in striate cortex has been observed in humans (Mercier et al., 2013). Moreover, it was shown that the spatial patterns of the BOLD response elicited in primary visual cortex by a simple auditory or tactile stimulus allowed to classify the type of sensory input (Liang et al., 2013). Taken together, these studies suggest that the striate cortex participates in multisensory audio-visual processing and that its activity can be modulated, at least to some extent, by unimodal auditory information as well.

Attempts have also been made to explore how in the absence of retinal stimulation, complex information like

auditory scenes translate from audition to the coding space of early visual cortex (Vetter et al., 2014). Results revealed that information specific to different categories of complex sound scenes could be decoded from early visual cortex activity of blindfolded participants. Provided that auditory scenes and task were designed to elicit imagination, the specific role of acoustic processing and imagery could not be disentangled.

The outstanding issue of whether and which specific acoustic features are mapped in human striate cortex has not however been addressed yet. Our goal was to model attributes of sounds properties and to assess their role in influencing V1 activity. To this aim, we combined stimulus modeling of different natural (or synthetically derived) sound categories (i.e., speech, speech-related and non-speech naturalistic sounds) with the measure of brain activity in absence of retinal input. In particular, we modelled the envelope power of sounds, that is their amplitude modulation (AM) over time. Natural sounds and vocalizations in particular are known to comprise profiles of high power at slow temporal amplitude modulations. Studies on animal models revealed that the statistical structure of natural sounds, such as their characteristic intensity fluctuations, matches the neural coding selectivity of the auditory system (Hsu et al., 2004; Riecke, 1995).

In humans, there is ample evidence suggesting that the activity in temporal cortex synchronizes to low (< 20 Hz) frequency modulations of the speech envelope (Di Liberto et al., 2015; Giraud and Poeppel, 2012; Luo and Poeppel, 2007). Moreover, in the visual cortex, a neural entrainment to visual signals associated to lip movements has also been recently highlighted (Bourguignon et al., 2020; Giordano et al., 2017; Hauswald et al., 2018). That is, input fluctuations over time are being tracked by both auditory and visual systems.

To assess whether acoustic features are represented in the visual cortex, sound amplitude changes occurring at typical speech envelope time-scales were exploited. Four categories of hierarchically organized acoustic stimuli were tested in isolation (Figure 1B). All of them comprised stimuli with a natural amplitude modulation. First, we measured the brain response to selected single words pertaining to the same semantic class. As control stimuli, from each word an associated pseudoword was generated. Pseudowords retained similar articulatory patterns and envelopes of each word from which they were derived, but did not convey any semantic content; these stimuli allowed to control for an association between V1 responses and semantic information (Figure 1A, see Vetter et al., 2014). As further control condition, for each word stimulus an artificial sound was also created by preserving the word sound envelope, but flattening the original spectral structure (Figure 1A, B). Thus, each artificial sound shared with the originating word only the information provided by its amplitude modulation. This control condition was conceived to isolate the role of sound amplitude modulations and to rule out the contribution of spectral properties occurring in speech, such as the fine harmonic structure.

Therefore, starting from a specific word, we had an associated pseudoword and an artificial sound, with comparable syllabic and phonemic time-scales of signal amplitude changes. Amplitude variations occurring within these timescales were used as descriptors in a fMRI multivoxel pattern analysis (MVPA). Finally, we introduced bird chirps as an additional control condition (Figure 1B). This stimulus category contained non-speech naturalistic sounds comprising envelope modulations in the low frequency bands resembling human speech. This category allowed to control for speech specificity of the neural responses.

Evoked hemodynamic activity was acquired during a listening task. Voxel-wise decoding, based on a parsimonious auditory model, evaluated across distinct sound categories whether envelope variations in the frequency ranges of interest were represented in the occipital cortex, and to validate results, in the temporal cortex as well (Figure 2).

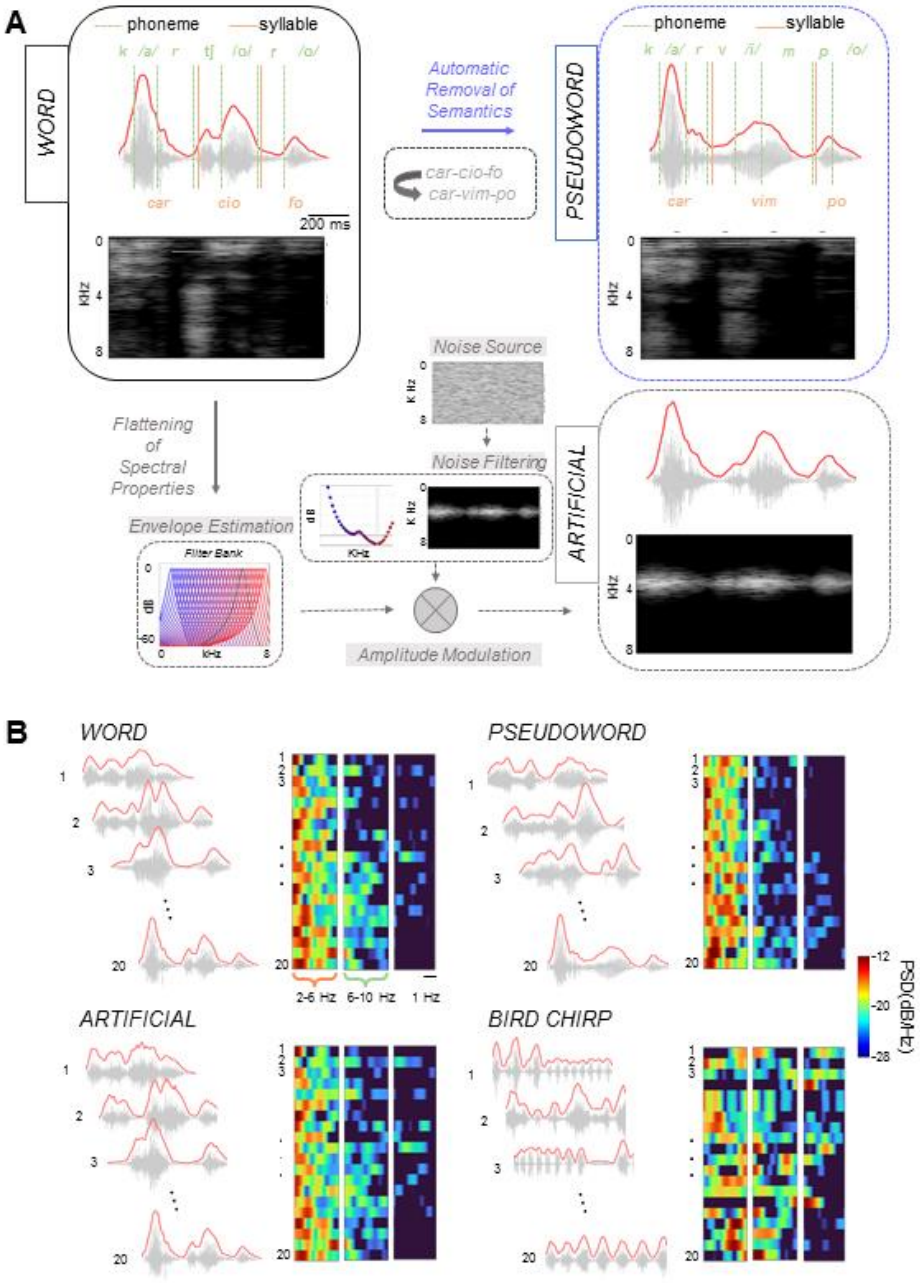

Figure 1. Stimuli. (A) The word sound category comprised 20 tri-/four- syllabic words pertaining to the same semantic category. From each word two control stimuli were created: (i) a pseudoword was obtained by using an automatic algorithm. A multilingual pseudoword generator able to produce polysyllabic stimuli that respect the phonotactic constraints of the Italian language was adopted. (ii) An artificial sound was created by preserving the sound envelope of

the word of origin but flattening its spectral structure. The original audio trace of each word was decomposed into 30 critical bands, using a gammatone filter-bank. Single sub-band envelopes were then linearly summed across these critical bands and then used to modulate the amplitude of a white band-passed Gaussian noise source whose central frequency was characterized by the lowest absolute threshold of hearing (f = 3.4 KHz). (B) For each stimulus of the four sound categories (word, pseudoword, artificial sounds and bird chirps) the envelope Power Spectral Density (PSD) was extracted. We defined two bins of interest between 2-6 and 6-10 Hz, representing the modulation power in a Low and High frequency ranges which were associated to phonemic and syllabic frequencies rates respectively, in speech-related categories. Note that speech-related categories retained only minimal energy above 10 Hz.

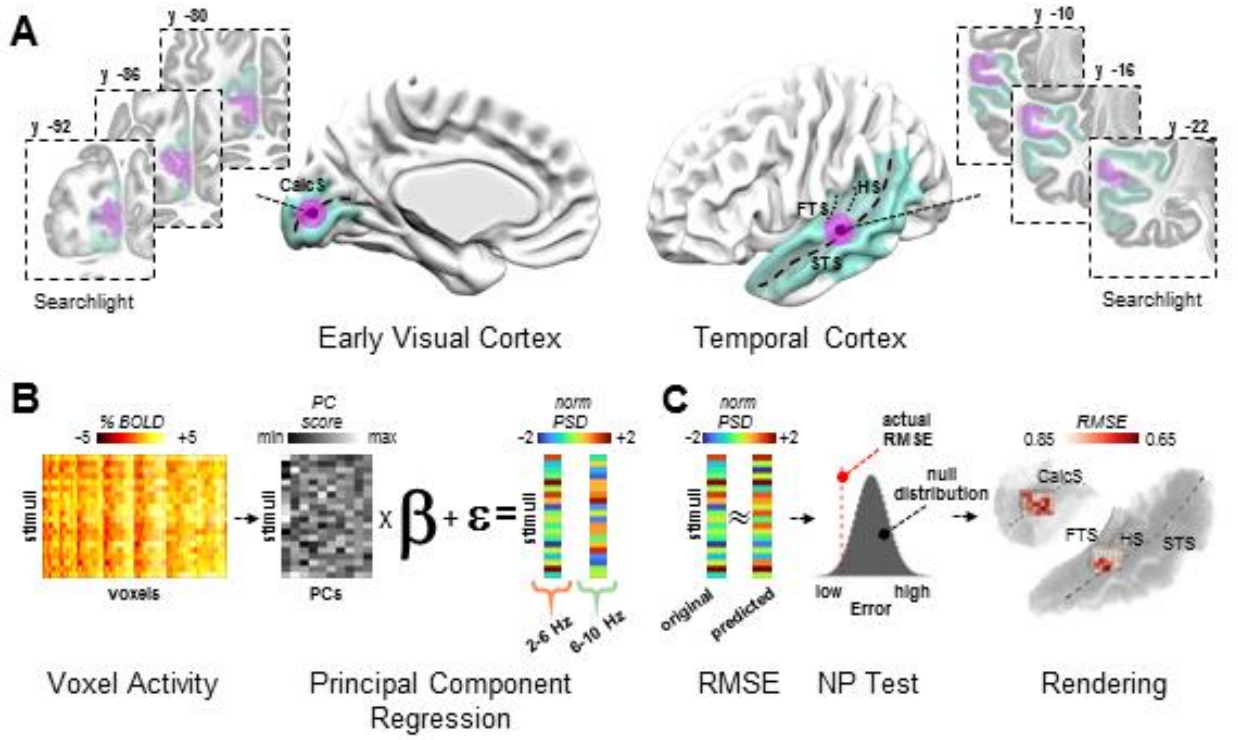

Figure 2. Analysis strategy. (A) fMRI analyses were carried out in temporal and occipital cortical areas. Temporal regions were defined using the AICHA atlas selecting the bilateral Superior Temporal Gyri and Sulci as well as the Middle Temporal Gyri. For the occipital regions, we isolated the striate Calcarine cortex by means of the probabilistic map by Wang and colleagues. A searchlight analysis was performed in the volumetric space respecting the cortical folding. Voxels were sampled along the cortical ribbon preserving the functional distinctions across adjacent sulcal walls and sulci. Calc S indicates the Calcarine Sulcus; FTS indicates First Transversal Sulcus; HS indicates Heschl's Sulcus; STS indicates the Superior Temporal Sulcus. FTS and HS define the anterior and posterior bounds of Heschl's Gyrus. (B) Hemodynamic responses were associated with the envelope modulation power using a machine learning algorithm. In each searchlight, we performed a Principal Components (PC) Analysis to extract orthogonal dimensions from normalized voxel responses within each experimental condition and subject. Afterwards, we performed a multiple regression analysis, using the PC scores as the independent variable and the power of Low (e.g., 2-6Hz) and High (e.g., 6-10Hz) modulation

frequencies as the dependent one. This procedure led to a reconstructed model of predicted power values in each experimental condition, subject and searchlight. (C) The reconstructed models were compared to the actual ones by calculating their root mean squared error (RMSE). To measure the statistical significance, we first averaged the reconstructed models across participants in each experimental condition and searchlight obtaining a group-level predicted set of acoustic features. Then, we performed a non-parametric (NP) permutation test, by shuffling the predictor matrix (i.e., PC scores) in each subject and experimental condition. This procedure ultimately provided a null distribution of group-level RMSE coefficients, against which the actual association was tested. Results were corrected for multiple comparisons using False Discovery Rate (FDR) and were mapped onto a 3D render of the temporal and occipital regions of interest.

