

Research Articles | Behavioral/Cognitive

Representation of emotional expressions across the face and voice brain networks

<https://doi.org/10.1523/JNEUROSCI.0366-26.2026>

Received: 2 March 2026

Revised: 30 June 2026

Accepted: 6 August 2026

Copyright © 2026 the authors

This Early Release article has been peer reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.jneurosci.org/alerts to receive customized email alerts when the fully formatted version of this article is published.

Representation of emotional expressions across the face and voice brain networks

Authors: Stefania Mattioni^{1,2}, Federica Falagiarda², Tanisha Arya², Rémi Gau², Mohamed Rezk², Ceren Battal^{2,3}, Alice Van Audenhaege², Olivier Collignon²

Affiliations:

¹Department of Experimental Psychology & Center for Cognitive Neuroscience (CCN), University of Gent, Belgium

²Institutes for research in Neuroscience (IoNS) , University of Louvain (UCLouvain), Belgium

³Faculty of Psychology and Educational Sciences, University of Coimbra, Portugal

Correspondence should be addressed to: Stefania Mattioni (stefania.mattioni@ugent.be).

CRedit Author Statement

Stefania Mattioni: Conceptualization, Methodology, Formal Analysis, Writing- Original Draft, Visualization

Federica Falagiarda: Conceptualization, Methodology, Software, Formal Analysis, Investigation, Writing- Original Draft, Visualization, Funding Acquisition

Tanisha Arya: Formal Analysis, Visualization

Olivier Collignon: Conceptualization, Methodology, Writing- Review and Editing, Supervision, Project Administration, Funding Acquisition

Remi Gau: Software, Data Curation

Mohamed Rezk: Software

Ceren Battal: Software, Data Curation

Alice Van Audenhaege: Investigation

Acknowledgements

This work was supported by the Belgian Excellence of Science program (EOS Project No. 30991544) attributed to OC, the Flag-ERA HBP PINT-MULTI (R.8008.19) attributed to OC, a “mandat d’impulsion scientifique” (MIS-FNRS) attributed to OC. FF was a research fellow, CB was a postdoctoral researcher and OC is a senior research associate at the National Fund for Scientific Research of Belgium (FRS-FNRS).

Abstract

Recognizing emotions from faces and voices is fundamental to successful social interaction, yet how the brain represents emotional information across sensory systems remains unclear. Here, we systematically mapped the contribution of individually-defined face- and voice-selective cortical regions in 24 human participants of either sex to unimodal and crossmodal emotion expressions. Using multivariate pattern classification, we show that emotional expressions are robustly decodable from the dominant sensory modality across all individually defined face- and voice-selective areas, and that these representations cannot be accounted for by low- and mid-level stimulus properties alone. Strikingly, emotional information from the non-dominant modality—vocal expressions within face networks and facial expressions within voice networks—was also represented throughout the network, with the exception of the occipital face areas. Crossmodal decoding further revealed a shared neural code for facial and vocal emotions in the temporal voice areas, posterior superior temporal sulcus, and precentral gyrus, indicating that these regions contain emotion representations that generalize across sensory modalities. Furthermore, in most face- and voice-selective regions, audiovisual emotional expressions could be distinguished from representations elicited by the dominant unimodal modality, indicating that bimodal signals are not simply represented as their strongest sensory component. Together, these findings reveal that emotion representations are distributed broadly across face- and voice-selective regions and identify temporo-frontal regions as key sites supporting modality-independent emotion representations.

Significance statement

Recognizing emotions from faces and voices is essential for everyday social interaction, yet the brain mechanisms that integrate these signals remain poorly understood. Using functional MRI (fMRI) and advanced pattern analyses, we show that emotional expressions are represented widely across both face- and voice-selective brain regions—even when conveyed through a non-preferred sensory modality. Strikingly, specific temporo-frontal regions encode facial and vocal emotions using a partially shared, modality-independent neural code. We further demonstrate that seeing and hearing emotions together reshapes how emotional information is organized in the brain, revealing a graded transition from sensory-specific to multisensory representations. These findings challenge strictly segregated models of emotion processing and provide a more integrated account of how the brain supports social understanding.

JNeurosci Accepted Manuscript

Introduction

Emotional states can be readily inferred from facial and vocal expressions. Specific configurations of facial muscle movements (Ekman & Friesen, 1971) and vocal acoustic cues (Sauter & Eimer, 2010) convey emotional states, and recognizing these expressions is fundamental for successful social interactions (Darwin, 1872).

In humans, face perception engages face-selective regions in occipital and temporal cortices (Duchaine & Yovel, 2015; Kanwisher & Yovel, 2006; Pitcher et al., 2011), organized into two processing streams (Finzi et al., 2021; Pitcher & Ungerleider, 2021; Weiner & Grill-Spector, 2013). The ventral stream, including the occipital (OFA) and fusiform (FFA) face areas, traditionally associated with invariant facial features and face identity (Andrews & Ewbank, 2004; Grill-Spector et al., 2004; Kanwisher et al., 1997; Winston et al., 2004), and the lateral stream, centered on the posterior superior temporal sulcus (pSTS), thought to process dynamic, social, and multimodal aspects of faces (Duchaine & Yovel, 2015; Pitcher et al., 2014; Pitcher & Ungerleider, 2021; Puce et al., 1998; Zhu & Beauchamp, 2017). Accordingly, emotional expressions are thought to be primarily encoded in the pSTS (Haxby et al., 2000; Pitcher & Ungerleider, 2021). However, emotional expressions also modulate activity in ventral face-selective regions (Ganel et al., 2005; Harry et al., 2013; Kawasaki et al., 2012; Monroe et al., 2013; Vuilleumier & Pourtois, 2007; Wegrzyn et al., 2015), suggesting that emotion representations are more broadly distributed across the face network (Jiahui et al., 2023).

Voice perception engages the temporal voice areas (TVAs; Belin et al., 2000). Although early models proposed that vocal emotions are processed mainly beyond the TVAs, particularly in the pSTS (Ethofer et al., 2006; Grandjean et al., 2005; Kotz et al., 2013; Wildgruber et al., 2005), later work suggests that emotional information might be represented throughout the TVA network (Ethofer et al., 2009).

Emotion perception is inherently multisensory, as emotional expressions are conveyed through congruent facial and vocal cues together with modality-independent information (Lewkowicz & Ghazanfar, 2009; Schirmer & Adolphs, 2018). Face and voice emotion recognition engage similar perceptual mechanisms (Kuhn et al., 2017). Integrating these signals is crucial for extracting coherent communicative meaning (Ghazanfar & Logothetis, 2003; Lewkowicz & Ghazanfar, 2009), improving both the speed and accuracy of emotion recognition (Collignon et al., 2008, 2010; Falagiarda & Collignon, 2019; Kreifelts et al., 2007; Vroomen et al., 2001). The integration of facial and vocal emotional expressions likely relies on core evolutionary mechanism (Darwin, 1872), as non-human primates (Ghazanfar et al., 2005) and young infants (Walker-Andrews & Lennon, 1991) also match their congruency.

Where facial and vocal emotions converge in the human brain, however, remains unclear (Klasen et al., 2012; Young et al., 2020). Hierarchical accounts propose that modality-specific regions extract emotional information before transmitting it to higher-order convergence regions, such as pSTS or medial prefrontal cortex, where abstract emotion representations emerge (Davies-Thompson et al., 2019; Hagan et al., 2009; Kreifelts et al., 2007; Peelen et al., 2010; Schirmer & Adolphs, 2018; Watson et al., 2014). Alternatively, emotional information may already be represented within modality-selective regions, consistent with electrophysiological recordings in macaques (Dureux et al., 2024; Ghazanfar et al., 2005; Khandhadia et al., 2021), voice-driven responses in the human FFA (Kriegstein et al., 2005; 2006; Lowe et al., 2021), structural connectivity between FFA and TVA (Benetti et al., 2018; Blank et al., 2011), and crossmodal emotion coding in the fusiform gyrus (Lettieri et al., 2024).

A comprehensive account of how categorical emotional expressions are represented across face and voice networks is still lacking. Here, we characterized neural representations of dynamic visual, auditory, and audiovisual emotional expressions from five emotions in individually defined face- and voice-selective regions. We asked whether emotional representations are distributed across face and voice network, extend to non-preferred sensory regions, share a common representational structure across modalities, and whether audiovisual expressions evoke neural representations distinct from the dominant unimodal modality. These questions were addressed using multivariate decoding and representational similarity analyses of fMRI data.

Methods

Participants

Twenty-four (twelve male, age range: 20-44) neurotypically developed adults and naïve to the experimental hypothesis, took part in the study. All participants were native French speakers, right-handed and had normal or corrected-to-normal vision (corrected in the scanner with MRI-compatible glasses when needed) and self-reported normal audition. They provided their written informed consent in the first experimental session, prior to the start of any data collection.

Stimuli

The experiment employs 20 audio-visual clips as stimuli. The stimuli were selected from a set of dynamic audio-visual recordings in which actors and actresses produced prototypical facial and vocal expressions (without linguistic content) of basic emotions (Belin et al., 2008; Simon et al., 2008). The original stimuli were bimodal (audio-visual) clips. These

clips were presented in their original format in the bimodal condition. For the unimodal conditions, the auditory and visual components were extracted from the same clips: the audio-only condition consisted of the auditory track of the original videos, while the visual-only condition consisted of the silent video track. Each video portrays one of four actors (two males, two females) expressing one of the five following expressions: disgusted, fearful, happy, sad or neutral. The clips last one second and aim at showing the emotional state through the facial expression as well as a concomitant non-linguistic vocalization. The stimuli were selected among a wider pool of actors and actresses, which were rated, separately for the face and voice expressions, by an independent group (N=20). The chosen stimuli did not induce any specific discrimination bias in the five selected expressions, neither in vision nor in audition. In addition to the face and voice stimuli, twenty videos and twenty audio tracks of everyday objects and scenes were used in the functional face and voice localizer respectively (see section *Localizers*). These stimuli were taken from the internet and were copyright free.

The use of dynamic stimuli in the investigation of visual perception in general, and facial perception specifically, is crucial for recreating the vivid experience that we typically encounter in everyday life. It has been shown that, compared to static images, dynamic stimulation produces higher discrimination performance, differential neural activation and successful decoding through machine learning (Davies-Thompson et al., 2019; Grèzes et al., 2007; Liang et al., 2018; Trautmann-Lengsfeld et al., 2013). Dynamic visual stimuli are even more important when investigating audio-visual perception and integration, where the auditory component cannot be un-embedded from its time course, and where a static visual stimulation would always have a limited generalizability.

Both the auditory and visual stimuli used in this experiment can be found online at [this OSF link](#).

Design and procedure

The experimental protocol was approved by the Ethical committee of the University Hospital of St Luc (Université Catholique de Louvain), Brussels, Belgium. The entirety of the experiment spanned two experimental sessions at St Luc Hospital, with a duration of about 2 hours each. The first session consisted in a brief phase of behavioral rating of the stimuli, followed by the MRI acquisition. The second session included only the MRI acquisition part. Each participant spent around 90 minutes in the scanner in each session, for a total of about 3 hours of MRI data acquisition for each participant.

In the two sessions inside the MRI scanner, each participant attended several acquisition runs. The runs included two block-design functional localizers and a total of 54 short event-related runs with three different modalities of stimuli presentation: visual only

(facial expression), auditory only (non-linguistic vocalization) or bimodal stimulation (congruent facial and vocal expression). One high resolution structural image was also acquired in the first session for each subject (see Acquisition parameters).

In each session, the participant completed one of the two functional localizers as well as nine runs in each of the three modalities. Across the two sessions, 18 runs per subject per modality were acquired. Each participant additionally completed a short behavioral experiment, which took place on the first day of the fMRI experiment, before the commencing of any MRI data acquisition. During this behavioral session, participants were asked to rate each unimodal stimulus on several scales (see *Supplemental Material* for more details) divided into two tasks and blocked by modality (20 visual stimuli and 20 auditory stimuli rated independently, tasks' order was counterbalanced across participants). These data are not analyzed or discussed in the current report, however it is relevant to note that participants were familiarized with the stimuli prior to the scanning, to ensure that the emotion expressed in the stimuli would be recognized as the emotion categories we expected.

Localizers

We used functional localizers for the definition of regions of interest in individual subjects. The aim of both face and voice localizers was to provide individually defined activation clusters of the core face and voice network regions. We were open to the inclusion of more areas that the localizers would consistently reveal across participants. The localizer designs (explained more in detail in the upcoming paragraphs), were initially piloted on six subjects that are not included in the final sample of participants.

Face localizer

The face localizer (see Figure 1A), which aims at the definition of the core face network in each participant (see Figure 1B and 1C), consists in one functional acquisition run with a block-design alternating ten blocks of faces to ten blocks of objects/scenes. Each block of either faces or objects consisted in the presentation of twenty 1s-long videos, with a 100ms inter-stimulus interval (ISI). The facial expressions videos were the same twenty clips used in the main experiment, while the twenty objects/scenes contained in the objects' block were similar to the ones utilized in a well-established dynamic face localizer (Fox et al., 2009), but downloaded anew from the internet and trimmed to a 1s length when needed (see the *Supplemental Material* for the exhaustive list of stimuli). We included ten seconds of presentation of the background only, at the beginning, middle and end of the run, to be used to estimate the baseline. All videos subtended a frame of 720 x 480 pixels on the screen, which corresponded to 8.5 degrees horizontally and 5.7 degrees vertically, of visual angle at a distance of about 175cm.

Voice localizer

Analogous to the face localizer, the voice localizer (see Figure 1A) aims at the definition of the core voice network (see Figure 1B and 1C) and consists in a functional acquisition with a block-design, alternating twelve blocks of voices to twelve blocks of auditory objects. Each block of either voices or objects consisted in the presentation of twenty 1s-long audio tracks, with 100ms ISI. All soundtracks were RMS-equalized and presented through in-ear MRI-compatible earphones (Sensimetric S14, SensiMetrics, USA). The vocal expressions were the same used in the main experiment, while the auditory objects' sounds (see the Supplemental Material for the exhaustive list of stimuli) were downloaded from the internet and trimmed to a 1s length when needed (copyright free sounds from freesounds.org). We included ten seconds of silence at the beginning, middle and end of the run for baseline. Participants were asked to set a loud but comfortable volume for the auditory stimuli at the beginning of each scanning session.

[Figure 1]

Main experiment

In the main experiment, only stimuli coming from the audio-visual emotional clips were presented. The experiment, conducted over two sessions, consists of fifty-four runs divided into three modality conditions: eighteen runs for the visual, eighteen runs for the auditory, and eighteen runs for the bimodal tasks. All runs of the main experiment have an event-related design and every run contains one repetition of each stimulus, i.e. 20 stimuli. In the visual task, the rendered videos were cropped to 332 pixels wide and 424 pixels high, sustaining 3.9 degrees horizontally and 5 degrees vertically of visual angles in the participants' visual field. For the auditory task, the audio track was extracted from the original clips as a wav file and presented through in-ear MRI-compatible earphones (Sensimetric S14, SensiMetrics, USA). During the auditory presentation, participants were asked to keep their eyes open and looked at a fixation cross that appeared on screen during the event. In the bimodal task, both the audio track and the video frames were presented and rendered as a synchronized, always congruent, audio-visual clip. In all tasks, each stimulus was presented twice in a row, for an event duration of 2 seconds (2 x 1-second file) and by an inter-stimulus interval (ISI) of 3 seconds (see Figure 2A for illustration of trial sequences of the main experimental runs).

In all the runs of the main experiment, participants were asked to press a button when the same emotion appeared in two events in a row, independently of the actor portraying the emotion. This ensured that their attention was on the emotional content of the expression,

rather than on lower-level features of the stimulus, since the emotion repetition was always portrayed across actor identity. One third of the runs in each modality condition contained two repeated emotions (two 1-back trials), a third of the runs contained one 1-back trial while a third of the runs contained none. Each run could therefore have 20, 21 or 22 stimuli. The order of the stimuli was pseudorandomized, so that the same emotion would *never* appear twice in a row, except in the case of the planned one-back trials.

