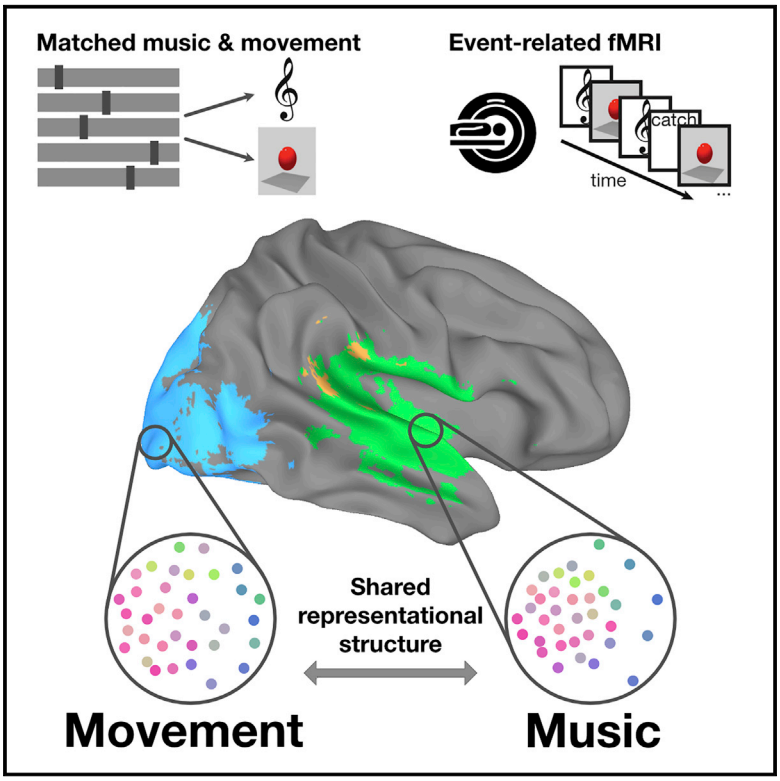
Article

Current Biology

Visual and auditory brain areas share a representational structure that supports emotion perception

Graphical abstract



Highlights

- Auditory and visual cortex represent emotional music and movement in comparable terms
- Temporal cortex uses a similar shared structure to represent both music and movement
- Visual cortex uses a similar shared structure during music listening alone
- Cross-sensory structure may arise to identify task-relevant feature combinations

Authors

Beau Sievers, Carolyn Parkinson, Peter J. Kohler, James M. Hughes, Sergey V. Fogelson, Thalia Wheatley

Correspondence

beau@beausievers.com (B.S.), thalia.p.wheatley@dartmouth.edu (T.W.)

In brief

Music and movement occur together across the world. Sievers et al. show a basis for this link: human auditory and visual cortex share a representational structure, putting music and movement in comparable terms. Shared crossmodal representations may arise to identify environmentally relevant feature combinations, such as those expressing emotion.

Sievers et al., 2021, Current Biology *31*, 5192–5203
December 6, 2021
https://doi.org/10.1016/j.cub.2021.09.043

CellPress

CellPress

Current Biology

Article

Visual and auditory brain areas share a representational structure that supports emotion perception

Beau Sievers,[1,8,*] Carolyn Parkinson,[2,3] Peter J. Kohler,[4,5] James M. Hughes,[6] Sergey V. Fogelson,[7] and Thalia Wheatley[8,9,10,*]

[1]Department of Psychology, Harvard University, Cambridge, MA 02138, USA
[2]Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095, USA
[3]Brain Research Institute, University of California, Los Angeles, Los Angeles, CA 90095, USA
[4]Department of Psychology, York University, Toronto, ON, Canada
[5]Centre for Vision Research, York University, Toronto, ON, Canada
[6]Wayne, NJ, USA
[7]Brooklyn, NY, USA
[8]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755, USA
[9]Santa Fe Institute, Santa Fe, NM 87501, USA
[10]Lead contact
*Correspondence: beau@beausievers.com (B.S.), thalia.p.wheatley@dartmouth.edu (T.W.)
https://doi.org/10.1016/j.cub.2021.09.043

SUMMARY

Emotionally expressive music and dance occur together across the world. This may be because features shared across the senses are represented the same way even in different sensory brain areas, putting music and movement in directly comparable terms. These shared representations may arise from a general need to identify environmentally relevant combinations of sensory features, particularly those that communicate emotion. To test the hypothesis that visual and auditory brain areas share a representational structure, we created music and animation stimuli with crossmodally matched features expressing a range of emotions. Participants confirmed that each emotion corresponded to a set of features shared across music and movement. A subset of participants viewed both music and animation during brain scanning, revealing that representations in auditory and visual brain areas were similar to one another. This shared representation captured not only simple stimulus features but also combinations of features associated with emotion judgments. The posterior superior temporal cortex represented both music and movement using this same structure, suggesting supramodal abstraction of sensory content. Further exploratory analysis revealed that early visual cortex used this shared representational structure even when stimuli were presented auditorily. We propose that crossmodally shared representations support mutually reinforcing dynamics across auditory and visual brain areas, facilitating crossmodal comparison. These shared representations may help explain why emotions are so readily perceived and why some dynamic emotional expressions can generalize across cultural contexts.

INTRODUCTION

Wherever there is music, there is movement.[1–3] Not only are music and dance pervasive across the anthropological and ethnographic record, some languages use a single word for both.[4,5] The link between music and movement is present from early in development, with infants as young as 7 months using movement to resolve ambiguities in musical rhythm.[6] Further, communication of emotion through music and movement occurs across a range of dissimilar cultures,[5,7,8] although there are also many important cross-cultural differences in emotion expression, perception, and conceptualization.[9–14] Here, we suggest that the link between music and movement may result from fundamental similarities in how music and movement are structured, perceived, and represented in the brain.

Supporting this account, preliminary research suggests that emotional music and movement can share structural features across cultures. In both the United States and a small-scale society in rural Cambodia, angry music and movement are both fast and move downward, peaceful music and movement are both slow and move upward, and so on.[7,15] Though suggestive, shared features do not fully explain the pervasive, experiential link between music and movement. Here, we examine a possible explanation: different sensory areas of the brain may share a representational geometry,[16] such that differences between sensory features and perceived emotions are represented by matched differences in patterns of neural activity, putting music and movement in comparable, task-relevant terms.

We tested two related main hypotheses concerning both *where* and *how* music and movement are represented in the brain: (H1) the separate regions, shared representations

hypothesis that separate, modality-specific auditory and visual areas use a shared representational geometry and (H2) the supramodal region hypothesis that a supramodal area (or areas) uses a single representational geometry for both auditory and visual stimuli. Note that H1 does not require patterns of activity in auditory and visual brain regions to be identical in every respect, as each sensory region likely represents modality-specific features. Evidence that the representation of music in auditory regions is very similar to the representation of movement in visual regions would support the separate regions, shared representations hypothesis (H1). By contrast, evidence of a single region that represents both music and movement using the same representational geometry would support the supramodal region hypothesis (H2). Importantly, H1 and H2 are not mutually exclusive, and while previous research has provided support for H2,[17] the status of H1 remains unknown.

Further, we asked how representations of perceived emotion in music and movement were organized, testing two auxiliary hypotheses: (A1) the simple features hypothesis that sensory brain regions represent emotional stimuli in terms of differences in simple stimulus features, without respect to how those features may later be inferentially processed to yield emotion judgments,[18,19] and (A2) the environmental conjunctions hypothesis that sensory representations of emotional stimuli closely track emotion judgments, suggesting that the human perceptual system may directly represent latent configurations of stimulus features associated with emotion content. These task-relevant representations may act as a shortcut, reducing the need for downstream inferential processing.[20] Evidence that sensory representations fit a model based on stimulus features would support the simple features hypothesis (A1), while evidence that sensory representations fit a model based on emotion judgments would support the environmental conjunctions hypothesis (A2). A1 and A2 are not mutually exclusive, as sensory regions may represent both stimulus features and environmentally relevant feature conjunctions.