Results

We first manually measured the syllabic and phonemic frequencies from our linguistic material (words and pseudowords) (median: word syllabic 3.80 Hz, CI-95th: 3.62-4.10; word phonemic 8.38 Hz, CI-95th: 7.48-9.12; pseudoword syllabic 3.80 Hz, CI-95th: 3.54-4.05; pseudoword phonemic 8.37 Hz, CI-95th: 7.47-9.67). As previously showed (Keitel et al., 2018), these results were congruent with the envelope power spectral density (PSD) calculated from sound waves (Figure 1B). Considering the dynamic frequency ranges of our linguistic stimuli, we defined in the envelope PSD two non-overlapping 4Hz-wide frequency ranges which were centered at 4 and 8 Hz for syllabic and phonemic rates, respectively. From now on, these intervals are identified as Low and High envelope frequency ranges. Note that, as shown in Figure 1B, signals did not comprise substantial energy above 10 Hz. Each artificial sound maintained the sound envelope of the word from which it was generated, but its spectral structure was flattened (Figure 1A). Although bird chirps comprised higher frequencies as compared to the human speech envelope, for consistency of analyses, their envelope PSD was extracted in the same frequency ranges.

Two regions of interest (ROIs) were defined to analyze the brain activity elicited by the four sound categories. (i) The *Occipital ROI* selectively included the Calcarine cortex (i.e., V1) using the probabilistic map by Wang and colleagues (Wang et al., 2015) (ii) The *Temporal ROI* comprised the Superior Temporal Gyrus and Sulcus as well as the Middle Temporal Gyrus (AICHA atlas, (Joliot et al., 2015). We aimed at reconstructing the acoustic features of our set of stimuli (i.e., the power of the envelope in the Low and High frequency ranges) by using brain activity as predictor (Figure 2C) (Pasley et al., 2012). This procedure yielded a voxel-based measure of the goodness-of-fit (i.e., root mean squared error, RMSE, (Poldrack et al., 2019) which was corrected for multiple comparisons using False Discovery Rate ($q<0.01$; additional cluster correction of 20 voxels, nearest neighbor). Moreover, to detail the sub-regions of the Calcarine cortex associated to the envelope modulations, we mapped the unthresholded RMSE (i.e. below $p<0.01$) onto a three-dimensional representation of the area (Figure 2C). To rule out the possibility that whole-brain hemodynamic fluctuations, associated to nuisance confounds (e.g., variations in arousal, physiological noise), could drive spurious correlations with our features of interest, we regressed out the global activity defined as the averaged brain response (Aguirre, 1998; Macey et al., 2004).

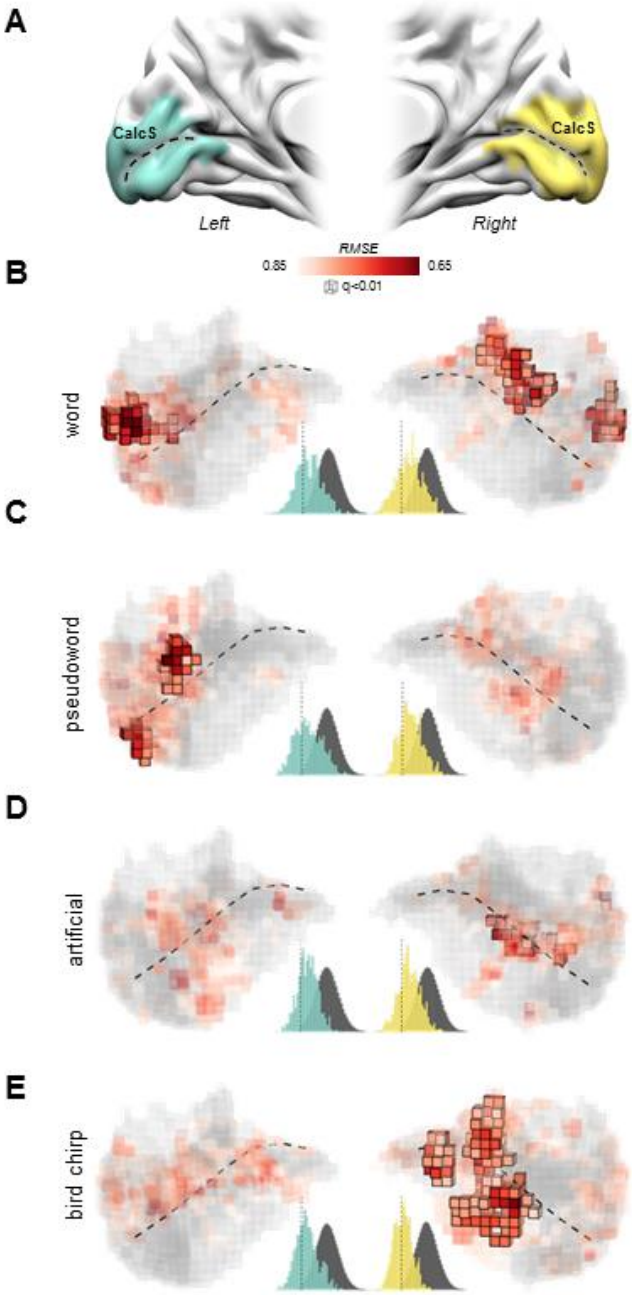

Figure 3. Reconstruction of the sound envelope power in the 6-10 frequency range in V1. (A) V1 ROI in the left (cyan) and right (yellow) hemispheres. (B, C, D, E) Reconstruction in V1 of the sound envelope power in the High (6-10 Hz) frequency range as a function of sound categories (i.e. word, pseudoword, artificial and bird chirps). Significant voxels surviving the multiple comparisons correction are represented by highlighted cubes ($q<0.01$; minimum cluster size = 20 voxels), whereas colored cubes displayed uncorrected threshold results at $p<0.01$. Population histograms represent the overall reconstruction performance in the left (i.e., cyan histogram) and right (i.e., yellow histogram) hemispheres for each ROI and sound category as compared to their null distributions (i.e., dark grey histogram; dashed vertical lines in histograms represents the 1st percentile). Results show that V1 represents sound envelope power for all sound categories with a lower RMSE as compared to the null distribution. The right hemisphere was increasingly engaged from speech stimuli to non-linguistic sounds. For anatomical landmarks please refer to Figure 2.

The envelope power measured in the High frequency range, associated to the phonemic range of our linguistic stimuli, represented the fastest time-scale of the sound amplitude variations we investigated. In a hierarchical feed-forward processing scheme the coding of higher frequencies would occur prior to the coding of lower frequencies (DeWitt and Rauschecker, 2012). Such conceptual framework has been suggested for both auditory and visual systems (DeWitt and Rauschecker, 2012; Hubel and Wiesel, 1962). Thus, results which emerged by assessing brain activity associated to the envelope power in the High frequency range represented

the main focus of the analysis. Note that comparable results emerged for the two frequency ranges we investigated. The results associated to the Low frequency range model (representing the syllabic rate) can be consulted in the Supplementary materials (Supplementary Figure 1).

First, it is noteworthy that for all sound categories a positive hemodynamic response was found in the temporal cortex and in striate cortex as well (see Supplementary Figure 5 for further details). On these premises, for each word we assessed whether the variation of the envelope power in the phonemic range was associated to the activity of V1 voxels. Results revealed a significant cluster in the posterior part of the left Calcarine sulcus which comprised the ~2.8% of the total volume of V1. Moreover, in the right hemisphere ~4.6% of the voxels survived to multiple comparisons, distributed across two patches of cortex, one in the middle of the Calcarine sulcus, and an additional smaller cluster in the lateral portion of the Cuneus (Figure 3A). Population histograms further emphasized the overall ability to reconstruct envelope power variations in the High frequency range, in both left and right V1 as compared to a null distribution. In line with previous evidence (DeWitt and Rauschecker, 2012) revealing the key role of temporal areas in speech processing, the envelope power in the High frequency range was successfully reconstructed in these regions. As expected, envelope variations were mainly linked to the brain activity measured in the left hemisphere as compared to the right (~2.7% of the total volume of the left Temporal ROI; ~0.8% of the total volume of the right Temporal ROI). Significant patches of cortex were identified in the left posterior part of the Superior Temporal Sulcus and Gyrus (pSTS/pSTG), which are pivotal regions in phonemic processing (DeWitt and Rauschecker, 2012) as well as in the left mid portions of Middle Temporal Gyrus (mMTG) and in the right mid portion of STS (mSTS) (Figure 4B).

The same analysis was performed for the variation of the envelope power in the Low frequency range. An involvement of the Calcarine cortex, as well as an asymmetry (left > right) in the temporal cortical areas were found. Detailed results were reported in Supplementary Figure 1. Overall, these results suggested that features extracted from the sound envelope of single words were not only traceable in temporal areas but in the Calcarine cortex as well. However, imagery processes, which are known to activate V1 (Cichy et al., 2012; Naselaris et al., 2015; Vetter et al., 2014) could explain such Calcarine cortex activations.