Behavioral performance on the one-back emotion detection task was high across all modalities (visual: $89.4 \pm 12.1\%$; auditory: $86.3 \pm 14.4\%$; bimodal: $92.0 \pm 8.1\%$). A Friedman test revealed a significant effect of modality ($\chi^2(2) = 7.27$, $p = 0.026$). However, pairwise Wilcoxon signed-rank tests did not reveal any significant differences between modalities after FDR correction for multiple comparisons (all $pFDR > 0.09$). Numerically, performance was highest in the bimodal condition and lowest in the auditory condition, suggesting a modest advantage for audiovisual over auditory-only stimuli. The raw data and a graph of these behavioral task results are reported in [OSF additional material section](#). Overall, these results indicate that participants performed the task accurately and attended to the emotional content of the stimuli across all modalities.

[Figure 2]

Acquisition parameters

Participants were scanned with a 3T GE (Signa™ Premier, General Electric Company, USA, serial number: 000000210036MR03, software version: 28\LX\MR, Software release : RX28.0_R04_2020.b) MRI scanner with a 48-channel head coil at the University Hospital of St Luc, Brussels, Belgium.

In the first scanning session, a whole-brain T1-weighted anatomical scan was acquired (3D-MPRAGE; 1.0 x 1.0 mm in-plane resolution; slice thickness 1mm; no gap; inversion time = 900 ms; repetition time (TR) = 2189,12 ms; echo time (TE) = 2.96 ms; flip angle (FA) = 8°; 156 slices; field of view (FOV) = 256 x 256 mm²; matrix size= 256 X 256).

All functional runs were T2*-weighted scans acquired with Echo Planar and gradient recalled (EP/GR) imaging (2.6 x 2.6 mm in-plane resolution; slice thickness = 2.6mm; no gap; Multiband factor = 2; TR = 1750ms, TE = 30ms, 58 interleaved ascending slices; FA = 75°; FOV = 220 x 220 mm²; matrix size= 84 X 84).

Data analyses

Data preprocessing and General Linear Model

The MRI data were pre-processed with bidsfm (v3.0.0; <https://github.com/cpp-lln-lab/bidsfm>; Gau et al., 2023) set up through the use of the in-house CPP_SPM repository, v0.1.0. (<https://doi.org/10.5281/zenodo.3554331>) and carried out through the Statistical Parametric Mapping toolbox (SPM12, RRID:SCR_007037) in Matlab, version 2016b. The preprocessing of the functional images was performed separately for each of the five tasks: the face localizer, the voice localizer, the main experiment visual, the main experiment auditory, the main experiment bimodal, and in the following order: removing of 7 dummy scans, slice timing correction, realignment, unwarping, normalization to MNI space, smoothing. Slice timing correction was performed using the 29th slice as a reference (interpolation: sinc interpolation). Functional scans from each participant were realigned and unwrapped using the first image of the first run as a reference (number of degrees of freedom: 6; cost function: least square). The mean image obtained from realignment was then co-registered to the anatomical T1 image (number of degrees of freedom: 6; cost function: normalized mutual information) (Friston et al., 1995). The transformation matrix from this co-registration was then applied to all the functional images. Functional MNI normalized images were then spatially smoothed using a 3D gaussian kernel (FWHM of 6mm for the runs of the functional localizers, 2mm for the runs of the main experiment).

The anatomical T1 image was bias field corrected, segmented and normalized to MNI space (target space: IxI549Space; target space resolution: 1 mm ; interpolation: 4th degree b-spline) using a unified segmentation. The deformation field obtained from this step was then applied to all the functional images (target space resolution equal that used at acquisition; interpolation: 4th degree b-spline).

For all functional data, at the subject level, we performed a mass univariate analysis with a linear regression at each voxel of the brain, using generalized least squares with a global FAST model to account for temporal auto-correlation (Corbin et al., 2018) and a drift fit with discrete cosine transform basis (128 seconds cut-off). Image intensity scaling was done run-wide before statistical modelling such that the mean image will have mean intracerebral intensity of 100. Regressors modelled the BOLD response following the onset of the stimuli (either blocks or events) and were convolved with SPM canonical hemodynamic response function (HRF). Nuisance covariates included the 6 realignment parameters to account for residual motion artefacts.

Regions of interest definition

Regions of interest (ROIs) were defined at the individual level using functional localizer contrasts. Face-selective regions were identified using the contrast *faces > images of objects*, and voice-selective regions using the contrast *voices > sounds of objects*. Statistical maps were thresholded at $p < 0.001$ (uncorrected), and clusters were identified within anatomically constrained search spaces based on prior literature.

Through the face localizer contrasting face and object responses, we aimed to define the core face network (Haxby et al., 2001), including bilateral FFA, OFA and pSTS. The network could be robustly identified in most of the participants: rFFA was found in 20 out of 24 participants, IFFA and rOFA were found in 21 out of 24 participants, IOFA and rpSTS in 23 and lpSTS in all 24 participants. In addition, the precentral gyrus (PCG) showed significant face-selective responses (*faces > objects*, $p < 0.001$ uncorrected) at the group level and was therefore included as an additional ROI, in line with previous reports (Rossion et al., 2012). The PCG could be identified in 22/24 participants in the right hemisphere and 21/24 in the left. We could not consistently identify any other area of the extended face network, like the amygdala or the anterior temporal lobe and therefore excluded those regions from our analyses (see Supplemental material 'Assessment of amygdala localization and signal quality' and [OSF project for tSNR analyses](#)).

In order to create individual masks for each ROI, we found the activation peak within each cluster of interest (identified with the contrast *faces > object*; all individual peak coordinates are reported in Table S1). We then created an expanding ROI starting as a sphere and that would follow the contours of the cluster until reaching 150 voxels (this was done through the in-house CPP_ROI repository). For activation clusters that did not have a minimum of 150 voxels from the start (at a threshold of .001 uncorrected), we drew a sphere of 10mm in radius, starting at the activation peak of each ROI in each subject (but without the boundaries of the activation cluster for that region). These masks have then been used to perform multivoxel pattern analyses (MVPA), with a feature selection of 100 voxels. In three cases, even the spherical ROI mask did not allow for the inclusion of 100 voxels for MVPA; those three areas have therefore been excluded.

Analogously, with our functional voice localizer we identified the core voice network in each participant, constituted mainly by cortical territory along the mid STS/STG that has been consistently shown to preferentially respond to voice, and that from here onwards will be referred to as TVA. We could robustly identify the TVAs with our localizer: left and right TVAs were identified in all analyzed subjects. An additional peak located in the precentral gyrus (PCG) emerged through the [*voice > objects*] contrast and could be identified in 19 out of 24

subjects, so we included the PCG in the areas defined through the voice localizer as well. Despite the notable overlap between the visually and auditorily defined prefrontal regions, these regions show different activation profiles at the univariate level (see Figure 2B), therefore we decided to keep them separate and we will refer to the visually defined pre-central gyrus as PCGv, and to the auditorily defined one as PCGa (all individual peak coordinates are reported in Table S2). The voice localizer acquisition run of one subject had huge signal distortion in the image, so we therefore discarded that acquisition and used the group-defined clusters to build the four masks used for that subject. As for the face localizer, when the individual-activation-cluster ROIs did not reach the desired size of 150vx at a lenient threshold, we drew a sphere of 10mm in radius, starting at the individual activation peak; in two cases, even the spherical ROI mask did not allow for the inclusion of 100 voxels, so those two areas have been excluded.

All the individual masks (see Figure 1B & 1C) for both the face and the voice localizer are also shared on the OSF project.

Univariate analysis

In order to characterize the response profile of the defined ROIs, we extracted the mean beta estimates for each regressor in each modality condition (across all voxels and all runs of each ROI, see Figure 2B for a representation of the mean beta across regressors, by modality). For each modality of stimulation, visual (faces), auditory (voices) or audio-visual, we tested whether the mean estimate of the group was different from zero through one-sample t-tests. We additionally tested whether, separately in each ROI, the emotion regressor and the modality of stimulation predicted the beta estimate, through a generalized linear mixed model (GLMM: $\text{beta} \sim \text{modality} * \text{emotion} + (1|\text{sub})$).

Figure 2B shows the beta weights extracted from all regions for the visual (blue), auditory (red), and bimodal (purple) conditions, averaged across emotion expressions. Beta estimates for each modality, presented separately for each emotion and ROI, are provided in Supplemental Figure 1.

Multivariate pattern classification: within-modality

In order to tackle the question of where in the face network we can find information about facial emotions, and where in the voice network we can find information about vocal emotions we performed within-modality multivariate pattern classification. We trained an algorithm to distinguish pairs of facial expressions in each face sensitive area and pairs of vocal expressions in each voice sensitive area, and evaluated its performance (decoding

accuracy, DA) in each of those ROIs. Given our 5 categories – neutral, disgust, fear, happiness and sadness – there are 10 possible pairs of expressions and the chance level for each pair is 50%. If the algorithm can distinguish pairs of expressions with an overall above chance DA, we conclude that in that ROI, information about e.g. facial emotion expressions, is represented.

This type of MVP-classification additionally allows us to answer the question of whether we can find traces of facial emotions in the voice network, and of vocal emotions in the face network.

All MVPA were implemented with a linear support vector machine, through the LIBSVM library (www.csie.ntu.edu.tw/~cjlin/libsvm/, (Chang & Lin, 2011), in CosmoMVPA (www.cosmomvpa.org, (Oosterhof et al., 2016) and Matlab, with a leave-one-run-out cross-validation procedure, and feature selection set at 100 voxels. Results from this analysis are reported in figure 3A.

Multivariate pattern classification: across visual & auditory modalities

This analysis is implemented in ROIs that present successful within-modality decoding with both the visual and auditory stimuli. It allows us to further understand whether the emotional information in these areas are represented in a way that is, at least partly, common between the two sensory modalities, and hence, partially independent of them. This is achieved through yet another multivariate pattern classification procedure, where the training set is constituted by the data from one modality of acquisition and the test set by the data from the opposite modality. The two modalities are both used as training and testing sets, and the results are averaged. Like in the previous analysis, this decoding analysis is carried out in CosmoMVPA, through a linear support vector machine, with a leave-one-run-out cross-validation procedure and a feature selection of 100 voxels.

Results from this analysis are reported in figure 3B.

Multivariate pattern classification, between unimodal and bimodal conditions

To anticipate part of the results, the within-modality MVP-classification showed a successful decoding of emotion expressions in the bimodal condition, without however any enhancement in the performance compared to the dominant modality (aka face in the face network, voices in the voice network- see figure 2A), made us wonder whether the representation in these areas comes down to the representation of the dominant modality when this is present, without distinction between a unisensory and a bisensory context. In the light of these results, a way to explore whether a multisensory context is distinguishable from

a unisensory one in these regions would be to perform MVP-classification within emotional category of different sensory contexts, specifically of the native modality of each ROI (e.g. visual facial expressions in visually-defined face network) and the bimodal condition. In order to do that, runs are taken from the bimodal condition and then either from the visual or the auditory condition; the classifier is then trained to distinguish e.g. visual disgust from bimodal disgust, visual fear from bimodal fear, and so on. If the representation of the emotional information in each ROI comes down to the dominant modality for that region, the classifier should not be able to tell the difference between e.g. auditory sadness and bimodal sadness in TVA. Importantly, although the unimodal and bimodal stimuli were presented in separate runs, this does not introduce a confound for the present analysis. The comparison between sensory contexts (e.g., visual vs. bimodal) was always performed using data from different runs for training and testing, ensuring that both conditions contributed information collected under equivalent circumstances. Any potential run-specific noise or baseline differences therefore affected the unimodal and bimodal data equally and cannot explain successful decoding. Because the classifier was trained to distinguish patterns corresponding to the same emotion across these two contexts (e.g., visual disgust vs. bimodal disgust) using independent runs for both conditions, any above-chance performance reflects genuine differences in the underlying neural representations rather than artefacts of the experimental design.

Results from this analysis are reported in figure 3C.

Representational Similarity Analysis (RSA)

To further investigate whether the representational geometry observed in our ROIs reflected emotional category rather than low- and mid-level stimulus properties, we performed a series of representational similarity analyses (RSA, Kriegeskorte et al., 2008). Unlike decoding analyses, RSA allows direct comparison between the representational structure of neural activity and theoretical or computational models. Importantly, by using partial correlation, RSA also enables testing whether the relationship between neural representations and emotion category remains after statistically controlling for variance explained by low- and mid-level visual or auditory features of the stimuli.

For each participant and ROI, neural representational dissimilarity matrices (RDMs) were computed from the condition-specific beta estimates using the cross-validated Mahalanobis (crossnobis) distance. For each pair of conditions, one run was iteratively held out as the test partition while the remaining runs were averaged to form the training partition. Pattern differences were computed independently in the training and test partitions and

compared across folds. Reliable differences between conditions resulted in positive distances, whereas unreliable differences yielded values close to zero (or occasionally slightly negative). Distances were averaged across cross-validation folds to obtain one RDM per participant and ROI.

Neural RDMs were compared with theoretical and computational dissimilarity models using Spearman rank correlations. Correlation coefficients were computed separately for each participant and ROI and subsequently averaged across participants.

To evaluate whether emotion-category information explained neural representational structure beyond low- and mid-level stimulus properties, we additionally performed partial RSA analyses. Specifically, we computed the correlation between each neural RDM and the emotion-category DSM after removing the variance explained by the corresponding stimulus-derived models. For the visual condition, these confound models included pixel motion, silhouette, and mouth configuration; for the auditory condition, they included amplitude envelope, Mel-frequency cepstral coefficients (MFCCs), and spectral centroid (see below for models' description). This analysis allowed us to test whether the neural representational structure remained related to emotion category after accounting for the portion of the emotion model that could be explained by some important low-/mid-level stimulus features.

Here is a description of each representational model.

The theoretical emotion-category model consisted of a 20×20 binary dissimilarity matrix in which pairs of stimuli belonging to the same emotional category (neutral, disgust, fear, happiness, sadness) were assigned a value of 0, whereas pairs belonging to different emotional categories were assigned a value of 1.

To characterize low- and mid-level visual properties, three computational models were constructed.

Pixel motion. Overall motion energy was estimated by computing the absolute pixel-wise difference between consecutive grayscale frames using OpenCV. Motion values were averaged across frames for each video and Euclidean distances were computed between all stimulus pairs.

Silhouette. To capture coarse body and facial shape independently of internal facial features, each frame was converted into a binary silhouette, assigning a value of 1 to the actor and 0 to the background. Binary masks were concatenated across time and Euclidean distances between stimulus pairs were used to construct the DSM.

Mouth configuration. Mouth dynamics were characterized using Action Units associated with orofacial movement (AU10, AU12, AU14, AU15, AU17, AU20, AU23, AU25 and AU26) extracted with OpenFace 2.0 (Baltrušaitis et al., 2018). Action Unit intensity vectors were concatenated across frames and Euclidean distances were computed between stimuli.

Three computational models were used to characterize low- and mid-level auditory properties. All features were extracted with Librosa (McFee et al., 2015), computed frame-wise, concatenated across time, and converted into DSMs using cosine distance.

Amplitude envelope. The temporal envelope of each audio signal was extracted to characterize overall amplitude fluctuations over time.

Mel-frequency cepstral coefficients (MFCCs). Thirteen MFCCs were extracted for each frame to capture timbre-related spectral envelope information commonly used to characterize speech and emotional prosody.

Spectral centroid. Spectral centroid was computed for each frame as an estimate of the spectral center of mass, reflecting the perceptual brightness of the sound.

Finally, all DSMs were rescaled to the [0,1] interval prior to the RSA analyses.

Brain data visualization: dissimilarity matrices, multidimensional scaling and dendrograms

To further characterize the nature of each ROI and examine in greater detail how different modalities coexist and interact within the visual and auditory networks, we focused on the neural representation of each emotion and modality. This approach allowed us to assess the similarity or differences in activity patterns across emotions and modalities.