By "perceived emotions," we refer only to participants' perceptions of the stimuli and their judgments of what emotions the stimuli expressed. We do not refer to emotional states evoked in the participants by the stimuli or to any other kind of emotion. Importantly, though we discuss the relevance of the findings to cross-cultural generalization, we did not test any hypotheses across cultures.

Testing both sets of hypotheses required comparing representations between brain areas. To accomplish this, we used model-based representational similarity analysis (RSA),[21,22] comparing representations evoked by separately presented auditory and visual stimuli to test H1 and H2. For detailed discussion of the limits and merits of this approach, see Roskies.[23] The model included predictors corresponding to both simple stimulus features and to participants' judgments of emotion content, supporting tests of A1 and A2. We performed an additional supporting test of H1 using a model-free approach that directly compared representational geometries across sensory areas without making any assumptions about representational content.

Previous research on neural representation of emotion

Emotion-related neural processes are distributed across a wide range of brain areas, with each area implicated in the production and/or perception of many emotions.[24,25] However, certain aspects of emotion processing are localized. Lesion and neuroimaging studies have demonstrated that some brain areas play an outsized role in the processing of specific emotions, for example, the amygdala for the conscious recognition of fearful stimuli[26,27] and the insula for recognizing disgust.[28,29] Because our hypotheses concern representations capable of distinguishing many different emotion expressions, we focus here on distributed representations of emotion and not on areas implicated in processing individual emotions.

Our hypotheses ask not only *where* in the brain emotions are represented but *how* those representations are structured. For example, a single brain area may distinguish stimulus classes using different spatial patterns of activity that all have the same mean. To characterize the representational properties of these areas, it is necessary to use techniques that are sensitive to such spatially distributed patterns, e.g., multivariate pattern classification[30] or representational similarity analysis.[16] For example, Peelen et al.[17] showed that medial prefrontal cortex (mPFC) and posterior superior temporal sulcus (pSTS) supramodally represent emotion identity by demonstrating that patterns of activity in these areas had greater within-emotion similarity than between-emotion similarity. Chikazoe et al.[31] used pattern analysis to locate supramodal valence (positive versus neutral versus negative) representations in medial and lateral orbitofrontal cortex and modality-specific valence representations in sensory cortices. Also investigating valence, Kim et al.[32] presented emotional movie clips and orchestral music and found a range of supramodal representations: valence direction in the precuneus; valence magnitude in mPFC, STS, and middle frontal gyrus (MFG); and both valence direction and magnitude in the STS, MFG, and thalamus. Skerry and Saxe[33] found that a model describing participants' appraisals of emotional narratives (e.g., "Did someone cause this situation intentionally or did it occur by accident?") fit activity in dorsal and middle medial prefrontal cortex, the temporoparietal junction, and a network of regions identified by a theory of mind localization task.

Importantly, where previous studies have focused on emotions evoked by narrative content while controlling for stimulus features,[31–33] the present study takes a different approach, focusing on emotion perceived solely from stimulus features without any contextualizing narrative. Emotions perceived from stimulus features make up a large and understudied part of human experience. For example, people often communicate emotion using only body language and tone of voice and actively seek out instrumental music and abstract visual art that communicates emotion only through variation in pitch, volume, shape, brightness, and so on. And although there are many cross-cultural differences in emotion experience, expression, and perception,[9–14] preliminary evidence suggests the use of shared stimulus features to express emotion can generalize across dissimilar cultures.[5,7] Despite its ubiquity and importance, the neural mechanisms supporting emotion perception from stimulus features remain poorly understood.

The present approach allows us to test the shared features (A1) and environmental conjunctions (A2) hypotheses, assessing whether sensory brain areas represent conjunctions of features associated with environmentally relevant stimuli, such as emotion expressions, or whether these areas represent simple features that may be used by separate, downstream areas to infer emotion content.

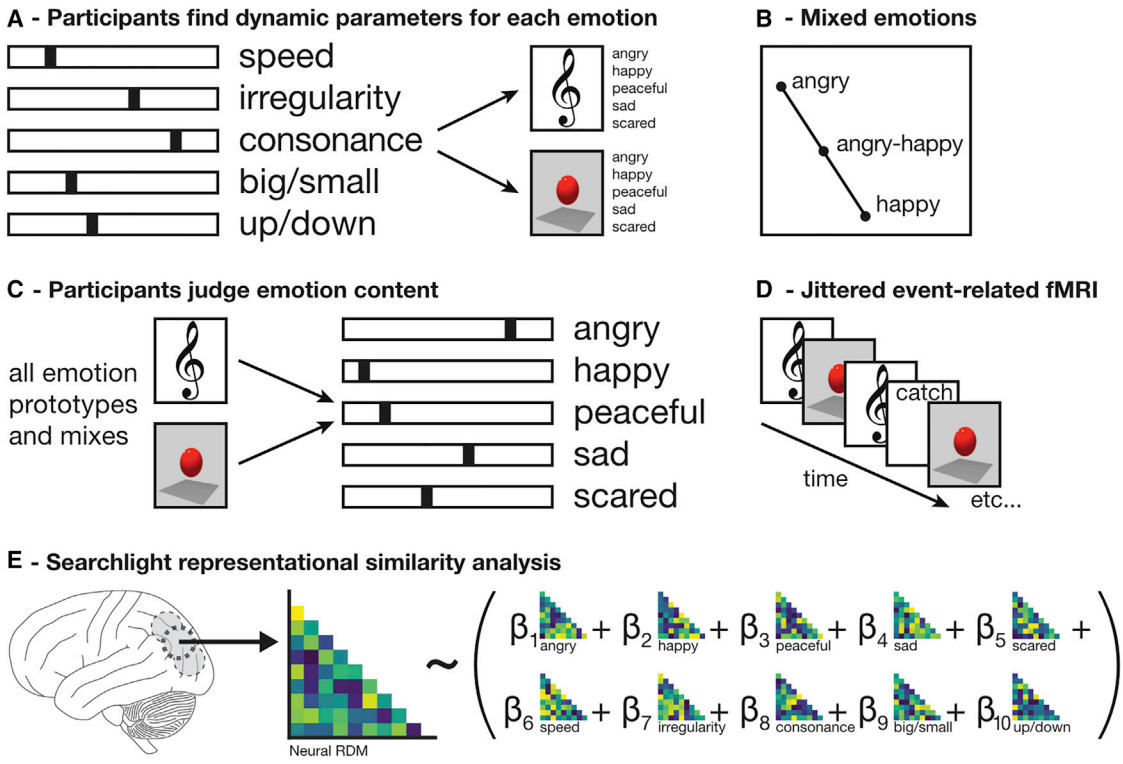


Figure 1. Experimental paradigm

(A) Participants in Sievers et al.[7] manipulated stimulus features to generate music and animation communicating five prototypical emotions: angry, happy, peaceful, sad, and scared.

(B) Mixed emotions were generated by linear interpolation between the stimulus features of prototypical emotions.

(C) Participants judged the emotion content of many prototypical and mixed emotions in music and animation.

(D) A subset of participants viewed many prototypical and mixed emotions in music and animation while undergoing jittered event-related fMRI scanning.

(E) Results were analyzed using searchlight representational similarity analysis.[16,21,22] For each searchlight sphere, the structure of the neural representational dissimilarity matrix (RDM) was predicted using a linear combination of stimulus feature and emotion judgment RDMs.

