Published in final edited form as:
Nat Hum Behav. 2021 September 01; 5(9): 1203–1213. doi:10.1038/s41562-021-01073-0.

The Representational Dynamics of Perceived Voice Emotions Evolve from Categories to Dimensions

Bruno L. Giordano[1,2,*], **Caroline Whiting**[2], **Nikolaus Kriegeskorte**[3], **Sonja A. Kotz**[4,5], **Joachim Gross**[2,6,*,†], **Pascal Belin**[1,7,*,†]

[1]Institute of Neuroscience of la Timone UMR 7289 Centre National de la Recherche Scientifique and Aix-Marseille University, Marseille, France

[2]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK

[3]Zuckerman Mind Brain Behavior Institute, Columbia University, New York, NY, USA

[4]Faculty of Psychology and Neuroscience, Department of Neuropsychology and Psychopharmacology, Maastricht University, Maastricht, The Netherlands

[5]Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

[6]Institute for Biomagnetism and Biosignalanalysis, University of Münster, Germany

[7]Department of Psychology, University of Montréal, Montréal, Canada

Abstract

Long-standing affective science theories conceive the perception of emotional stimuli either as discrete categories (e.g., an angry voice) or continuous dimensional attributes (e.g., an intense and negative vocal emotion). Which position provides a better account is still widely debated. Here, we contrast them to account for acoustics-independent perceptual and cerebral representational geometry of perceived voice emotions. We combined multimodal imaging of the cerebral response to heard vocal stimuli (functional magnetic resonance imaging – fMRI – and magneto-encephalography – MEG) with post-scanning behavioral assessment of voice emotion perception. By using representational similarity analysis (RSA), we find that categories prevail in perceptual and early (< 200ms) fronto-temporal cerebral representational geometries and that dimensions impinge predominantly on a later limbic-temporal network (240ms and > 500ms). These results reconcile the two long-opposing views by reframing the perception of emotions as

[*]Corresponding authors: Bruno L. Giordano (bruno.giordano@univ-amu.fr); Joachim Gross (joachim.gross@wwu.de); Pascal Belin (pascal.belin@univ-amu.fr).
[†]Joint senior authors

Author contributions

Conceptualization: BLG, PB; Methodology: BLG, CW, NK, SAK, PB, JG; Software: BLG; Validation: BLG; Formal Analysis: BLG, CW, JG; Investigation: BLG, CW; Resources: BLG, PB; Data Curation: BLG, CW; Writing - Original Draft: BLG, CW, SAK, PB, JG; Writing – Review & Editing: BLG, CW, NK, SAK, PB, JG; Visualization: BLG; Supervision: BLG, PB, JG; Project Administration: JG; Funding Acquisition: BLG, SAK, PB, JG.

Competing interests

The authors declare no competing interests.

the interplay of cerebral networks with different representational dynamics that emphasize either categories or dimensions.

A persistent and controversial debate in affective sciences is whether emotions are better conceptualized as discrete categories or continuous dimensions. Categorical theories argue that emotion is best described by a small number of discrete basic categories such as fear, happiness, or anger [1,2]. Dimensional theories instead postulate the underpinning of continuous dimensions such as valence (reflecting the degree of pleasantness ranging from negative to positive) or arousal (reflecting the degree of intensity ranging from calm to excited) [3,4].

Despite decades of continuous effort, this question is unresolved [5]. Neuroimaging research on the cerebral bases of emotion, either felt or perceived in others, has not settled this debate as meta-analyses of large bodies of neuroimaging data support both the notion of categories [6] and that of large-scale networks representing dimensional attributes [7,8]. Multivoxel pattern analyses of functional magnetic resonance imaging (fMRI) data have identified distributed patterns of cerebral activity that allow machine learning algorithms to classify emotions into discrete categories, but also to provide estimates of valence and arousal indicating the relevance of both dimensions and categories in explaining brain response to affective stimuli [9–13]. Recent studies using time-sensitive methods such as electro- or magneto-encephalography (MEG) instead suggest complex dynamics in both categorical and dimensional accounts of cerebral activity, although with little neuroanatomical detail [14,15]. Thus, a comprehensive understanding of the precise spatio-temporal dynamics of the cerebral networks involved in representing emotional dimensions and categories is still missing but appears crucial to reconciling the two accounts.

We address this issue by combining comprehensive behavioral assessments of affective vocal bursts (Fig. 1) with brain-activity measurements of the same participant at high spatial and temporal resolution (Fig. 2). Healthy adult participants (n=10) were scanned in 8 alternating fMRI and MEG (4 fMRI and 4 MEG sessions) while they listened to affective bursts (Fig. 1) and performed a 1-back task (they had to press a button when they detected that a stimulus had been played twice in succession): this allowed to maintain and monitor attention to the auditory stimuli while avoiding directing attention to a particular stimulus dimension (e.g., emotional). After scanning, participants rated the perceived dissimilarity of all pairs of stimuli, again in the absence of instructions that would bias their judgment toward a specific stimulus feature. In the final session they then explicitly evaluated emotional attributes of the stimuli by rating their perceived valence, arousal, and intensity along four emotional categories and performed a 4-alternative forced-choice categorization for Anger, Fear, Disgust, or Pleasure.

Emotional stimuli consisted of brief (796ms) synthetic vocalizations generated by morphing between recordings from a validated database of affective bursts [16] that portrayed angry, fearful, disgusted, pleased, and neutral expressions of the vowel /a/. Morphing combined pairs of expressions with weights varying in 25% steps from 0 to 100% as well as each emotional expression with the neutral expression weighted between 100% (neutral) to 0 (original emotion) to -25% (emotional caricature). This resulted in 39 stimuli that sampled densely the space of perceived emotions (Fig. 1a). Morphing was performed independently on recordings from two different actors (one male, one female) to dissociate general emotional from identity- and gender-related acoustics, resulting in 78 stimuli overall. Morphing weights of the original expressions reliably modulated perceived emotion: Fig. 1b illustrates the clear variation in perceived emotional category by morphing weights (note the large inter-individual variability for 50%-50% morphs) and Fig. 1c shows that each perceived emotional attribute was selectively modulated by at least two morphing weights.

We used representational similarity analysis (RSA) [17] to relate the emotion categories and dimensions perceived by each participant to their own multivariate cerebral responses to the affective bursts (Fig. 2). With RSA, we examine the representational geometry of the cerebral response to emotional voices as a window onto the cerebral representation of vocal emotion.

This is performed by abstracting across participants over the specific structure of the cerebral response pattern that carries information about perceived emotions (e.g., potential lower latency evoked response for fear stimuli). By mapping each brain with its own perceptions, we simultaneously account for and generalize across idiosyncrasies in perception by identifying regions and response latencies that specify emotions reliably regardless of how much individual perceptions might deviate from the group average. The large amount of multimodal imaging data for each individual (8 sessions each) was crucial to adjudicate between competing but largely overlapping emotion models of their own perception (Fig. 3d) with robust analyses. As low-level acoustical properties could plausibly influence both perceptual and cerebral responses to the stimuli, we also considered their acoustic structure as reflected by their spectro-temporal modulations. For this we analyzed stimuli using banks of spectro-temporal modulation filters that can intuitively be described as the auditory analogue of Gabor-like representations in the early visual system [18]. We built acoustic 39x39 representational dissimilarity matrices (RDM) capturing overall acoustic differences for each post-onset time window (Fig. 1d) and used them in subsequent analyses to remove acoustic confounds.

Results

Modulation of perceived emotions through voice morphing

We initially sought to assess the presence of effects of voice morphing on perceived emotions, and potential asymmetries between the modulatory effect of morphing for the perception of categories and dimensions. We assessed the selective modulatory effect of each of the five morphing weights (one for each of the four original emotions plus one for the emotionally neutral vocalization) on each of the emotion rating scales (four category-intensity rating scales, and valence and arousal rating scales) via semi-partial correlations (s.p.r). Data from each of the emotion rating tasks were selectively modulated by at least two of the morphing weights ($p \leq 0.05$ FWE-corrected across rating tasks and morphing weights (absolute significant $T(9) \geq 4.343$, permutation-based two-tailed $p \leq 0.039$ corrected for multiple comparisons across pairs of rating scales with morphing weights, absolute Fisher Z scale chance-corrected s.p.r ≥ 0.042, SEM ≤ 0.077, widest percentile bootstrap 95% CI = -0.522/-0.394; absolute non-significant $T(9) \leq 3.801$, $p \geq 0.078$, absolute s.p.r

≤ 0.123, SEM ≥ 0.016, narrowest 95% CI = -0.010/0.022; N permutation and bootstrap samples = 100,000, see Supplementary Table 1 for full results), showing that our morphing manipulation reliably modulated perceived emotions (Fig. 1c). We then sought to assess whether the morphing modulated more strongly perceived categories or dimensions. To this purpose, we tested for significant differences between the speaker- and rating scale averaged variance explained by the five morphing weights together for the categories and dimensions tasks. We observed a descriptively smaller effect of morphing on perceived categories than dimensions that however failed to reach significance (T(9) for categories minus dimensions RSQ contrast = -0.125, two-tailed p = 0.906, Fisher Z scale RSQ categories – dimensions difference =-0.007, SEM = 0.053, bootstrap 95% CI = -0.088/0.104). Accordingly, we found no credible support for a potential asymmetry between the strength of the perceptual modulation of categories and dimensions via our morphing manipulation.

Perceptual representational geometry of voice emotions

To address our main question about emotional dimensions and categories, we first asked whether the perceived structure of the stimulus set, as captured by ratings of perceived voice dissimilarity, resembled the emotion ratings provided in the final session, while accounting for acoustic confounds. For this, we built different behavioral RDMs and investigated their inter-relation (T(9) for acoustics-independent pairwise correlation r between categories and valence RDMs = 10.57, p < 0.001, Fisher Z r = 0.367, SEM = 0.035, percentile bootstrap 95% CI = 0.318/0.440; T(9) for categories and arousal RDMs = 3.18, p = 0.016, Fisher Z r = 0.154, SEM = 0.048, bootstrap 95% CI = 0.069/0.246; T(9) for valence and arousal RDMs = 4.63, p = 0.002, Fisher Z r = 0.334, SEM = 0.072, bootstrap 95% CI = 0.209/0.476, N permutation samples and bootstrap samples for CI computation = 100,000). Visual inspection of Fig. 3a, 3b, and 3c suggests a strong resemblance between the dissimilarity RDM and the categories RDMs, which we confirmed by examining the variance in dissimilarity ratings selectively explained by categories or dimensions after accounting for acoustics (semi-partial correlation – s.p.r – tests and unique explained variance contrasts; one- and two-tailed cluster and permutation-based inference for s.p.r and unique explained variance, respectively; p for s.p.r tests FWE corrected across the three emotion RDMs). As shown in Fig 3d, perceived stimulus dissimilarity represented selectively both categories and dimensions (T(9) for categories s.p.r. = 14.96, p < 0.001, Fisher Z s.p.r = 0.373, SEM = 0.025, percentile bootstrap 95% CI = 0.327/0.420; T(9) for arousal s.p.r = 6.34, p = < 0.001, Fisher Z s.p.r = 0.089, SEM = 0.014, bootstrap 95% CI = 0.060/0.110; T(9) for valence s.p.r = 2.76, p = 0.063, Fisher Z s.p.r = 0.046, SEM = 0.017, bootstrap 95% CI = 0.060/0.077), but was more strongly modulated by categories than by both dimensions together (T(9) for the categories minus dimensions contrast = 14.057, p < 0.001, contrast of Fisher Z unique variances = 0.251, SEM = 0.018, bootstrap 95% CI = 0.066/0.286; cf. Supplementary Fig. 1-2 for cerebral representational geometry of perceived dissimilarity). Thus, behavioral data indicate that both categories and dimensions influence the perception of the emotional voice stimuli, but that categories have a stronger influence, even after removing acoustic confounds (see below and Fig. 4 for perceptual and cerebral representational geometry of acoustics).