First, for each subject, we computed a dissimilarity matrix (DSM) for each ROI, incorporating the five emotions across the three modalities. Using Spearman's correlation, we calculated the 1 - similarity between the activity patterns for each emotion and the other emotions, both within the same modality and across different modalities. To account for potential confounds due to varying noise across runs (since each modality was presented in separate runs), we used a split-half approach. Specifically, we divided the data from each modality into two halves (i.e., odd and even runs) and averaged the activity patterns for each of the five emotions in each half. This resulted in two activity patterns for each emotion in each modality. We then computed the 1 - Spearman's correlation between the patterns from one half and the corresponding patterns from the other half. This approach generated non-symmetrical 15x15 matrices (5 emotions x 3 modalities). To convert these matrices into

symmetrical dissimilarity matrices, we averaged the values from the upper and lower triangles and set the diagonal values to zero (see Figure 5).

In addition to DSMs, we also visualize the data using 3D multidimensional scaling.

The emotions from the 3 modalities of presentation have been arranged such that their pairwise distances approximately reflect response pattern similarities. Items placed close together show similar response patterns. Items arranged far apart generated different response patterns (Mattioni et al., 2020). Finally we also visualize the brain data in the form of a dendrogram. We used MATLAB's 'dendrogram' function to create a tree-like diagram that shows how the brain data is grouped into clusters. Different colors are used to highlight these clusters, based on how closely the data points are related. We used the default setting, where the colors change when the similarity between clusters drops below 70% of the highest possible value. This helps us easily see which emotions/modalities are grouped together.

Searchlight decoding analyses

Whole-brain searchlight

For the decoding analyses previously described, we complemented the ROI approach with a whole-brain searchlight analysis. The primary objectives of these additional analyses were to: 1) confirm that the face and voice networks identified by the localizer were the main regions involved in the representation of facial and vocal emotions and 2) investigate whether other brain regions, outside of our functionally defined face and voice networks, were also involved in representing these emotions.

We repeated the three within modality decoding analyses (within visual modality, within auditory modality & within bimodal modality) and the crossmodal audio-visual decoding analysis described above (Figure 6), but this time the same procedure was implemented in each 100-voxels sphere using a searchlight approach (Tong & Pratte, 2012). In all the classification analyses we ran multiclass decoding analyses, testing the discriminability of patterns for the five emotions using a linear support vector machine (SVM), through the LIBSVM library (www.csie.ntu.edu.tw/~cjlin/libsvm/), (Chang & Lin, 2011), in CosmoMVPA (www.cosmomvpa.org), (Oosterhof et al., 2016) and Matlab, with a leave-one-run-out cross-validation procedure. Classification accuracy for each sphere was assigned to the central voxel of the sphere, in order to produce accuracy maps for each subject. The resulting accuracy maps were then smoothed with an 8-mm Gaussian kernel.

ROI-based searchlight

In a second step, we further focused on the three regions that exhibited a supramodal profile (i.e. significant decoding within- and across- modalities) in the ROI results—namely, the TVA, pSTS, PCG— to examine their internal spatial organization. Specifically, we aimed to investigate the presence of spatial gradients of modality sensitivity within these regions, assessing whether distinct subregions preferentially represented visual, auditory, or crossmodal emotional information. Importantly, these regions were not defined from the whole-brain searchlight analyses. Instead, they were identified from the ROI analyses performed on independently localized face- and voice-selective regions derived from the functional localizer tasks. To do this, we first combined the individual localizer-defined ROIs from all participants to create average group-level masks for each of the six regions (left and right TVA, pSTS, and PCG). These independently defined masks were then used to constrain the searchlight analyses, allowing us to characterize the spatial distribution of visual, auditory, and crossmodal decoding within each ROI. The goal of this analysis was therefore not to establish the presence of decoding in these regions, but rather to examine their internal functional organization.

Because the goal of this analysis was to characterize the internal spatial organization of modality-specific and crossmodal decoding within regions already identified as multimodal (in previous ROIs analysis), rather than to establish the presence of significant decoding, the resulting maps are displayed at a threshold of $p < 0.001$ uncorrected (see Figure 7). Within these constrained regions, we examined whether subregions showed preferential decoding for visual or auditory modality, or instead exhibited crossmodal (audio–visual) integration. In the case of the TVA, we also compared the within-auditory, within-visual, and crossmodal decoding maps with the anterior, middle, and posterior TVA peaks previously identified by Pernet et al., 2015 (The exact MNI coordinates are reported in Figure 7), to assess potential correspondence with known functional subdivisions.

Statistical analyses

Significance of the decoding accuracies in the ROIs analyses is evaluated through a non-parametric procedure that involves permutations and bootstrapping of the data (Stelzer et al., 2013) in the following fashion: for each subject, an additional one-hundred MVP-classification procedures are performed, but where the labels for the 5 emotions are randomly assigned. Then, one value from the obtained one-hundred is sampled from each subject, ten thousand times, in order to obtain a group-level null distribution. The statistical significance of the classification results is assessed by calculating the proportion of accuracy values in the

group null distribution that exceed the observed value. All p-values are then FDR corrected for multiple comparisons (number of comparisons equal to the number of ROIs tested).

For the RSA analyses, statistical significance of the correlation coefficients was assessed using a non-parametric permutation procedure. For each ROI and participant, the association between the neural representational dissimilarity matrix (RDM) and the target model was recomputed 10,000 times after randomly permuting the labels of the target model, generating a null distribution of correlation values. The observed group-average correlation was then compared against this null distribution to estimate a p-value. The same procedure was applied to both the standard and partial RSA analyses. P-values were corrected for multiple comparisons across ROIs using the false discovery rate (FDR).

For the searchlight decoding analyses we used the brain maps obtained from the searchlight analyses for each subject to run group-level statistics in SPM. Specifically, we performed a voxel-wise one-sample t-test to identify brain regions where the decoding performance was significantly above chance across participants. We report results corrected for multiple comparisons using a $p < 0.05$ (FWE) threshold. In addition, to transparently provide a comprehensive view of the networks involved we also report the results at an uncorrected threshold of $p < 0.005$ in the Supplemental Information (see figure S2).

Results

Univariate results

The analyses of the beta estimates show that the average response in each region is significantly above zero for the stimuli in the preferred modality of each ROI, as well as for the bimodal stimuli. The FFAs and the TVAs do not show significant positive estimates for the stimulation in their non-dominant modality, vocal and facial stimulation respectively, while the OFAs show significant negative response to the voices (deactivation). The visually-defined pSTS shows significant positive response to vocal as well as facial stimuli bilaterally, with a significantly higher response to faces, and an even higher response to bimodal stimuli. The IPCGv also shows significant positive estimates for all modalities, but without differences among them. Similarly, the bilateral PCGa shows significant responses to all modalities, with higher estimates for the bimodal stimulation but no difference for the two unimodal ones (see Figure 2B and Supplemental Figure 1).

Finally, in only a few regions, we found a main effect of emotion (e.g., rOFA: fear > neutral; bilTVA: sadness > disgust, fear, and neutral; bilpSTS: disgust > neutral), but overall, different emotions elicited similar brain responses across all ROIs at the univariate level.

Distinct responses for separate emotions were therefore mostly explored in the following multivariate analyses.

Multivariate pattern classification, within-modality

The results of the within-modality decoding accuracies are shown in Figure 3A. We observed that emotion expressions could be robustly classified in the dominant modality of any area (facial expressions in visually defined regions and vocal expressions in auditorily defined regions, all $p_{FDR} < 0.001$) as well as from the bimodal conditions (all $ps_{FDR} < 0.001$), with generally large effect sizes (see Supplementary Table S3). Additionally, vocal expressions could be selectively decoded from face-selective regions in IFFA ($p_{FDR} = 0.032$) and rFFA ($p_{FDR} = 0.026$), pSTS and PCGv, bilaterally (all $ps_{FDR} < 0.001$); and facial expressions could be decoded in the voice selective regions ITVA ($p_{FDR} = 0.012$), rTVA ($p_{FDR} = 0.003$) and PCGa bilaterally ($p_{FDR} < 0.001$). In the FFAs and in the TVAs, the decoding accuracy is higher in their native modality than in the opposite modality, while pSTS and all the PCG (a and v) exhibit similar levels of decoding for facial and vocal expressions. In all ROIs, the decoding accuracy in the bimodal condition was not statistically different from the decoding accuracy of the dominant modality. The dominant (or native) modality of a ROI refers to the sensory modality used to define the ROI and that yields to the highest unimodal decoding accuracy, e.g. visual (facial expressions) in FFAs and audition (vocal expressions) in TVAs. Effect sizes (Cohen's d), together with mean decoding accuracies and pFDR values, are reported in Supplementary Table S3.

[Figure 3]

Multivariate pattern classification across audio-visual modalities

We proceeded to run MVP-classification across modalities in the ROIs that showed significant within-modality decoding both for facial and vocal emotions. These areas are: bil-FFA, bil-pSTS, bil-PCGv, bil-TVA and bil-PCGa. Results are shown in Figure 3B. The pSTS, the TVA and both the PCGv and PCGa showed significant cross-modal decoding bilaterally (all $ps_{FDR} \leq 0.0015$), while the crossmodal decoding accuracy in the FFAs do not differ from chance level (significance of the decoding accuracies is evaluated through the same procedure described in method section *Statistical analysis*, with FDR-correction carried out here for 10 ROIs). Effect sizes (Cohen's d), together with mean decoding accuracies and pFDR values, are reported in Supplementary Table S4.

Decoding between unimodal and bimodal conditions

The within-modality MVP-classification showed a successful decoding of emotion expressions in the bimodal condition, without however any enhancement in the decoding accuracies. To follow up on these results, we performed a cross-decoding analysis between the dominant modality and the bimodal modality (e.g. visual-bimodal in FFA and auditory-bimodal in TVA) to see whether a multisensory context is distinguishable from a unisensory one in these regions. If the representation of the emotional information in each ROI comes down to the dominant modality for that region, the classifier should not be able to tell the difference between e.g. auditory sadness and bimodal sadness in TVA. The results for this analyses are shown in Figure 3C and show that overall, the classifier could distinguish unimodal vs bimodal emotions significantly in rFFA ($p_{\text{FDR}} = 0.041$), IOFA ($p_{\text{FDR}} = 0.027$), bil-pSTS ($p_{\text{FDR}} < 0.001$), rPCGv ($p_{\text{FDR}} = 0.007$), bil-TVA ($p_{\text{FDR}} = 0.019$), bil-PCGa ($p_{\text{FDR-left}} = 0.041$, $p_{\text{FDR-right}} < 0.001$), but only a trend is found in rOFA ($p_{\text{FDR}} = 0.059$), while decoding is not significantly different from chance in IFFA ($p_{\text{FDR}} = 0.291$), and IPCGv ($p_{\text{FDR}} = 0.146$). Although not completely homogeneous across regions, these findings show that there is a differential activation in the face network when voices are concurrently perceived, and the same is true in the voice network when faces are concurrently perceived; even though this never translated in a better ability of the classifier to distinguish emotion categories in the bimodal condition compared to the respective native unimodal condition in any region. These results, together with the other MVP-classification analyses carried out in this report, speak for the multimodal nature of these networks.

Representational Similarity Analysis (RSA)

To determine whether the neural representational structure primarily reflected emotion category rather than low- and mid-level stimulus properties, we compared each ROI with a categorical emotion model and a series of computational visual and auditory models (Figure 4).

For the visual condition, the neural representational structure correlated significantly with the emotion-category model in the bilateral FFA (left: $r = 0.028$, $p_{\text{FDR}} = 0.014$; right: $r = 0.067$, $p_{\text{FDR}} < 0.001$), bilateral OFA (left: $r = 0.067$, $p_{\text{FDR}} < 0.001$; right: $r = 0.076$, $p_{\text{FDR}} < 0.001$), bilateral PCGv (left: $r = 0.031$, $p_{\text{FDR}} = 0.007$; right: $r = 0.039$, $p_{\text{FDR}} = 0.001$), bilateral pSTS (left: $r = 0.075$, $p_{\text{FDR}} < 0.001$; right: $r = 0.078$, $p_{\text{FDR}} < 0.001$), and bilateral PCGa (left: $r = 0.028$, $p_{\text{FDR}} = 0.012$; right: $r = 0.051$, $p_{\text{FDR}} < 0.001$). No significant correlation was

observed in either TVA. These results are reported in supplemental material figure S4A and in [additional material on OSF in the file RSA Tables \(Table 1-OSF\)](#).

After controlling for the visual low- and mid-level models (pixel motion, silhouette, and mouth configuration), significant correlations with the emotion-category model (Figure 4; Table 1) remained in the bilateral FFA (left: $r = 0.046$, $pFDR = 0.041$; right: $r = 0.056$, $pFDR = 0.016$), left OFA ($r = 0.056$, $pFDR = 0.016$), left PCGv ($r = 0.056$, $pFDR = 0.016$), left pSTS ($r = 0.086$, $pFDR = 0.004$) and the right pSTS showed a trend ($pFDR = 0.054$). The FDR corrected p-values for this analysis are reported with Choen's D in Table 1-left side.

For the auditory condition, significant correlations with the emotion-category model were observed in the right FFA ($r = 0.025$, $pFDR = 0.019$), bilateral PCGv (left: $r = 0.044$, $pFDR < 0.001$; right: $r = 0.079$, $pFDR < 0.001$), bilateral pSTS (left: $r = 0.062$, $pFDR < 0.001$; right: $r = 0.044$, $pFDR < 0.001$), bilateral PCGa (left: $r = 0.075$, $pFDR < 0.001$; right: $r = 0.104$, $pFDR < 0.001$), and bilateral TVA (left: $r = 0.212$, $pFDR < 0.001$; right: $r = 0.202$, $pFDR < 0.001$). These results are reported in supplemental material figure S4B and [additional material on OSF in the file RSA Tables \(Table 5-OSF\)](#).

Importantly, these effects remained after controlling for the auditory low- and mid-level models (amplitude envelope, MFCCs, and spectral centroid) (figure 4). Significant correlations were observed in the bilateral PCGv (left: $r = 0.049$, $pFDR = 0.022$; right: $r = 0.052$, $pFDR = 0.015$), left pSTS ($r = 0.063$, $pFDR = 0.004$), bilateral PCGa (left: $r = 0.070$, $pFDR = 0.004$; right: $r = 0.087$, $pFDR < 0.001$), and bilateral TVA (left: $r = 0.151$, $pFDR < 0.001$; right: $r = 0.151$, $pFDR < 0.001$). The FDR corrected p-values for this analysis are reported with Choen's D in Table 1-right side.

[Table 1]

Finally, we examined the extent to which the computational low- and mid-level stimulus models accounted for the neural representational structure. For the visual condition (figure S4A), the mouth model correlated significantly with the right OFA, right PCG and right pSTS, whereas the pixel-motion model showed significant correlations in the right FFA, bilateral OFA, right PCGv and right pSTS. In contrast, the silhouette model was only significantly correlated with the left OFA. For the auditory condition (figure S4B), the amplitude envelope model was represented in the left pSTS, right PCGa, and bilateral TVA, the MFCC model in the right PCG and bilateral TVA, whereas the spectral centroid model was selectively represented in bilateral TVA.

Overall, the RSA analyses confirmed that emotion-category information could not solely be explained by the low- and mid-level stimulus properties examined here. However,

RSA did not reproduce all of the within-modality decoding effects observed in the non-dominant modality (e.g., auditory emotions in the FFA or facial emotions in the TVA). This difference is not unexpected, as RSA and decoding quantify complementary properties of neural representations. While decoding is optimized to detect discriminative information between conditions, RSA assesses the overall similarity structure of the representational space. Consequently, more subtle representational differences may be detectable with decoding but not necessarily reflected in RSA (Haxby et al., 2014). Importantly, the purpose of these analyses was not to reproduce every decoding effect, but rather to evaluate whether the dominant-modality emotion representations could be accounted for by low- and mid-level stimulus properties. The exact mean, p and Cohen's D values for every ROI and analysis are reported in the [additional material on OSF in the file RSA Tables](#).