Stimuli and experimental paradigm

The stimuli consisted of short piano melodies and animations of a bouncing ball generated by participants in a previous study. This study showed that emotions were expressed the same way in music and in movement in both the United States and a small-scale society in rural Cambodia.[7] The participants used a computer program to create examples of five emotions (angry, happy, peaceful, sad, and scared) by manipulating five stimulus features (speed, irregularity, consonance and spikiness, ratio of big-to-small movements, and ratio of upward-to-downward movements). Participants were split into separate music and movement groups, each of which had no knowledge of the other. This approach did not presuppose what combinations of features would be used for each emotion, and participants were not instructed to use any specific features or feature combinations. Instead, they were encouraged to explore the entire possibility space. Critically, this method allowed us to vary what emotions were communicated while holding the depicted objects constant (i.e., each emotion was communicated using only the piano or the bouncing ball). This guaranteed that emotion content could only be communicated by variation in stimulus features and that processing requirements were consistent across the stimulus set.

Note that this approach differs from previous research where emotion was communicated using narrative stories or emotionally charged images, e.g., the International Affective Picture System.[34] Such studies often control for stimulus features, guaranteeing that emotion judgments are based solely on the content depicted in the stimuli. For example, a study of perceived emotion in spoken narrative might control for the speaker's tone of voice, focusing on what the speaker said rather than how they said it. The present study takes the opposite approach, controlling for the content depicted in the stimuli, guaranteeing that participants' emotion judgments are based solely on variation in stimulus features. This is analogous to holding a speaker's words constant, so that emotion can only be communicated by tone of voice.

Because many emotions are perceived as mixes of other emotions,[35] the stimulus set was augmented by linearly mixing the features of each emotion pair, creating mixed emotions (e.g., happy-sad). Emotions were mixed at 25%, 50%, and 75%. Three additional, "neutral" emotions were identified by searching for points in the stimulus feature possibility space that were distant from all other emotional feature combinations. For each set of stimulus features, or stimulus class, many individual stimuli were probabilistically generated (STAR Methods). This ensured the results were not dependent on the idiosyncrasies of single stimuli but were instead generalizable to all stimuli that shared the same features. Further, this prevented participants from

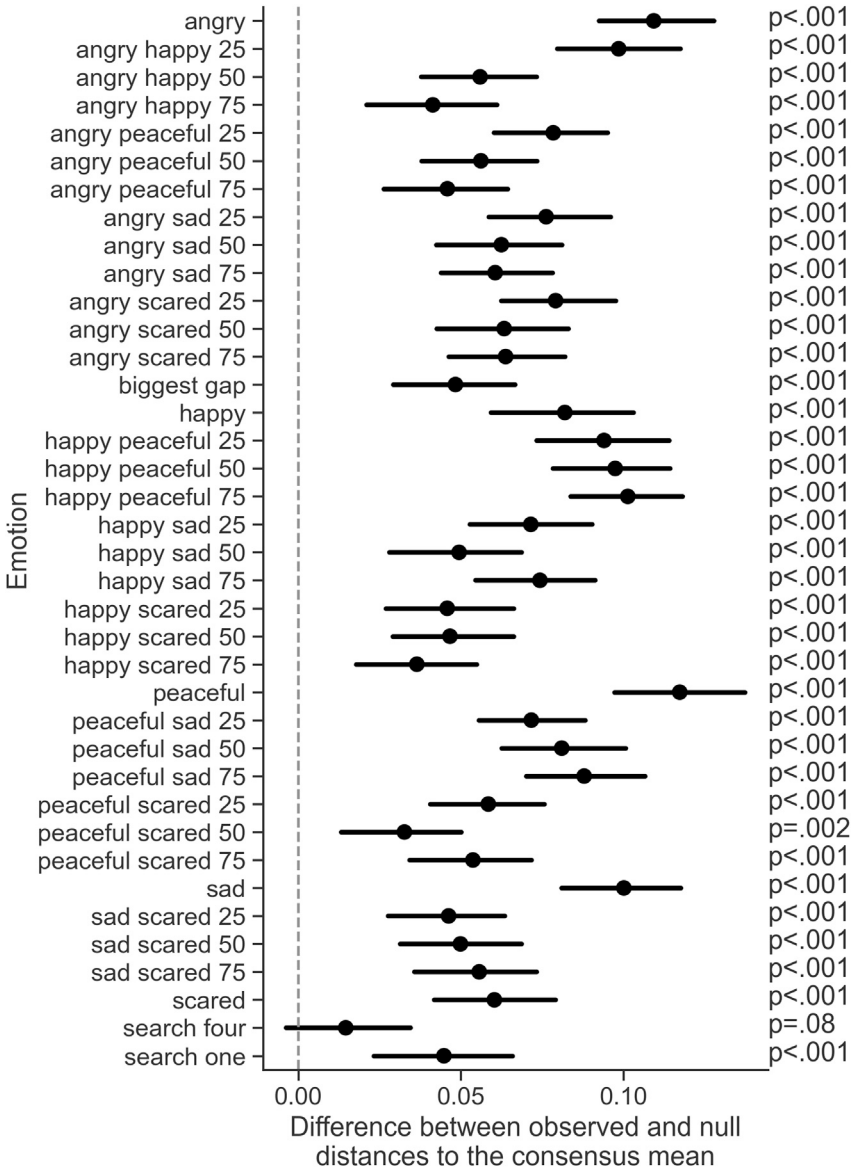


Figure 2. Emotion judgment agreement

Agreement between participants was assessed by measuring the distance of participants' individual emotion judgments from the class mean. Significance was assessed using permutation testing (STAR Methods). Values above 0 indicate more agreement than expected by chance. Error bars show 95% confidence intervals.

memorizing arbitrary associations between individual stimuli and emotion labels. Music and animation were matched such that, for each musical stimulus class, there was an animation stimulus class with the same features. This process yielded 76 total emotional stimulus classes, including both music and animation. All stimuli are available at https://osf.io/kvbqm/.

A separate set of participants judged how well each stimulus fit all five emotion labels, and a subset of these participants viewed many music and animation stimuli while undergoing functional magnetic resonance imaging (fMRI) (Figure 1).

RESULTS

Emotion judgments

Participants broadly agreed about the emotion content of each stimulus class (Figure 2). Agreement was assessed by measuring the distance of participants' individual emotion judgments from the class mean, scaled by the maximum possible distance, and significance was assessed using permutation testing (STAR Methods). For all 76 stimulus classes except one "neutral" emotion, participants' judgments were closer to the class mean than would be expected by chance (mean $t = -4.37$; mean difference from null = 0.07; mean $p < 0.001$). Importantly, this agreement rules out the possibility that participants invented and then memorized arbitrary associations between combinations of stimulus features and combinations of emotion labels.

Shared representational geometry

Auditory and visual brain regions shared a representational geometry. A single model of representational similarity[16,21,22] explained patterns of activity in visual brain regions during animation trials and auditory brain regions during music trials, providing strong support for the separate regions, shared representations hypothesis (H1) (Figure 3; Table 1). The model used 10 representational dissimilarity matrices (RDMs) as predictors: five based on the mean parameter settings used to create the stimuli (speed, irregularity and jitter, consonance and spikiness, ratio of big-to-small movements, and ratio of upward-to-downward movements), and five based on the mean emotion judgments of the behavioral participants (angry, happy, peaceful, sad, and scared; Figure 4). The model included no information specific to either vision or audition.

The peak of the average model fit across participants was in the left medial lingual gyrus for animation trials (mean R^2_{adj} = 0.15; 95% confidence interval [CI]: 0.08–0.21; t(19) = 4.68; p = 0.005 corrected) and in right anterior superior temporal gyrus for music trials (mean R^2_{adj} = 0.15; 95% CI: 0.1–0.2; t(19) = 6.08; p = 0.01 corrected; Figure 3). Critically, a direct, model-free test of similarity between these areas showed that they were more similar to each other than would be expected by chance ($\rho = 0.68$; $p < 0.001$), further supporting the separate regions, shared representations hypothesis (H1) and making it unlikely that the results reported above are an artifact of model misspecification (STAR Methods).