Cerebral representational geometry of voice emotions

Next, we asked where and when in the brain was neural activity associated with the categorical or dimensional models. We first built fMRI RDMs at each cerebral location (voxel) reflecting the pairwise stimulus-evoked blood oxygenation level signal differences measured via fMRI within a local sphere centered on that voxel. Each fMRI RDM was tested for a significant correlation with the categorical, arousal or valence RDMs derived from the behavioral ratings from the same participant (Supplementary Fig. 2). We used these fMRI correlations maps to constrain spatially subsequent MEG analyses and built time-varying MEG RDMs of pairwise signal differences projected in the brain only at locations that yielded significant fMRI-emotion RDM correlations (123ms before stimulus onset to 1037ms after stimulus onset). This procedure developed a spatio-temporal analysis of emotion representation that took advantage, simultaneously, of the strengths of each imaging modalities in terms of spatial (fMRI) and temporal (MEG) resolution. As for the behavioral data, we examined the representation of MEG RDM variance unique to each perceived emotion via semi-partial correlation tests that also removed contributions of acoustic structure (Fig. 5; Supplementary Table 2 for peak coordinates and statistics; Supplementary Fig. 3 for pairwise emotion RDM contrasts; Supplementary Fig. 4 for individual results at peak coordinates).

As shown in Fig. 5a, selective representations of the categories model in the cerebral representational geometry (cluster and permutation-based $p \leq 0.05$, FWE corrected across voxels and latencies, N permutations = 10,000; one-tailed inference to ascertain an expected increase in the dissimilarity of cerebral response patterns for increasingly diverse voice emotions) occurred as early as 117ms after stimulus onset in bilateral perisylvian areas, peaking in left supramarginal gyrus and then in left inferior frontal gyrus pars orbitalis 40ms later ($T(9) \geq 4.139$, Fisher Z s.p.r ≥ 0.051, SEM ≤ 0.012, widest percentile bootstrap 95% CI = 0.028/0.076, N bootstrap samples for CI computation = 100,000). A second period of significant category representation occurred at around 400ms post-onset in a bilateral auditory network peaking in right anterior superior temporal gyrus ($T(9) = 7.870$, Fisher Z s.p.r = 0.049, SEM = 0.006, bootstrap 95% CI = 0.037/0.060). In contrast, selective representations of dimensions in the cerebral representational geometry occurred later and in a strongly right-lateralized network (Fig. 5b; one-tailed cluster and permutation-based $p \leq 0.05$, FWE corrected across voxels and latencies, N permutations = 10,000). They occurred first briefly at 237ms post-onset in right temporal cortex (arousal; $T(9) = 4.468$, Fisher Z s.p.r = 0.078, SEM = 0.017, bootstrap 95% CI = 0.048-0.113), then dominated the later post-stimulus period with selective association peaks in the right amygdala at 557ms (arousal; $T(9) = 3.561$, Fisher Z s.p.r = 0.043, SEM = 0.012, bootstrap 95% CI = 0.021-0.066), then in right precentral gyrus at 717ms (arousal; $T(9) = 8.102$, Fisher Z s.p.r = 0.228, SEM = 0.028, bootstrap 95% CI = 0.179-0.283) and the right insula at 757ms (valence; $T(9) = 5.493$, Fisher Z s.p.r = 0.080, SEM = 0.015, bootstrap 95% CI = 0.050-0.105). Differences between the representational dynamics of the categories and dimensions models were confirmed by directly contrasting their unique explained variance while accounting for the confounding effects of acoustics (two-tailed cluster and permutation-based $p \leq 0.05$, FWE corrected across voxels and latencies, N permutations = 10,000). Figure 5c shows that categories initially dominated over dimensions at around 157ms post-onset in right mid-STG (T(9) = 6.801, Fisher Z unique variance contrast = 0.080, SEM = 0.012, bootstrap 95% CI = 0.062-0.106). Nearly the same area of right temporal cortex showed the opposite pattern 560ms later, with significantly greater uniquely explained variance for dimensions than categories at 717ms (T(9) = 6.450, Fisher Z unique variance contrast = 0.050, SEM = 0.008, bootstrap 95% CI = 0.037-0.066).

Perceptual and cerebral representational geometry specifying acoustics

We finally verified the role of acoustics in the perceived dissimilarity and emotions of the voice stimuli, and in their cerebral representational geometry (Fig. 4; Supplementary Fig. 4 for individual results at peak coordinates). We observed a reliable association between perceived dissimilarity and acoustics throughout the entire sound duration peaking during the first 200ms of the heard voice (Fig. 4a, left panel). An analysis of the selective acoustical specification of perceived emotions revealed that acoustics include strong unique information about categories throughout the entire length of the sound signal, and of emotion dimensions at around sound onset (before 250ms from sound onset) or close to the sound offset (starting from 750ms). Critically, a direct contrast of the unique acoustics variance about categories and dimensions in acoustics reveals predominant categorical information at around 300ms and 650ms after sound onset, and of dimensions information at around 750ms. Significant representations of acoustics in the cerebral representational geometry finally emerged in the bilateral temporal cortex starting at 117ms from sound onset, and covering the initial 200ms of the evoked response (Fig. 4b; one-tailed cluster and permutation-based $p \leq 0.05$, FWE corrected across voxels and latencies, N permutations = 10,000; Peaks $T(9) \geq 8.134$, Fisher Z $r \geq 0.090$, $SEM \leq 0.010$, widest bootstrap 95% CI = 0.082-0.126). These results show that early acoustical structure was, as largely expected, reflected in the perceived dissimilarity and cerebral representational geometry of the heard voices, and that strong acoustical information about categories and dimensions emerged at mid- and very late sound latencies for categories and dimensions, respectively. Considering that the evoked cerebral response lags relative to sound onset, none of these results can easily account for the transition from an early cerebral dominance of categories (157ms) to a late dominance of dimensions (717ms) be it for an implicit attentional bias to focus on the acoustical structure that best differentiates the two (starting at 300ms for categories and at 750ms for dimensions).

Discussion

Converging evidence from three modalities (behavior, fMRI, and MEG) contrasting directly the categorical and dimensional models shows that both explain the perceptual and cerebral representational geometry of emotions in the voice—but with markedly different spatio-temporal dynamics. Our results indicate progressive refinement of emotional stimulus representations from the formation of emotional categories well suited to trigger fast adaptive reactions to increasingly fined-grained representations modulated by valence and arousal. Selective representation of categories impinged on the early cerebral representational geometries in a bilateral temporal network extending to the left inferior prefrontal cortex, the latter potentially reflecting early activation of verbal categorical labels [19]. The representation of dimensions instead relied on late cerebral representational geometries in a right-lateralized temporal network extending to two regions, the insula and amygdala, part of a “salience” network [4] that links the processing of emotional states and events across species [20,21] and thought to represent a phylogenetic precursor for communicative behavior in primates and humans [10].

The representational dynamics observed for the right auditory cortex suggests a transition from an early dominance of feed-forward sensory processing to late attentional modulations potentially resulting from feedback signals transmitted through lateral and medial cortical connections from the amygdala. Fig. 5c (right panel) shows the 2D representations of the stimulus set obtained by interindividual difference scaling of the Cerebral RDMs in the right temporal cortex at the two different time points, and exemplifies how in the same right-temporal area representational geometries related to perceived categories evolve in time to subsequently emphasize dimensions instead.

Our study aimed at disentangling the perceptual and cerebral representational geometry of categories and dimensions in perceived voice emotions, but left unanswered several factors and important theoretical distinctions considered in the affective science literature. Our results are consistent with those of previous neuroimaging studies of emotions recognized in linguistic materials [14,15] and of behavioral studies of felt emotions while watching video clips or hearing prosodically rich speech [22,23]. However, future studies will be required to disentangle the influence of emotion-evoking materials and of the distinction between perceived and felt emotions on the spatio-temporal dynamics of emotion representation. Future studies will also be required to ascertain the role of context on perceived voice emotions, factor central to appraisal theories of emotion [11]. Finally, while factoring out the role of acoustics from perceived and cerebral representational geometries, we considered a strictly feed-forward signal-processing model of sound representation in the auditory cortex [18]. By doing so, we did not contemplate the effects of feedback projections of acoustic representations [24] that have a potentially determinant role on the cerebral representation of sound stimuli.

In sum, converging evidence from behavior and neuroimaging thus demonstrates that both the categorical and dimensional models explain the representational geometry of behavioral and cerebral response to emotions in the voice. This is consistent with evidence in support of either one or the other model provided by many studies [9–15]. Our results however also shed significant light on this debate by showing that categorical and dimensional representations develop along different timescales in different cerebral regions. Our fine-grained characterization of the dynamics of perceived emotion in the voice thus reframes a debate of long-opposing theories as the interplay of partially overlapping large-scale cortico-subcortical systems with different representational dynamics [25].

Methods

Participants

Ten right-handed healthy adults (5 females; age from 19 to 38, mean = 25.1) participated in this study. All participants had normal hearing as assessed by an audiogram, provided written informed consent, and received financial compensation of £6/hour for their participation. The study was conducted in accordance with the Declaration of Helsinki (version 2013) and was approved by the local ethics committee (College of Science and Engineering, University of Glasgow). No statistical methods were used to pre-determine sample sizes. The number of participants used is similar to that reported in [15] but smaller than that reported in [11,14]. However, the number of neuroimaging trials acquired for each participant is up to one order of magnitude larger than that reported in previous publications [11,14,15], leading to more accurate estimates of single-participant cerebral responses, and a corresponding decrease of the contribution of single-participant estimation error on between-participants variability. The analyses of this study did not consider replicates as done in e.g., biological research on animal models. In other words, we did not assess statistically our questions on a separate group of participants investigated with the same experimental design as all others that was then discarded from the analyses reported in this study to prevent distortions in the inferential framework [26]. However, we aimed to boost the replicability of our main conclusions by relying on the analysis of three different datasets collected on each of the experiment participants (behavior, fMRI and MEG).