[Figure 4]

Brain data visualization: dissimilarity matrices, multidimensional scaling and dendrograms

In Figure 5, we show the group average dissimilarity matrix, multidimensional scaling, and a dendrogram for each ROI, allowing us to accurately visualize how the five categories, presented in three different modalities, cluster (or do not cluster) together. Although more qualitative in nature, data visualization provides a complementary perspective for exploring our brain data and can reveal additional nuances beyond the general results found through classification analyses.

The first key observation concerns the contribution of the non-dominant modality to emotion representation: does the bimodal condition cluster with the dominant modality, or does it form a distinct cluster? The guiding idea is that the more the bimodal representations separate from the dominant modality within a given region, the stronger the contribution of the non-dominant modality to that region's representational space.

The multidimensional scaling plots in Figure 5 suggest that in OFA the bimodal condition largely overlaps with the visual (dominant) modality. In FFA and TVA, the bimodal representations appear to become visually distinguishable from the dominant unimodal modality, whereas in pSTS and PCG—particularly in the auditorily defined regions—the separation between bimodal and dominant-modality representations appears even more pronounced.

The hierarchical clustering visualizations similarly suggest that in OFA and FFA for visual processing and TVA for auditory processing the bimodal condition tends to be grouped together with the dominant modality. In contrast, in the bilateral pSTS and PCG, the bimodal

representations appear less associated with the dominant modality and form more distinct clusters. Together with the decoding results reported above (Fig.3C), these visualizations are consistent with the idea that, in these regions, the combined presentation of visual and auditory information may give rise to representational patterns that are more distinct from those elicited by the unimodal conditions. Interestingly, the dissimilarity matrices (DSMs; Figure 5) for the left pSTS, the right PCGv and the right PCGa—and even more prominently for the left PCGv and PCGa—show visible intermediate diagonals. This pattern suggests that representations of the same emotion (e.g., happiness) may share similarities across sensory modalities (visual, auditory, and bimodal). Thus, despite the three modalities maintaining distinct overall representational structures, individual emotions appear to share a common neural signature that transcends the modality of presentation (in line with the results from crossmodal audio-visual decoding, Fig 3B).

[Figure 5]

Searchlight decoding analyses

Results from the searchlight decoding analysis are presented in Figure 6.

For decoding the five emotions within the visual modality, we observed significant decoding ($p < 0.05$ FWE) in a portion of the the bilateral pSTS and a part of the infero-temporal cortex (Figure 6A). For the auditory modality, we observed significant decoding ($p < 0.05$ FWE) in the bilateral superior and middle temporal cortex and in the left postcentral gyri (Figure 6B). Finally, for the cross-modal audio-visual decoding (Figure 6C and 6D) results, we observed significant decoding ($p < 0.05$ FWE) in a portion of the pSTS.

In Supplementary Figure S2, a wider network of regions involved in within- visual decoding, within- auditory decoding ,and between audio-visual cross-decoding is reported at a less stringent threshold ($p < 0.005$, uncorrected).

It should be noted that the searchlight and ROI analyses are complementary and are not expected to yield identical results. In particular, searchlight analyses are more susceptible to the effects of inter-individual anatomical and functional variability because decoding is assessed at the voxel level across participants, whereas subject-specific ROI analyses reduce this source of variability and may therefore reveal effects that are not detectable in the group-level searchlight analysis.

For decoding the five emotions within the visual modality at this more lenient threshold, we observed significant decoding in a bilateral region of the cortex, encompassing the superior and middle temporal cortex and extending to the inferior occipital cortex and to the fusiform gyrus. This network also included the bilateral postcentral gyri (PCG), inferior temporal cortex, superior and middle occipital gyri, inferior parietal cortex, supplementary motor area, medial and inferior frontal cortex, and left insula (Figure S2A).

For decoding the five emotions within the auditory modality, we observed significant decoding ($p < 0.005$, uncorrected) in the bilateral superior and middle temporal cortex, the left infero-temporal cortex, small bilateral portions of the fusiform gyrus, the bilateral postcentral gyri, the bilateral infero-parietal cortex, the bilateral medial and infero-frontal cortex and insula (Figure S2B).

Finally, for the cross-modal audio-visual decoding (Figure S2C and S2D) results, we observed bilateral middle/superior temporal cortex, the bilateral postcentral gyri, a small portion of the left fusiform gyrus, a left portion of the inferior frontal cortex, the medial frontal cortex and the left insula.

[Figure 6]

Beyond the whole-brain analyses, we conducted a more focused examination of the searchlight results within six regions showing a supramodal profile: the left and right TVA, pSTS, and PCG. The goal of this analysis was to assess whether these regions contain spatially organized subregions exhibiting differential sensitivity to sensory modality. In particular, we tested whether distinct subportions preferentially responded to visual or auditory information, and whether intermediate areas displayed more integrative, multimodal patterns of activation.

In the case of the TVA, we directly compared the within-auditory, within-visual, and between-audio-visual decoding results with the peaks of the anterior, middle, and posterior TVA previously identified by Pernet et al., 2015 (exact coordinates from this study are reported in figure 7). We found that the group peak for within-auditory decoding was located between the middle and anterior TVA, whereas the peak for within-visual decoding was closer to the posterior TVA. The crossmodal (audio-visual) decoding peak was situated in the region of overlap between these two areas.

A similar spatial organization was observed in both the left and right pSTS. Specifically, the within-auditory decoding peaked in the more anterior portion of the ROI, the within-visual

decoding peaked in the more posterior portion, and the crossmodal decoding emerged in the intermediate region connecting the two.

In contrast, no such gradient was evident in the left or right PCG. In this region, the group peaks for the within- and between-modality decoding were spatially close, suggesting a more general modality-independent representation of emotion in this region. See Figure 7 for the MNI coordinates of the group peaks.

[Figure 7]

Discussion

We investigated how emotional expressions are represented across the human face and voice processing networks using multivariate decoding and representational similarity analyses.

Emotional facial expressions were decoded in all face-selective ROIs, and vocal emotions were decoded in all voice-selective ROIs (Figure 3A). Importantly, RSA showed that emotion-category representations remained significant after controlling for multiple low- and mid-level visual or auditory features, indicating that they cannot be explained solely by these stimulus properties (Figure 4). Together, these findings argue against models proposing that emotional information is confined to a limited subset of face- or voice-selective regions (e.g., Kriegeskorte et al., 2007). Instead, they support a broadly distributed representation of emotion across both processing networks. Our findings extend previous evidence that emotional expressions modulate ventral face-selective regions (Bernstein et al., 2018; Liang et al., 2018; Wegrzyn et al., 2015), challenging the classical view that the FFA primarily represents invariant facial properties such as identity while dynamic information is processed mainly within the pSTS (Pitcher et al., 2011).

Further, we found that face- and voice-selective regions are not strictly unimodal. With the exception of the OFA, emotional expressions could be decoded in all investigated regions even when presented through their non-preferred sensory modality (Figure 3A), showing for the first time that crossmodal responses can be reliably encoded in distributed patterns of activity in TVAs and FFAs, but not in OFA. These results strengthen the view that OFA may encode more visual features of faces while FFA encodes more social aspects of faces beyond their visual features (Powell et al., 2018; Tsantani et al., 2020).

Multisensory stimulation also altered the representational geometry of emotional information throughout the face and voice networks (Figure 5). Although the brain representations of audiovisual expressions frequently clustered near the dominant unimodal

condition (Figure 5), decoding analyses nevertheless distinguished unimodal from bimodal responses in most ROIs (Figure 3C). This apparent discrepancy likely reflects the different sensitivity of the two approaches: multivariate decoding identifies subtle distributed differences between conditions, whereas RSA primarily captures the overall geometry of neural representations. Together, these analyses indicate that audiovisual emotional expressions are represented as more than a simple copy of the dominant sensory input (Figure 5).

Faces and voices provide complementary information about socially relevant attributes, making interactions between their respective cortical networks advantageous for extracting emotional meaning (Froesel et al., 2022). Neuropsychological studies have shown that impairments in emotion recognition often affect both facial and vocal expressions (Kuhn et al., 2017; Young et al., 2020). Young infants already match congruent facial and vocal signals (Lewkowicz & Hansen-Tift, 2012). Functional and structural imaging further demonstrates direct interactions between face- and voice-selective regions (Benetti et al., 2018; Blank et al., 2011; Kriegstein et al., 2005). Similar principles emerge from sensory deprivation (Lettieri et al., 2024), where the TVA responds preferentially to faces in deaf individuals (Benetti et al., 2017), whereas face-selective regions respond preferentially to voices in congenitally blind individuals (Fairhall et al., 2017; Mattioni et al., 2020, 2022; Van Den Hurk et al., 2017). Together, these findings suggest that face and voice regions constitute interacting networks specialized for socially meaningful information rather than strictly unimodal sensory systems. Although each network preferentially processes its dominant sensory modality, their continuous interaction to solve a common perceptual problem—recognizing emotions conveyed by talking faces—may allow crossmodal inputs to evoke distributed activity in the non-preferred network, consistent with our crossmodal decoding results.

These findings are compatible with predictive coding accounts of multisensory perception, in which facial and vocal expressions provide complementary evidence about the same emotional state through reciprocal interactions between sensory systems (Talsma, 2015). The widespread representation of emotional information across face- and voice-selective regions, together with modality-independent representations in the TVA, pSTS, and PCG, supports this framework.

Although vocal emotions could be decoded in the FFA, crossmodal decoding between facial and vocal expressions was absent, whereas the TVA supported partially shared representations across modalities. This asymmetry parallels the stronger decoding of facial emotions in the TVA than vocal emotions in the FFA and may reflect differences in the reliability of emotional signals across the senses. Facial expressions are generally recognized

more accurately than vocal expressions and dominate perceptual judgments when the two modalities convey conflicting emotional information (Collignon et al., 2008, 2010), making visual signals more likely to be incorporated into voice-selective regions than vice versa. This interpretation is consistent with evidence showing that vocal emotion perception is influenced by unconsciously perceived facial expressions but not by unrelated affective visual stimuli (De Gelder et al., 2002).

Crossmodal decoding analyses demonstrated that the TVA, pSTS, and PCG represent facial and vocal emotions using partially shared neural codes (Figure 3B). Within both the TVA and pSTS, crossmodal decoding peaked between areas preferentially representing auditory and visual information (Figure 7), suggesting a transition from modality-specific to multisensory representations. This organization closely matches previous descriptions of the TVA (Pernet et al., 2015), where anterior and middle portions respond preferentially to acoustic voice features, whereas posterior regions contribute more strongly to audiovisual integration (Charest et al., 2013; Ethofer et al., 2013; Giordano et al., 2014; Kreifelts et al., 2007, 2009; Latinus et al., 2013; Watson et al., 2014). Accordingly, the peak of crossmodal decoding was located near the pSTS, a region implicated in the integration of voices and visual speech information such as lip movements (Karthik et al., 2024; Zhu & Beauchamp, 2017), reinforcing the role of posterior temporal cortex in combining auditory and visual emotional cues. In contrast, the PCG showed no comparable gradient, with within- and crossmodal decoding largely overlapping, suggesting a more homogeneous multimodal representation, consistent with multisensory processing in frontal cortex in both humans and non-human primates (Guntupalli et al., 2017; Romanski, 2007; Sugihara et al., 2006).

This organization resembles that reported for person identity processing, where representations progress from modality-specific coding in posterior cortex to increasingly invariant representations in the pSTS and frontal cortex (Guntupalli et al., 2017; Tsantani et al., 2020; Visconti di Oleggio Castello et al., 2021). Emotional expressions may therefore follow a similar organizational principle. Future studies comparing the encoding of familiar identities (all identities here were unfamiliar) and emotional expressions could further clarify how these multisensory processes interact within face- and voice-responsive networks.

Although the present study demonstrates widespread emotion representations across the face and voice networks, several questions remain open. Recent work has identified additional face-selective regions in the anterior fusiform gyrus and anterior temporal lobe that support higher-level aspects of face perception, including identity and conceptual representations (Collins et al., 2016; Mende-Siedlecki et al., 2013; Nasr & Tootell, 2012; Rossion et al., 2024). These regions could not be reliably localized with our functional localizer, likely because of their susceptibility to signal loss (see maps on the OSF project, in the [tSNR](#)

[section](#)). Although the amygdala was also not consistently localized at the individual-subject level, supplementary analyses indicated that this absence is unlikely to reflect reduced signal quality alone (see Figure S3 and [tSNR section](#), OSF project).

Because mental imagery recruits sensory-selective cortices (O'Craven & Kanwisher, 2000; Sunday et al., 2018), could it have contributed to the crossmodal representations observed in classical unisensory cortices? This explanation is unlikely, as imagery studies typically require explicit instructions and extensive training (Lee et al., 2012), whereas participants here performed only an orthogonal one-back task without imagery instructions. Secondly, evidence from person-identity studies suggests that crossmodal activations in face-selective regions arise during perception rather than explicit mental imagery (Schall et al. (2013). Thirdly, although both facial and vocal emotions could be decoded in the FFA, crossmodal decoding between the two modalities was absent. If vocal emotion responses in the FFA primarily reflected visual imagery, successful crossmodal decoding would be expected (Dijkstra et al., 2019). Instead, the two modalities were represented by distinct neural patterns despite both containing emotion information, arguing against a purely imagery-based account (see Van Audenhaege et al., 2025; Cerpelloni et al., 2025 for similar findings in the visual word form area).

Although our RSA analyses indicate that the observed representations cannot be explained solely by the low- and mid-level visual and auditory features examined here (Figure 4), future studies using larger and more diverse stimulus sets together with computational models capturing a broader range of stimulus properties will be necessary to further disentangle emotion-category representations from correlated stimulus features.

Taken together, our results suggest a broadly distributed encoding of emotion expressions across the face and voice networks. We find that both the face and the voice networks show widespread representation of emotional information, even across sensory modalities. Furthermore, the posterior superior temporal sulci, the temporal voice areas, and the precentral gyri, not only represent emotional information both in vision and audition, but it does so through a code that is shared between, and hence partially independent of, the sensory modalities of the input.

References

- Andrews, T. J., & Ewbank, M. P. (2004). Distinct representations for facial identity and changeable aspects of faces in the human temporal lobe. *NeuroImage*, 23(3), 905–913. <https://doi.org/10.1016/j.neuroimage.2004.07.060>

977 Baltrušaitis, T., Mahmoud, M., & Robinson, P. (2015, May). Cross-dataset learning and
 978 person-specific normalisation for automatic action unit detection. In 2015 11th IEEE
 979 international conference and workshops on automatic face and gesture recognition
 980 (FG) (Vol. 6, pp. 1-6). IEEE.

981 Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal Affective Voices: A
 982 validated set of nonverbal affect bursts for research on auditory affective processing.
 983 *Behavior Research Methods*, 40(2), 531–539. <https://doi.org/10.3758/BRM.40.2.531>

984 Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in
 985 human auditory cortex. *Nature*, 403(6767), 309–312.
 986 <https://doi.org/10.1038/35002078>

987 Benetti, S., Novello, L., Maffei, C., Rabini, G., Jovicich, J., & Collignon, O. (2018). White matter
 988 connectivity between occipital and temporal regions involved in face and voice
 989 processing in hearing and early deaf individuals. *NeuroImage*, 179, 263–274.
 990 <https://doi.org/10.1016/j.neuroimage.2018.06.044>

991 Benetti, S., Van Ackeren, M. J., Rabini, G., Zonca, J., Foa, V., Baruffaldi, F., Rezk, M., Pavani,
 992 F., Rossion, B., & Collignon, O. (2017). Functional selectivity for face processing in the
 993 temporal voice area of early deaf individuals. *Proceedings of the National Academy of
 994 Sciences*, 114(31). <https://doi.org/10.1073/pnas.1618287114>

995 Bernstein, M., Erez, Y., Blank, I., & Yovel, G. (2018). An Integrated Neural Framework for
 996 Dynamic and Static Face Processing. *Scientific Reports*, 8(1), 7036.
 997 <https://doi.org/10.1038/s41598-018-25405-9>

998 Blank, H., Anwender, A., & Von Kriegstein, K. (2011). Direct Structural Connections between
 999 Voice- and Face-Recognition Areas. *The Journal of Neuroscience*, 31(36), 12906–
 1000 12915. <https://doi.org/10.1523/JNEUROSCI.2091-11.2011>

1001 Cerpelloni, F., Van Audenhaege, A., Matuszewski, J., Gau, R., Battal, C., Falagiarda, F., ... &
 1002 Collignon, O. (2025). How learning to read visual Braille co-opts the reading
 1003 brain. *Journal of Cognitive Neuroscience*, 37(11), 2010-2052.