Model fit was driven by both stimulus feature *and* emotion judgment predictors and was not dominated by a small number

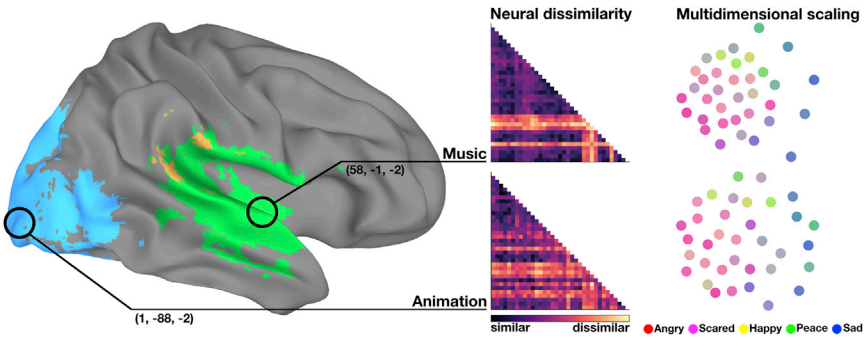


Figure 3. Main result

Highlighted brain areas were identified using a model including both stimulus features and emotion judgments as predictors, which was separately fit to animation trials (blue) and music trials (green). A significant proportion of participants' model fits overlapped for both trial types (yellow). Neural dissimilarity matrices show pairwise distances between activity patterns evoked by each stimulus at the locations of best model fit (circled)—medial lingual gyrus (animation) and lateral superior temporal gyrus (music). Labels are as in Figure 4. Multidimensional scaling flattens these matrices to two dimensions, so the distance between dots reflects the similarity of patterns of neural activity. Dots are colored by mixing the legend colors based on participants' judgments of the emotion content of each stimulus. Error bars for Spearman's ρ and β weights show 95% confidence intervals.

of predictors, providing support for both the simple features and environmental conjunctions hypotheses (A1 and A2). Individual predictors were assessed by mapping Spearman's ρ across the brain. Spearman's ρ was significant for all 10 predictors at the location of peak model fit (Figure 5) and was distributed similarly across the brain (Figure S4). β weight maps for each predictor were also calculated (Figure S5), reflecting only the unique contribution of each predictor, whereas Spearman's ρ reflects both unique and shared contributions. See Figure S6 for an assessment of model multicollinearity, including variance inflation factors for each predictor.

The model accounted for 51% of the variance for animation trials and 31% of the variance for music trials, relative to the lower bound of the noise ceiling (STAR Methods). Note that, because of small differences in functional anatomy across participants, the peak of the average model fit underestimates individual model fit. The mean of the individual peak model fits was in bilateral anterior superior temporal gyrus for music trials (mean individual R^2_{adj} = 0.26; 95% CI: 0.21–0.31; t(19) = 10.95; p < 0.001 uncorrected) and in the lingual gyrus for animation trials (mean individual R^2_{adj} = 0.31; 95% CI: 0.24–0.38; t(19) = 9.2; p < 0.001 uncorrected; Figures S1 and S2).

Overlapping auditory and visual model fit

Brain regions where music and animation were both represented were found in bilateral posterior superior temporal gyrus (pSTG) in 60% of participants (95% CI 36%–84%; p < 0.001 corrected), supporting the *supramodal region* hypothesis (H2) (Figure 6A; see Figure S3 for per-participant maps). To locate such supramodal representations, we created binary overlap masks, selecting voxels where both music and animation model fits were significant at the individual level (permutation p < 0.05 uncorrected). Multiple comparisons correction of these overlap maps was performed at the group level, testing the proportion of individuals with overlap in a region against the null hypothesis that no participants had overlap in that region. Critically, this analysis is insensitive to the magnitude of R^2_{adj} at the individual level, allowing detection of overlapping signals that have low magnitude but reappear across a significant proportion of the participants. The model fit for music trials was also significant at this location, though the model fit for animation trials was not (music mean R^2_{adj} = 0.03, 95% CI: 0.02–0.05, t(19) = 5.78, p = 0.01 corrected; animation mean R^2_{adj} = 0.03, 95% CI: 0.01–0.05, t(19) = 2.94, p = 0.13 corrected). The model accounted for 26% of the variance for animation trials and 31% of the variance for music trials, relative to the lower bound of the noise ceiling. Due to individual differences in functional anatomy, this procedure underestimates the proportion of participants with supramodal representations.

Exploratory intermodal RSA

To find brain areas that represented stimuli presented in that area's non-preferred modality, we performed an exploratory intermodal RSA (STAR Methods). Intermodal RSA revealed a bilateral set of areas across occipital, superior parietal, temporal, cingulate, and frontal cortex that represented stimuli presented in their non-preferred modality (Figure 6B; Table 1). Note that some of these areas did not show significant unimodal model fit. Peak intermodal model fit was in left lingual gyrus (mean R^2_{adj} = 0.28; 95% CI: 0.20–0.37; t(19) = 6.9; p < 0.001 corrected). Notably, the peak intermodal model fit exceeded the peak within-modality model fit for both music and animation and also exceeded the lower bound of the noise ceiling, explaining 40% of the variance relative to the upper bound. This suggests intermodal activity in left lingual gyrus was dominated by representations of model features. However, the lower bound of the intermodal noise ceiling was relatively low (0.07), suggesting that most reliable neural activity in this region was modality specific.

DISCUSSION

Music and movement are subjectively linked, and both use similar features to communicate emotion content. We examined a possible explanation for this link: that the brain represents music and movement using a shared representational geometry. To investigate this, we tested two primary hypotheses: (H1) the separate regions, shared representations hypothesis, where separate auditory and visual regions use the same representational geometry, and (H2) the supramodal region hypothesis, where some region(s) represent both auditory and visual stimuli.

Table 1. Peak model fits

Analysis	x, y, z	Nearest atlas label[36]	R^2_{adj}	%	95% CI	p
Music	58, −2, −2	R lateral aspect of the superior temporal gyrus	0.15		0.10–0.20	0.011
Music	−62, −16, 7	L lateral aspect of the superior temporal gyrus	0.09		0.05–0.12	0.011
Anim.	2, −88, −2	L lingual gyrus, lingual part of the medial occipito-temporal gyrus (O5)	0.15		0.08–0.21	0.005
Anim.	46, −68, 1	R inferior occipital gyrus (O3) and sulcus	0.04		0.01–0.07	0.005
Anim.	22, −82, 31	R superior occipital gyrus (O1)	0.03		0.01–0.06	0.005
Overlap	64, −28, 22	R supramarginal gyrus		60%	0.36–0.84	<0.001
Overlap	−58, −34, 19	L supramarginal gyrus		40%	0.16–0.64	0.008
Inter.	2, −88, −2	L lingual gyrus, lingual part of the medial occipito-temporal gyrus (O5)	0.28		0.20–0.37	<0.001
Inter.	64, −28, 22	R supramarginal gyrus	0.09		0.06–0.12	<0.001
Inter.	−56, −40, 22	L planum temporale or temporal plane of the superior temporal gyrus	0.08		0.05–0.10	<0.001
Inter.	32, −56, 61	R superior parietal lobule (lateral part of P1)	0.07		0.05–0.09	<0.001
Inter.	−32, −56, 61	L superior parietal lobule (lateral part of P1)	0.06		0.04–0.08	<0.001
Inter.	−16, −22, 40	L marginal branch (or part) of the cingulate sulcus	0.05		0.03–0.07	<0.001
Inter.	−28, −58, −53	L lateral occipito-temporal gyrus (fusiform gyrus, O4-T4)	0.04		0.03–0.06	<0.001
Inter.	−46, 44, 22	L middle frontal gyrus (F2)	0.03		0.02–0.04	<0.001
Inter.	−4, 64, 22	L superior frontal gyrus (F1)	0.03		0.02–0.04	<0.001

Labels determined programmatically using the atlas of Destrieux, Fischl, Dale, and Halgren.[36] For results per model predictor, see Figures S4 and S5. Anim., model fit to animation trials; inter., intermodal regions that fit the model even when the stimulus was presented in the non-preferred modality; music, model fit to music trials; overlap, percentage of participants with overlapping music and animation model fits.