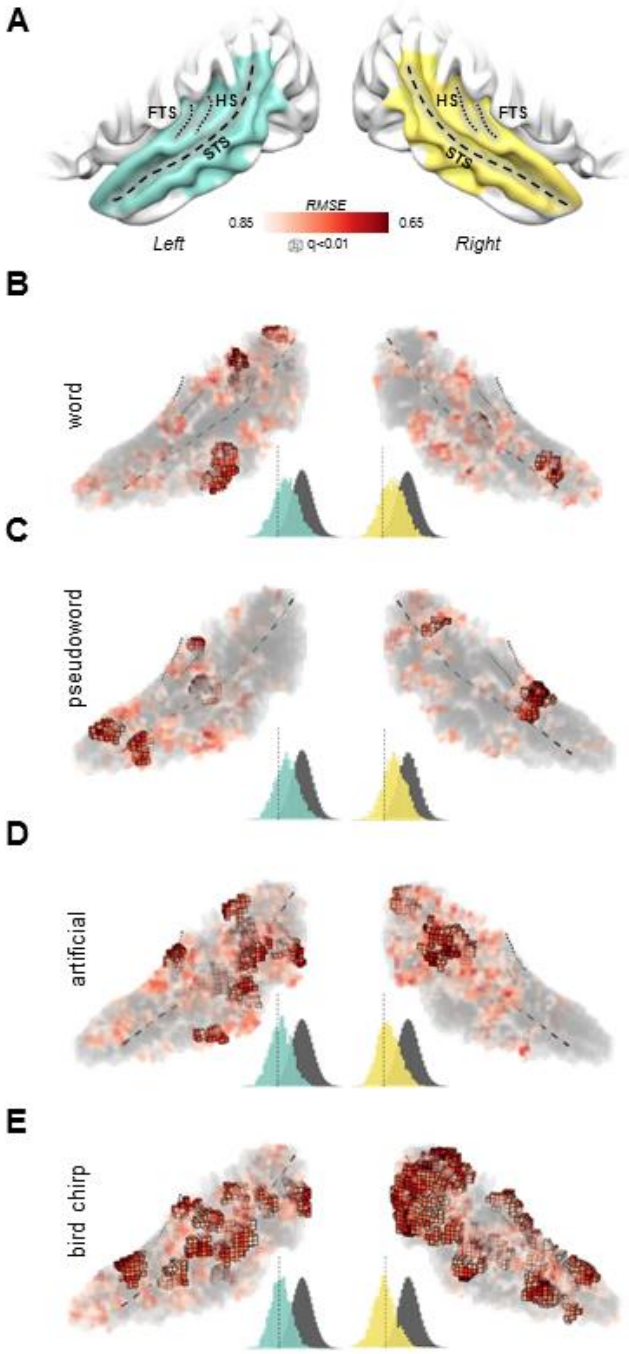

Figure 4. Reconstruction of the sound envelope power in the 6-10 Hz frequency range in the Temporal ROI. (A) Temporal ROI in the left (cyan) and right (yellow) hemispheres. (B, C, D, E) Reconstruction in the temporal ROI of the sound envelope power in the High (6-10 Hz) frequency range as a function of sound categories (i.e. word, pseudoword, artificial and bird chirps). Significant voxels surviving the multiple testing correction ($q<0.01$; minimum cluster size = 20 voxels) are represented by highlighted cubes, whereas colored cubes displayed uncorrected threshold results at $p<0.01$. Population histograms represent the overall reconstruction performance for each ROI and sound category as compared to null distributions. Significant clusters emerged in Superior Temporal Sulcus and Gyrus (pSTS/pSTG) in both left and right hemispheres. For anatomical landmarks please refer to Figure 2.

Test of the role of semantic content

Does the observed Calcarine engagement depend on word semantic processing and imaginability of semantic content? To answer this question, we tested whether the association between brain activity and envelope power variations was still present in the absence of semantic information. Therefore, we measured the brain response

to meaningless pseudowords, but retained similar articulatory patterns and phonotactic constraints of the original words from which they were derived (see Figure 1A).

Results in the *Occipital ROI* were consistent between pseudoword and word categories. Envelope variations in the High frequency range were associated to two significant clusters in the posterior part of the left Calcarine cortex, extending from the superior bank of the Calcarine sulcus to the Cuneus, and from the inferior bank to the Lingual Gyrus, respectively. Overall ~3.2% of the total volume of left V1 was engaged, whereas no voxels survived to multiple comparisons correction in the right hemisphere (Figure 3C). Population histograms resembled the word category ones, when considering both the goodness-of-fit of the reconstruction and the hemispheric asymmetry (left > right; see Figure 3C).

As expected, the variations of the envelope power in the High frequency range were also successfully reconstructed in the *Temporal ROI*. Specifically, ~1.9% of the voxels in the left hemisphere were significantly modulated by the envelope in the phonemic range, while the total recruitment in the right hemisphere was ~1.5%. Significant patches of temporal cortex were identified in the left anterior part of the Superior Temporal Sulcus and Gyrus (aSTS/aSTG), as well as the left mMTG and in the right mid portion of STG (mSTG) and pSTS (Figure 4C).

Significant clusters in Calcarine cortex were found for variations of the envelope power in the Low frequency range as well (corresponding to the syllabic rate; see Supplementary Figure 1), suggesting that the modulation of V1 activity was not selective for the 6-10 Hz frequency range.

Taken together, this evidence demonstrated that features of the envelope modulation were represented in Calcarine cortex even in the absence of semantic content. However, sound amplitude modulation, as conveyed by the envelope power variation in both Low and High frequency ranges are known to be associated to changes in sound spectral properties (Di Liberto et al., 2015).

Test of the role of spectral properties

Do the observed results depend specifically on amplitude modulations? Do they rather rely on spectral properties of speech sounds? Or, perhaps, on their combination? Speech comprises envelope variations with specific higher frequency spectral features. Thus, to assess the specificity of the information conveyed by envelope variations, we measured the brain response to an additional set of control stimuli. From each word, an artificial sound was generated reproducing in detail the low-level features (duration and estimated envelope), but removing original spectral properties. To this aim, a white noise was filtered within the most sensitive hearing frequency band (3.4 kHz) and modulated according to the envelope of the original word (see Figure 1A). We tested whether the artificial stimuli retained their intelligibility by means of an additional control experiment. Participants listened to each artificial sound and were then asked to identify the original sound from which it was generated by means of a two-alternative forced choice task. For an artificial sound derived from a word, participants had to identify the original sound choosing between the word and its associated pseudoword.

The results demonstrated chance level accuracy (accuracy ± SE: 51.5% ± 3%, t(9)=0.52, p=0.3069, one-tail t-test, see Supplementary Figure 4B), indicating that intelligibility was not retained despite the envelope modulations were maintained. Results of the fMRI experiment using artificial sounds highlighted a cluster in the mid-posterior part of the sulcus in the right Calcarine cortex (~1.9%), whereas the results did not show an involvement in the left hemisphere (Figure 3D).

The envelope power in the High frequency range was also successfully reconstructed in temporal cortex. Specifically, ~7.2% of the voxels in the left hemisphere were significantly modulated by the envelope power, while the total recruitment in the right hemisphere was ~2.6%. Significant patches of cortex were identified in the bilateral pSTS/pSTG, the left mSTS/mSTG as well as the left Heschl Gyrus (HG; Figure 4D). Comparable results were found in the Low frequency range and were detailed in Supplementary Figure 1.

Using artificial sounds, we showed that even in the absence of spectral properties, the amplitude modulation of a sound (in both frequency ranges tested here) is represented in early visual cortex. These results suggested that fine spectral details associated to envelope variations could not account for the specific response in V1. Overall, these findings supported the central role of the envelope in modulating occipital (and temporal) activity. However, we did not rule out the possibility that a specific speech-like modulation of the envelope was responsible of such an effect.

Test of speech specificity

Is Calcarine activity specifically associated to speech envelope variations, or can the same phenomenon be observed with non-speech natural sounds as well? In order to better comprehend if there is a speech specificity of our effects, we examined the brain response to other natural sounds, bird chirps. These sounds had a strong envelope power modulation in the same frequency bands as human speech. Thus, the MVPA analysis previously described for other sound categories was performed on these sounds to clarify if V1 recruitment is specific to speech-related material.

Significant clusters associated to the High envelope variation frequency range were found in the middle part of the right, but not the left Calcarine cortex (Figure 3E). Specifically, results comprised both the superior and inferior bank of the right Calcarine sulcus. Results revealed the engagement of ~9.4% of the total volume of right V1, whereas left V1 was not involved.

The envelope power in the High frequency range was reconstructed in temporal cortex as well. Specifically, a large extent of cortex successfully retained envelope characteristics: ~8.2% of the voxels in the left hemisphere and ~21.5% in the right hemisphere. Significant cortical areas encompassed the whole Superior Temporal Sulcus and Gyrus (STS/STG), bilaterally (Figure 4E). No significant results were found in V1 for the Low frequency range.

Envelope amplitude modulation across frequencies and categories

Do High and Low frequency ranges share a common spatial organization in the left hemisphere across linguistic materials? Previous results showed that the sound envelope is represented in speech-related cortical areas within the left hemisphere (Giraud, 2000; Oganian and Chang, 2019). An antero-posterior gradient from low to high modulation frequencies emerged (DeWitt and Rauschecker, 2012; Hullett et al., 2016). Thus, to assess the degree of spatial correlation, we first tested RMSE pattern similarities between High and Low frequency ranges within each sound category (Figure 5A). No significant associations between Low and High frequency ranges were found in early visual cortex (word $\rho = 0.015$, CI-99th: -0.098 - 0.108; pseudoword $\rho = -0.094$, CI-99th -0.098 - 0.107; artificial $\rho = -0.061$, CI-99th -0.116 - 0.116). Results in temporal cortex demonstrated that the two frequency ranges were not spatially associated for word and artificial sound categories (word $\rho = 0.048$, CI-99th -0.042 - 0.056; artificial $\rho = -0.026$, CI-99th -0.042 - 0.055), and they were only negatively correlated in the pseudoword category (pseudoword $\rho = -0.065$, CI-99th -0.042 - 0.056). Note that the PSD of the two frequency ranges were uncorrelated for each of the three sound categories (see Methods). Overall, these results suggested that the same voxels could not represent either frequency ranges.

On these premises, we performed a conjuction analysis (i.e. non-parametric combination) across the linguistic material (i.e. word, pseudoword and artificial stimuli), independently for each frequency range (High and Low), selectively for the two left ROIs ($q<0.01$; additional cluster correction of 60 voxels, nearest neighbor; see Figure 5B). As depicted in Figure 5, the conjunction analysis demonstrated a significant common involvement of the three categories in representing the envelope power modulation. In details, the Low frequency range (i.e., the syllabic range for word and pseudoword categories) was successfully reconstructed in a wide patch of cortex spreading from anterior to mid portions of STG/STS. Conversely, pSTS/pSTG encoded higher frequency information (i.e., 6-10 Hz). Overall, these results were in line with previous studies which indicated a topographical representation of envelope characteristics in STG, with a distribution from anterior to posterior temporal cortex mapping respectively from lower (~2 Hz) to higher (~10 Hz) modulation frequencies (Hullett et al., 2016). These findings were also consistent with results of previous literature which showed that left temporal cortex was associated to the processing of linguistic information where words, syllables and phonemes were represented with an anterior-posterior organization in left STG (DeWitt and Rauschecker, 2012).

Crucially, a similar topographical organization was found for linguistic stimuli in left Calcarine cortex as well. As depicted in Figure 5, high amplitude modulation frequencies were represented in the posterior part of the Calcarine cortex, whereas low frequencies were preferentially encoded more anteriorly. Population histograms of anterior and posterior temporal patches, as well as anterior and posterior portions of Calcarine cortex highlighted the differences in the goodness of envelope reconstruction for the two frequency ranges (Figure 5).