1004 Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM
 1005 Transactions on Intelligent Systems and Technology*, 2(3), 1–27.
 1006 <https://doi.org/10.1145/1961189.1961199>

1007 Charest, I., Pernet, C., Latinus, M., Crabbe, F., & Belin, P. (2013). Cerebral Processing of
 1008 Voice Gender Studied Using a Continuous Carryover fMRI Design. *Cerebral Cortex*,
 1009 23(4), 958–966. <https://doi.org/10.1093/cercor/bhs090>

1010 Collignon, O., Girard, S., Gosselin, F., Roy, S., Saint-Amour, D., Lassonde, M., & Lepore, F.
 1011 (2008). Audio-visual integration of emotion expression. *Brain Research*, 1242, 126–
 1012 135. <https://doi.org/10.1016/j.brainres.2008.04.023>

1013 Collignon, O., Girard, S., Gosselin, F., Saint-Amour, D., Lepore, F., & Lassonde, M. (2010).
 1014 Women process multisensory emotion expressions more efficiently than men.
 1015 *Neuropsychologia*, 48(1), 220–225.
 1016 <https://doi.org/10.1016/j.neuropsychologia.2009.09.007>

1017 Collins, J. A., Koski, J. E., & Olson, I. R. (2016). More than meets the eye: the merging of
 1018 perceptual and conceptual knowledge in the anterior temporal face area. *Frontiers in
 1019 human neuroscience*, 10, 189.

1020 Cooke, D. F., & Graziano, M. S. A. (2004). Sensorimotor Integration in the Precentral Gyrus:
1021 Polysensory Neurons and Defensive Movements. *Journal of Neurophysiology*, 91(4),
1022 1648–1660. <https://doi.org/10.1152/jn.00955.2003>

1023 Corbin, N., Todd, N., Friston, K. J., & Callaghan, M. F. (2018). Accurate modeling of temporal
1024 correlations in rapidly sampled fMRI time series. *Human Brain Mapping*, 39(10), 3884–
1025 3897. <https://doi.org/10.1002/hbm.24218>

1026 Darwin, C. (1872). *The expression of the emotions in man and animals*.
1027 <https://psycnet.apa.org/fulltext/2004-16316-000-FRM.pdf>

1028 Davies-Thompson, J., Elli, G. V., Rezk, M., Benetti, S., Van Ackeren, M., & Collignon, O.
1029 (2019). Hierarchical Brain Network for Face and Voice Integration of Emotion
1030 Expression. *Cerebral Cortex*, 29(9), 3590–3605.
1031 <https://doi.org/10.1093/cercor/bhy240>

1032 De Gelder, B., Pourtois, G., & Weiskrantz, L. (2002). Fear recognition in the voice is modulated
1033 by unconsciously recognized facial expressions but not by unconsciously recognized
1034 affective pictures. *Proceedings of the National Academy of Sciences*, 99(6), 4121–
1035 4126. <https://doi.org/10.1073/pnas.062018499>

1036 Dijkstra, N., Bosch, S. E., & van Gerven, M. A. (2019). Shared neural mechanisms of visual
1037 perception and imagery. *Trends in cognitive sciences*, 23(5), 423-434.

1038 Duchaine, B., & Yovel, G. (2015). A Revised Neural Framework for Face Processing. *Annual*
1039 *Review of Vision Science*, 1(1), 393–416. [https://doi.org/10.1146/annurev-vision-](https://doi.org/10.1146/annurev-vision-082114-035518)
1040 082114-035518

1041 Dureux, A., Zanini, A., & Everling, S. (2024). Mapping of facial and vocal processing in
1042 common marmosets with ultra-high field fMRI. *Communications Biology*, 7(1), 317.
1043 <https://doi.org/10.1038/s42003-024-06002-1>

1044 Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion.
1045 *Journal of Personality and Social Psychology*, 17(2), 124–129.
1046 <https://doi.org/10.1037/h0030377>

1047 Ethofer, T., Bretscher, J., Wiethoff, S., Bisch, J., Schlipf, S., Wildgruber, D., & Kreifelts, B.
1048 (2013). Functional responses and structural connections of cortical areas for
1049 processing faces and voices in the superior temporal sulcus. *NeuroImage*, 76, 45–56.
1050 <https://doi.org/10.1016/j.neuroimage.2013.02.064>

1051 Ethofer, T., Pourtois, G., & Wildgruber, D. (2006). Investigating audiovisual integration of
1052 emotional signals in the human brain. In *Progress in Brain Research* (Vol. 156, pp.
1053 345–361). Elsevier. [https://doi.org/10.1016/S0079-6123\(06\)56019-4](https://doi.org/10.1016/S0079-6123(06)56019-4)

1054 Ethofer, T., Van De Ville, D., Scherer, K., & Vuilleumier, P. (2009). Decoding of Emotional
1055 Information in Voice-Sensitive Cortices. *Current Biology*, 19(12), 1028–1033.
1056 <https://doi.org/10.1016/j.cub.2009.04.054>

1057 Fairhall, S. L., Porter, K. B., Bellucci, C., Mazzetti, M., Cipolli, C., & Gobbini, M. I. (2017).
1058 Plastic reorganization of neural systems for perception of others in the congenitally
1059 blind. *NeuroImage*, 158, 126–135. <https://doi.org/10.1016/j.neuroimage.2017.06.057>

1060 Falagiarda, F., & Collignon, O. (2019). Time-resolved discrimination of audio-visual emotion
1061 expressions. *Cortex*, 119, 184–194. <https://doi.org/10.1016/j.cortex.2019.04.017>

1062 Finzi, D., Gomez, J., Nordt, M., Rezai, A. A., Poltoratski, S., & Grill-Spector, K. (2021).
1063 Differential spatial computations in ventral and lateral face-selective regions are
1064 scaffolded by structural connections. *Nature Communications*, 12(1), 2278.
1065 <https://doi.org/10.1038/s41467-021-22524-2>

1066 Fox, C. J., Iaria, G., & Barton, J. J. S. (2009). Defining the face processing network:
1067 Optimization of the functional localizer in fMRI. *Human Brain Mapping*, 30(5), 1637–
1068 1651. <https://doi.org/10.1002/hbm.20630>

1069 Friston, K. J., Ashburner, J., Frith, C. D., Poline, J. B., Heather, J. D., & Frackowiak, R. S.
1070 (1995). Spatial registration and normalization of images. *Human brain mapping*, 3(3),
1071 165-189.

1072 Froesel, M., Gacoin, M., Clavagnier, S., Hauser, M., Goudard, Q., & Ben Hamed, S. (2022).
1073 Socially meaningful visual context either enhances or inhibits vocalisation processing
1074 in the macaque brain. *Nature Communications*, 13(1), 4886.
1075 <https://doi.org/10.1038/s41467-022-32512-9>

1076 Ganel, T., Valyear, K. F., Goshen-Gottstein, Y., & Goodale, M. A. (2005). The involvement of
1077 the “fusiform face area” in processing facial expression. *Neuropsychologia*, 43(11),
1078 1645–1654. <https://doi.org/10.1016/j.neuropsychologia.2005.01.012>

1079 Gau, R., Barilari, M., Battal, C., Caron-Guyon, J., Cerpelloni, F., Falagiarda, F., MacLean, M.,
1080 Mattioni, S., Rezk, M., & Shahzad, I. (2023). Bidspm: An spm-centric bids app for
1081 flexible statistical analysis. *In Organization for Human Brain Mapping 2023*.

1082 Ghazanfar, A. A., & Logothetis, N. K. (2003). Facial expressions linked to monkey calls.
1083 *Nature*, 423(6943), 937–938. <https://doi.org/10.1038/423937a>

1084 Ghazanfar, A. A., Maier, J. X., Hoffman, K. L., & Logothetis, N. K. (2005). Multisensory
1085 Integration of Dynamic Faces and Voices in Rhesus Monkey Auditory Cortex. *The*
1086 *Journal of Neuroscience*, 25(20), 5004–5012.
1087 <https://doi.org/10.1523/JNEUROSCI.0799-05.2005>

1088 Giordano, B. L., Pernet, C., Charest, I., Belizaire, G., Zatorre, R. J., & Belin, P. (2014).
1089 Automatic domain-general processing of sound source identity in the left posterior
1090 middle frontal gyrus. *Cortex*, 58, 170–185.
1091 <https://doi.org/10.1016/j.cortex.2014.06.005>

1092 Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M. L., Scherer, K. R., &
1093 Vuilleumier, P. (2005). The voices of wrath: Brain responses to angry prosody in
1094 meaningless speech. *Nature Neuroscience*, 8(2), 145–146.
1095 <https://doi.org/10.1038/nn1392>

1096 Graziano, M. S. A., & Gandhi, S. (2000). Location of the polysensory zone in the precentral
1097 gyrus of anesthetized monkeys. *Experimental Brain Research*, 135(2), 259–266.
1098 <https://doi.org/10.1007/s002210000518>

1099 Grèzes, J., Pichon, S., & De Gelder, B. (2007). Perceiving fear in dynamic body expressions.
1100 *NeuroImage*, 35(2), 959–967. <https://doi.org/10.1016/j.neuroimage.2006.11.030>

1101 Grill-Spector, K., Knouf, N., & Kanwisher, N. (2004). The fusiform face area subserves face
1102 perception, not generic within-category identification. *Nature Neuroscience*, 7(5), 555–
1103 562. <https://doi.org/10.1038/nn1224>

1104 Guntupalli, J. S., Wheeler, K. G., & Gobbini, M. I. (2017). Disentangling the representation of
 1105 identity from head view along the human face processing pathway. *Cerebral*
 1106 *Cortex*, 27(1), 46-53.

1107 Hagan, C. C., Woods, W., Johnson, S., Calder, A. J., Green, G. G. R., & Young, A. W. (2009).
 1108 MEG demonstrates a supra-additive response to facial and vocal emotion in the right
 1109 superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 106(47),
 1110 20010–20015. <https://doi.org/10.1073/pnas.0905792106>

1111 Harry, B., Williams, M. A., Davis, C., & Kim, J. (2013). Emotional expressions evoke a
 1112 differential response in the fusiform face area. *Frontiers in Human Neuroscience*, 7.
 1113 <https://doi.org/10.3389/fnhum.2013.00692>

1114 Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001).
 1115 Distributed and Overlapping Representations of Faces and Objects in Ventral
 1116 Temporal Cortex. *Science*, 293(5539), 2425–2430.
 1117 <https://doi.org/10.1126/science.1063736>

1118 Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system
 1119 for face perception. *Trends in Cognitive Sciences*, 4(6), 223–233.
 1120 [https://doi.org/10.1016/S1364-6613\(00\)01482-0](https://doi.org/10.1016/S1364-6613(00)01482-0)

1121 Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational
 1122 spaces using multivariate pattern analysis. *Annual review of neuroscience*, 37(1), 435-
 1123 456.

1124 Jiahui, G., Feilong, M., Visconti di Oleggio Castello, M., Nastase, S. A., Haxby, J. V., &
 1125 Gobbini, M. I. (2023). Modeling naturalistic face processing in humans with deep
 1126 convolutional neural networks. *Proceedings of the National Academy of*
 1127 *Sciences*, 120(43), e2304085120.

1128 Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in
 1129 Human Extrastriate Cortex Specialized for Face Perception. *The Journal of*
 1130 *Neuroscience*, 17(11), 4302–4311. [https://doi.org/10.1523/JNEUROSCI.17-11-](https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997)
 1131 [04302.1997](https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997)

1132 Kanwisher, N., & Yovel, G. (2006). The fusiform face area: A cortical region specialized for
 1133 the perception of faces. *Philosophical Transactions of the Royal Society B: Biological*
 1134 *Sciences*, 361(1476), 2109–2128. <https://doi.org/10.1098/rstb.2006.1934>

1135 Karthik, G., Cao, C. Z., Demidenko, M. I., Jahn, A., Stacey, W. C., Wasade, V. S., & Brang,
 1136 D. (2024). Auditory cortex encodes lipreading information through spatially distributed
 1137 activity. *Current Biology*, 34(17), 4021-4032.e5.
 1138 <https://doi.org/10.1016/j.cub.2024.07.073>

1139 Kawasaki, H., Tsuchiya, N., Kovach, C. K., Nourski, K. V., Oya, H., Howard, M. A., & Adolphs,
 1140 R. (2012). Processing of facial emotion in the human fusiform gyrus. *Journal of*
 1141 *Cognitive Neuroscience*, 24(6), 1358–1370. https://doi.org/10.1162/jocn_a_00175

1142 Khandhadia, A. P., Murphy, A. P., Romanski, L. M., Bizley, J. K., & Leopold, D. A. (2021).
 1143 Audiovisual integration in macaque face patch neurons. *Current Biology*, 31(9), 1826-
 1144 1835.e3. <https://doi.org/10.1016/j.cub.2021.01.102>

1145 Kuhn, L. K., Wydell, T., Lavan, N., McGettigan, C., & Garrido, L. (2017). Similar
 1146 representations of emotions across faces and voices. *Emotion*, 17(6), 912.