We also tested two auxiliary hypotheses: (A1) the simple features hypothesis, where sensory areas represent individual stimulus features that are not directly associated with emotion content, and (A2) the environmental conjunctions hypothesis, where sensory areas represent conjunctions of features that directly track differences in emotion judgment.

We found that brain activity in separate auditory and visual areas shared a representational geometry, supporting the separate regions, shared representations hypothesis (H1). Providing additional support for H1, representations in auditory and visual brain areas were more similar to each other than to any randomly chosen pair of brain areas. Further, music and animation were represented in pSTG, suggesting the pSTG uses a supramodal representation, supporting the supramodal region hypothesis (H2).

Stimulus feature predictors (speed, jitter, consonance and spikiness, ratio of upward-to-downward movements, and ratio of big-to-small movements) were significant in both auditory and visual regions, supporting the simple features hypothesis (A1). A shared, crossmodal representation of simple stimulus features would support downstream comparison of auditory and visual stimuli, including inferential assessment of emotion content by, e.g., simulation theory[18] or theory theory[19] systems. On such an account, there may be nothing emotional per se about representations in sensory brain regions. However, predictors based on participants' emotion judgments were also significant in both auditory and visual regions, even when controlling for the stimulus feature predictors, supporting the environmental conjunctions hypothesis (A2). On this account, sensory regions represent conjunctions of task-relevant environmental features, such as those associated with emotion expressions, supporting direct perception of social information.[20,37]

Other possible parameters, such as valence and arousal,[38] value,[39,40] Fourier features,[15] HMAX,[41] and motion energy,[42] certainly covary with the stimulus feature and emotion judgment predictors. Because these and similar measures would be fully dependent on the model parameters, including them as controls would introduce collinearity and create post-treatment bias. Identifying exactly how the features used in the reported model map to the true dimensions on which emotion, audition, and vision are organized will require future research.

An exploratory intermodal representational similarity analysis found that visual areas represented both stimulus feature and

Figure 4. Representational dissimilarity matrices
Columns and rows share labels. "Biggest gap," "search one," and "search four" are "neutral" emotions.

emotion judgment predictors when musical stimuli were presented. However, most reliable neural activity in these areas was modality specific, as indicated by a low intermodal noise ceiling. Previous studies have shown multimodal processing in unimodal areas,[43–45] which may depend on projections between unimodal areas.[46–48] The reported results extend this account by showing that crossmodal perception is the product not only of operations in association cortices or activity dependent on inter-areal projections but of the use of a representational geometry that is shared across modalities.

The reported findings in pSTG are near previously reported pSTS activation during action understanding,[49,50] emotion perception,[51–53] affective and linguistic prosody recognition,[54] and crossmodal perception and recognition tasks.[55,56] Interestingly, the reported results were right lateralized, similar to previous findings on prosody recognition.[54] Damage to the pSTS does not impair voice recognition,[57] suggesting its representations are downstream of feature detectors. Alongside these results, the reported findings are consistent with the hypothesis that the pSTG or pSTS acts as a hub for transforming unimodal inputs into a common supramodal representation.[58]

Evoked emotion

Although our participants perceived emotions in our stimuli, it is unlikely that our stimuli evoked emotions in our participants. This disjunction highlights the complex and sometimes paradoxical relationship between perceived and evoked emotion. For example, perceiving sadness in music can evoke feelings of romance or pleasure.[59] The gap between perception and feeling has been theorized in terms of direct versus vicarious emotions[59] and in terms of emotion modules serving complementary functions.[60,61] Another possibility is that perceptual representations of stimulus features and emotion content interact with regions that produce context-sensitive appraisals and emotion experiences, such as those identified by Skerry and Saxe,[33] and subcortical regions sensitive to emotion content, including the amygdala.[62] Activation of these appraisal- and experience-related regions may not be necessary for making simple judgments of emotion content from stimulus features, possibly accounting for their absence in our results.

However, perceptual representations of emotion may also be linked to evoked emotions. Saarimäki et al.[63] showed that emotions evoked by listening to short stories produced activity in visual cortex, suggesting that evoked emotions can activate associated sensory representations. This may be a special case of the more general principle that mental imagery and episodic memory depend in part on activity in sensory regions associated with similar experiences.[64] Accordingly, perceptual representations of emotion content may form over development in a process similar to memory consolidation. This developmental process may be guided by language, supporting culture-specific particularity.[65,66] Activation of perceptual representations of emotion by imagined emotion experience could play an important role in art and music by allowing artists and composers to iteratively check whether their artistic products correspond with their perceptual representations.

Systematicity, iconicity, and conceptual scope

The neural representational system identified here is likely involved in phenomena beyond emotion perception, raising an interesting question: what concepts can and cannot be communicated via combinations of crossmodal stimulus features? If feature combinations in music and movement are symbolic, like words in natural language, then we would expect stimulus feature combinations that refer to abstract, non-emotional concepts. Just as arbitrary sequences of phonemes can point to "the housing market" or "editorial policy," arbitrary combinations of stimulus features should be able to do the same.

But strikingly, music and movement do not operate wholly on an arbitrary, symbolic basis. Music and movement systematically use variation in the magnitudes of stimulus features to communicate variation in the magnitudes of concepts which the stimulus iconically resembles.[7,15,67] For example, participants in the present study perceived mixes of the features for "happy" and the features for "sad" as expressing emotions on a continuum between happiness and sadness, with this pattern