Stimuli

Stimuli consisted of nonverbal emotionally expressive vocalizations from the Montreal Affective Voices database [16] and were produced by two actors (one male, one female). Each actor produced five vocalizations (vowel /a/) either with an emotionally neutral intention, or expressing anger, disgust, fear or pleasure. Vocalizations normalized in root mean square (RMS) amplitude were then used to generate the stimulus set by morphing between each pair of vocalizations from the same speaker (each stimulus morphed to the same duration of 796 ms, corresponding to the average duration of the sound stimuli in the ten selected unmorphed samples from the MAV database).

Voice morphing was performed using STRAIGHT [27] in Matlab (Mathworks, Inc, Natick, USA). STRAIGHT performs an initial time-by-time pitch-adaptive spectral smoothing in each stimulus to separate the contributions to the voice signal arising from the glottal source or from the supra-laryngeal filtering. A voice stimulus is decomposed by STRAIGHT into five parameters (f0, frequency structure, duration, spectro-temporal density, and aperiodicity) that can be manipulated and combined independently across stimuli. Time-frequency landmarks that aid correspondences across voices during morphing were manually identified in each stimulus, and corresponded to the frequencies of the first three formants at the onset and offset of phonation. Morphed stimuli were then generated by resynthesis based on the linear (time and aperiodicity) and logarithmic (f0, the frequency structure and spectro-temporal density) interpolation at these time-frequency landmarks.

Two types of morphing continua were produced: 1) between neutral and each of the four emotions (neutral-anger, neutral-disgust, neutral-fear, and neutral-pleasure), and 2) between pairs of emotions (anger-disgust, anger-fear, anger-pleasure, disgust-fear, disgust-pleasure, and fear-pleasure). The morphing continuum between neutral and each emotion consisted of 6 stimuli, progressing in acoustically equal steps of 25% (e.g., 100% neutral; 75% neutral/25% anger; 50% neutral/50% anger; 25% neutral/75% anger; 100% anger; 125% anger). The 125% emotion morph was generated by extrapolating along the neutral-to-emotion dimension to create a caricatured emotion. The morphing continuum between pairs of emotions consisted of 5 stimuli, again progressing in acoustically equal steps of 25%. In total, 78 stimuli were used in the experiment, consisting of 39 stimuli for each speaker (cf. Supplementary Audio Files 1-78). They were low-pass filtered at 5 kHz to account for the spectral bandwidth of the MEG stimulation system (same stimuli used in all imaging and behavioral sessions; see below) and finally RMS normalized once more.

Experimental design

Each individual took part in 11 experimental sessions. Neuroimaging data were collected during the first 8 sessions (4 fMRI and 4 MEG; imaging modalities alternated with fMRI first for a random selection of half of the participants; MEG at least 3 days after prior fMRI session to avoid magnetization artifacts). Behavioral data were collected during the last three sessions, perceived emotion categories and dimensions being estimated only during the last session to avoid introducing explicitly an attentional focus towards either during the rest of the experiment. Data collection and analysis were not performed blind to the conditions of the experiments.

On each run of the fMRI and MEG acquisition (20 runs per fMRI session and at least 78 runs per participant across all of the MEG sessions), participants were presented with all of the stimuli from one speaker (random speakers order on each pair of subsequent blocks; inter-stimulus interval – ISI – jittered uniformly between 3 and 5 s) while carrying out a one-back repetition detection task (1 repetition per run; random selection of repeated stimulus; group averaged p correct = 98%; SEM = 0.2%). Throughout the session, participants were instructed to fixate a black cross presented against a white background (RGB values = [0,0,0] and [255,255,255], respectively; screen field of view = 19° x 80° and 26° x 19° degrees for fMRI and MEG, respectively).

On each of the first two behavioral sessions, participants rated the dissimilarity between all of the stimuli from the same speaker (speaker order counterbalanced across participants). On each trial, they were presented with one of the possible 741 pairs of sounds (within-pair ISI = 250ms; random within-pair order) and were asked to rate how dissimilar they were by placing a slider along a visual analogue scale marked “very similar” and “very dissimilar” at the two extremes. They could listen to the pair of stimuli as many times as necessary before giving a response. This experimental phase was preceded by an initial familiarization phase during which participants were presented with all of the sound stimuli two times (ISI = 250 ms; random order). In this phase, they were instructed to estimate the maximum and minimum between-sound dissimilarity, so as to optimize the usage of the rating scale in the subsequent experimental phase. The dissimilarity rating task was initially practiced with a set of 10 vocalizations not included in the main experiment.

During the last behavioral session, participants performed two tasks – categorical and dimensional ratings as well as emotion categorization. In the rating task, participants rated each stimulus on arousal (low to high), valence (negative to positive), and emotional intensity for four emotions (anger/disgust/fear/pleasure, low to high) using an on-screen slider. In the categorization task, they identified the emotion as being anger, disgust, fear, or pleasure. Before the experiment began, participants were given 10 practice trials for both tasks on a set of vocal stimuli not included in the main experiment. Participants were then familiarized to the entire stimulus set before the first block. On each block of trials, participants carried out either the rating or categorization tasks (alternated across blocks) for all of the stimuli from the same speaker (pseudo-random order of speaker gender with not more than two subsequent same-gender blocks). Throughout the session, each of the two tasks was repeated three times for each of the speakers, for a total of 12 blocks of trials.

Sound stimuli (sampling rate = 48 kHz; bit depth = 16 bit) were presented through electrostatic headphones (NordicNeuroLab, Bergen, Norway) for fMRI, Etymotic ER-30 tubephone for MEG, and during the behavioral sessions through BeyerDynamic DT 770 Pro headphones receiving the audio signal from the Audiophile 2496 sound card amplified with a Mackie 1604-VLZ PRO monitor system. The MEG tubephone system introduced strong spectral coloring of the sound stimuli and suppressed heavily frequencies > 6 kHz. Stimuli for all sessions were consequently low-pass filtered at 5 kHz. Flat-frequency response for the MEG audio stimulation chain was achieved through inverse filtering methods.

Neuroimaging data acquisition

fMRI scans were acquired with a Siemens 3T Trio scanner, using a 32-channel head coil. Functional multiband echo planar imaging (EPI) volumes were collected with a repetition time (TR) of 1s (echo time TE = 26 ms; flip angle = 60; multiband factor = 4; GRAPPA = 2). Each functional volume included 56 slices of 2.5 mm thickness (inter-slice gap = 2.5 mm; interleaved even acquisition order in an axial orientation along the direction of the temporal lobe, providing nearly whole-brain coverage. The in-plane voxel size was 2.5 mm2 (78 × 78 matrix). A whole-brain, high-resolution, structural T1-weighted MP-RAGE image (192 sagittal slices, 256 × 256 matrix size, 1 mm3 voxel size) was also acquired to characterize the participant's anatomy. In each of the fMRI sessions, we also collected a field map to correct for geometric distortions in the EPI volumes caused by magnetic field inhomogeneities [28].

MEG recordings were acquired with a 248-magnetometers whole-head MEG system (MAGNES 3600 WH, 4-D Neuroimaging) at a sampling rate of 1017.25 Hz. Participants were seated upright. The position of five coils, marking fiducial landmarks on the head of the participants, was acquired at the beginning and at the end of each block.

Measurement of acoustic dissimilarity

We modelled the time-varying acoustic structure of the sound stimuli by considering most accurate acoustics-driven computational model of the cortical representation of complex sounds currently available: the modulation transfer function (MTF [182930]. The MTF was computed on each of 100 post-onset temporal windows (0-800 ms), resulting in a 5-dimensional complex-number representation with dimensions temporal window (ms), frequency (Hz), modulation rate (ω), modulation-rate direction (up vs. down) and scale (Ω). For each temporal window independently, we then computed the acoustic RDM by computing the Euclidean distance between the stimulus-specific MTF representations in the complex plane (Fig 1d).

Preliminary analysis of behavioral data

We initially assessed whether block averaged emotion categorization proportions (Fig. 1b) and category intensity ratings were, as expected positively correlated. We computed the Spearman rank correlation between ratings and categorizations for each speaker and scale and then used a permutation approach to assess whether the speaker- and group-averaged correlation was significantly larger than zero (one-tailed inference, 100,000 independent stimulus permutations for each participant). As expected, categorization proportions and category intensity ratings were strongly correlated with each other (T(9) for anger categorization/ratings correlation = 8.77, p < 0.001 corrected for multiple comparisons across emotions, participant averaged Fisher Z correlation = 0.948, SEM = 0.108, percentile bootstrap 95% CI = 0.784/1.176; T(9) for disgust = 12.95, p < 0.001, Fisher Z correlation = 1.024, SEM = 0.079, bootstrap 95% CI = 0.887/1.180; T(9) for fear = 12.72, p < 0.001, Fisher Z correlation = 0.898, SEM = 0.071, bootstrap 95% CI = 0.753/1.011; T(9) for pleasure = 11.45, p < 0.001, Fisher Z correlation = 0.829, SEM = 0.072, bootstrap 95% CI = 0.694/0.964). Following previous studies, all subsequent analyses considered the category intensity ratings to avoid confounding genuine differences between perceived category and dimensions with differences in the measurement scale and reliability of categorization and rating data [14,15]. We subsequently assessed the effect of voice morphing on perceived emotion attributes (five morphing % parameters describing each experimental stimulus – one for each expressed emotion and one for the neutral vocalization; six measures of perceived emotion – four emotion category intensity ratings and valence and arousal ratings). The five morphing parameters were not orthogonal because for each stimulus only two at best had a non-zero value (average Spearman correlation between morph parameters = -0.24; STD = 0.02). For this reason, the selective perceptual effect of morph parameters was assessed independently of their shared variance by measuring their Spearman semi-partial correlation (s.p.r) with the perceived emotion ratings. Significance testing for all of the analyses in this study relied on a permutation-based group-level random effects (RFX) approach [2431]. Here, we: [1] estimated independently for each participant and speaker the null s.p.r distribution for each of the 30 morph/emotion pairs by permuting randomly the stimulus labels (N permutations = 100,000; same permutations across speakers and morph/emotion pairs, but not across participants); [2] averaged across speaker genders permuted and unpermuted s.p.r converted to the Fisher Z scale; [3] chance-corrected both the permuted and unpermuted s.p.r by subtracting the median of the null s.p.r distributions under permutation; [4] computed the T(9) test for the group-average permuted and unpermuted s.p.r; [5] finally established significance thresholds for the unpermuted T(9) tests as the 95th percentile of the distribution of the maximum of the absolute value of the permuted T(9) statistics (two-tailed inference) across pairs of morph parameters with perceived emotion measures, thus controlling for family-wise error (FWE) at a 0.05 level. We used a similar strategy to assess whether our morphing manipulation affected more the categories or dimension ratings. To this purpose, we measured the Fisher Z scale RSQ in a rank GLM predicting independently for each of the participant, speaker and rating scale the perceived emotions from the morphing weights, and subsequently averaged the resulting RSQ across speakers and across either the category intensity or dimension rating scales. We then used the same permutation-based approach specified above to assess whether the RSQ for the category intensity rating scales differed significantly from the RSQ for the dimension rating scales (two-tailed inference; N permutations = 100,000).