- 1147 Klasen, M., Chen, Y.-H., & Mathiak, K. (2012). Multisensory emotions: Perception,
1148 combination and underlying neural processes. *Reviews in the Neurosciences*, 23(4).
1149 <https://doi.org/10.1515/revneuro-2012-0040>
- 1150 Kotz, S. A., Kalberlah, C., Bahlmann, J., Friederici, A. D., & Haynes, J. (2013). Predicting vocal
1151 emotion expressions from the human brain. *Human Brain Mapping*, 34(8), 1971–1981.
1152 <https://doi.org/10.1002/hbm.22041>
- 1153 Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., & Wildgruber, D. (2007). Audiovisual integration
1154 of emotional signals in voice and face: An event-related fMRI study. *NeuroImage*,
1155 37(4), 1445–1456. <https://doi.org/10.1016/j.neuroimage.2007.06.020>
- 1156 Kreifelts, B., Ethofer, T., Shiozawa, T., Grodd, W., & Wildgruber, D. (2009). Cerebral
1157 representation of non-verbal emotional perception: fMRI reveals audiovisual
1158 integration area between voice- and face-sensitive regions in the superior temporal
1159 sulcus. *Neuropsychologia*, 47(14), 3059–3066.
1160 <https://doi.org/10.1016/j.neuropsychologia.2009.07.001>
- 1161 Kriegeskorte, N., Formisano, E., Sorger, B., & Goebel, R. (2007). Individual faces elicit distinct
1162 response patterns in human anterior temporal cortex. *Proceedings of the National*
1163 *Academy of Sciences*, 104(51), 20600–20605.
1164 <https://doi.org/10.1073/pnas.0705654104>
- 1165 Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-
1166 connecting the branches of systems neuroscience. *Frontiers in systems*
1167 *neuroscience*, 2, 249.
- 1168 Kuhn, L. K., Wydell, T., Lavan, N., McGettigan, C., & Garrido, L. (2017). Similar
1169 representations of emotions across faces and voices. *Emotion*, 17(6), 912–937.
1170 <https://doi.org/10.1037/emo0000282>
- 1171 Latinus, M., McAleer, P., Bestelmeyer, P. E. G., & Belin, P. (2013). Norm-Based Coding of
1172 Voice Identity in Human Auditory Cortex. *Current Biology*, 23(12), 1075–1080.
1173 <https://doi.org/10.1016/j.cub.2013.04.055>
- 1174 Lee, S.-H., Kravitz, D. J., & Baker, C. I. (2012). Disentangling visual imagery and perception
1175 of real-world objects. *NeuroImage*, 59(4), 4064–4073.
1176 <https://doi.org/10.1016/j.neuroimage.2011.10.055>
- 1177 Lettieri, G., Handjaras, G., Cappello, E. M., Setti, F., Bottari, D., Bruno, V., Diano, M., Leo, A.,
1178 Tinti, C., Garbarini, F., Pietrini, P., Ricciardi, E., & Cecchetti, L. (2024). Dissecting
1179 abstract, modality-specific and experience-dependent coding of affect in the human
1180 brain. *Science Advances*, 10(10), eadk6840. <https://doi.org/10.1126/sciadv.adk6840>
- 1181 Lewkowicz, D. J., & Ghazanfar, A. A. (2009). The emergence of multisensory systems through
1182 perceptual narrowing. *Trends in Cognitive Sciences*, 13(11), 470–478.
1183 <https://doi.org/10.1016/j.tics.2009.08.004>
- 1184 Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth
1185 of a talking face when learning speech. *Proceedings of the National Academy of*
1186 *Sciences*, 109(5), 1431–1436. <https://doi.org/10.1073/pnas.1114783109>
- 1187 Liang, Y., Liu, B., Li, X., & Wang, P. (2018). Multivariate Pattern Classification of Facial
1188 Expressions Based on Large-Scale Functional Connectivity. *Frontiers in Human*
1189 *Neuroscience*, 12, 94. <https://doi.org/10.3389/fnhum.2018.00094>

1190 Lowe, M. X., Mohsenzadeh, Y., Lahner, B., Charest, I., Oliva, A., & Teng, S. (2021). Cochlea
 1191 to categories: The spatiotemporal dynamics of semantic auditory representations.
 1192 *Cognitive Neuropsychology*, 38(7–8), 468–489.
 1193 <https://doi.org/10.1080/02643294.2022.2085085>

1194 Mattioni, S., Rezk, M., Battal, C., Bottini, R., Cuculiza Mendoza, K. E., Oosterhof, N. N., &
 1195 Collignon, O. (2020). Categorical representation from sound and sight in the ventral
 1196 occipito-temporal cortex of sighted and blind. *eLife*, 9, e50732.
 1197 <https://doi.org/10.7554/eLife.50732>

1198 Mattioni, S., Rezk, M., Battal, C., Vadlamudi, J., & Collignon, O. (2022). Impact of blindness
 1199 onset on the representation of sound categories in occipital and temporal cortices.
 1200 *eLife*, 11, e79370. <https://doi.org/10.7554/eLife.79370>

1201 McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., & Nieto, O. (2015).
 1202 librosa: Audio and music signal analysis in python. *SciPy*, 2015(18–24), 7.

1203 Mende-Siedlecki, P., Verosky, S. C., Turk-Browne, N. B., & Todorov, A. (2013). Robust
 1204 selectivity for faces in the human amygdala in the absence of expressions. *Journal of*
 1205 *cognitive neuroscience*, 25(12), 2086–2106.

1206 Monroe, J. F., Griffin, M., Pinkham, A., Loughhead, J., Gur, R. C., Roberts, T. P. L., &
 1207 Christopher Edgar, J. (2013). The fusiform response to faces: Explicit versus implicit
 1208 processing of emotion. *Human Brain Mapping*, 34(1), 1–11.
 1209 <https://doi.org/10.1002/hbm.21406>

1210 Nasr, S., & Tootell, R. B. (2012). Role of fusiform and anterior temporal cortical areas in facial
 1211 recognition. *Neuroimage*, 63(3), 1743–1753.

1212 O’Craven, K. M., & Kanwisher, N. (2000). Mental Imagery of Faces and Places Activates
 1213 Corresponding Stimulus-Specific Brain Regions. *Journal of Cognitive Neuroscience*,
 1214 12(6), 1013–1023. <https://doi.org/10.1162/08989290051137549>

1215 Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-Modal
 1216 Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. *Frontiers*
 1217 *in Neuroinformatics*, 10. <https://doi.org/10.3389/fninf.2016.00027>

1218 O’Toole, A. J., Roark, D. A., & Abdi, H. (2002). Recognizing moving faces: A psychological
 1219 and neural synthesis. *Trends in Cognitive Sciences*, 6(6), 261–266.
 1220 [https://doi.org/10.1016/S1364-6613\(02\)01908-3](https://doi.org/10.1016/S1364-6613(02)01908-3)

1221 Peelen, M. V., Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal Representations of
 1222 Perceived Emotions in the Human Brain. *The Journal of Neuroscience*, 30(30), 10127–
 1223 10134. <https://doi.org/10.1523/JNEUROSCI.2161-10.2010>

1224 Pernet, C. R., McAleer, P., Latinus, M., Gorgolewski, K. J., Charest, I., Bestelmeyer, P. E. G.,
 1225 Watson, R. H., Fleming, D., Crabbe, F., Valdes-Sosa, M., & Belin, P. (2015). The
 1226 human voice areas: Spatial organization and inter-individual variability in temporal and
 1227 extra-temporal cortices. *NeuroImage*, 119, 164–174.
 1228 <https://doi.org/10.1016/j.neuroimage.2015.06.050>

1229 Pitcher, D., Dilks, D. D., Saxe, R. R., Triantafyllou, C., & Kanwisher, N. (2011). Differential
 1230 selectivity for dynamic versus static information in face-selective cortical regions.
 1231 *NeuroImage*, 56(4), 2356–2363. <https://doi.org/10.1016/j.neuroimage.2011.03.067>

1232 Pitcher, D., Duchaine, B., & Walsh, V. (2014). Combined TMS and fMRI Reveal Dissociable
1233 Cortical Pathways for Dynamic and Static Face Perception. *Current Biology*, 24(17),
1234 2066–2070. <https://doi.org/10.1016/j.cub.2014.07.060>

1235 Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a Third Visual Pathway Specialized for
1236 Social Perception. *Trends in Cognitive Sciences*, 25(2), 100–110.
1237 <https://doi.org/10.1016/j.tics.2020.11.006>

1238 Pitcher, D., Walsh, V., & Duchaine, B. (2011). The role of the occipital face area in the cortical
1239 face perception network. *Experimental Brain Research*, 209(4), 481–493.
1240 <https://doi.org/10.1007/s00221-011-2579-1>

1241 Powell, L. J., Kosakowski, H. L., & Saxe, R. (2018). Social Origins of Cortical Face Areas.
1242 *Trends in Cognitive Sciences*, 22(9), 752–763.
1243 <https://doi.org/10.1016/j.tics.2018.06.009>

1244 Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal Cortex
1245 Activation in Humans Viewing Eye and Mouth Movements. *The Journal of*
1246 *Neuroscience*, 18(6), 2188–2199. [https://doi.org/10.1523/JNEUROSCI.18-06-](https://doi.org/10.1523/JNEUROSCI.18-06-02188.1998)
1247 [02188.1998](https://doi.org/10.1523/JNEUROSCI.18-06-02188.1998)

1248 Romanski, L. M. (2007). Representation and Integration of Auditory and Visual Stimuli in the
1249 Primate Ventral Lateral Prefrontal Cortex. *Cerebral Cortex*, 17(suppl 1), i61–i69.
1250 <https://doi.org/10.1093/cercor/bhm099>

1251 Rossion, B., Hanseeuw, B., & Dricot, L. (2012). Defining face perception areas in the human
1252 brain: A large-scale factorial fMRI face localizer analysis. *Brain and Cognition*, 79(2),
1253 138–157. <https://doi.org/10.1016/j.bandc.2012.01.001>

1254 Rossion, B., Jacques, C., & Jonas, J. (2024). The anterior fusiform gyrus: The ghost in the
1255 cortical face machine. *Neuroscience & Biobehavioral Reviews*, 158, 105535.

1256 Sauter, D. A., & Eimer, M. (2010). Rapid Detection of Emotion from Human Vocalizations.
1257 *Journal of Cognitive Neuroscience*, 22(3), 474–481.
1258 <https://doi.org/10.1162/jocn.2009.21215>

1259 Schall, S., Kiebel, S. J., Maess, B., & von Kriegstein, K. (2013). Early auditory sensory
1260 processing of voices is facilitated by visual mechanisms. *Neuroimage*, 77, 237–245.

1261 Schirmer, A., & Adolphs, R. (2017). Emotion perception from face, voice, and touch:
1262 comparisons and convergence. *Trends in cognitive sciences*, 21(3), 216–228.

1263 Simon, D., Craig, K. D., Gosselin, F., Belin, P., & Rainville, P. (2008). Recognition and
1264 discrimination of prototypical dynamic expressions of pain and emotions. *Pain*, 135(1),
1265 55–64. <https://doi.org/10.1016/j.pain.2007.05.008>

1266 Stelzer, J., Chen, Y., & Turner, R. (2013). Statistical inference and multiple testing correction
1267 in classification-based multi-voxel pattern analysis (MVPA): Random permutations and
1268 cluster size control. *NeuroImage*, 65, 69–82.
1269 <https://doi.org/10.1016/j.neuroimage.2012.09.063>

1270 Sugihara, T., Diltz, M. D., Averbeck, B. B., & Romanski, L. M. (2006). Integration of Auditory
1271 and Visual Communication Information in the Primate Ventrolateral Prefrontal Cortex.
1272 *The Journal of Neuroscience*, 26(43), 11138–11147.
1273 <https://doi.org/10.1523/JNEUROSCI.3550-06.2006>

- 1274 Sunday, M. A., McGugin, R. W., Tamber-Rosenau, B. J., & Gauthier, I. (2018). Visual imagery
1275 of faces and cars in face-selective visual areas. *PLOS ONE*, 13(9), e0205041.
1276 <https://doi.org/10.1371/journal.pone.0205041>
- 1277 Talsma, D. (2015). Predictive coding and multisensory integration: an attentional account of
1278 the multisensory mind. *Frontiers in integrative neuroscience*, 9, 19.
- 1279 Tong, F., & Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annual review of*
1280 *psychology*, 63(1), 483-509.
- 1281 Trautmann-Lengsfeld, S. A., Domínguez-Borràs, J., Escera, C., Herrmann, M., & Fehr, T.
1282 (2013). The Perception of Dynamic and Static Facial Expressions of Happiness and
1283 Disgust Investigated by ERPs and fMRI Constrained Source Analysis. *PLoS ONE*,
1284 8(6), e66997. <https://doi.org/10.1371/journal.pone.0066997>
- 1285 Tsantani, M., Kriegeskorte, N., Storrs, K., Williams, A. L., McGettigan, C., & Garrido, L. (2021).
1286 FFA and OFA encode distinct types of face identity information. *Journal of*
1287 *Neuroscience*, 41(9), 1952-1969.
- 1288 Van Audenhaege, A., Mattioni, S., Cerpelloni, F., Gau, R., Szmalec, A., & Collignon, O. (2025).
1289 Phonological representations of auditory and visual speech in the occipito-temporal
1290 cortex and beyond. *Journal of Neuroscience*, 45(26).
- 1291 Van Den Hurk, J., Van Baelen, M., & Op De Beeck, H. P. (2017). Development of visual
1292 category selectivity in ventral visual cortex does not require visual experience.
1293 *Proceedings of the National Academy of Sciences*, 114(22).
1294 <https://doi.org/10.1073/pnas.1612862114>
- 1295 Visconti di Oleggio Castello, M., Haxby, J. V., & Gobbini, M. I. (2021). Shared neural codes
1296 for visual and semantic information about familiar faces in a common representational
1297 space. *Proceedings of the National Academy of Sciences*, 118(45), e2110474118.
- 1298 Von Der Heide, R. J., Skipper, L. M., & Olson, I. R. (2013). Anterior temporal face patches: a
1299 meta-analysis and empirical study. *Frontiers in human neuroscience*, 7, 17.
- 1300 Von Kriegstein, K., & Giraud, A.-L. (2006). Implicit Multisensory Associations Influence Voice
1301 Recognition. *PLoS Biology*, 4(10), e326. <https://doi.org/10.1371/journal.pbio.0040326>
- 1302 Von Kriegstein, K., Kleinschmidt, A., Sterzer, P., & Giraud, A.-L. (2005). Interaction of Face
1303 and Voice Areas during Speaker Recognition. *Journal of Cognitive Neuroscience*,
1304 17(3), 367–376. <https://doi.org/10.1162/0898929053279577>
- 1305
- 1306 Vroomen, J., Driver, J., & De Gelder, B. (2001). Is cross-modal integration of emotional
1307 expressions independent of attentional resources? *Cognitive, Affective, & Behavioral*
1308 *Neuroscience*, 1(4), 382–387. <https://doi.org/10.3758/CABN.1.4.382>
- 1309 Vuilleumier, P., & Pourtois, G. (2007). Distributed and interactive brain mechanisms during
1310 emotion face perception: Evidence from functional neuroimaging. *Neuropsychologia*,
1311 45(1), 174–194. <https://doi.org/10.1016/j.neuropsychologia.2006.06.003>
- 1312 Walker-Andrews, A. S., & Lennon, E. (1991). Infants' discrimination of vocal expressions:
1313 Contributions of auditory and visual information. *Infant Behavior and Development*,
1314 14(2), 131–142. [https://doi.org/10.1016/0163-6383\(91\)90001-9](https://doi.org/10.1016/0163-6383(91)90001-9)

- Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., & Belin, P. (2014). Crossmodal Adaptation in Right Posterior Superior Temporal Sulcus during Face–Voice Emotional Integration. *The Journal of Neuroscience*, 34(20), 6813–6821. <https://doi.org/10.1523/JNEUROSCI.4478-13.2014>
- Wegrzyn, M., Riehle, M., Labudda, K., Woermann, F., Baumgartner, F., Pollmann, S., Bien, C. G., & Kissler, J. (2015). Investigating the brain basis of facial expression perception using multi-voxel pattern analysis. *Cortex*, 69, 131–140. <https://doi.org/10.1016/j.cortex.2015.05.003>
- Weiner, K. S., & Grill-Spector, K. (2013). Neural representations of faces and limbs neighbor in human high-level visual cortex: Evidence for a new organization principle. *Psychological Research*, 77(1), 74–97. <https://doi.org/10.1007/s00426-011-0392-x>
- Wildgruber, D., Riecker, A., Hertrich, I., Erb, M., Grodd, W., Ethofer, T., & Ackermann, H. (2005). Identification of emotional intonation evaluated by fMRI. *NeuroImage*, 24(4), 1233–1241. <https://doi.org/10.1016/j.neuroimage.2004.10.034>
- Winston, J. S., Henson, R. N. A., Fine-Goulden, M. R., & Dolan, R. J. (2004). fMRI-Adaptation Reveals Dissociable Neural Representations of Identity and Expression in Face Perception. *Journal of Neurophysiology*, 92(3), 1830–1839. <https://doi.org/10.1152/jn.00155.2004>
- Xia, M., Wang, J., & He, Y. (2013). BrainNet Viewer: a network visualization tool for human brain connectomics. *PloS one*, 8(7), e68910.
- Young, A. W., Frühholz, S., & Schweinberger, S. R. (2020). Face and Voice Perception: Understanding Commonalities and Differences. *Trends in Cognitive Sciences*, 24(5), 398–410. <https://doi.org/10.1016/j.tics.2020.02.001>
- Zhu, L. L., & Beauchamp, M. S. (2017). Mouth and Voice: A Relationship between Visual and Auditory Preference in the Human Superior Temporal Sulcus. *The Journal of Neuroscience*, 37(10), 2697–2708. <https://doi.org/10.1523/JNEUROSCI.2914-16.2017>

Figures' captions

Figure 1. A) Experimental design implemented in the fMRI for the face (in blue) and voice (in red) localizers. **B)** Results from the localizers, representing each individual peak for the contrasts [faces>objects] in dark to light blue and for [voices>objects] in dark to light red (see all individual peaks in Table S1 and Table S2). **C)** Example of the 12 ROIs built around one subject's individual peaks.