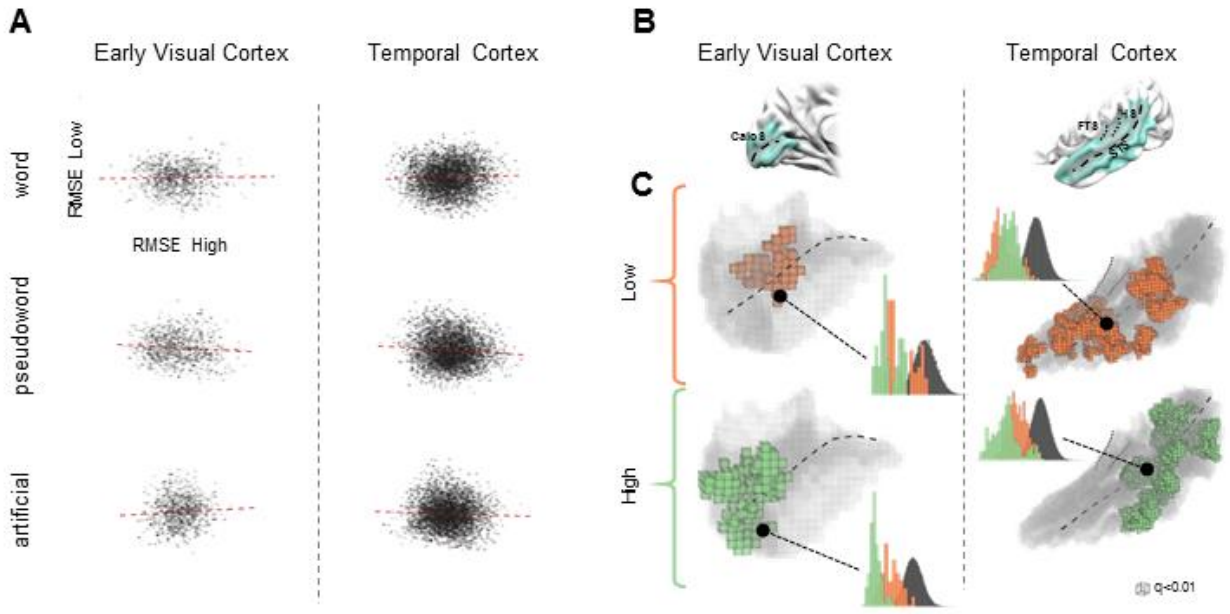

Figure 5. Sound Envelope in the two frequencies ranges across speech-related sound categories in both ROIs of the left hemisphere. (A) RMSE pattern similarities between High and Low frequency ranges for speech-related sound categories. No positive correlation between the two frequency ranges of interest emerged in either visual and temporal ROIs. Results suggested that voxel tuning was not overlapping across High and Low frequency ranges. (B) V1 and Temporal ROIs in the left hemisphere. (C) Reconstruction of sound envelope power in both ROIs (non-parametric combination analysis). Significant voxels surviving the multiple comparisons correction are represented by highlighted cubes. Population histograms represent the average reconstruction performance in each ROI of the two frequency ranges (High frequency in light green and Low frequency in light orange) across speech-related sound categories (ie. word, pseudoword, artificial) as compared to null distributions. In the posterior part of the superior temporal cortex sound envelope in the 6-10 Hz (High) frequency was reconstructed with lower RMSE as compared to the 2-6 Hz (Low) range, viceversa occurred for the anterior part of temporal cortex. A similar organization emerged in V1 with an antero-posterior mapping of 6-10 and 2-6 Hz frequency ranges, respectively. For anatomical landmarks please refer to Figure 2.

Discussion

Here we investigated whether a functional sensitivity to sound properties could be observed in striate cortex. To this aim, we measured the brain response to natural or derived auditory stimuli in absence of visual input. We built an acoustic model based on the envelope power modulation in specific frequency ranges (Low, 2-6 and High, 6-10 Hz) and evaluated the cortical mapping of such acoustic features. Sound stimuli of different categories were used to investigate the degree of the response specificity and to control for possible confounds.

Amplitude modulations of sounds, in both frequency ranges, were traceable not only in temporal areas classically involved in acoustic processing, but notably in the Calcarine cortex as well, revealing a crossmodal representation of sound features. We identified a common pattern of activation in classical left temporal areas for all speech-related sound categories. Consistently with the existing literature, a topographical representation of envelope power variations was evident in STG and STS, with an antero-posterior gradient from Low to High modulation frequencies (Hullett et al., 2016). Strikingly, a topographic organization for words and pseudowords was found in the left Calcarine cortex as well, with High and Low amplitude modulation frequencies encoded respectively in its posterior and mid parts.

The analysis of each sound category revealed that the crossmodal mapping in striate cortex occurred regardless of the semantic content, as it was found following speech stimuli having no meaning such as pseudowords. V1 involvement was dissociated by the spectral properties of sounds: a representation of sound amplitude modulation emerged for stimuli in which the spectral properties had been flattened, i.e. artificial sounds. Moreover, the tuning of this region was not speech-specific, since similar results were found even for sound amplitude modulations of bird chirps. Overall these results demonstrated that human striate cortex can represent sound attributes.

Semantic content and imagery

Evidence that sounds elicit specific functional activations in primary visual cortex already existed. Muckli and colleagues (Muckli et al., 2015; Vetter et al., 2014) analyzed the BOLD signal of blindfolded individuals listening to complex auditory stimuli, such as the traffic noise or a forest auditory-scene. The authors showed that the category of each sound scene could be identified, by using decoding techniques, from early visual cortex activity. Abstract conceptualization and visual imagery were suggested to represent the driving mechanisms of the visual cortex engagement and our results relative to isolated words could well be interpreted in a similar way. All stimuli of the word category comprised imaginable, graspable objects. Similar modality-independent assessments have also been performed adopting visual stimuli. The sight of scenes or written text carrying acoustic abstract features (e.g., a man playing a trumpet or the word “telephone”) were found to elicit responses in acoustic regions (such as pSTG/MTG) as a function of the amount of acoustic relevance (Kiefer et al., 2008; Proverbio et al., 2011).

In order to control for the role of semantic information, we analyzed the brain mapping of pseudowords in our regions of interest. While any auditory stimulus can elicit visual representations in principle, these stimuli prevented a coherent semantic-based imagery across participants. The observed responses to pseudowords provided evidence that semantic-related imaginability alone could not explain the mapping of amplitude modulation of natural sounds observed in the striate cortex.

Amplitude Modulation specificity

Brain areas such as STS are functionally tuned for the processing of speech-specific temporal structures. A recent study generated sound quilts by shuffling segments of a foreign language and other natural or artificial sounds, to assess which regions were tuned for temporal properties (Overath et al., 2015). This approach allowed to preserve original sound properties only at short timescales and disrupted them on longer timescales. While primary auditory cortex was not sensitive to quilt durations, possibly due to the narrow frequency tuning which characterizes its response (Moerel et al., 2015; Thwaites et al., 2017), only bilateral portion of STS displayed an activity which parametrically varied with quilt lengths. These regions have been recognized as a major hub to encode the temporal structure of speech. The mapping was specific for STS, which apparently operated at a timescale near the one of syllables and words. Moreover, control quilts whose amplitude modulation profiles matched those of speech signals did not elicit responses selective to quilt segment duration, suggesting that subregions of STS are particularly tuned to spectro-temporal structures (Overath et al., 2015). Conversely, it was shown that portions of STG encode the amplitude variations of speech as well as amplitude-modulated tone stimuli independent from other spectro-temporal features (Oganian and Chang, 2019). Taken together, these results reveal that in the superior temporal cortex, regions coding speech-specific spectro-temporal features and non-speech-specific envelope variations coexist.

Results of the present study do not exclude a role of spectral properties for the representation of sound features in striate cortex, but reveal that natural, low-frequency modulations of the envelope power are sufficient to elicit selective responses in V1. Each artificial stimulus retained duration and estimated envelope of the word of origin, but its spectral structure was flattened. This approach allowed to exclude that congruent variations across amplitude modulation and spectral properties of sounds could explain the observed effects. Indeed, amplitude variations in speech signals are known to be highly associated to spectral changes occurring at the phonemic level. One of the reasons concerns the different acoustic intensity provided by consonants and vowels. Our results confirmed that even for the artificial sound category, in which spectral properties characterizing consonants and vowels were removed, both frequency ranges of interest were functionally mapped in striate cortex.

Finally, the sound envelope mapping observed in striate cortex in response to artificial stimuli allowed to exclude that articulatory-based imagery could explain the results (Hauswald et al., 2018). While pseudowords are still reproducible from an articulatory point of view, artificial stimuli were not. Although each artificial sound was generated by a word, the flattening of spectral properties eliminated its intelligibility. Participants performed at chance level when asked to discriminate from two alternatives whether an artificial sound was derived by the original word or by its corresponding pseudoword (see Supplementary Figure 4B).

The assessment of speech specificity

The human auditory system is functionally tuned for the processing of language since the early stages of development. Evidence exist that the auditory pathway is more tuned and adaptive to language sounds, as

compared to environmental noise even before it has reached full-term maturation (Webb et al., 2015). However, the degree to which temporal lobe regions are specific to speech processing remains to be ascertained. Evidence of speech specific processing subregions of superior temporal cortex were reported, mostly referring to the phonemic processing (Formisano, 2008; Mesgarani, 2014; Rampinini et al., 2019). STS neural tuning for spectro-temporal timescales appears to be selective for speech material. In case natural, non-linguistic sounds were employed, such selectivity did not emerge (Overath et al., 2015). Nevertheless, examples suggesting that the temporal cortex is functionally tuned for sound features shared by both speech and non-speech material are many. The tuning of STG for amplitude modulation of natural sounds was observed for both speech and non-speech sounds (Oganian and Chang, 2019).

To clarify whether the observed effects in V1 were functionally specialized for speech material, non-linguistic natural sounds characterized by a rich amplitude modulated profile, such as bird chirps, were presented as well. V1 was recruited upon acoustic processing of this sound category as well, demonstrating a non-specificity for speech. We ensured that all our participants were not experts in ornithology, thus specific imaginability related to bird species was prevented. Indeed, while visual representations of birds could occur in each participant, we observed a mapping of specific frequency ranges of the amplitude modulation.

Lateralization

Evidence that a left hemispheric functional specialization for speech processing exist already in three-month infants (Dehaene-Lambertz, 2002). Reviews of the literature have suggested that in adults left hemispheric dominant responses progressively emerge as the acoustic materials increasingly provide speech-like input (Hickok, 2012; Peelle, 2012; Rauschecker and Scott, 2009). If amplitude-modulated noises activate bilaterally the primary auditory cortex, the processing of isolated phonemes and syllables results in activity typically along the left but not the right STS and MTG (DeWitt and Rauschecker, 2012). Moreover, the response tends to be increased in the left hemisphere as compared to the right when words and pseudowords are contrasted, suggesting a clear role of the left hemisphere for lexical processing (Davis and Gaskell, 2009). Moreover, the hemispheric dominance in the superior temporal lobes seems to depend on the type of processing performed on speech signals. On the one hand, temporal details of the speech signals primarily elicit brain activations of the left anterolateral STS and STG, whereas the right homolog is more sensitive to the spectral parameters (Obleser et al., 2008; Schönwiesner et al., 2005).

Consistently, our results in the temporal cortex showed a shift from a major involvement of the left with respect to the right hemisphere, the more the material becomes strictly linguistic. The recruitment of the right temporal cortex exceeded the one of the left temporal cortex only for bird chirps. In the other sound categories, the left and right temporal cortex were similarly recruited, albeit with a slight prevalence of the left hemisphere.

Even the pattern of activity of the striate cortex was not independent of the sound category. Indeed, changes in the lateralization of the sound envelope mapping could be observed across different sound stimuli. The more we move from speech stimuli to non-linguistic sounds, the more the activation involved the right hemisphere. The observed patters of results suggest a similar organization for the sound amplitude modulation in both

temporal lobes and occipital cortex.

Functional significance in a multisensory frame of reference

Multisensory audio-visual stimulations were observed to subtly modulate the striate cortex activity in rodents (Ibrahim et al., 2016) and in the awake monkey (Wang et al., 2008). The anatomical scaffolding for early audio-visual interactions has been provided by tracing studies. Sparse monosynaptic (i.e. direct) anatomical projections originating from auditory areas and terminating in primary visual cortex have been consistently found (Cappe and Barone, 2005; Charbonneau et al., 2012; Clavagnier et al., 2004; Falchier et al., 2002; Kim et al., 2015; Rockland and Ojima, 2003). In humans, the audio-visual convergence and integration between primary visual cortices were demonstrated at functional level (Martuzzi et al., 2007; Mercier et al., 2013; Molholm et al., 2004; Romei et al., 2009). The existence of intrinsic functional coupling between primary visual and primary auditory areas was found even in absence of external stimulation or task (Eckert et al., 2008). Modulations of BOLD activity measured in V1 have been associated to low-level audio-visual interactions (Watkins et al., 2006). Moreover, intracranial recordings revealed clear signs of audio-visual integration in striate cortex (Mercier et al., 2013). These consistent auditory modulations of visual cortices activity have provided a conceptual expansion of the range of potential activity patterns of the visual system.