We then assessed the acoustic specification of perceived emotions by carrying out s.p.r tests of the association of the time-by-time acoustic RDMs with the perceived emotion representational dissimilarity matrices (emotion RDMs). We subsequently contrasted the variance in the time-by-time acoustic RDM variance uniquely explained by the category vs. dimensions model (valence and arousal together). S.p.r and unique variance measures were computed independently for each participant and speaker. The valence and arousal RDMs measured the absolute pairwise difference in valence and arousal ratings, respectively. The category RDM was defined as the Euclidean distance between stimulus-specific category ratings response profiles (e.g., category intensity ratings of 1, 0.2, 0.2 and 0.05 for anger, disgust, fear and pleasure, respectively). Significance testing relied on a similar approach as for the analysis of the effect of morph parameters on perceived emotions (N permutations = 100,000; reshuffling of rows and columns of distance matrices traditionally known as Mantel test [32]; one- and two-tailed inference for s.p.r and for the unique variance contrasts, respectively; p ≤ 0.05 adjusted for multiple comparisons across acoustic RDM time points at family-wise error rate (FWER) = 0.05; square root of unique explained variances Fisher Z transformed prior to contrast; Fig S2a). Note that the permutation-based chance-level adjustment at the basis of our statistics framework captures and corrects for expected chance-level differences between the variance explained by one category RDM vs. two dimension RDMs.

Analysis of dissimilarity ratings

We then assessed the s.p.r of perceived stimulus dissimilarity (pairwise dissimilarity ratings) with the emotion RDMs (one-tailed inference to assess an expected increase of perceived dissimilarity for increasingly diverse voice emotions; p ≤ 0.05 FWE-corrected across emotion-RDM pairs; Fig 2d) and subsequently contrasted the dissimilarity RDM variance uniquely explained by the category vs. dimension RDMs (two-tailed inference; p ≤ 0.05 as based on the percentiles of the null permutation distribution). The results of these two tests also partialled out from the emotion RDM variance what accounted for by acoustics dissimilarity. Critically, and mirroring the procedure followed for the cerebral RDM analyses (see below), partialling from the emotion RDMs the perceived dissimilarity variance jointly explained by the acoustic RDM required establishing which of the 100 time-by-time acoustic MTF RDMs best explained the dissimilarity RDM. Indeed, within the traditional non-cross-validated association framework adopted in this study, partialling out all of the 100 acoustic RDMs from each of the emotion RDMs would have resulted in overfitting and overestimation of the emotion RDM variance accounted for by acoustics. To estimate which MTF RDM latency best explained perceived dissimilarity we thus adopted a procedure inspired by a recent study on the cerebral representation of speech [24] and relying on a leave-one-participant out scheme. In practice, and independently for each participant, we initially computed the speaker-average Fisher Z correlation between each of the MTF RDMs and the perceived dissimilarity RDM in the rest of the participants, and used as optimal MTF RDM latency estimate for a given participant that which maximized the group-average correlation between acoustics and perceived dissimilarity (resulting optimal latency = 0 ms for all participants). This optimal latency acoustic RDM was then partialled out from each of the emotion RDMs prior to the computation of their s.p.r with perceived dissimilarity or of the uniquely explained perceived dissimilarity variances. These analyses of perceived dissimilarity were finally complemented with the assessment of the correlation between perceived dissimilarity on the one hand, and the time-by-time MTF RDM, on the other (one-tailed inference; $p \leq 0.05$ FWE-corrected across MTF RDM time points; Fig S2a), and with the assessment of the correlation between the (time-invariant) emotion RDMs, after partialling out from each of them the variance they shared with the optimal latency acoustic RDM (one-tailed inference; $p \leq 0.05$ FWE-corrected across emotion-RDM pairs; Fig 3d).

Preprocessing of neuroimaging data

Analyses were carried out in Matlab using SPM12, Fieldtrip [33], GLMdenoise [34] and custom code. The initial preprocessing of fMRI and MEG data produced for each participant the stimulus-specific responses further analyzed to assess the representation of perceived emotions (see below). Functional MRI images from all runs were realigned to the first image in the first run, unwarped to correct for movement-by-distortion interactions (full width at half maximum – FWHM = 5 and 4 mm for realignment and unwarp, respectively; for both 7th degree B-spline for interpolation), and slice time corrected to the onset of the temporally central slice. Anatomical volumes were co-registered to the grand-average of the preprocessed functional volumes and segmented into grey matter, white matter, and cerebro-spinal fluid. Diffeomorphic Anatomical Registration using Exponentiated Lie algebra (DARTEL [35]) was used to create a common brain template for all of the participants. An initial group DARTEL grey-matter mask was created by considering all non-cerebellum voxels with a grey-matter probability > 0.1. The final analysis mask for each individual was given by the 6-connected voxels within the conjunction of the group mask deformed to native space with the voxels associated with a participant-specific grey-matter probability > 0.25.

For each participant, the 80 fMRI runs (40 for each of the two speaker genders) were divided into 5 mixed-gender groups of 16 runs each (interleaved assignment of runs to groups). Unsmoothed native-space data within the analysis mask for each group of runs were analyzed within a massively univariate general linear model (GLM) that estimated the fMRI response specific to each stimulus. Stimulus-specific regressors were created by convolving a sound on-off binary time-series with the canonical hemodynamic response function (HRF). The GLM included a high-pass discrete cosine transform (DCT) filter (cut off = 128 s), the head motion regressors estimated during the realignment step and a run-specific intercept. The GLM also included additional noise regressors that modeled temporal effects unrelated to the stimulus condition (e.g., blood pulse). These noise regressors were estimated independently for each of the groups of runs and participants using GLMdenoise [34] (default polynomial detrending replaced with DCT filter), resulting in N noise regressors = 6 on average (across-participant STD = 2).

Several initial steps of the preprocessing of MEG data were carried out on the unsegmented data from each run. Infrequent SQUID jumps (observed in 2.3% of the channels, on average) were repaired using piecewise cubic polynomial interpolation. For each participant independently, we then removed channels that consistently deviated from the median spectrum (shared variance < 25%) on at least 25% of the runs (N removed channels = 8.4 on average; STD = 2.2). Runs associated with excessive head movements or MEG channels noise or containing reference channel jumps were finally discarded, leaving on average 75.9 runs per participant (range = 65–84; average maximum coil movement across blocks and participants = 5 mm; STD = 1 mm). Environmental magnetic noise was removed using regression based on principal components of reference channels. Both the MEG and reference data were then filtered using a forward-reverse 70 Hz FIR low-pass (-40 dB at 72.5 Hz), a 0.2 Hz elliptic high-pass (-40 dB at 0.1 Hz) and a 50 Hz FIR notch filter (-40 dB at 50 ± 1Hz), and were subsequently resampled to 150 Hz. Residual magnetic noise was then removed applying once more the same method as for the full-resolution signal. ECG and EOG artifacts were removed using ICA (runica on 30 components) and were identified based on the time course and topography of IC components [36]. MEG data from each run was finally segmented into trials (-0.2 to 1.3 s after sound onset).

A native-space source-projection grid with a resolution of 3.5 mm was prepared for each participant by resampling the native-space analysis mask for the fMRI data. Depth-normalized lead fields were computed based on a single shell conductor model. Source-projection filters were then computed for each run using LCMV beamformers (regularization = 5%; sensor covariance across all trials excluding repetitions) and reduced to the maximum-variance orientation across all runs. Source-projected stimulus-specific time courses were finally averaged within 5 independent mixed-gender groups of runs (interleaved assignment of runs to groups), leading to a reduction of the computational burden for subsequent data-analysis steps.

Analysis of cerebral representational geometries

We implemented a whole-brain searchlight representational similarity analysis (RSA [17]; Fig. 2) to assess the spatiotemporal representational dynamics of perceived emotions. To this purpose, we relied on a method of statistical multimodal fusion of representations in multivariate spatial (fMRI) and spatiotemporal (MEG) cerebral response patterns. Importantly, we combined the complementary strengths of fMRI (high spatial resolution) and MEG (high temporal resolution) by adopting a method that builds on recent work on the fusion of same-participant fMRI and MEG data [3738]. For this purpose, we used an initial mask of fMRI correlations to constrain spatially subsequent tests on time-varying MEG RDMs at searchlight-center locations within the fMRI masks (see below for details; note that the constraining procedure operates only on the selection of the RDM-specific searchlights, i.e., MEG RDMs are computed at all brain locations before the fMRI masking takes place). As compared to previous work, our approach: [1] considered spatial information in the MEG data; [2] measured representations of stimulus and perceived attributes independently in MEG and fMRI instead of taking fMRI as the golden representational standard; [3] imposed on MEG only mild spatial fMRI constraints by considering lenient non-selective fMRI representation masks instead of stringent selective fMRI representation masks. This approach also mitigated the multiple comparison problem for source-based MEG analysis compared to what otherwise faced in the absence of fusion with fMRI results. The statistical framework followed in general the same approach adopted for the analysis dissimilarity rating data, and measured here the association between emotion and cerebral response RDMs.

For fMRI, cerebral RDMs were computed in native space within a spherical region (6 mm diameter) centered at each grey-matter mask location (at least 50% in-mask voxels). In particular, we computed the cross-validated Mahalanobis (Crossnobis) distance between stimulus-specific response patterns (Mahalanobis whitening of stimulus-specific GLM estimates using the GLM residuals within the searchlight) by cross-validating the response pattern covariance across the 5 groups of mixed-gender runs, and finally converting it to a (whitened) Euclidean distance [39,40]. For MEG, cerebral RDMs were computed within a native-space spatiotemporal searchlight of 10 mm diameter and 50 ms temporal window from -0.15 to 1.1 seconds from onset with 15 ms of overlap between subsequent temporal windows. For each searchlight, we derived the cross-validated Euclidean distance between stimulus-specific beamformed time-courses from the covariance between stimulus-specific response patterns cross-validated between the 5 groups of mixed-gender runs.