Figure 2. A) Experimental design implemented in the fMRI for the main experiment. One example of a trial sequence from each condition (only visual in blue, only auditory in red and audio-visual in purple. **B)** Univariate responses to facial, vocal and bimodal expressions in all defined ROIs. Beta weights extracted in individually defined ROIs from independent face and voice localizers. Regions whose name is reported in blue on the x-axis are defined through the face localizer, in red through the voice localizer. The beta weights are extracted in all regions from the visual data (in blue), the auditory data (in red), and the bimodal data (in purple). Beta estimates are represented here averaged across emotion expressions. Differences from zero are evaluated through a one sample t-test, p-values are Bonferroni corrected for 12 ROIs. Differences between modalities within each ROI were assessed using a generalized linear mixed model (GLMM), including modality and emotion as fixed effects and subject as a random effect.

Figure 3: Decoding results. A) Within-modality classification accuracies. Regions whose name is reported in blue on the x-axis are defined through the face localizer, in red through the voice localizer. The decoding is performed in all regions with visual data (in blue), auditory data (in red), and bimodal data (in purple). Each dot represents one participant and the average of the decoding accuracies of the ten possible emotion pairs. Error bars are standard deviations. Significance against chance level is assessed with permutations ($n = 100$) + bootstrap ($n = 10k$), with p-values FDR corrected for 12 ROIs. **B)** Crossmodal decoding accuracies in the ten ROIs where within-modality MVPA is above chance for both vision and audition. Significance assessed with permutations ($n = 100$) + bootstrap ($n = 10k$), p-values FDR corrected for 10 ROIs. **C)** Decoding accuracies of the MVP-classification of within-emotion pairs from the data of the dominant modality and the multimodal data. In the regions identified through the face localizer (blue labels), same-emotion pairs are classified between the visual and bimodal data; in the regions identified through the voice localizer (red labels), same-emotional pairs are classified between the auditory and bimodal data. Each dot represents one subject and it is the average of the decoding accuracies of the five within-emotion pairs. Error bars are standard deviations. Significance against chance level is assessed with permutations ($n = 100$) + bootstrap ($n = 10k$), with p-values FDR corrected for 12 ROIs.

Figure 4: Representational similarity analysis (RSA) results. Partial Spearman correlations between the neural representational dissimilarity matrices (RDMs) and the emotion-category model after regressing out low- and mid-level stimulus models. For visual data (blue), the visual models (pixel motion, silhouette, and mouth configuration) were regressed out; for auditory data (red), the auditory models (amplitude envelope, MFCCs, and spectral centroid) were regressed out. Each dot represents one participants. Group mean (horizontal bars) $\pm$ SEM are also reported. Asterisks denote significant correlations after FDR correction for multiple comparisons across ROIs ($p < 0.05$, $p < 0.01$, $p < 0.001$).

Figure 5: Brain data visualization within face- and voice-preferential regions. Representational dissimilarity matrices (DSM), multidimensional scaling (MDS) plots, and dendrograms illustrating the neural representational structure of emotions across modalities within face-preferential areas (top panels) and voice-preferential areas (bottom panels). For each region of interest (ROI), the DSMs display the pairwise dissimilarity ($1 - \text{Spearman's } r$) between the neural patterns associated with the five emotion categories (N = neutral, D = disgust, F = fear, H = happiness, S = sadness) across the three modalities (visual, auditory, and bimodal). In each panel, the MDS plots (3D insets) visualize the relative distances between emotion representations, where items placed closer together correspond to more similar neural patterns. Dendrograms summarize the hierarchical clustering of emotion–modality representations, with color-coded branches indicating clusters that share $\geq 70\%$ similarity. Blue, red, and purple labels correspond to visual, auditory, and bimodal modalities, respectively.

Figure 6: Whole-brain searchlight decoding analyses for emotion representation at a stringent threshold of $p < 0.05$ FWE corrected. Statistical maps show regions with significant ($p < 0.05$ FWE) decoding accuracy for emotional categories across four analyses: (A) within visual modality (blue/green map), (B) within auditory modality (red/yellow map), (C) within bimodal modality (dark/light purple map), and (D) the overlap between the three. Each voxel represents the center of a 100-voxel searchlight sphere in which multiclass decoding was performed using a linear SVM classifier (LIBSVM; Chang & Lin, 2011) implemented in CoSMoMVPA (Oosterhof et al., 2016). Maps display group-level t-values thresholded at $p < 0.05$ FWE corrected. Results in panels A, B, and C are projected on the inflated cortical surface for visualization using BrainNet Viewer (Xia et al., 2013). Color bars indicate t-value ranges for each decoding type.

Figure 7: ROI-constrained searchlight decoding analyses within supramodal regions. Searchlight decoding results within the group masks of the temporal voice area (TVA), posterior superior temporal sulcus (pSTS), and precentral gyrus (PCG) for the four decoding analyses: within auditory modality (red), within visual modality (blue), within bimodal modality (purple), and between (crossmodal) audio–visual modalities (pink). Each map shows regions displaying significant decoding accuracy for emotional categories ($p < 0.001$ uncorrected). Reported coordinates next to each mask indicate the group decoding peaks in MNI space for each region. Orange circles with black outlines mark the anterior, middle, and posterior TVA peaks reported by Pernet et al. (2015), shown for reference to illustrate the correspondence between our crossmodal peaks and previously described TVA subregions. The orange-outlined boxes summarize the MNI coordinates of the aTVA, mTVA, and pTVA peaks from Pernet et al. (2015). Boxes at the bottom left and right of the figure summarize the group peak MNI y-coordinates for each decoding analysis, specifically along the anterior–posterior axis. Together, these results reveal a posterior-to-anterior gradient within the TVA and pSTS—posterior regions showing stronger visual decoding, anterior regions stronger auditory decoding, and highest crossmodal decoding emerging at their intersection—whereas the PCG exhibited overlapping peaks for all decoding types, consistent with a more integrated, supramodal representation.

1435 **Tables' captions**

1436

1437 **Table1: (left)** P-values of non parametric permutation test for Emotional category, regressed with low-
1438 level models (Mouth, Pixel Motion, Silhouette) - Visual condition. **(right)** P-values of non parametric
1439 permutation test for Emotional category, regressed with low-level models (ENV, Spectral Centroid,
1440 MFCCs) - Auditory condition. Asterisks next to the percentage decoding indicates a $p < 0.05$ after FDR-
1441 correction.

1442

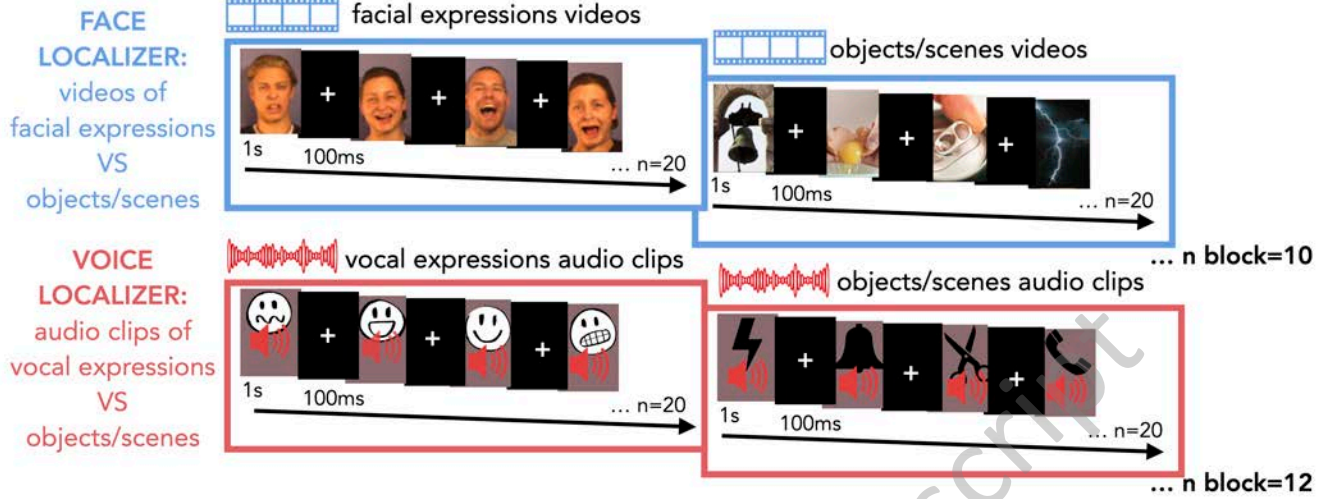
JNeurosci Accepted Manuscript

VISUAL DATA					AUDITORY DATA				
ROI (visual data)	Mean Spearman's rho	UncP	pFDR	Cohen's D	ROI (auditory data)	Mean Spearman's rho	UncP	pFDR	Cohen's D
leftFFA	0.0461*	0.0170	0.0408	2.0963	leftFFA	0.0037	0.4306	0.4697	0.1702
rightFFA	0.0556*	0.0048	0.0165	2.5416	rightFFA	0.0267	0.1110	0.1480	1.2286
leftOFA	0.0560*	0.0055	0.0165	2.5755	leftOFA	0.0045	0.4185	0.4697	0.2080
rightOFA	0.0373	0.0439	0.0658	1.7006	rightOFA	-0.0040	0.5707	0.5707	-0.1724
leftPCGv	0.0558*	0.0050	0.0165	2.5615	leftPCGv	0.0491*	0.0129	0.0221	2.2640
rightPCGv	0.0091	0.3364	0.3364	0.4198	rightPCGv	0.0520*	0.0075	0.0150	2.4155
leftpSTS	0.0858*	0.0003	0.0036	4.0377	leftpSTS	0.0632*	0.0018	0.0043	3.0250
rightpSTS	0.0425*	0.0272	0.0544	1.9213	rightpSTS	0.0388*	0.0343	0.0514	1.8106
leftPCGa	0.0416	0.0360	0.0617	1.7897	leftPCGa	0.0699*	0.0013	0.0039	3.1280
rightPCGa	0.0309	0.0974	0.1062	1.2978	rightPCGa	0.0868*	0.0001	0.0004	3.9223
leftTVA	0.0340	0.0533	0.0711	1.6184	leftTVA	0.1511*	0.0001	0.0004	7.6190
rightTVA	0.0297	0.0853	0.1023	1.3917	rightTVA	0.1510*	0.0001	0.0004	7.4947

1443

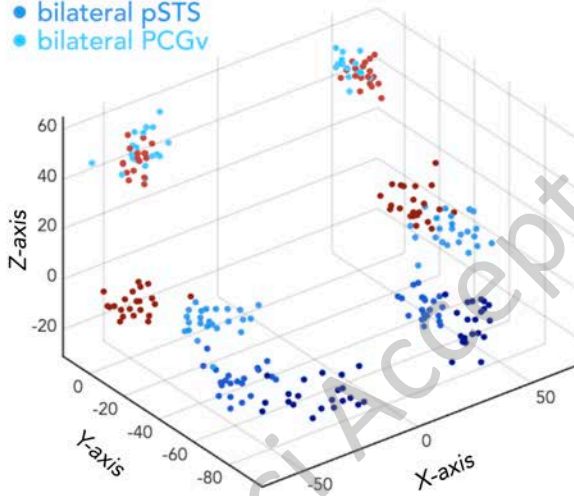
1444

A. Localizers fMRI design

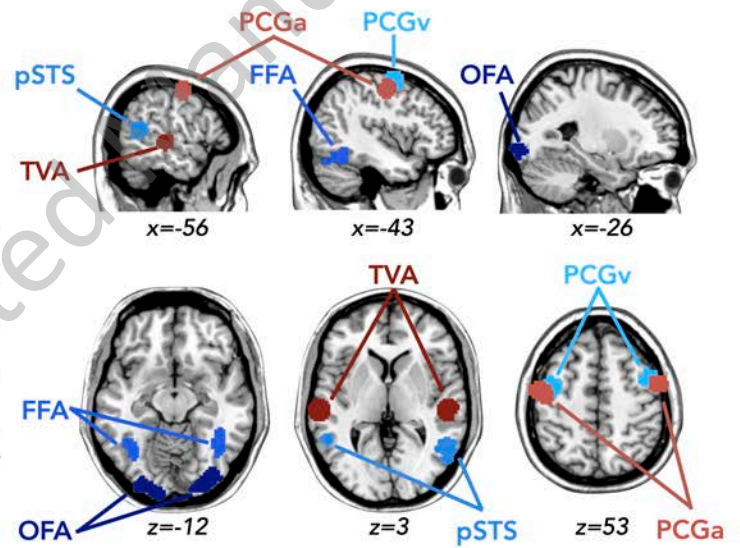


B. Individual ROIs' peaks from localizers

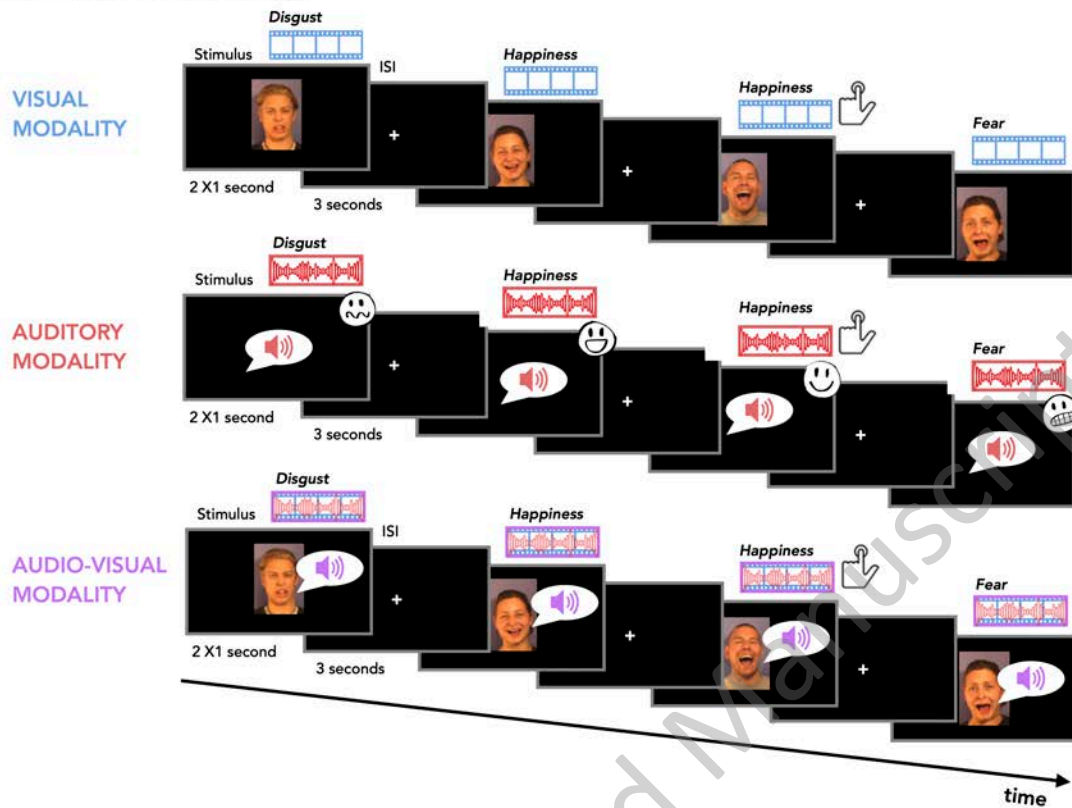
- FACE AREAS**
- bilateral OFA
 - bilateral FFA
 - bilateral pSTS
 - bilateral PCGv
- VOICE AREAS**
- bilateral TVA
 - bilateral PCGa