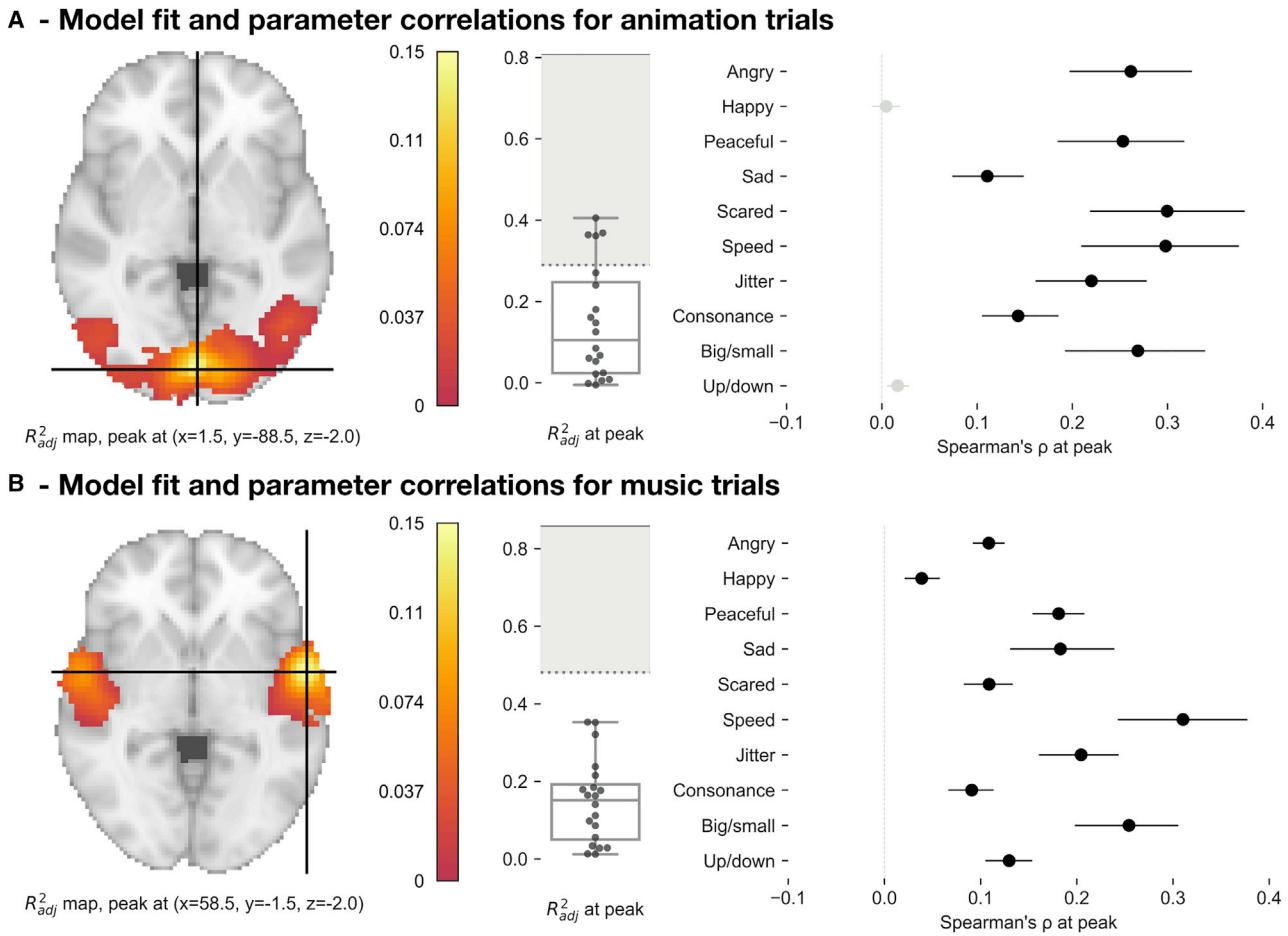


Figure 5. Model fits

Maps of the mean coefficient of determination (R^2_{adj}) across participants. Maps thresholded at FWER = 0.05. Boxplots show the median, quartiles, and range of the per-participant R^2_{adj} values at the location of best model fit at the group level. The dotted line indicates the lower bound of the noise ceiling and the solid line the upper bound. For per-parameter Spearman's ρ and β weights, see Figures S1, S2, S4, and S5 and Table S1. For an assessment of multicollinearity and variance inflation factors, see Figure S6.

generalizing across emotion pairs. The present results suggest this systematic mixing is made possible in part by a crossmodally shared neural representational geometry.

The systematic and iconic properties of musical communication may partially account for its use in expressing emotion, even across different cultural contexts. For example, the stimulus generator for the present study was previously used to show that the same combinations of music and movement features express the same emotions in both the United States and a small-scale society in rural Cambodia.[7] It may be that the forms of emotional music and movement are fixed by iconic, functional relationships that are shared across cultures.[61,68] This may be why lullabies are slow and consonant across a global sample of ethnographic reports and recordings[3] and why high emotional arousal is expressed using harsh-sounding, high spectral centroid sounds, even across species of terrestrial vertebrates.[69]

Music, movement, and crossmodal neural representations may be well suited to communicate iconic concepts that vary in magnitude, whereas language may be well suited to communicate symbolic concepts that vary in kind. Similarly, it may be that iconic communication tends to generalize across cultures while symbolic communication tends to be more culture specific. For example, previous research has shown cross-cultural generalization of valence perception, but not categorical emotion (though categorical emotions can also be shared across cultures[70]).[9] Further, musical narrative built from contrasting sets of stimulus features[71] shows large cross-cultural variability in interpretation,[10] unlike the present stimuli,[7] which contained no such contrasts.

Importantly, the present study did not test any hypotheses across cultures. Future research will need to explore broad areas of concept space across many cultures, collect free responses from participants, characterize culture-specific emotion concepts, and contrast concepts that vary in magnitude with concepts that vary in kind.

Direct perception

The results support the environmental conjunctions hypothesis (A2) that sensory brain regions represent task-relevant combinations of stimulus features, reducing the need for downstream

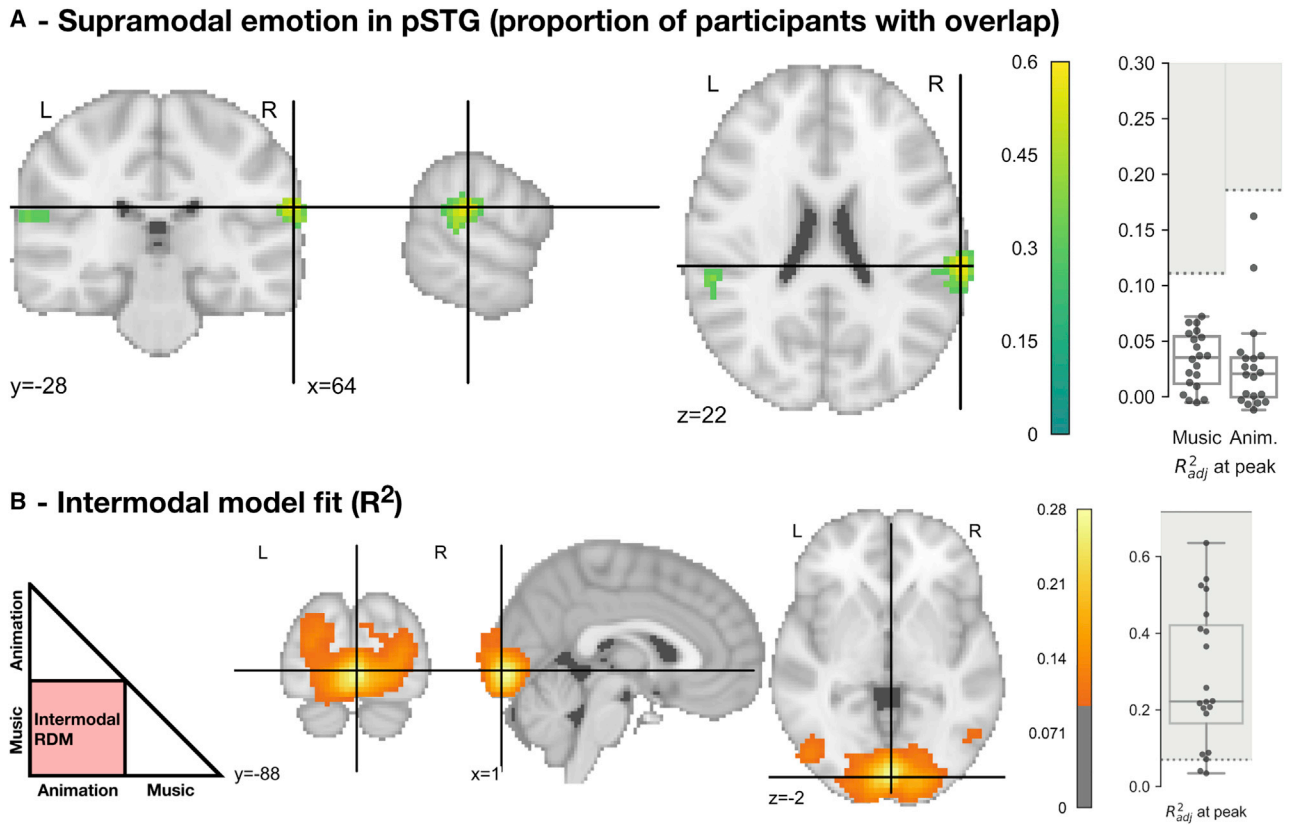


Figure 6. Results across modalities

(A) Supramodal emotion in pSTG. Maps show the proportion of participants representing emotion in music and animation in the same brain areas, thresholded at voxelwise FWER = 0.05. Boxplots show the median, quartiles, and range of R^2_{adj} for music and animation trials at the marked peak. See also Figure S3 and Table S1.