RSA analyses for fMRI assessed the Spearman correlation r between cerebral RDMs and each of the three emotion RDMs or the perceived dissimilarity RDM (non-selective representation; one-tailed inference to ascertain an expected increase in the dissimilarity of cerebral response patterns for increasingly diverse voice emotions; $p \leq 0.05$ FWE-corrected across fMRI searchlights; insets of Fig 3a-b). For MEG, RSA analyses assessed representation within progressively nested significance masks and included, in order: [1] the Spearman correlation between cerebral and emotion RDMs (non-selective representation; one-tailed inference; mask = significant fMRI r for categories, valence and arousal for MEG representation of categories, valence and arousal, respectively; $p \leq 0.05$ FWE-corrected within mask); [2] the Spearman s.p.r between cerebral and emotion RDMs (selective representation; one-tailed inference; mask = significant r for a given emotion RDM; $p \leq 0.05$ FWE-corrected within mask); [3] the pairwise contrasts of the unique cerebral RDM variance explained by each of three pairs of emotion RDMs (pairwise representational dominance contrasts; two-tailed inference; mask = union of significant s.p.r for the emotion RDMs in the pairwise contrast; $p \leq 0.05$ FWE-corrected within mask); [4] the explained cerebral RDM variance contrasts between the categorical and dimensional models (valence and arousal together; two-tailed inference; mask = significant s.p.r with category RDM in union with significant unique MEG RDM variance explained jointly by the valence and arousal RDMs; $p \leq 0.05$ FWE-corrected within mask); [5] whether the emotion RDM s.p.rs (Fig 3a-b), the pairwise emotion RDM contrasts (Fig S5) and the category vs. dimensions contrast (Fig 3c) survived after partialling out of the cerebral RDM variance explained by the acoustic RDM (mask = significance of same test prior to partialling out acoustics; $p \leq 0.05$ FWE-corrected within mask). Importantly, at each step we: [1] computed all measures of representation (r, s.p.r and unique variance contrasts) in native space and carried out group-level RFX inference (T tests) on the representation maps transformed to the group DARTEL space (FWHM of Gaussian smoothing of native-space encoding maps = 8 for both fMRI and MEG; 500 native-space permutations for each of the participants; 10,000 group-level DARTEL-space permutations each computed by selecting at random one of the 500 permutations for each of the participants; median-permutation chance-level correction of native-space statistics prior to DARTEL normalization notably leading to the elimination of chance-level differences between unique variance terms based on one – categories – vs. two – dimensions – RDMs; chance levels computed independently for each participant and searchlight); [2] used cluster mass enhancement of the group-level statistics, permutations included (permutation of rows and columns of RDMs, as for analysis of perceptual dissimilarity; 3D and 4D spatiotemporal cluster mass enhancement for fMRI and MEG, respectively; cluster-forming threshold of T(9) = 1.83 and 2.26 for one- and two-tailed inference, respectively [41]; [3] mitigated the multiple comparison problem by constraining analysis masks at each testing step within the significance mask from the previous step (see above). Finally note that as a statistical instrument, semi-partial correlations are not biased towards finding segregated cerebral networks specifying different emotion attributes because they do not exclude the possibility that variance unique to each of the features is represented in the same cerebral searchlight (e.g., see simultaneous selective representation of categories and arousal RDM in perceived dissimilarity).

Partialling out acoustics variance from each of the measures of the representation of emotions in the cerebral representational geometry was achieved using a similar method as the leave-one-participant-out method outlined for perceived dissimilarity but now focusing on the MEG RDMs. Here, and building on previous methods [24], we optimized independently for each participant, and cerebral location (MEG projection grid points) the lag between the entire MTF RDM time series and the MEG RDM time series (acoustics-to-brain lags from 0 to 250 ms, modelling MEG acoustic representations subsequent to the presentation of the sound stimulus) so as to maximize the Spearman correlation between the MEG RDM time series (window = 800 ms, corresponding to the temporal window of latencies for the MTF RDMs) on the one hand, and the MTF RDM time series, on the other (MEG/MTF correlation computed independently for each time window and then averaged across MEG latencies). The final acoustics-independent tests of the MEG representation of emotions at a given MEG latency were finally carried out by partialling out from the emotion RDMs the MTF RDM at the latency resulting from the combination of the MEG latency and the optimal acoustics-to-brain lag. Critically, this leave-one-participant-out approach also made it possible to assess the cerebral representation of the optimal-latency RDM in the absence of circularities in the analysis and subsequent artificial boosting of the statistical power. This final supplementary analysis of cerebral acoustic representations was carried out by measuring the Spearman correlation between the latency-optimized MTF RDM and the time-varying MEG RDM within a mask of significant fMRI correlations with any among the emotion RDMs, and the perceived dissimilarity RDM ($p \leq 0.05$ FWE-corrected within mask). For all analyses of fMRI and MEG data, we corrected for multiple comparisons within the entire analysis mask by establishing significance thresholds for the non-permuted cluster-mass enhanced T statistics as the 95th percentile of the permutation distribution for within-mask CM enhanced maxima for one-tailed tests and as the 2.5th and 97.5th percentiles of within-mask minima and maxima, respectively for two-tailed inference (maximum-statistic approach; FWER = 0.05).

The global and local peaks of the significant-statistic MEG maps were identified based on an automatic local-peak detection algorithm applied to the Dartel-space results (negated for significant negative effects). These were further filtered to select only latencies associated to a peak effect for a specific spatial location, separated in space by a minimum of 20 mm from peaks at the same latency, and associated with an absolute T(9) value of >3.25. The anatomical label for the selected peak was established by selecting the most frequent label within 10 mm from the identified location of the Dartel-deformed Automated Anatomical Labelling (AAL) atlas shipped with SPM12 [42]. The Dartel coordinates of the identified peaks were finally transformed to MNI space.

Visualization

The non-metric 2D MDS models in Fig. 1, 3 and 4 were computed using the R-package SMACOF [43]. We modeled the dissimilarity ratings and the emotion RDMs (Fig 3) and the cerebral RDMs (Fig 4) using an inter-individual difference scaling model (INDSCAL; participant- and speaker-specific RDMs as inputs), avoiding well known distortions of the representational geometry associated with group averaging of distance data [44]. Importantly, before being modelled with INDSCAL the cerebral RDMs were biased towards the cerebral variance diagnostic of perception (emotion and perceived dissimilarity RDMs) using the leave-one-participant-out cross-validation scheme described in [31] meant to minimize the impact of perception-independent variance (e.g., measurement noise) on the INDSCAL solution. To this purpose, and separately for each speaker and ROI, we first estimated the regression coefficients for predicting the cerebral RDM of all participants together minus one (all ranked RDMs stacked in one vector) using their emotion and perceived dissimilarity RDMs as predictors. The perception-diagnostic cerebral RDM input to the INDSCAL model was than estimated by predicting the cerebral RDM of the left-out participant from her own perception RDMs, but using the regression coefficients estimated during the first “training” step of the cross-validated procedure.

The 3D glass brains in Fig. 4–5 and Supplementary Fig. 1-3 comprised two components: [1] a mesh of the ICBM 152 2009c Nonlinear Asymmetric template [45]; [2] the functional blobs, rendered by first modeling the surface of each blob with a 3D mesh, and then projecting onto it the volumetric statistical map it circumscribed (maximum projection within 7 mm radius sphere centered at mesh vertex). All meshes and projections were computed within BrainVISA (http://brainvisa.info/), and were rendered using a custom OpenGL shader for the transparency effect.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Supported by UK’s Biotechnology and Biological Sciences Research Council (grants BB/M009742/1 to JG, BLG, SAK, and PB, and BB/L023288/1 to PB and JG), by the French Fondation pour la Recherche Médicale (grant AJE201214 to PB), and by Research supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI), and the Excellence Initiative of Aix-Marseille University (A*MIDEX). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank Olivier Coulon and Oliver Garrod for help with the development of the 3D glass brain, as well as Yinan Cao, Ian Charest, Carlos Crivelli, Beatrice De Gelder, Guillaume Masson, Robin A.A. Ince, Flor Kusnir, Stephen McAdams, and Robert Zatorre for useful comments on previous manuscript versions.

Data availability

The following materials are available at a Dryad repository (https://datadryad.org/stash/dataset/doi:10.5061/dryad.m905qfv0k): single-trial behavioural data, single-cross-validation fold fMRI data, and single-trial MEG data for all participants; anonymized anatomical information required to reconstruct the MEG sources and deform native-space statistical maps to Dartel and MNI space; sound stimuli and modulation transfer function representations.

Code availability

The Matlab code for reconstructing the MEG sources, carrying out a group-level RSA analysis of the fMRI and MEG representation of perceived emotions, and generating MNI-space statistical maps is available at the following Dryad repository: https://datadryad.org/stash/dataset/doi:10.5061/dryad.m905qfv0k.

References

1. Ekman, P. The Science of Facial Expression. Fernandez-Dols, JM, Russell, JA, editors. Oxford University Press; 2017. 39–56.
2. Sauter DA, Eimer M. Rapid detection of emotion from human vocalizations. J Cogn Neurosci. 2010; 22 :474–481. [PubMed: 19302002]
3. Russell JA. Core affect and the psychological construction of emotion. Psychol Rev. 2003; 110 :145–172. [PubMed: 12529060]
4. Barrett LF. The theory of constructed emotion: an active inference account of interoception and categorization. Soc Cogn Affect Neurosci. 2017; 12 :1–23. [PubMed: 27798257]
5. Hamann S. Mapping discrete and dimensional emotions onto the brain: controversies and consensus. Trends Cogn Sci. 2012; 16 :458–466. [PubMed: 22890089]
6. Vytal K, Hamann S. Neuroimaging support for discrete neural correlates of basic emotions: a voxel-based meta-analysis. J Cogn Neurosci. 2010; 22 :2864–2885. [PubMed: 19929758]
7. Lindquist KA, Wager TD, Kober H, Bliss-Moreau E, Barrett LF. The brain basis of emotion: a meta-analytic review. Behav Brain Sci. 2012; 35 :121–143. [PubMed: 22617651]
8. Kober H, et al. Functional grouping and cortical-subcortical interactions in emotion: a meta-analysis of neuroimaging studies. Neuroimage. 2008; 42 :998–1031. [PubMed: 18579414]
9. Rolls ET, Grabenhorst F, Franco L. Prediction of subjective affective state from brain activations. J Neurophysiol. 2009; 101 :1294–1308. [PubMed: 19109452]
10. Kotz SA, Kalberlah C, Bahlmann J, Friederici AD, Haynes JD. Predicting vocal emotion expressions from the human brain. Hum Brain Mapp. 2013; 34 :1971–1981. [PubMed: 22371367]
11. Skerry AE, Saxe R. Neural representations of emotion are organized around abstract event features. Curr Biol. 2015; 25 :1945–1954. [PubMed: 26212878]
12. Saarimaki H, et al. Discrete Neural Signatures of Basic Emotions. Cereb Cortex. 2016; 26 :2563–2573. [PubMed: 25924952]
13. Kragel PA, LaBar KS. Decoding the Nature of Emotion in the Brain. Trends Cogn Sci. 2016; 20 :444–455. [PubMed: 27133227]
14. Briesemeister BB, Kuchinke L, Jacobs AM. Emotion word recognition: discrete information effects first, continuous later? Brain Res. 2014; 1564 :62–71. [PubMed: 24713350]
15. Grootswagers T, Kennedy BL, Most SB, Carlson TA. Neural signatures of dynamic emotion constructs in the human brain. Neuropsychologia. 2017
16. Belin P, Fillion-Bilodeau S, Gosselin F. The “Montreal Affective Voices”: a validated set of nonverbal affect bursts for research on auditory affective processing. Behav Brain Res. 2008; 40 :531–539.