MVPA of fMRI signals first revealed that evoked responses of the primary visual cortex could be found even when auditory stimuli were applied transiently and in isolation. Distinguishable spatial patterns of neuronal responses could be elicited not only in the primary auditory cortex, but in V1 as well, showing that auditory inputs elicit a characteristic pattern of activation also in other primary sensory cortices (Liang et al., 2013). Sound-driven crossmodal activations of the striate cortex have been recently confirmed also by measurements of intracranial stereotactic electroencephalographic (SEEG) recordings. High gamma neural oscillations were measured in striate cortex following the presentation of brief white noise stimuli. Such activity was found within the first 100 ms after stimulus onset and suggested the existence of populations of neurons performing auditory processing in striate cortex (Ferraro et al., 2020).

Which neurophysiological mechanisms could explain the crossmodal modulation of V1 by an auditory input? Studies on the animal model have revealed that auditory stimulations change the firing and selectivity of V1 neurons responses. It was shown that a sound can modulate the responses to visual input by impacting inhibitory neurons in V1. In anaesthetized mice, brief noise bursts (50 ms) generated activations in the auditory cortex, which in turn, had a modulatory role on inhibitory circuits in V1 and on the visually driven spike activities (Iurilli et al., 2012). In particular, GABAergic inhibitory piramydal neurons of the infragranular layers of V1 were activated by cortico-cortical connections originating in the auditory cortex. Recordings in mice revealed that sound sharpens orientation selectivity of pyramidal neurons in layers 2/3 (L2/L3) of V1. These sharpening effects are made possible by L1 inhibitory neurons that are the more directly innervated by A1 axons (Ibrahim et al., 2016). L1 neurons could inhibit other inhibitory neurons in L2/L3, generating a disinhibitory effect that appeared to have the functional role of increasing the firing rate at the preferred orientation of pyramidal cells. More generally, these works constitute an important step not only in

characterizing the neural bases of multisensory processes across circuits and synaptic levels, but also in linking physiology with behavior.

One of the mechanisms which have been suggested as possible explanations for the crossmodal interactions in primary sensory cortices is the phase resetting of slow oscillatory activity (Kayser et al., 2008; Lakatos et al., 2008). Strong positive correlations are known between high frequency (40-130 Hz) local field potentials (LFP) and the fMRI signal. Moreover, negative correlations are known between low frequency (5-15 Hz) local field potentials (LFP) and the fMRI signal (Mukamel et al., 2005). Both alpha and gamma oscillations are known to be linked to the activity of inhibitory GABAergic interneurons (such as parvalbumin fast-spiking interneurons (Hensch, 2005). The observed pattern in V1 in the present study could well be a byproduct of oscillatory activity occurring in striate cortex. Whether such responses are mostly linked to slow, fast oscillations or their combination remains to be ascertained.

One possibility is that sound features like AM are conveyed to the visual cortex to prompt analyses on multisensory objects, as occurs in the case of audio-visual associations and cross-modal correspondences.

Audio-visual associations are found at the earliest developmental stages. Already at two months, infants show reliable matching of vowel information extracted in faces and voices (Patterson and Werker, 2003). It could be wondered whether the observed patterns of activation in V1 unveil common representations in auditory and visual cortices. The term cross-modal correspondence has been used to indicate the tendency to preferentially associate certain features or dimensions of stimuli across different sensory modalities. For instance, humans systematically associate higher-pitched sounds with angular contours and smaller or brighter objects, whereas lower-pitched sounds are preferentially associated with smooth contours and larger or darker objects (Parise and Spence, 2013; Spence, 2011). Even the correlation between the phonetic representation of a word and its meaning ("*bouba-kiki* effect", Köhler, 1929) can be considered an example of a crossmodal correspondence. However, sound-shape associations seem to entail a rather protracted sensitive period (Sourav et al., 2019).

More than fifty years ago Barlow (1961) suggested that sensory processing would exploit redundancies and the correlation structure existing within the input. Sensory signals are typically dynamic, multimodal streams. Correlation detection, such as lag and synchrony, has been suggested as a general mechanism behind multisensory integration as well (Parise et al., 2016). In speech, signal amplitude changes originate from the rhythmic movements of the mouth and the other phonatory apparatus. Importantly, coherent temporal modulations have been found between 2-7 Hz for both voice envelope and mouth openings (Chandrasekaran et al. 2009). The unity at the source level is reflected at the neural level. Speech signals conveyed by the sound envelope and rhythmic variations of lip movements are known to be tracked by the neural oscillatory activity in both auditory and visual cortices (Giraud et al., 2012; Park et al., 2016). Moreover, auditory and visual tracking is known to interact to maximize efficient processing (Crosse et al., 2016; Park et al., 2018). It is thus reasonable to speculate that the sound envelope mapping in V1 might represent input regularities which could contribute to an efficient visual processing and ultimately to a unified A-V percept. In this respect, auditory AM sounds have been reported to boost the visual spatial frequency analysis, suggesting that a type of

crossmodal mapping between these low-level auditory and visual features exists (Guzman-Martinez et al., 2012; Orchard-Mills et al., 2013).

It could also be wondered whether the observed patterns of activation in V1 unveil common representations between auditory and visual cortices. The fact that V1 mapping of sounds envelope power displayed a non-random spatial representation (see Figure 5) supports such idea. In particular, the modulation power in High and Low frequency ranges was mapped in posterior and more anterior portions of striate cortex, respectively. This organization, despite being on a completely different scale, follows the typical spatial frequency organization of V1 (Carandini et al., 2005). It is well known that amplitude modulations of natural sounds follow a 1/f relationship, which implies that low frequencies of amplitude modulation (i.e. slow intensity fluctuations) are the most represented. It is noteworthy that also natural visual scenes respect the power law relationship of 1/f, and in particular, the spatial and temporal luminance contrast (e.g. Srivastava et al., 2003). While the physical causes of the power law relationship in natural sounds and natural visual scenes occurs for different reasons, it is possible that a sort of cross-modal mapping, between the two sensory systems occurs already at the earliest stages of processing.

Limitations

It could be wondered whether the associations between the acoustic features and the V1 activity could be explained by whole-brain hemodynamic fluctuations linked to arousal or alertness induced by sound onsets (Pisauro et al., 2016). It is worth mentioning that the global signal was responsible up to 30% of the variance of the data. However, to exclude this confound the BOLD signal was corrected with a Global Signal Regression (Aguirre, 1998; Macey et al., 2004). Thus, the observed pattern seems more likely to genuinely represents auditory sensory information mapping in V1. Nevertheless, further investigations are required to provide a finer characterization of the envelope PSD (e.g., with shorter frequency bins) associated to the Calcarine cortex hemodynamic activity. By adopting natural sounds, we had the opportunity to assess the hemodynamic response to stimuli whose AMs vary over time across multiple timescales. We do not know whether sound AM representations in V1 occur selectively for these types of sounds or whether they would occur for purely periodic AM sounds (Leaver and Rauschecker, 2016).

Finally, we selected relatively short (~1 sec) stimuli in isolation as these represented a compromise to estimate the elementary dynamic properties of natural sounds; in speech typical syllabic and phonemic frequencies occur < 20 Hz. Ultimately, this allowed to avoid the risks associated to the averaging of acoustic features of longer non-stationary sounds which would be represented in the physiological hemodynamic response (de Heer et al., 2017). Moreover, this choice was made to prevent confounds associated to neural summations which potentially derive from the presentation of more continuous stimuli. However, further experiments are required to assess whether and to which extent the envelope of continuous stimuli would be mapped in V1 as well.

Conclusions

The results of the present study clearly reveal that, in absence of visual input, the human striate cortex maps sound attributes and, in particular, sound envelopes. Ultimately, this finding suggests that not only multisensory processes occur in V1 but demonstrates that even unimodal sound properties are represented within the brain area that has always been conceived as being purely visual in its function.

Materials and Methods

As described in Figure 1 and 2, we aimed at investigating the activity elicited by sounds in early visual cortex by means of fMRI. First, we recorded four sets of sounds, three were speech-specific (i.e., words, pseudowords and artificial stimuli, which were built upon the word category) and one included non-speech natural sounds (i.e., bird chirps). Then, from each sound wave, we calculated its envelope and the modulation power in the High (6-10 Hz) and Low (2-6 Hz) frequency ranges. Hence, we exploited voxel-wise modelling to measure the association between these acoustic features and the hemodynamic activity in occipital and temporal areas.

Subjects

Twenty right-handed healthy individuals were recruited for the fMRI (10F; mean age ± standard deviation: 34.5 ± 6.5 years; Edinburgh Handedness Inventory: 16.4 ± 2.5 (Oldfield, 1971). All participants were native Italian speakers, recruited from the local area of Lucca. None of the participants was diagnosed with language-related developmental disorders (e.g., dyslexia, specific language impairment, delay of language onset). They gave their written informed consent and received 50 euros at the end of the experiment as a reimbursement. Ethical approval was obtained from Area Vasta Nord Ovest Ethics Committee (Protocol number 1485/2017) and conducted according to the Declaration of Helsinki (2013).

Stimuli

First, we selected 20 tri- and quadri-syllabic commonly used Italian words belonging to the same semantic category (i.e., vegetables; word length: 8 ± 2; CoLFIS frequency: 1.25 ± 1.74 (Bambini and Trevisan, 2012). Words pertained to graspable objects with comparable size. Starting from these 20 words, we created 20 pseudowords using “Wuggy”, a multilingual pseudoword generator able to produce polysyllabic stimuli that respect the phonotactic constraints of the Italian language (Keuleers and Brysbaert, 2010). All the pseudowords had no meaning and matched the corresponding word stimuli in the sub-syllabic structure, the letter length and the length of subsyllabic segments, besides matching transition frequencies of the reference word (see Supplementary Table 1). Our speech-specific stimuli were read by a trained Italian actress, and were acquired in a recording studio, using a microphone (Behringer C-1U; 40-20,000 Hz, 130db max SPL) connected to an iMac™ (Apple Inc.). Sampling frequency of the recording was 44100 Hz. Resulting waveforms were edited using the Audacity software (©Audacity Team) to remove silence periods before and after each stimulus, and

were slightly changed in tempo (within ±15%) to set the duration at 1 s for all stimuli. As results, sounds lasted about 1.00 ± 0.06 s for words, and 1.03 ± 0.06 s for pseudowords. Afterwards, we manually tagged the duration of syllables and phonemes composing our words and pseudowords (medians and confidence intervals estimated by bootstrapping method, 1000 iterations).

We then constructed a third class of stimuli (i.e. artificial sounds), reproducing in detail the low-level features (duration and estimated envelope) of the above-mentioned words, but with a completely different spectral structure. As detailed below, we aimed at creating a set of stimuli which retained the prosodic pitch modulation of the original words, but, at the same time, fine spectral details were flattened (Figure 1A). First of all, the original audio trace of each word was decomposed into 30 critical bands linearly spaced in the Equivalent Rectangular Bandwidth scale between 100 and 8000 Hz, using a zero-phase gammatone filter-bank (Hohmann, 2002). Single sub-band envelopes were then evaluated as the amplitude of the corresponding Hilbert transform and linearly summed across critical bands. After being smoothed with a 10ms-moving average filter, the obtained global envelope was used to modulate the amplitude of a spectrally-homogeneous source. This source was generated by filtering white gaussian noise in a single critical band of the gammatone filter-bank whose central frequency was characterized by the lowest absolute threshold of hearing (f = 3.434 KHz; see Figure 1A). The absolute threshold of hearing within the 30 central frequencies was evaluated considering a diffuse binaural field, according to the ISO 389-7: 2005 standard. All the described operations were performed within the Auditory Modeling Toolbox (Søndergaard and Majdak, 2013) in the Matlab environment.