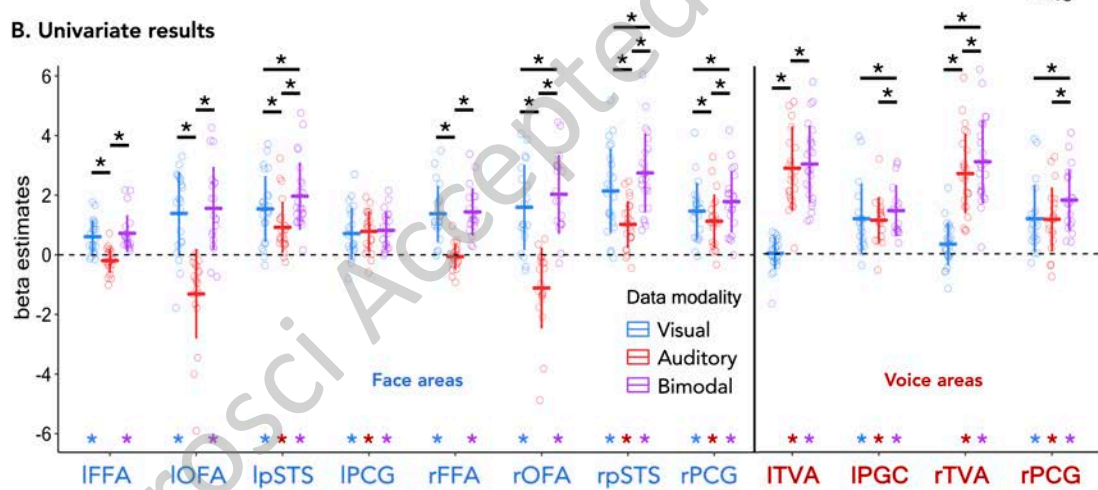
C. ROIs from localizers: example from one subject



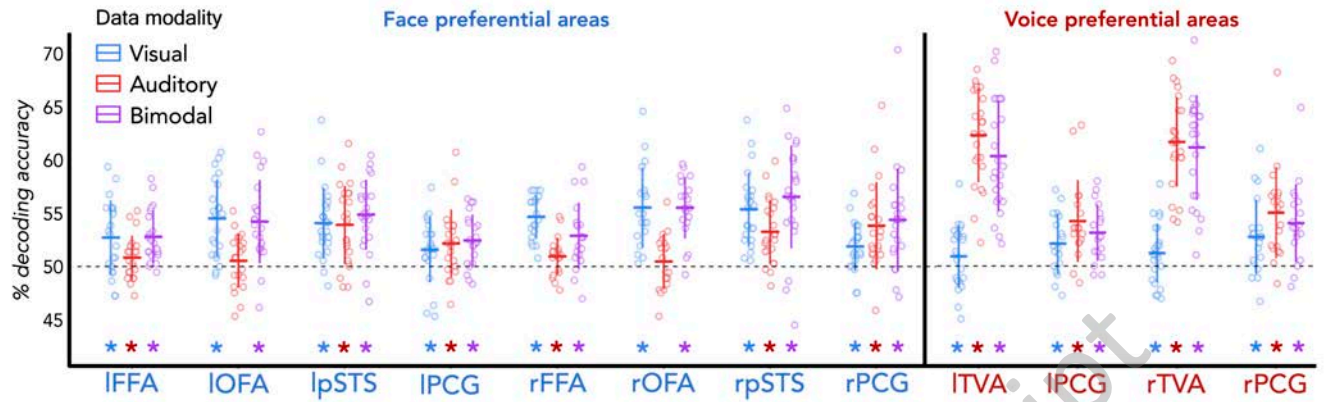
A. Main Experiment fMRI design



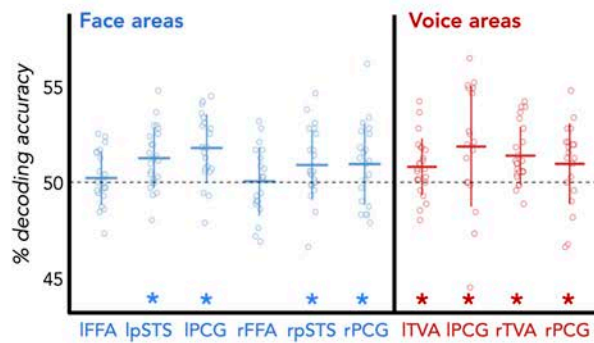
B. Univariate results



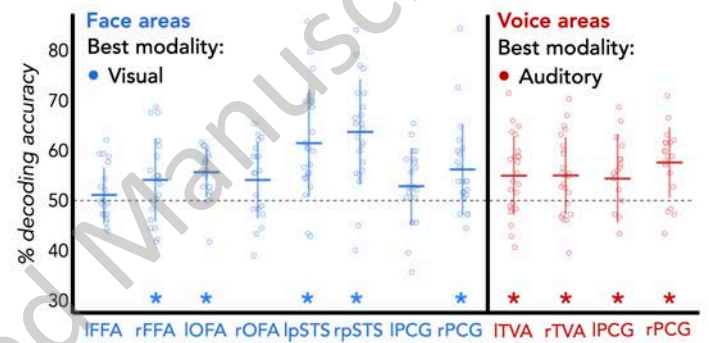
A. Within modality decoding

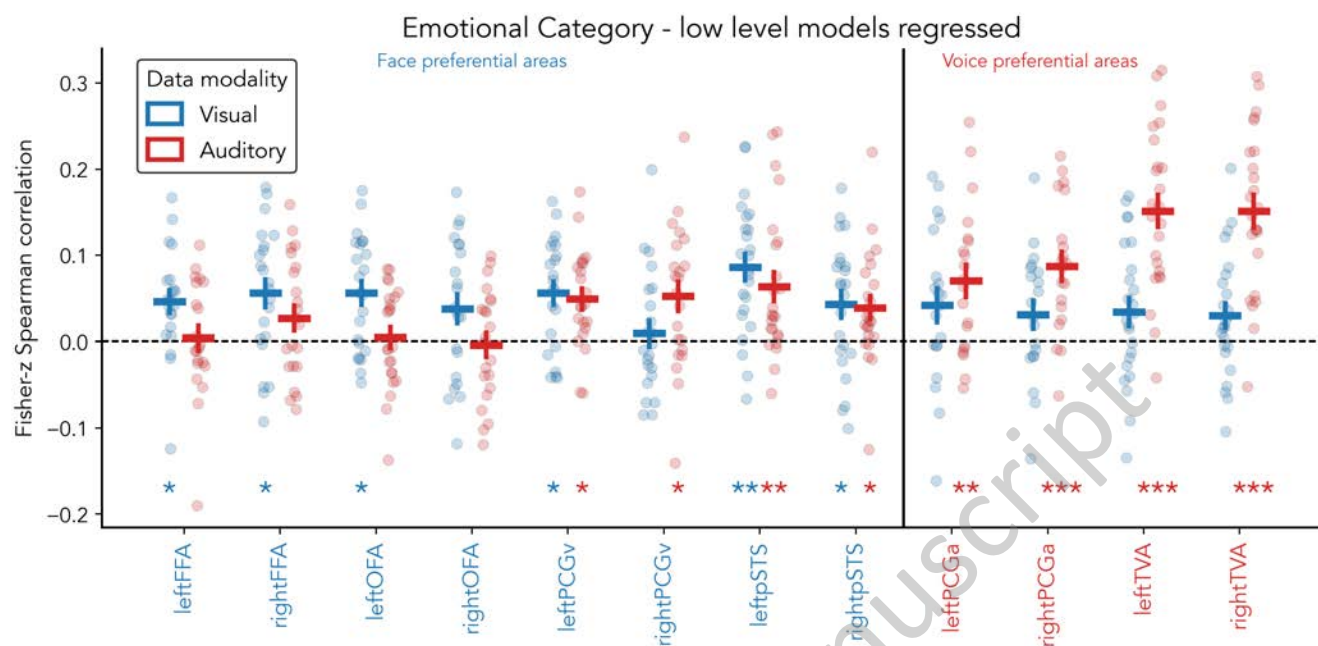


B. Cross-modal (audio-visual) decoding

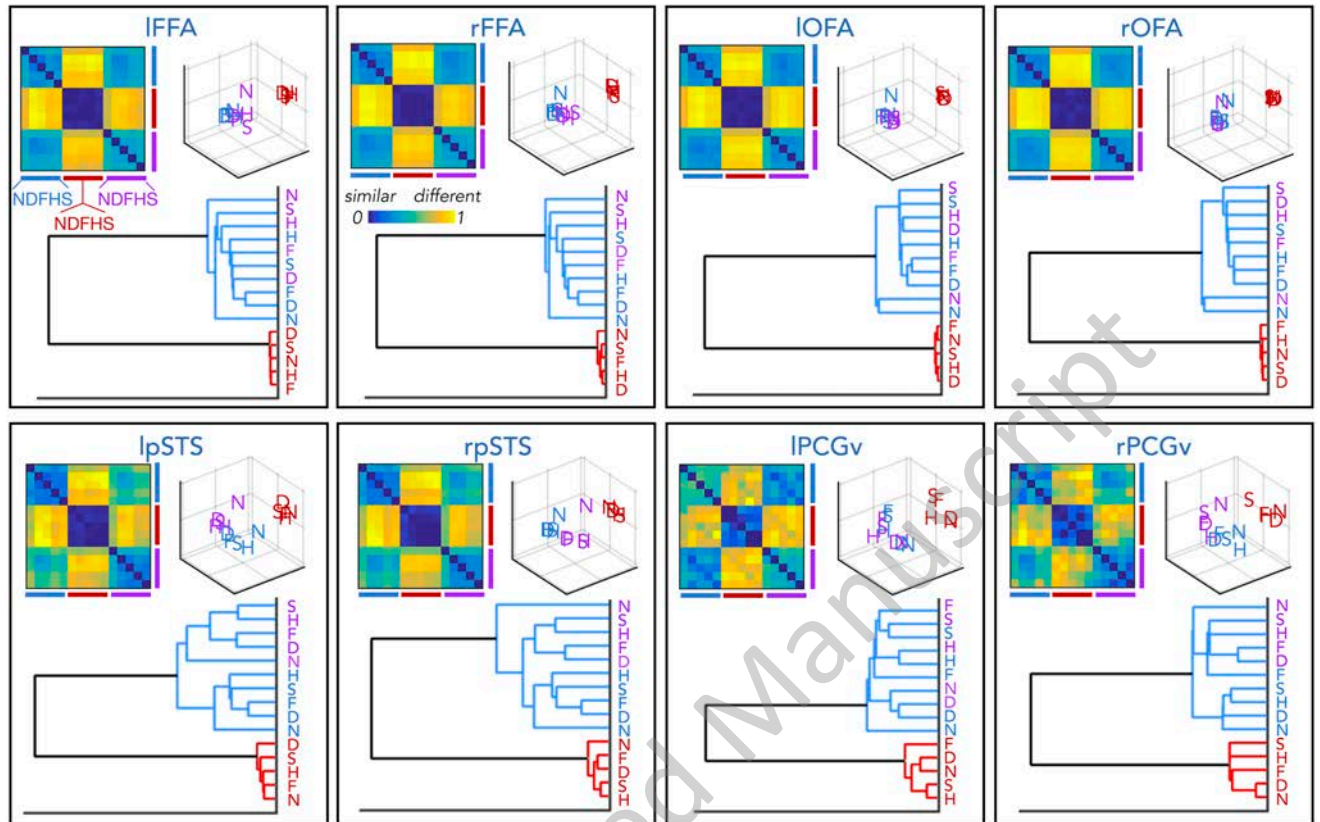


C. Decoding between bimodal & dominant modality

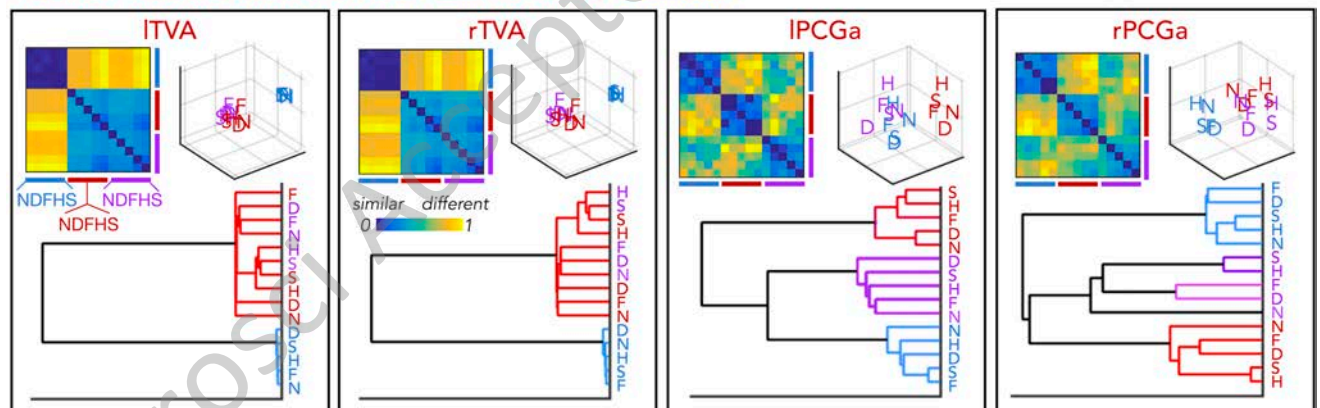




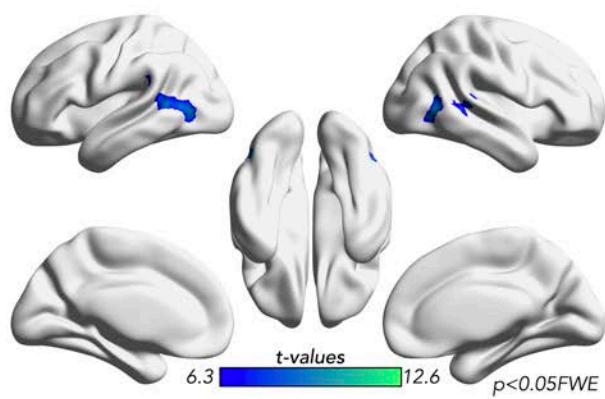
Face Preferential Areas



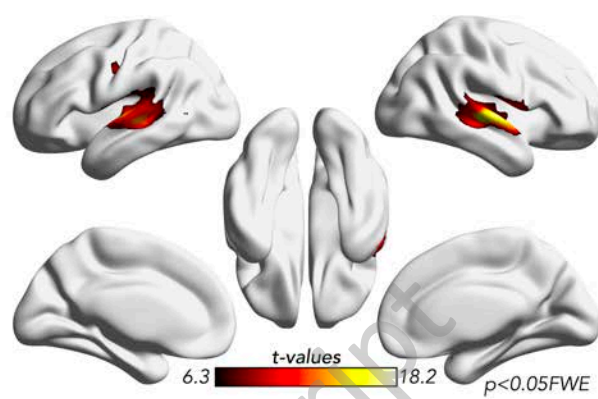
Voice Preferential Areas



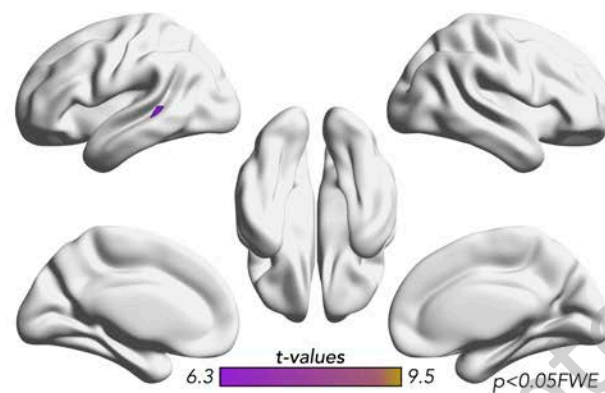
A. Within Visual Decoding



B. Within Auditory Decoding



C. Between Audio-Visual Decoding



D. Overlap W/Visual, W/Audio & B/Audio-Vis