17. Kriegeskorte N, Mur M, Bandettini P. Representational similarity analysis – connecting the branches of systems neuroscience. Front Syst Neurosci. 2009; 2 :1–28.

18. Chi T, Ru P, Shamma SA. Multiresolution spectrotemporal analysis of complex sounds. J Acoust Soc Am. 2005; 118 :887–906. [PubMed: 16158645]

19. Belyk M, Brown S, Lim J, Kotz SA. Convergence of semantics and emotional expression within the IFG pars orbitalis. Neuroimage. 2017; 156 :240–248. [PubMed: 28400265]

20. Touroutoglou A, et al. A ventral salience network in the macaque brain. Neuroimage. 2016; 132 :190–197. [PubMed: 26899785]

21. Anderson DJ, Adolphs R. A framework for studying emotions across species. Cell. 2014; 157 :187–200. [PubMed: 24679535]

22. Cowen AS, Keltner D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proc Natl Acad Sci U S A. 2017; 114 :E7900–E7909. [PubMed: 28874542]

23. Cowen AS, Laukka P, Elfenbein HA, Liu R, Keltner D. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. Nat Hum Behav. 2019; 3 :369–382. [PubMed: 30971794]

24. Giordano BL, et al. Contributions of local speech encoding and functional connectivity to audio-visual speech perception. Elife. 2017; 6 :e24763. [PubMed: 28590903]

25. Pessoa L. Understanding emotion with brain networks. Curr Opin Behav Sci. 2018; 19 :19–25. [PubMed: 29915794]

26. Vaux DL, Fidler F, Cumming G. Replicates and repeats--what is the difference and is it significant? A brief discussion of statistics and experimental design. EMBO Rep. 2012; 13 :291–296. [PubMed: 22421999]

27. Kawahara, H; Matsui, H. Auditory morphing based on an elastic perceptual distance metric in an interference-free time-frequency representation. Proc. 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003); 2003. 256–259.

28. Hutton C, et al. Image distortion correction in fMRI: A quantitative evaluation. Neuroimage. 2002; 16 :217–240. [PubMed: 11969330]

29. Santoro R, et al. Encoding of natural sounds at multiple spectral and temporal resolutions in the human auditory cortex. PLoS Comput Biol. 2014; 10 :e1003412. [PubMed: 24391486]

30. Kell AJE, Yamins DLK, Shook EN, Norman-Haignere SV, McDermott JH. A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. Neuron. 2018; 98 :630–644. [PubMed: 29681533]

31. Cao Y, Summerfield C, Park H, Giordano BL, Kayser C. Causal Inference in the Multisensory Brain. Neuron. 2019; 102 :1076–1087. [PubMed: 31047778]

32. Mantel N. The detection of disease clustering and a generalized regression approach. Cancer Research. 1967; 27 :209–220. [PubMed: 6018555]

33. Oostenveld R, Fries P, Maris E, Schoffelen JM. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. Comput Intell Neurosci. 2011; 2011

34. Kay KN, Rokem A, Winawer J, Dougherty RF, Wandell BA. GLMdenoise: a fast, automated technique for denoising task-based fMRI data. Front Neurosci. 2013; 7 :247. [PubMed: 24381539]

35. Ashburner J. A fast diffeomorphic image registration algorithm. Neuroimage. 2007; 38 :95–113. [PubMed: 17761438]

36. Hipp JF, Siegel M. Dissociating neuronal gamma-band activity from cranial and ocular muscle activity in EEG. Front Hum Neurosci. 2013; 7 :338. [PubMed: 23847508]

37. Cichy RM, Pantazis D, Oliva A. Resolving human object recognition in space and time. Nat Neurosci. 2014; 17 :455–462. [PubMed: 24464044]

38. Cichy RM, Pantazis D. Multivariate pattern analysis of MEG and EEG: A comparison of representational structure in time and space. Neuroimage. 2017; 158 :441–454. [PubMed: 28716718]

39. Walther A, et al. Reliability of dissimilarity measures for multi-voxel pattern analysis. Neuroimage. 2016; 137 :188–200. [PubMed: 26707889]

40. Diedrichsen J, Kriegeskorte N. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. PLoS Comput Biol. 2017; 13 :e1005508. [PubMed: 28437426]

41. Maris E, Oostenveld R. Nonparametric statistical testing of EEG- and MEG-data. J Neurosci Methods. 2007; 164 :177–190. [PubMed: 17517438]

42. Rolls ET, Joliot M, Tzourio-Mazoyer N. Implementation of a new parcellation of the orbitofrontal cortex in the automated anatomical labeling atlas. NeuroImage. 2015; 122 :1–5. [PubMed: 26241684]

43. De Leeuw J, Mair P. Multidimensional Scaling Using Majorization: SMACOF in R. Journal of Statistical Software. 2009; 31 :1–30.

44. Ashby FG, Boynton G, Lee WW. Categorization response time with multidimensional stimuli. Percept Psychophys. 1994; 55 :11–27. [PubMed: 8036090]

45. Fonov V, et al. Unbiased average age-appropriate atlases for pediatric studies. Neuroimage. 2011; 54 :313–327. [PubMed: 20656036]

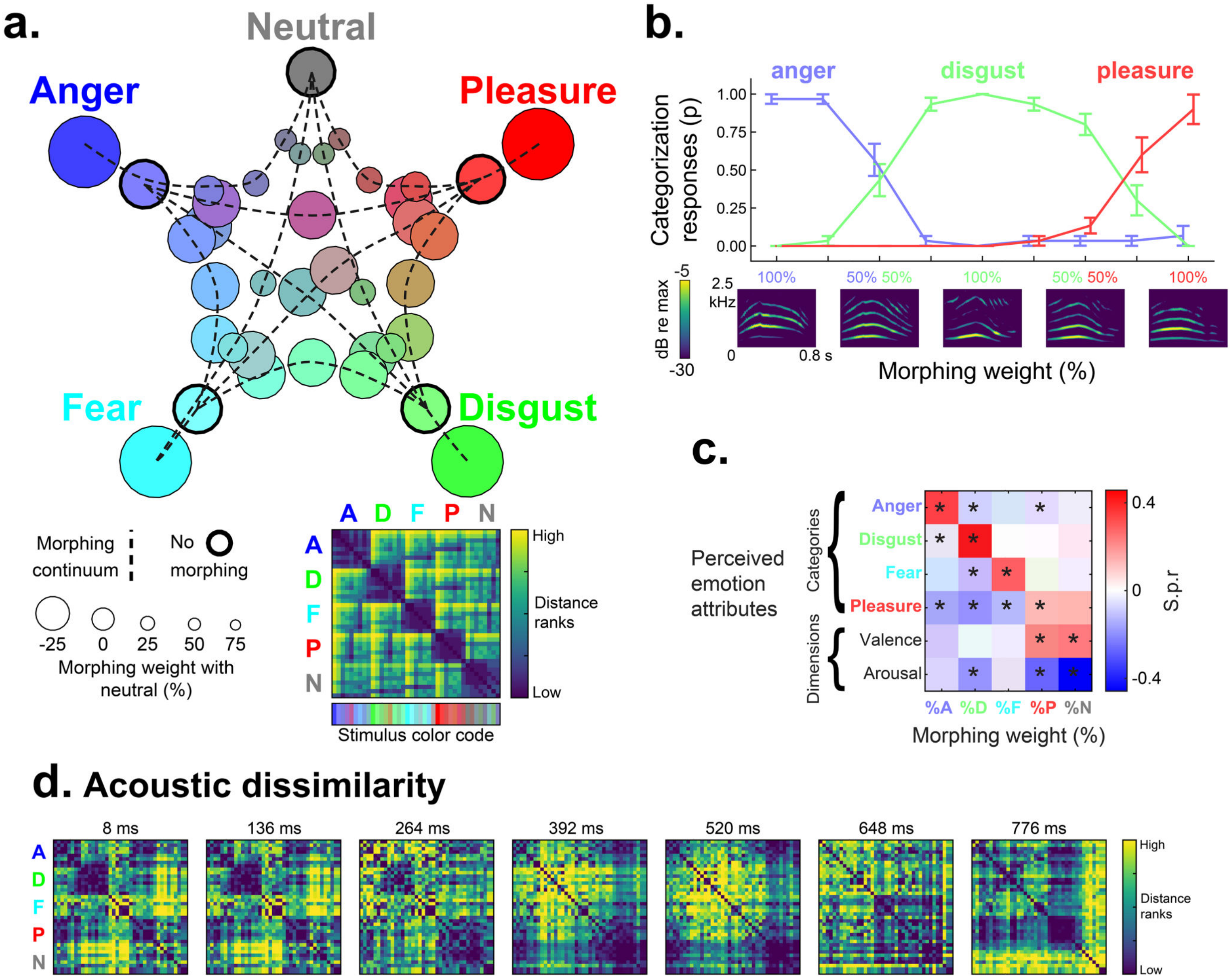

Figure 1. Emotional voice stimuli.

Auditory stimuli consist of synthetic voice samples generated by morphing between brief affective bursts from a validated database (Montreal Affective Voices [16]). **a**. 2-dimensional multi-dimensional scaling representation of the morphing-weight distance between stimuli, and corresponding distance matrix. Stimuli include both morphs between one emotion and the neutral expression (including caricatures, largest circles) and morphs between pairs of emotions. **b**. Example of morphing between three vocalizations expressing different emotions and categorization responses (error-bar = bootstrap SE; N participants = 10; N bootstrap samples = 100,000). Spectrograms of example stimuli are shown below. **c**. Effect of morphing on perceived emotional attributes (semi-partial correlation = s.p.r of emotion ratings with morphing weights; * two-sided p ≤ 0.05 FWE-corrected across morphing weight/rating scale pairs, see Supplementary Table 1 for full results). Each emotional attribute is selectively modulated by at least two morphing weights, and each morphing weight modulates at least two emotional attributes. **d**. Acoustic RDMs showing time-varying acoustic dissimilarity of the sound stimuli (modulation transfer function, MTF [17]) as a function of peri-stimulus time.