As further control condition, we selected bird chirps. They represent naturalistic sounds, comprising envelope modulations in the low frequency bands (<20 Hz) resembling human speech. In particular, we extracted 20 bird chirps of 10 different bird species from video documentaries retrieved from YouTube® (128 kbps, compressed using advanced audio coding and converted to 44100 Hz) (see Supplementary Table 1). Bird chirps matched the duration of the other sound categories (1.01 +0.02 s).

Loudness was normalized across all stimuli by imposing a fixed root mean square value on each raw signal.

Feature extraction: modulation power in the Low and High frequency ranges

In order to describe each stimulus, two separate features were extracted from the signal envelope, conveying information about the Low (2-6 Hz) and High (6-10 Hz) frequency ranges of the modulation power, which were associated to phonemic and syllabic phonological properties in the speech material. In particular, the signal envelope was calculated using the procedure described above to generate artificial stimuli (Biesmans et al., 2015; Hohmann, 2002). Syllabic and phonemic power were then defined as the integral of the envelope power spectral density (PSD) in the two ranges of interest. These two intervals were centered on the manually estimated syllabic and phonemic frequencies, reported in the Results section. For consistency, the amplitude variations in the same Low and High frequency ranges were also extracted for bird chirps. Importantly, the

PSD in Low and High frequency ranges were uncorrelated within each sound category (word: Pearson's r=0.213, CI-95th: -0.128 - 0.534; pseudoword: r=0.257, CI-95th: -0.294 - 0.616; artificial: r=- 0.025, CI-95th: -0.359 - 0.342; bird chirps: r= -0.222, CI-95th: -0.523 - 0.175).

Following the stimuli generation procedure words and pseudowords had comparable envelope power modulations in the Low (Pearson's r=0.375, CI-95th: 0.061-0.640) frequency range, whereas they differed in the High frequency range (r=0.189, CI-95th: -0.259-0.543). This was due to the preservation of the syllabic rate and to the alteration of the phonemic features. Conversely since words and artificial sounds shared the same envelope, the two ranges of the envelope PSD were highly collinear (2-6 Hz range: Pearson's r=0.963, CI-95th: 0.909-0.986; 6-10 Hz range: r=0.886, CI-95th: 0.721-0.982).

Experimental procedures

A slow event-related paradigm was implemented using E-prime 3.0 software (Psychology Software Tools, Sharpsburg, Pennsylvania), and comprised 80 stimuli (equally divided across the four categories, i.e., 20 words, 20 pseudowords, 20 artificial sounds and 20 bird chirps). Each event comprised a ~1 s stimulus followed by 9 s of rest. Each stimulus was randomly presented three times across six runs lasting 8 minutes each (Mumford et al., 2014). Participants laid down blindfolded in the scanner. To ensure participant's attention, they were instructed to detect by a button press rare target sound. These rare deviant sounds (30 out of 270) were generated adding to our original stimuli a silent period (lasting 550 ms, starting at 150 ms after stimulus onset), and were distributed in time in a pseudo-random order to ensure the presence of ten targets across two sequential runs. The aim of the task was also to ensure that each participant maintained the attentional focus on the temporal dynamics of sound waves. Behavioral responses were analyzed to calculate precision -i.e., True Positive: TP; False Positive: FP; False Negative: FN; TP/(TP+FP)- and recall –i.e., TP/(TP+FN)- measures. The behavioral task resulted in high level of precision (87% ± 3%) and recall (90% ± 3%), suggesting that participants attended to the stimuli. During the scanning session, prior to the fMRI acquisition, participants underwent a brief session (~12 minutes) to familiarize with the task.

MRI Data Acquisition and Preprocessing

Brain activity was recorded using Philips 3T Ingenia scanner, equipped with a 12 channels phased-array coil, and a gradient recall echo-planar (GRE-EPI) sequence with the following acquisition parameters: TR/TE = 2000/30ms, FA = 75°, FOV = 256 mm, acquisition matrix = 84 × 82, reconstruction matrix = 128 × 128, acquisition voxel size = 3 × 3 × 3 mm, reconstruction voxel size = 2 × 2 × 3 mm, 38 interleaved axial slices, 240 volumes. Twenty seconds of rest preceded in each run the first stimulus onset and followed the last one. Stimuli were delivered with MR-compatible on-ear headphones (VisuaStim, 30 dB noise-attenuation, 40 Hz to 40 kHz frequency response).

Three-dimensional high-resolution anatomical image of the brain was also acquired using a magnetization-prepared rapid gradient echo (MPRAGE) sequence (TR/TE = 7/3.2ms, FA = 9°, FOV = 224 mm, acquisition matrix = 224 × 224, voxel size = 1 × 1 × 1 mm, 156 sagittal slices).

The fMRI preprocessing was performed with the AFNI software package (Cox, 1996). All volumes within each run first underwent spike removal (*3dDespike*), were temporally aligned (*3dTshift*) and corrected for head motion (*3dvolreg*). The transformation matrices were also used to compute the frame-wise displacement (Power et al., 2012) that identified time points affected by excessive motion (threshold: 0.5 mm). Afterward, a spatial smoothing was performed with a Gaussian kernel having 4 mm Full Width at Half Maximum (FWHM). In this regard, we adopted the AFNI's *3dBlurToFWHM* tool, which first estimated and then iteratively increased the smoothness of data until a specific FWHM level was reached. Considering that the original smoothness was above 3 mm (*3dFWHMx*), this procedure generated a final homogeneous smoothness of 4 mm across voxels and subjects, far less than the one obtained by simply adding a smoothing filter of the same width to the data and aimed at preserving the functional distinctions across adjacent sulcal walls and sulci. Runs were normalized by dividing the intensity of each voxel for its mean over the time series. Normalized runs were then concatenated and a multiple regression analysis was performed (*3dDeconvolve*). Each stimulus comprised three repetitions and was modeled by a standard hemodynamic function (i.e., BLOCK), lasting 1 s. Deviant sounds and button presses were similarly modeled but were excluded from the multivariate procedure detailed below. Movement parameters, frame-wise displacement and polynomial signal trends were included in the analysis as regressors of no interest. The t-score maps of the 80 stimuli were used as input in the multivariate analysis. Single subject data were also registered to the MNI152 standard space (Fonov et al., 2009) using a nonlinear registration (AFNI *3dQWarp*) and resampled to a final resolution of 2x2x2 mm.

Regions of interest

We focused our analysis on two regions of interest in temporal and occipital cortical areas. For temporal areas, we adopted the AICHA atlas (Joliot et al., 2015), which takes into account brain hemispheric specialization and is widely used to identify language responsive areas. The selected regions were the bilateral Superior Temporal Gyri and Sulci as well as the Middle Temporal Gyri. These regions were selected to isolate the portions of the auditory stream activated during the processing of fine spectral features (Santoro et al., 2017), from short-timescales (e.g., phonemes) up to more complex sound patterns (e.g., syllables and words) (DeWitt and Rauschecker, 2012; Keitel et al., 2018; Pernet et al., 2015). Since the primary aim of our study was to investigate the activity elicited by sounds in early visual cortex, we selected the Calcarine cortex as a region of interest, using the probabilistic map (V1, threshold > 10% ; Gau et al., 2020) by Wang and colleagues (Wang et al., 2015). A spatial mask was applied to temporal and occipital areas to select gray matter voxels only (gray matter probabilistic threshold > 0.25).

Global activity regression

To exclude that brain-wide hemodynamic fluctuations could impact the results obtained in temporal and occipital cortices, we opted for a global signal regression procedure (Aguirre, 1998; Macey et al., 2004). The rationale behind this approach is related to the fact that global hemodynamic activity represents motion,

vascular, cardiac and respiratory confounds, as well as sources of neural activity (Liu et al., 2017). Intriguingly, the latter components were described as task-related anticipatory mechanisms (Cardoso et al., 2012; Sirotin and Das, 2009), variations in arousal or alertness (Pisauro et al., 2016) and produced large-scale hemodynamic responses also in resting state, particularly in the eyes-closed condition (Scholvinck et al., 2010). Thus, to avoid possible effects of these physiological confounds, we removed the global signal from our data. Global activity of the full set of stimuli was obtained by averaging hemodynamic responses across gray and white matter voxels. Prior to the regression procedure, we first assessed whether the four categories significantly retained a different whole brain average activity (see Supplementary Figure 3A). The results of this procedure did not highlight specific category-based differences, suggesting that global activity retained a common, positive, hemodynamic average response. Second, we measured the correlation (Spearman's ρ) of the global signal between each pair of subjects (see Supplementary Figure 3B). The results demonstrated that global activity had a subject-specific pattern (ρ=0.002, CI-99th: -0.258, 0.292). We directly measured the association between global activity and envelope modulation power in the High and Low frequency ranges using Spearman's ρ across participants (see Supplementary Figure 3C). The results of this analysis demonstrated that global activity was, on average, uncorrelated with the modulation power of our stimuli. To rule out a possible correlation at the single subject level, global responses were regressed out across all voxels (Macey et al., 2004). The association between global and single voxel activity, measured using R^2, was reported in Supplementary Figure 3D, and showed a relatively high impact of the global signal removal in primary sensory regions, ranging on average between $R^2 \approx 0.2$ to $R^2 \approx 0.3$ in the occipital ROI.

Multivariate analysis

After the removal of global brain signal, the obtained hemodynamic responses were associated with the envelope modulation power using a searchlight approach (Kriegeskorte, 2006) and a machine learning algorithm based on principal component regression (Thirion, 2014). Specifically, for each stimulus category, a searchlight analysis was performed in the above-mentioned regions of interest to predict envelope features using principal components derived from brain activity (see Figure 2A).

It is important to note that, despite the analysis was performed in volumetric space, we respected the cortical folding by sampling voxels (radius: 8 mm) along the cortical ribbon (i.e., the space between pial surface and gray-to-white matter boundary) (Yu, 2019). Specifically, we ran the Freesurfer recon-all analysis pipeline (Reuter et al., 2012) on the standard space template (Fonov et al., 2009), used as reference for the nonlinear alignment of single-subject data. This procedure provided a reconstruction of the gray-matter ribbon, which has been used to isolate searchlight voxels taking into account the cortical folding. Voxel proximity was evaluated using the Dijkstra metric as it represents a computationally efficient method to estimate cortical distance (Fischl, 1999; Lettieri et al., 2019).

For each searchlight, we performed a Principal Components (PC) Analysis to extract orthogonal dimensions from normalized voxel responses within each experimental condition and subject. The retained PCs explained on average 95% of the total variance across the regions of interest and subjects. Afterwards, we performed a

multiple regression analysis, using the PC scores as the independent variable and the power of Low (e.g., 2-6Hz) and High (e.g., 6-10Hz) modulation frequencies as the dependent one. To avoid overfitting and to obtain a robust estimate of these associations (Varoquaux et al., 2017), we performed the regression using a 10-fold cross-validation procedure on both envelope frequency ranges separately. This procedure, ultimately led to a reconstructed model of predicted power values in each experimental condition, subject and searchlight. The reconstructed models were compared to the actual ones by calculating their root mean squared error (RMSE; Poldrack et al., 2019).

To measure the statistical significance, w**e** first averaged the reconstructed models across participants in each experimental condition and searchlight, thus to obtain a group-level predicted set of acoustic features. Secondly, we performed a non-parametric test (1,000 iterations), by shuffling rows and columns of the independent variable (i.e., PC scores) in the training set of each k-fold iteration, keeping the permutation scheme fixed across voxels (Winkler, 2016). This procedure led to a set of 1,000 group-level reconstructed models for each experimental condition and searchlight and ultimately provided a null distribution of RMSE coefficients, against which the actual association was tested. To compute a more accurate estimate (i.e., beyond the number of iterations used in the non-parametric test) of the p-value, we modeled the left tail of the RMSE null-distribution ($p<0.10$) using a generalized Pareto distribution (Knijnenburg et al., 2009; Winkler, 2016). Results were corrected for multiple comparisons using False Discovery Rate (FDR) for each hemisphere separately to control for test dependency (Benjamini and Yekutieli, 2001).