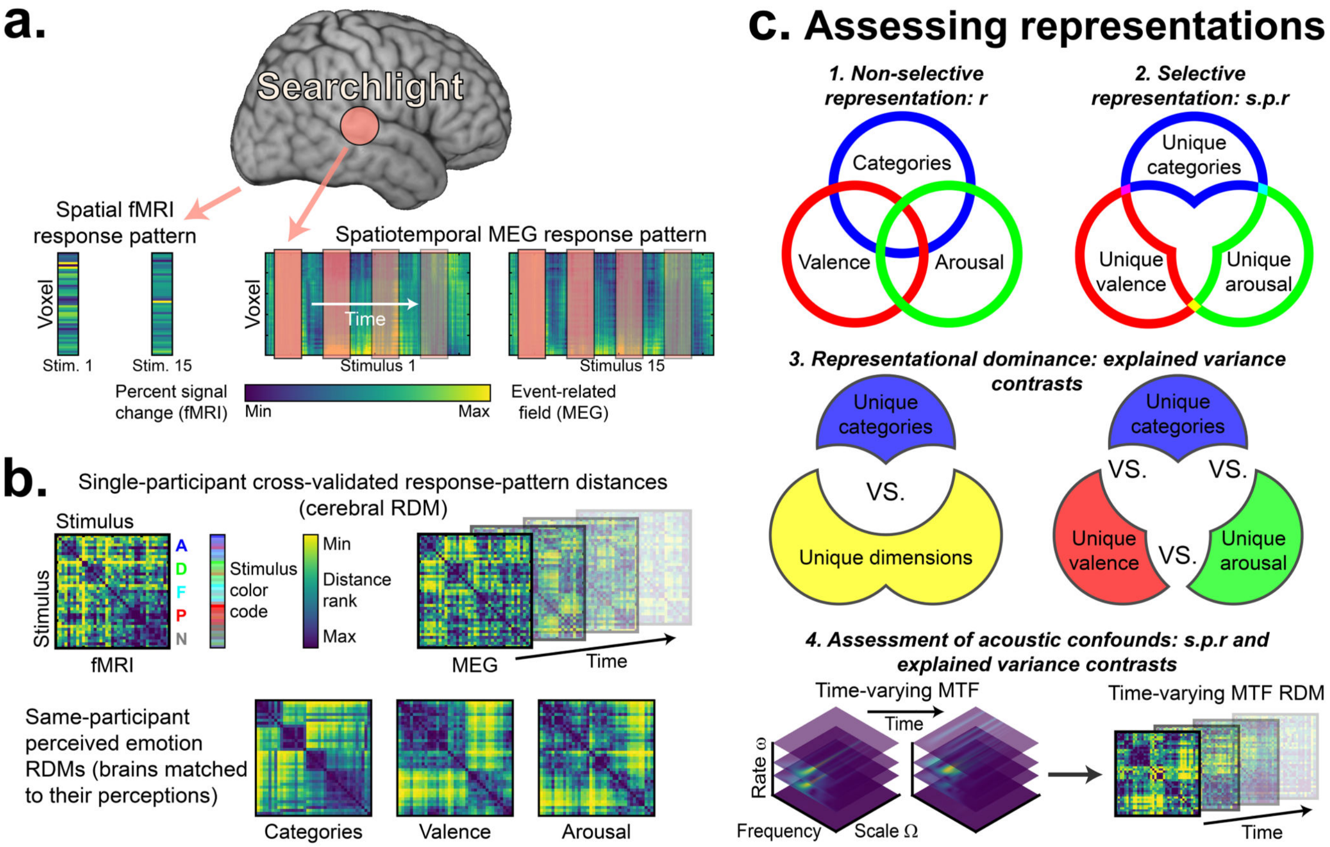
a.
Searchlight
Spatial fMRI response pattern
Voxel
Stim. 1
Stim. 15
Spatiotemporal MEG response pattern
Voxel
Time
Stimulus 1
Stimulus 15
Percent signal change (fMRI)
Min
Max
Event-related field (MEG)
b.
Single-participant cross-validated response-pattern distances (cerebral RDM)
Stimulus
Stimulus
A
D
F
P
N
Stimulus color code
Min
Distance rank
Max
fMRI
MEG
Time
Same-participant perceived emotion RDMs (brains matched to their perceptions)
Categories
Valence
Arousal
c. Assessing representations
1. Non-selective representation: r
Categories
Valence
Arousal
2. Selective representation: s.p.r
Unique categories
Unique valence
Unique arousal
3. Representational dominance: explained variance contrasts
Unique categories
VS.
Unique dimensions
Unique categories
VS.
VS.
Unique valence
VS.
Unique arousal
4. Assessment of acoustic confounds: s.p.r and explained variance contrasts
Time-varying MTF
Time
Rate ω
Frequency
Scale Ω
Time-varying MTF RDM
Time

Figure 2. Spatio-temporal representational similarity analysis [17] of the representation of perceived emotion attributes in cerebral representational geometries.

a. Spatial (fMRI) and spatio-temporal (MEG) stimulus-specific response patterns (percent signal change and source-space event-related fields for fMRI and MEG, respectively) were extracted within a grey-matter spherical searchlight (fMRI radius: 6mm. MEG radius: 10mm) for each of the participants. The MEG spatio-temporal searchlight had a duration of 50ms (hop size = 35ms). **b**. Cerebral representation analyses measured the association between the pairwise distance of stimulus-specific cerebral responses (cerebral representational dissimilarity matrices – RDMs, measuring the cerebral representational geometry) and the pairwise distance of stimulus-specific perceived emotion attributes derived from the behavioral data of the same participant (emotion RDMs; matching of each brain to the perceptions from the same participant). **c**. MEG representations were assessed in four subsequent steps within nested significance masks. First, we assessed the non-selective representation of perceived emotions by measuring the correlation (r) between same-participant cerebral and emotion RDMs. The masks of significant fMRI correlations derived from this step were used to constrain spatially the MEG correlation tests which were carried out within the fMRI correlation masks. Note that different emotion RDMs are correlated (see Fig. 2d) and explain overlapping portions of the cerebral RDM variance. The second step focused on portions of the cerebral RDM variance explained uniquely by each emotion RDM, and assessed their selective representation by means of their semi-partial correlations (s.p.r) with the cerebral RDMs. Third, we assessed the representational dominance of each emotion RDM or of the categories vs. dimensions models by contrasting directly the explained cerebral RDM variance specific to the category- or dimension-encoding model, and the unique explained cerebral RDM variance explained by each of the three perceived emotion RDMs (three pairwise contrasts). Finally, we ruled out the acoustic confound hypothesis for the selective representation and representational dominance tests by repeating these tests after having partialled out form the emotion RDMs what explained by the time-varying dissimilarity of the modulation transfer function of the voice stimuli (MTF [18]). This same approach was adopted to assess emotion representations in the perceived dissimilarity of the voice stimuli (Fig. 3).

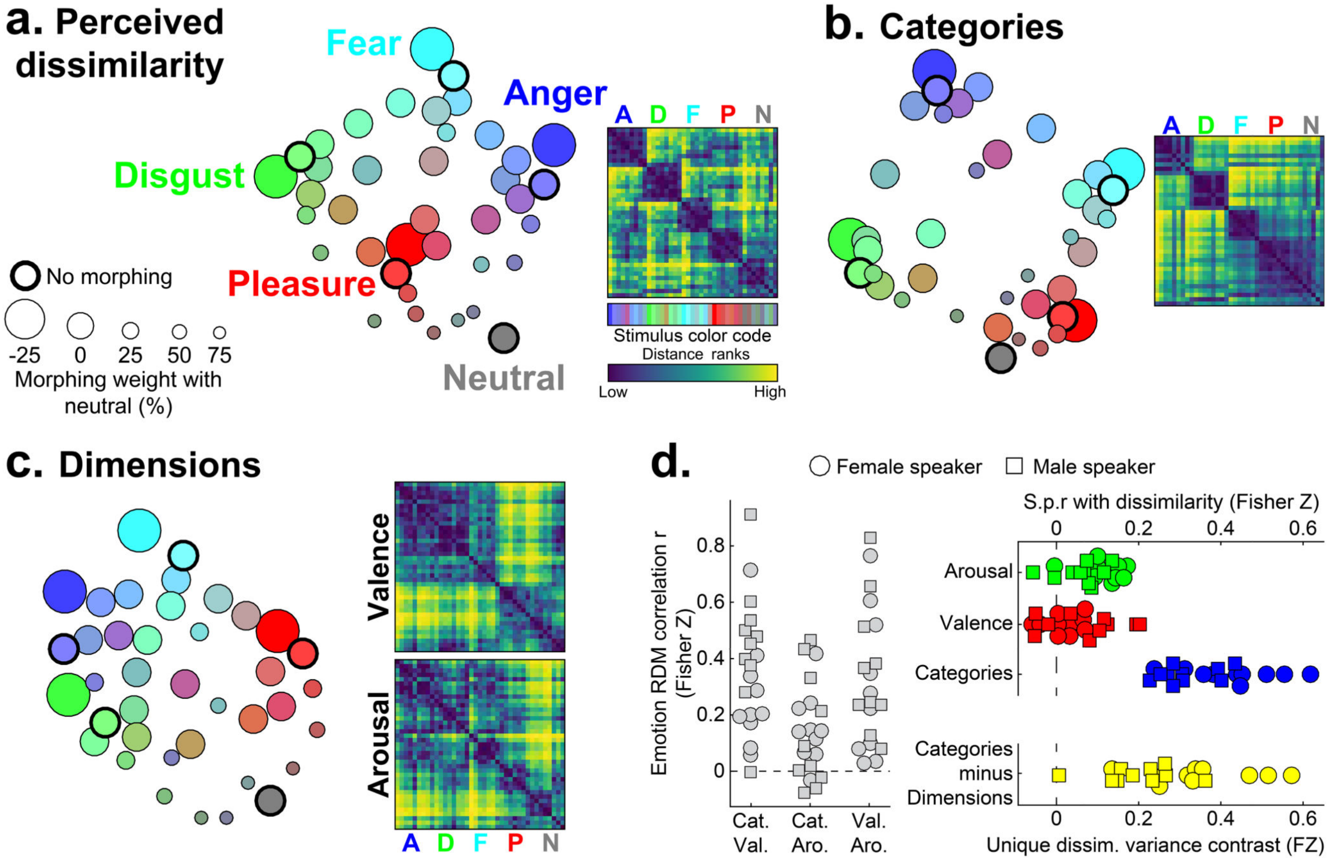

Figure 3. Categories better account for perceived dissimilarity than Valence and Arousal.

a–c. Perceptual RDMs averaged across speakers and participants (right) and 2D non-metric inter-individual difference multidimensional scaling model (left). **a**. Dissimilarity RDM: note the clear clustering of each emotion and their separation from morphs with the neutral expression. **b**. Categories RDM. Note the strong resemblance with the perceived dissimilarity RDM. **c**. Perceived dimensions RDMs. d. Left: Correlation (r) between perceived emotional attributes for each participant and speaker after partialling out the acoustic RDM. Right: semi-partial correlations (s.p.r) between emotional attributes and dissimilarity (selective representation; red, green and blue plots), and categories vs. dimensions variance contrast (representational dominance; yellow plot; both after partialling out the acoustic RDM). One-tailed inference for r and s.p.r, FWE corrected across emotion RDM pairs or emotion RDMs, respectively, and two-tailed inference for unique variance contrasts. N participants = 10.