Sound envelope is known to be represented in speech-related cortical areas within the left hemisphere, with antero-posterior gradient from low to high modulation frequencies (Giraud, 2000; Oganian and Chang, 2019; DeWitt and Rauschecker, 2012; Hullett et al., 2016). To assess the degree of spatial overlap of the goodness of fit for the High and Low frequency ranges, we measured the correlation of the RMSE between them, for each speech-related sound category (words, pseudowords, artificial), across all voxels in the occipital and temporal ROIs (Figure 5A). Confidence intervals (99^{th} percentile) were evaluated using the null distribution of RMSE coefficients obtained in the non-parametric procedure described above. Representations of High and Low frequency ranges were not positively correlated for all tested sound categories. We then performed a conjunction analysis (i.e. non-parametric combination; Winkler, 2016) across the linguistic material (i.e. word, pseudoword and artificial stimuli), independently for the two frequency ranges of interest (High and Low), selectively for the two left ROIs. Raw p-values of the three conditions were combined using the Tippett method (Winkler, 2016), and the resulting p-value maps were corrected for multiple comparisons using FDR (Figure 5B).

All the multivariate analyses were performed using MATLAB R2016b (MathWorks Inc., Natick,MA, USA).

Data availability

The data that support the findings of this study will be provided to all readers upon reasonable request.

Code availability

All relevant MATLAB code is available from the corresponding author upon reasonable request.

Acknowledgements

We thank Russel Poldrack for suggestions concerning the data analysis. We also thank Davide Crepaldi for proving access to the Italian version of the Wuggy software.

References

Aguirre, G.K.Z., E.; and D'Esposito, M. (1998). The Variability of Human, BOLD Hemodynamic Response. Neuroimage *8*, 360-369.

Bambini, V., and Trevisan, M. (2012). EsploraCoLFIS: Un'interfaccia web per le ricerche sul Corpus e Lessico di Frequenza dell'Italiano Scritto (CoLFIS). Quaderni del Laboratorio di Linguistica *11*, 1-16.

Barlow H (1961) Possible principles underlying the transformation of sensory messages. In: Rosenblith W, ed. Sensory Communication. Cambridge, MA: MIT Press. 217–234.

Benjamini, Y., and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. Annals of statistics, 1165-1188.

Biesmans, W., Vanthornhout, J., Wouters, J., Moonen, M., Francart, T., and Bertrand, A. (2015). Comparison of speech envelope extraction methods for EEG-based auditory attention detection in a cocktail party scenario. In In 2015 37th annual international conference of the ieee engineering in medicine and biology society (embc) (IEEE), pp. pp. 5155-5158.

Bourguignon, M., Baart, M., Kapnoula, E.C., and Molinaro, N. (2020). Lip-Reading Enables the Brain to Synthesize Auditory Features of Unknown Silent Speech. J Neurosci *40*, 1053-1065.

Cappe, C., and Barone, P. (2005). Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. Eur J Neurosci *22*, 2886-2902.

Carandini, M., Demb, J.B., Mante, V., Tolhurst, D.J., Dan, Y., Olshausen, B.A., Gallant, J.L., and Rust, N.C. (2005). Do we know what the early visual system does? J Neurosci *25*, 10577-10597.

Cardoso, M.M., Sirotin, Y.B., Lima, B., Glushenkova, E., and Das, A. (2012). The neuroimaging signal is a linear sum of neurally distinct stimulus- and task-related components. Nat Neurosci *15*, 1298-1306.

Chanauria, N., Bharmauria, V., Bachatene, L., Cattan, S., Rouat, J., and Molotchnikoff, S. (2019). Sound induces change in orientation preference of V1 neurons: Audio-visual cross-influence. Neuroscience *404*, 48-61.

Chandrasekaran, C. et al. (2009) The natural statistics of audiovisual speech. PLoS Comput. Biol. *5, 1-18.*

Charbonneau, V., Laramée, M.E., Boucher, V., Bronchti, G., and Boire, D. (2012). Cortical and subcortical projections to primary visual cortex in anophthalmic, enucleated and sighted mice. European Journal of Neuroscience *36*, 2949-2963.

Cichy, R.M., Heinzle, J., and Haynes, J.D. (2012). Imagery and perception share cortical representations of content and location. Cereb Cortex *22*, 372-380.

Clavagnier, S., Falchier, A., and Kennedy, H. (2004). Long-distance feedback projections to area V1: Implications for multisensory integration, spatial awareness, and visual consciousness. Cognitive, Affective & Behavioral Neuroscience *4*, 117-126.

Cox, R.W. (1996). AFNI: Software for Analysis and Visualization of Functional Magnetic Resonance Neuroimages. Computers and Biomedical Reasearch *29*, 162-173.

Crosse, M. J., Di Liberto, G. M., and Lalor, E. C. (2016). Eye can hear clearly now: inverse effectiveness in natural audiovisual speech processing relies on long-term crossmodal temporal integration. J. Neuroscience, *36(38), 9888-9895.*

Davis, M.H., and Gaskell, M.G. (2009). A complementary systems account of word learning: neural and behavioural evidence. Philos Trans R Soc Lond B Biol Sci *364*, 3773-3800.

de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L., and Theunissen, F.E. (2017). The Hierarchical Cortical Organization of Human Speech Processing. J Neurosci *37*, 6539-6557.
Dehaene-Lambertz, G.D., S. and Hertz-Pannier, L. (2002). Functional Neuroimaging of Speech Perception in Infants. Science *298*, 2013-2015.
DeWitt, I., and Rauschecker, J.P. (2012). Phoneme and word recognition in the auditory ventral stream. Proc Natl Acad Sci U S A *109*, E505-514.
Di Liberto, G.M., O'Sullivan, J.A., and Lalor, E.C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. Curr Biol *25*, 2457-2465.
Eckert, M.A., Kamdar, N.V., Chang, C.E., Beckmann, C.F., Greicius, M.D., and Menon, V. (2008). A cross-modal system linking primary auditory and visual cortices: evidence from intrinsic fMRI connectivity analysis. Hum Brain Mapp *29*, 848-857.
Falchier, A., Clavagnier, S., Barone, P., and Kennedy, H. (2002). Anatomical Evidence of Multimodal Integration in Primate Striate Cortex. The Journal of Neuroscience *22*, 5749-5759.
Ferraro, S., Van Ackeren, M.J., Mai, R., Tassi, L., Cardinale, F., Nigri, A., Bruzzone, M.G., D'Incerti, L., Hartmann, T., Weisz, N., *et al.* (2020). Stereotactic electroencephalography in humans reveals multisensory signal in early visual and auditory cortices. Cortex *126*, 253-264.
Fischl, B., Sereno, M. I., & Dale, A. M. (1999). Cortical surface-based analysis: II: inflation, flattening, and a surface-based coordinate system. Neuroimage *9*, 195-207.
Fonov, V.S., Evans, A.C., McKinstry, R.C., Almli, C.R., and Collins, D.L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage *47*.
Formisano, E.D.M.F.B.M.G.R. (2008). "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. Science *322*, 970-973.
Gau, R., Bazin, P.L., Trampel, R., Turner, R., and Noppeney, U. (2020). Resolving multisensory and attentional influences across cortical depth in sensory cortices. Elife *9*.
Ghazanfar, A.A., and Schroeder, C.E. (2006). Is neocortex essentially multisensory? Trends Cogn Sci *10*, 278-285.
Giard, M.H., and Peronnet, F. (1999). Auditory-Visual Integration during Multimodal Object Recognition in Humans: A Behavioral and Electrophysiological Study. Journal of Cognitive Neuroscience *11*, 473-490.
Giordano, B.L., Ince, R.A.A., Gross, J., Schyns, P.G., Panzeri, S., and Kayser, C. (2017). Contributions of local speech encoding and functional connectivity to audio-visual speech perception. Elife *6*.
Giraud, A.L., and Poeppel, D. (2012). Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci *15*, 511-517.
Giraud, A.L.L., C.; Ashburner, J.; Wable, J.; Johnsrude, I.; Frackowiak, R. and Kleinschmidt, A. (2000). Representation of the Temporal Envelope of Sounds in the Human Brain. Journal of Neurophysiology *84*, 1588-1598.
Guzman-Martinez, E., Ortega, L., Grabowecky, M., Mossbridge, J., and Suzuki, S. (2012). Interactive coding of visual spatial frequency and auditory amplitude-modulation rate. Curr Biol *22*, 383-388.
Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E., and Weisz, N. (2018). A Visual Cortical Network for Deriving Phonological Information from Intelligible Lip Movements. Curr Biol *28*, 1453-1459 e1453.
Hensch, T.K. (2005). Critical period plasticity in local cortical circuits. Nat Rev Neurosci *6*, 877-888.
Hickok, G. (2012). The cortical organization of speech processing: feedback control and predictive coding the context of a dual-stream model. J Commun Disord *45*, 393-402.
Hohmann, V. (2002). Frequency analysis and synthesis using a Gammatone filterbank. Acta Acustica united with Acustica *88*, 433-442.
Hsu, A., Woolley, S.M., Fremouw, T.E., and Theunissen, F.E. (2004). Modulation power and phase spectrum of natural sounds enhance neural encoding performed by single auditory neurons. J Neurosci *24*, 9201-9211.
Hubel, D.H., and Wiesel, T.N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Phisiology *160*.
Hullett, P.W., Hamilton, L.S., Mesgarani, N., Schreiner, C.E., and Chang, E.F. (2016). Human Superior Temporal Gyrus Organization of Spectrotemporal Modulation Tuning Derived from Speech Stimuli. The Journal of Neuroscience *36*, 2014-2026.
Ibrahim, L.A., Mesik, L., Ji, X.Y., Fang, Q., Li, H.F., Li, Y.T., Zingg, B., Zhang, L.I., and Tao, H.W. (2016). Cross-Modality Sharpening of Visual Cortical Processing through Layer-1-Mediated Inhibition and Disinhibition. Neuron *89*, 1031-1045.

Iurilli, G., Ghezzi, D., Olcese, U., Lassi, G., Nazzaro, C., Tonini, R., Tucci, V., Benfenati, F., and Medini, P. (2012). Sound-driven synaptic inhibition in primary visual cortex. Neuron *73*, 814-828.

Joliot, M., Jobard, G., Naveau, M., Delcroix, N., Petit, L., Zago, L., Crivello, F., Mellet, E., Mazoyer, B., and Tzourio-Mazoyer, N. (2015). AICHA: An atlas of intrinsic connectivity of homotopic areas. J Neurosci Methods *254*, 46-59.

Kayser, C., Petkov, C.I., and Logothetis, N.K. (2008). Visual modulation of neurons in auditory cortex. Cereb Cortex *18*, 1560-1574.

Keitel, A., Gross, J., and Kayser, C. (2018). Perceptually relevant speech tracking in auditory and motor cortex reflects distinct linguistic features. PLoS Biol *16*, e2004473.

Keuleers, E., and Brysbaert, M. (2010). Wuggy: a multilingual pseudoword generator. Behav Res Methods *42*, 627-633.

Kiefer, M., Sim, E.J., Herrnberger, B., Grothe, J., and Hoenig, K. (2008). The sound of concepts: four markers for a link between auditory and conceptual brain systems. J Neurosci *28*, 12224-12230.

Kim, E.J., Juavinett, A.L., Kyubwa, E.M., Jacobs, M.W., and Callaway, E.M. (2015). Three Types of Cortical Layer 5 Neurons That Differ in Brain-wide Connectivity and Function. Neuron *88*, 1253-1267.

Knijnenburg, T.A., Wessels, L.F., Reinders, M.J., and Shmulevich, I. (2009). Fewer permutations, more accurate P-values. Bioinformatics *25*, i161-168.

Köhler, W. (1929). Gestalt psychology (New York, NY: Liveright).

Kriegeskorte, N.G., R.; and Bandettini, P. (2006). Information-based functional brain mapping. PNAS *103*, 3863–3868.

Lakatos, P., Karmos, G., Mehta, A.D., Ulbert, I., and Schroeder, C.E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. science *320*, 110-113.

Leaver, A.M., and Rauschecker, J.P. (2016). Functional Topography of Human Auditory Cortex. J Neurosci *36*, 1416-1428.

Lettieri, G., Handjaras, G., Ricciardi, E., Leo, A., Papale, P., Betta, M., Pietrini, P., and Cecchetti, L. (2019). Emotionotopy in the human right temporo-parietal cortex. Nat Commun *10*, 5568.

Liang, M., Mouraux, A., Hu, L., and Iannetti, G.D. (2013). Primary sensory cortices contain distinguishable spatial patterns of activity for each sense. Nat Commun *4*, 1979.

Liu, T.T., Nalci, A., and Falahpour, M. (2017). The global signal in fMRI: Nuisance or Information? Neuroimage *150*, 213-229.

Luo, H., and Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron *54*, 1001-1010.

Macey, P.M., Macey, K.E., Kumar, R., and Harper, R.M. (2004). A method for removal of global effects from fMRI time series. Neuroimage *22*, 360-366.

Martuzzi, R., Murray, M.M., Michel, C.M., Thiran, J.P., Maeder, P.P., Clarke, S., and Meuli, R.A. (2007). Multisensory interactions within human primary cortices revealed by BOLD dynamics. Cereb Cortex *17*, 1672-1679.

Massopust, L.C., Barnes, H.W., and Verdura, J. (1965). Auditory frequency discrimination in cortically ablated monkeys. Journal of Auditory Research.

McGurk, H.a.M., J. (1976). Hearing lips and seeing voices. Nature *264*, 746-748.

Mercier, M.R., Foxe, J.J., Fiebelkorn, I.C., Butler, J.S., Schwartz, T.H., and Molholm, S. (2013). Auditory-driven phase reset in visual cortex: human electrocorticography reveals mechanisms of early multisensory integration. Neuroimage *79*, 19-29.

Mesgarani, N.C., C.; Johnson, K. and Chang, E. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. Science *349*, 1006-1010.

Moerel, M., De Martino, F., Ugurbil, K., Yacoub, E., and Formisano, E. (2015). Processing of frequency and location in human subcortical auditory structures. Sci Rep *5*, 17048.

Molholm, S., Ritter, W., Javitt, D.C., and Foxe, J.J. (2004). Multisensory visual-auditory object recognition in humans: a high-density electrical mapping study. Cereb Cortex *14*, 452-465.

Muckli, L., Vizioli, L., Petro, L., De Martino, F., and Vetter, P. (2015). Predictive coding of auditory and contextual information in early visual cortex-evidence from layer specific fMRI brain reading. Journal of vision *15*, 720-720.

Mukamel, R., Gelbard, H., Arieli, A., Hasson, U., Fried, I., and Malach, R. (2005). Coupling between neuronal firing, field potentials, and FMRI in human auditory cortex. Science *309*, 951-954.

Mumford, J.A., Davis, T., and Poldrack, R.A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. Neuroimage *103*, 130-138.
Naselaris, T., Olman, C.A., Stansbury, D.E., Ugurbil, K., and Gallant, J.L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. Neuroimage *105*, 215-228.
Obleser, J., Eisner, F., and Kotz, S.A. (2008). Bilateral speech comprehension reflects differential sensitivity to spectral and temporal features. J Neurosci *28*, 8116-8123.
Oganian, Y., and Chang, E.F. (2019). A speech envelope landmark for syllable encoding in human superior temporal gyrus. Science Advances *5*.
Oldfield, R.C. (1971). The assessment and analysis of handedness: The Endinburgh Inventory. Neuropsychologia *9*, 97-113.
Orchard-Mills, E., Van der Burg, E., and Alais, D. (2013). Amplitude-modulated auditory stimuli influence selection of visual spatial frequencies. J Vis *13*.
Overath, T., McDermott, J.H., Zarate, J.M., and Poeppel, D. (2015). The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat Neurosci *18*, 903-911.
Parise, C., and Spence, C. (2013). Audiovisual cross-modal correspondences in the general population. The Oxford handbook of synaesthesia, 790-815.
Parise, C. V., & Ernst, M. O. (2016). Correlation detection as a general mechanism for multisensory integration. Nature comm, *7(1)*, 1-9.
Park, H., Ince, R. A., Schyns, P. G., Thut, G., & Gross, J. (2018). Representational interactions during audiovisual speech entrainment: Redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. PLoS biology, *16(8)*
Pascual-Leone, A., and Hamilton, R. (2001). The metamodal organization of the brain. Progress in Brain Research *134*, 1-19.
Pasley, B.N., David, S.V., Mesgarani, N., Flinker, A., Shamma, S.A., Crone, N.E., Knight, R.T., and Chang, E.F. (2012). Reconstructing speech from human auditory cortex. PLoS Biol *10*, e1001251.
Patterson, M.L., and Werker, J.F. (2003). Two-month-old infants match phonetic information in lips and voice. Developmental Science *6*, 191-196.
Peelle, J.E. (2012). The hemispheric lateralization of speech processing depends on what "speech" is: a hierarchical perspective. Front Hum Neurosci *6*, 309.
Pernet, C.R., McAleer, P., Latinus, M., Gorgolewski, K.J., Charest, I., Bestelmeyer, P.E., Watson, R.H., Fleming, D., Crabbe, F., Valdes-Sosa, M., *et al.* (2015). The human voice areas: Spatial organization and inter-individual variability in temporal and extra-temporal cortices. Neuroimage *119*, 164-174.
Pick H. L., W.D.H.a.H.J.C. (1969). Sensory conflict in judgments of spatial direction. Perception & Psychophysics *6*, 203-205.
Pisauro, M.A., Benucci, A., and Carandini, M. (2016). Local and global contributions to hemodynamic activity in mouse cortex. J Neurophysiol *115*, 2931-2936.
Poldrack, R.A., Huckins, G., and Varoquaux, G. (2019). Establishment of best practices for evidence for prediction: a review. JAMA psychiatry.
Power, J.D., Barnes, K.A., Snyder, A.Z., Schlaggar, B.L., and Petersen, S.E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. Neuroimage *59*, 2142-2154.
Proverbio, A.M., D'Aniello, G.E., Adorni, R., and Zani, A. (2011). When a photograph can be heard: vision activates the auditory cortex within 110 ms. Sci Rep *1*, 54.
Rampinini, A.C., Handjaras, G., Leo, A., Cecchetti, L., Betta, M., Marotta, G., Ricciardi, E., and Pietrini, P. (2019). Formant Space Reconstruction From Brain Activity in Frontal and Temporal Regions Coding for Heard Vowels. Front Hum Neurosci *13*, 32.
Rauschecker, J.P., and Scott, S.K. (2009). Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci *12*, 718-724.
Reuter, M., Schmansky, N.J., Rosas, H.D., and Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage *61*, 1402-1418.
Riddoch, G. (1917). Dissociation of visual perceptions due to occipital injuries, with especial reference to appreciation of movement. Brain *40*, 15-57.
Riecke, F.B., D. A.; and Bialek, W. (1995). Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. Proceedings: Biological Sciences *262*, 259-265.
Rockland, K.S., and Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. International Journal of Psychophysiology *50*, 19-26.

Rohe, T., and Noppeney, U. (2016). Distinct Computational Principles Govern Multisensory Integration in Primary Sensory and Association Cortices. Curr Biol *26*, 509-514.
Romei, V., Murray, M.M., Cappe, C., and Thut, G. (2009). Preperceptual and stimulus-selective enhancement of low-level human visual cortex excitability by sounds. Curr Biol *19*, 1799-1805.
Santoro, R., Moerel, M., De Martino, F., Valente, G., Ugurbil, K., Yacoub, E., and Formisano, E. (2017). Reconstructing the spectrotemporal modulations of real-life sounds from fMRI response patterns. Proc Natl Acad Sci U S A *114*, 4799-4804.
Scholvinck, M.L., Maier, A., Ye, F.Q., Duyn, J.H., and Leopold, D.A. (2010). Neural basis of global resting-state fMRI activity. Proc Natl Acad Sci U S A *107*, 10238-10243.
Schönwiesner, M., Rübsamen, R., and Von Cramon, D.Y. (2005). Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. European Journal of Neuroscience *22*, 1521-1528.
Schroeder, C.E., and Foxe, J. (2005). Multisensory contributions to low-level, 'unisensory' processing. Curr Opin Neurobiol *15*, 454-458.
Sirotin, Y.B., and Das, A. (2009). Anticipatory haemodynamic signals in sensory cortex not predicted by local neuronal activity. Nature *457*, 475-479.
Søndergaard, P.L., and Majdak, P. (2013). The auditory modeling toolbox. In The technology of binaural listening Springer, ed. (Berlin, Heidelberg.), pp. 33-56.
Sourav, S., Kekunnaya, R., Shareef, I., Banerjee, S., Bottari, D., and Röder, B. (2019). A Protracted Sensitive Period Regulates the Development of Cross-Modal Sound– Shape Associations in Humans. Psychological Science *30*, 1437-1482.
Spence, C. (2011). Crossmodal correspondences: a tutorial review. Atten Percept Psychophys *73*, 971-995.
Stein, B.E., and Meredith, M.A. (1993). The merging of the senses.
Srivastava, A., Lee, A. B., Simoncelli, E. P. & Zhu, S. C. (2003). On advances in statistical modeling of natural images. J. Math. Imaging Vis. 18, 17–33.
Thirion, B.V., G.; Grisel, O.; Poupon, C.; Pinel, P. (2014). Principal Component Regression predicts functional responses across individuals. International Conference on Medical Image Computing and Computer-Assisted Intervention, 741-748.
Thwaites, A., Schlittenlacher, J., Nimmo-Smith, I., Marslen-Wilson, W.D., and Moore, B.C. (2017). Tonotopic representation of loudness in the human cortex. Hear Res *344*, 244-254.
Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017). Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. NeuroImage *145*, 166-179.
Vetter, P., Smith, F.W., and Muckli, L. (2014). Decoding sound and imagery content in early visual cortex. Curr Biol *24*, 1256-1262.
Wang, L., Mruczek, R.E., Arcaro, M.J., and Kastner, S. (2015). Probabilistic Maps of Visual Topography in Human Cortex. Cereb Cortex *25*, 3911-3931.
Wang, Y., Celebrini, S., Trotter, Y., and Barone, P. (2008). Visuo-auditory interactions in the primary visual cortex of the behaving monkey: electrophysiological evidence. BMC Neurosci *9*, 79.
Watkins, S., Shams, L., Tanaka, S., Haynes, J.D., and Rees, G. (2006). Sound alters activity in human V1 in association with illusory visual perception. Neuroimage *31*, 1247-1256.
Webb, A.R., Heller, H.T., Benson, C.B., and Lahav, A. (2015). Mother's voice and heartbeat sounds elicit auditory plasticity in the human brain before full gestation. Proc Natl Acad Sci U S A *112*, 3152-3157.
Winans, S.S. (1967). Visual form discrimination after removal of the visual cortex in cats. Science *158*, 944-946.
Winkler, A.W., M. A.; Brooks, J. C.; Tracey, I.; Smith, S. M.; and Nichols, T. E. (2016). Non-parametric combination and related permutation tests for neuroimaging. Hum Brain Mapp *37*, 1486-1511.
Yu, Y.H., L.; Yang, J.; Jangraw, D. C.; Handwerker, D. A.; Molfese, P. J.; Chen, G.; Ejima, Y.; Wu, J.; and Bandettini P. A. (2019). Layer-specific activation of sensory input and predictive feedback in the human primary somatosensory cortex. Science Advances *5*.
Zangenehpour, S., and Zatorre, R.J. (2010). Crossmodal recruitment of primary visual cortex following brief exposure to bimodal audiovisual stimuli. Neuropsychologia *48*, 591-600.