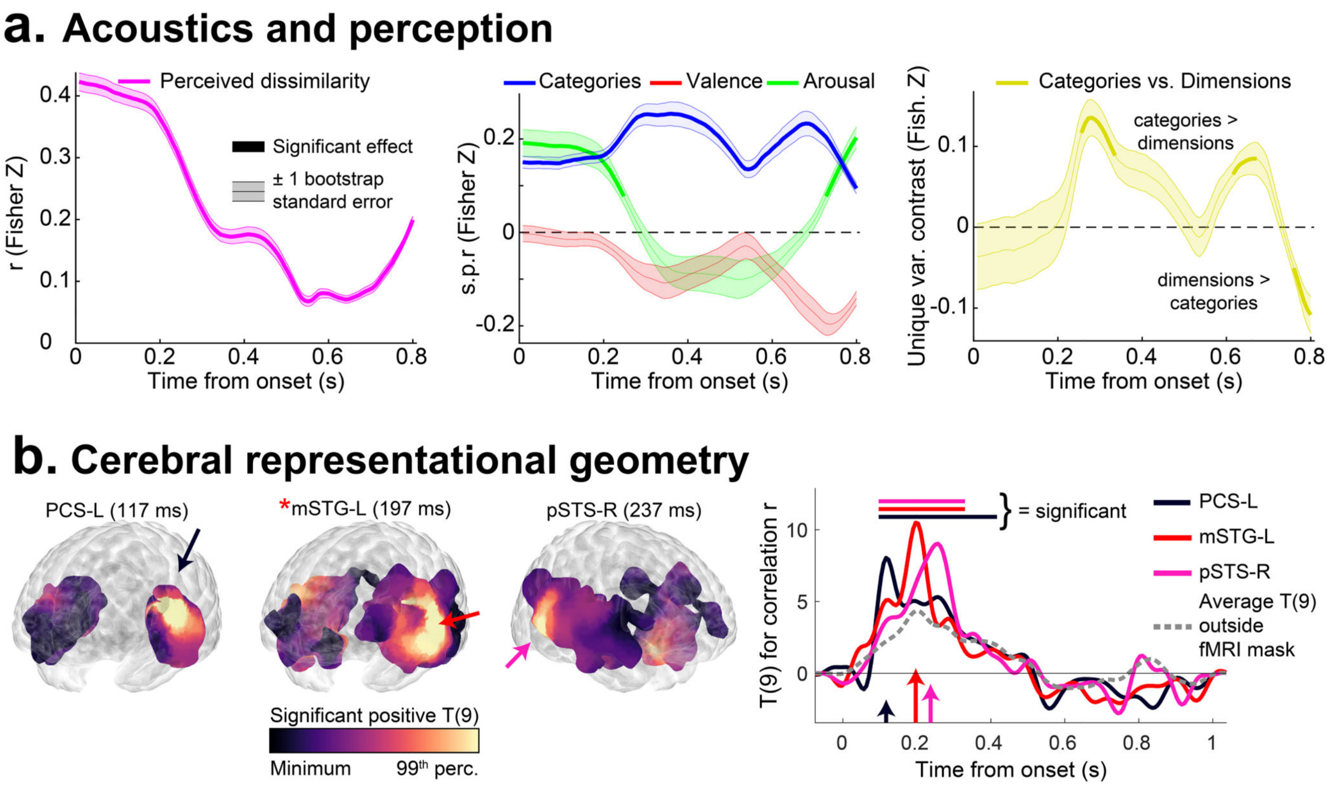

Figure 4. Perceptual correlates and cerebral representational geometry of acoustics.

a. Left: Correlation (r) between the time-varying RDM of the modulation transfer function (MTF [18]) and the perceived dissimilarity RDM. Note the stronger correlation of perceived dissimilarity with the initial acoustical structure. Middle: selective acoustical specification of emotion RDMs measured with the semi-partial correlation (s.p.r) of the emotion RDMs with the time-varying MTF RDM. Right: unique variance (var.) contrast for the acoustical specification of emotion categories and dimensions (Fish. = Fisher). Thick lines = significant at $p \leq 0.05$ FWE-corrected across MTF time points (one-tailed inference for r and s.p.r and two-tailed inference for unique variance contrasts; N participants = 10; N bootstrap samples for SE computation = 100,000). Note how both categories and dimensions are specified in acoustics around the onset and offset of the sound stimulus, and how stronger acoustic information about acoustics emerges only late in the sound stimulus (> 200ms).

b. Significant cerebral representation of time-varying acoustics in MEG representational geometries as a function of peri-stimulus time; one-tailed $p \leq 0.05$ FWE-corrected across voxels and time points. Cerebral acoustic representations were assessed within a mask of significant MEG correlation with any among the emotion RDMs and perceived dissimilarity RDM. PCS-L = left precentral sulcus; mSTG = middle superior temporal gyrus; pSTS-R = right posterior superior temporal sulcus; * = global T(9) peak; arrows = peak-effect latency; horizontal lines above graph indicate significant latencies (see Supplementary Table 2 for peaks statistics).

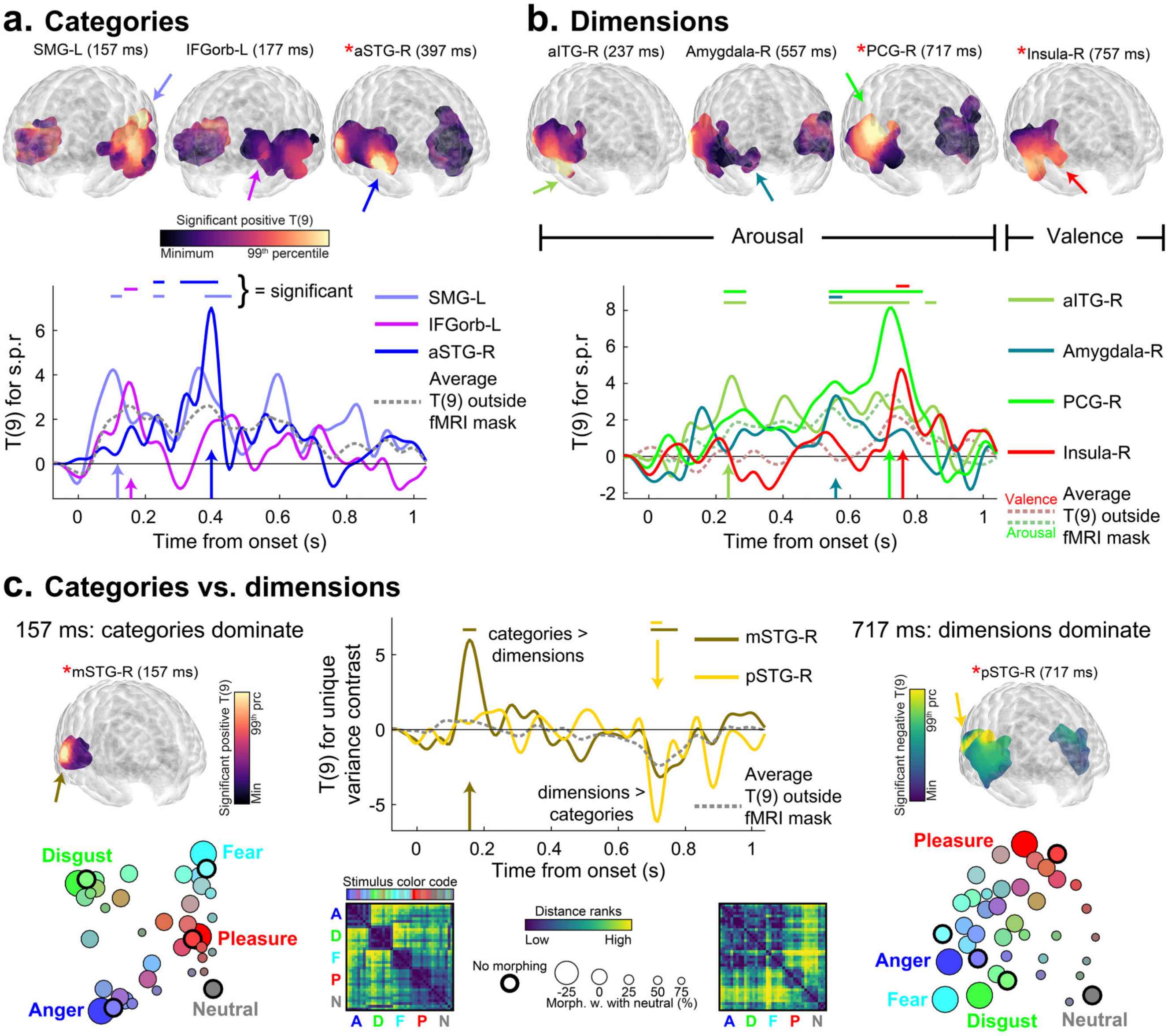

Figure 5. Cerebral representational geometries initially dominated by categories subsequently emphasize dimensions.

a. Top: Selective representation of emotion categories at selected latencies. Statistical maps are thresholded for significance (one-tailed p ≤ 0.05 FWE-corrected across latencies and/or voxels; N participants = 10; * = global peak) and represented within a transparent MNI template. Bottom: time-varying T(9) statistics at selected peaks (see Supplementary Table 2 for peaks statistics and anatomical abbreviations). Arrows = peak-effect latency; horizontal lines above graph indicate significant latencies. Dashed lines indicate the average value of the T(9) statistics outside the fMRI mask. **b**. Selective representation of dimensions (one-tailed p ≤ 0.05 FWE-corrected across latencies and/or voxels). **c**. Contrast for the cerebral-RDM variance uniquely explained by categories or dimensions (representational dominance; two-tailed p ≤ 0.05 FWE-corrected across latencies and voxels). Bottom: average cerebral RDMs and their 2D multidimensional inter-individual differences model.

Note how an initial temporal lobe representation with clear clustering of emotion categories (left) subsequently unfolds into a representation (left) where voices are arranged along a vertical valence dimension (bottom/top = negative/positive) and an almost horizontal arousal dimension (left/right = high/low; morph intensity of emotions increases with symbols size).