(B) Intermodal RSA model fit. Maps show areas that represented emotional stimuli even when presented in the area's non-preferred modality (STAR Methods), thresholded at voxelwise FWER = 0.05. R^2_{adj} values below 0.1 are hidden for visual clarity. Boxplot shows the median, quartiles, and range of per-participant R^2_{adj} values at the marked peak. Dotted lines indicate the lower bound of the noise ceiling, while solid lines indicate the upper bound.

See also Table S1.

inferential processing and acting as a shortcut for making important judgments. These representations may provide a neural basis for the direct perception of social information[20,37]—exemplified here by emotion judgment and potentially covering a range of other phenomena. Importantly, the simple features hypothesis (A1) was also supported, suggesting that direct perception and inferential processing systems coexist and may interact.

Without contextualizing narrative, judgments of emotion content in music and movement depend on configurations of stimulus features[7,15] in much the same way that the solution to a puzzle depends on the configuration of the pieces. In other words, stimulus features and emotion judgments are naturally confounded. The crux of the environmental conjunctions hypothesis (A2) is that any combination of features that is sufficiently confounded with a target is useful for identifying that target. We argue that the brain uses such natural confounds as a shortcut to make task-relevant judgments: if sensory regions represent feature combinations that are perfectly confounded with a target's identity, downstream inferential processing may not be necessary to identify the target.

In demonstrating support for the environmental conjunctions hypothesis (A2), we do not mean to suggest that sensory brain areas alone support purely conceptual, symbolic, or cognitive labeling of emotions. Previous studies using context-dependent and narrative stimuli have demonstrated the importance of inferential processing for emotion perception.[33,72] Further, inferential processing may play a role in the gradual tuning of perceptual systems for direct perception across development. Simple adaptations for perceiving cross-sensory magnitude or position information[73,74] or for adaptive signaling[75–78] may work in concert with learning,[79–82] language,[66] and cultural evolution[83] processes to support the development of task-relevant representations. This arrangement could flexibly accommodate culture-specific emotion concepts and display rules.[11–13]

Such tuning of sensory representations to the features used to communicate and categorize emotions shows that the need to identify such signals has exerted a profound shaping force on perceptual processes. We do not see or hear the actions of others as raw sense impressions first, later decode their conceptual content, and finally make an abstract emotion judgment. Rather, we begin accumulating evidence for emotion judgments from the lowest levels of sensory processing.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
- METHOD DETAILS
  - Stimuli
  - Experiment 1 (emotion evaluation)
  - Experiment 2 (fMRI of emotion viewing)
  - fMRI acquisition
  - fMRI preprocessing
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Post hoc power analysis
  - Representational similarity analysis
  - Intermodal RSA
  - Model-free similarity analysis
  - Noise ceiling
  - Overlap maps
  - Multiple comparisons correction
  - Emotion judgments permutation procedure

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cub.2021.09.043.

ACKNOWLEDGMENTS

We thank Sam Nasatase, Matteo Visconti di Oleggio Castello, J. Swaroop Guntupalli, and Joshua Greene for helpful comments during the writing process and Paulina Calcaterra, Rebecca Drapkin, Caitlyn Lee, Elizabeth Reynolds, Tshibambe Nathanael Tshimbombu, and Kelsey Wheeler for assistance collecting fMRI data. This research was supported in part by the Nelson A. Rockefeller Center for Public Policy at Dartmouth, the John Templeton Foundation, the Neukom Institute for Computational Science, the Vision Science to Applications (VISTA) program funded by the Canada First Research Excellence Fund (CFREF, 2016–2023), and the Natural Sciences and Engineering Research Council of Canada.

AUTHOR CONTRIBUTIONS

Conceptualization, B.S. and T.W.; data curation, B.S.; formal analysis, B.S.; funding acquisition, B.S. and T.W.; investigation, B.S. and C.P.; methodology, B.S., C.P., P.J.K., J.M.H., and S.V.F.; project administration, B.S.; software, B.S.; visualization, B.S.; writing – original draft, B.S.; writing – review & editing, B.S., C.P., P.J.K., J.M.H., S.V.F., and T.W.; resources, T.W.; supervision, T.W.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: March 2, 2021
Revised: July 7, 2021
Accepted: September 16, 2021
Published: October 12, 2021

REFERENCES

1. Kaeppler, A.L. (1978). Dance in anthropological perspective. Annu. Rev. Anthropol. *7*, 31–49.
2. Savage, P.E., Brown, S., Sakai, E., and Currie, T.E. (2015). Statistical universals reveal the structures and functions of human music. Proc. Natl. Acad. Sci. USA *112*, 8987–8992.
3. Mehr, S.A., Singh, M., Knox, D., Ketter, D.M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A.A., Hopkins, E.J., et al. (2019). Universality and diversity in human song. Science *366*, eaax0868.
4. Baily, J. (1985). Music structure and human movement. In Musical Structure and Cognition, P. Howell, I. Cross, and R. West, eds. (Academic), pp. 237–258.
5. Trehub, S.E., Becker, J., and Morley, I. (2015). Cross-cultural perspectives on music and musicality. Philos. Trans. R. Soc. Lond. B Biol. Sci. *370*, 20140096.
6. Phillips-Silver, J., and Trainor, L.J. (2005). Feeling the beat: movement influences infant rhythm perception. Science *308*, 1430.
7. Sievers, B., Polansky, L., Casey, M., and Wheatley, T. (2013). Music and movement share a dynamic structure that supports universal expressions of emotion. Proc. Natl. Acad. Sci. USA *110*, 70–75.
8. Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., Friederici, A.D., and Koelsch, S. (2009). Universal recognition of three basic emotions in music. Curr. Biol. *19*, 573–576.
9. Gendron, M., Roberson, D., van der Vyver, J.M., and Barrett, L.F. (2014). Perceptions of emotion from facial expressions are not culturally universal: evidence from a remote culture. Emotion *14*, 251–262.
10. Margulis, E.H., Wong, P.C.M., Simchy-Gross, R., and McAuley, J.D. (2019). What the music said: narrative listening across cultures. Palgrave Commun. *5*, 146.
11. Yuki, M., Maddux, W.W., and Masuda, T. (2007). Are the windows to the soul the same in the East and West? Cultural differences in using the eyes and mouth as cues to recognize emotions in Japan and the United States. J. Exp. Soc. Psychol. *43*, 303–311.
12. Jack, R.E., Caldara, R., and Schyns, P.G. (2012). Internal representations reveal cultural diversity in expectations of facial expressions of emotion. J. Exp. Psychol. Gen. *141*, 19–25.
13. Jack, R.E., Sun, W., Delis, I., Garrod, O.G.B., and Schyns, P.G. (2016). Four not six: revealing culturally common facial expressions of emotion. J. Exp. Psychol. Gen. *145*, 708–730.
14. Jackson, J.C., Watts, J., Henry, T.R., List, J.M., Forkel, R., Mucha, P.J., Greenhill, S.J., Gray, R.D., and Lindquist, K.A. (2019). Emotion semantics show both cultural variation and universal structure. Science *366*, 1517–1522.
15. Sievers, B., Lee, C., Haslett, W., and Wheatley, T. (2019). A multi-sensory code for emotional arousal. Proc. Biol. Sci. *286*, 20190513.
16. Kriegeskorte, N., and Kievit, R.A. (2013). Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. *17*, 401–412.
17. Peelen, M.V., Atkinson, A.P., and Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. J. Neurosci. *30*, 10127–10134.
18. Gordon, R.M. (1986). Folk psychology as simulation. Mind Lang. *1*, 158–171.
19. Gopnik, A., and Wellman, H.M. (1994). The theory theory. In Mapping the Mind: Domain Specificity in Cognition and Culture, L.A. Hirschfeld, and S.A. Gelman, eds. (Cambridge University), pp. 257–293.
20. Gallagher, S. (2008). Direct perception in the intersubjective context. Conscious. Cogn. *17*, 535–543.
21. Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. Proc. Natl. Acad. Sci. USA *103*, 3863–3868.

22. Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. *2*, 4.

23. Roskies, A.L. (2021). Representational similarity analysis in neuroimaging: proxy vehicles and provisional representations. Synthese. Published online February 9, 2021. https://doi.org/10.1007/s11229-021-03052-4.

24. Lindquist, K.A., Wager, T.D., Kober, H., Bliss-Moreau, E., and Barrett, L.F. (2012). The brain basis of emotion: a meta-analytic review. Behav. Brain Sci. *35*, 121–143.

25. Wager, T.D., Kang, J., Johnson, T.D., Nichols, T.E., Satpute, A.B., and Barrett, L.F. (2015). A Bayesian model of category-specific emotional brain responses. PLoS Comput. Biol. *11*, e1004066.

26. Adolphs, R., Tranel, D., Damasio, H., and Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. Nature *372*, 669–672.

27. Tsuchiya, N., Moradi, F., Felsen, C., Yamazaki, M., and Adolphs, R. (2009). Intact rapid detection of fearful faces in the absence of the amygdala. Nat. Neurosci. *12*, 1224–1225.

28. Phillips, M.L., Young, A.W., Senior, C., Brammer, M., Andrew, C., Calder, A.J., Bullmore, E.T., Perrett, D.I., Rowland, D., Williams, S.C.R., et al. (1997). A specific neural substrate for perceiving facial expressions of disgust. Nature *389*, 495–498.

29. Calder, A.J., Lawrence, A.D., and Young, A.W. (2001). Neuropsychology of fear and loathing. Nat. Rev. Neurosci. *2*, 352–363.

30. Norman, K.A., Polyn, S.M., Detre, G.J., and Haxby, J.V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. *10*, 424–430.

31. Chikazoe, J., Lee, D.H., Kriegeskorte, N., and Anderson, A.K. (2014). Population coding of affect across stimuli, modalities and individuals. Nat. Neurosci. *17*, 1114–1122.

32. Kim, J., Shinkareva, S.V., and Wedell, D.H. (2017). Representations of modality-general valence for videos and music derived from fMRI data. Neuroimage *148*, 42–54.

33. Skerry, A.E., and Saxe, R. (2015). Neural representations of emotion are organized around abstract event features. Curr. Biol. *25*, 1945–1954.

34. Lang, P.J., Bradley, M.M., and Cuthbert, B.N. (2008). International Affective Picture System (IAPS): Affective Ratings of Pictures and Instruction Manual. Technical Report A-8 (University of Florida).

35. Cowen, A.S., and Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. Proc. Natl. Acad. Sci. USA *114*, E7900–E7909.

36. Destrieux, C., Fischl, B., Dale, A., and Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. Neuroimage *53*, 1–15.

37. Chemero, A. (2006). Information and direct perception: a new approach. In Advanced Issues in Cognitive Science and Semiotics, P. Farias, and J. Queiroz, eds. (Shaker Verlag), pp. 59–72.

38. Russell, J.A. (1980). A circumplex model of affect. J. Pers. Soc. Psychol. *39*, 1161–1178.

39. Levy, D.J., and Glimcher, P.W. (2012). The root of all value: a neural common currency for choice. Curr. Opin. Neurobiol. *22*, 1027–1038.

40. Shuster, A., and Levy, D.J. (2018). Common sense in choice: the effect of sensory modality on neural value representations. eNeuro *5*, 1–14.

41. Riesenhuber, M., and Poggio, T. (1999). Hierarchical models of object recognition in cortex. Nat. Neurosci. *2*, 1019–1025.

42. Nishimoto, S., Vu, A.T., Naselaris, T., Benjamini, Y., Yu, B., and Gallant, J.L. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. Curr. Biol. *21*, 1641–1646.

43. Bulkin, D.A., and Groh, J.M. (2006). Seeing sounds: visual and auditory interactions in the brain. Curr. Opin. Neurobiol. *16*, 415–419.

44. Ghazanfar, A.A., and Schroeder, C.E. (2006). Is neocortex essentially multisensory? Trends Cogn. Sci. *10*, 278–285.

45. Kayser, C., and Logothetis, N.K. (2007). Do early sensory cortices integrate cross-modal information? Brain Struct. Funct. *212*, 121–132.

46. Falchier, A., Clavagnier, S., Barone, P., and Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. J. Neurosci. *22*, 5749–5759.

47. Rockland, K.S., and Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. Int. J. Psychophysiol. *50*, 19–26.

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

49. Beauchamp, M.S., Lee, K.E., Argall, B.D., and Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. Neuron *41*, 809–823.

50. Wyk, B.C.V., Hudac, C.M., Carter, E.J., Sobel, D.M., and Pelphrey, K.A. (2009). Action understanding in the superior temporal sulcus region. Psychol. Sci. *20*, 771–777.

51. Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., and Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: an event-related fMRI study. Neuroimage *37*, 1445–1456.

52. Robins, D.L., Hunyadi, E., and Schultz, R.T. (2009). Superior temporal activation in response to dynamic audio-visual emotional cues. Brain Cogn. *69*, 269–278.

53. Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., and Belin, P. (2014). Crossmodal adaptation in right posterior superior temporal sulcus during face-voice emotional integration. J. Neurosci. *34*, 6813–6821.

54. Belyk, M., and Brown, S. (2014). Perception of affective and linguistic prosody: an ALE meta-analysis of neuroimaging studies. Soc. Cogn. Affect. Neurosci. *9*, 1395–1403.

55. Wright, T.M., Pelphrey, K.A., Allison, T., McKeown, M.J., and McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. Cereb. Cortex *13*, 1034–1043.

56. Werner, S., and Noppeney, U. (2010). Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. Cereb. Cortex *20*, 1829–1842.

57. Jiahui, G., Garrido, L., Liu, R.R., Susilo, T., Barton, J.J.S., and Duchaine, B. (2017). Normal voice processing after posterior superior temporal sulcus lesion. Neuropsychologia *105*, 215–222.

58. Schirmer, A., and Adolphs, R. (2017). Emotion perception from face, voice, and touch: comparisons and convergence. Trends Cogn. Sci. *21*, 216–228.

59. Kawakami, A., Furukawa, K., Katahira, K., and Okanoya, K. (2013). Sad music induces pleasant emotion. Front. Psychol. *4*, 311.

60. Gelstein, S., Yeshurun, Y., Rozenkrantz, L., Shushan, S., Frumin, I., Roth, Y., and Sobel, N. (2011). Human tears contain a chemosignal. Science *331*, 226–230.

61. Mehr, S.A., Krasnow, M.M., Bryant, G.A., and Hagen, E.H. (2020). Origins of music in credible signaling. Behav. Brain Sci. Published August 26, 2020. https://doi.org/10.1017/S0140525X20000345.

62. Wang, S., Tudusciuc, O., Mamelak, A.N., Ross, I.B., Adolphs, R., and Rutishauser, U. (2014). Neurons in the human amygdala selective for perceived emotion. Proc. Natl. Acad. Sci. USA *111*, E3110–E3119.

63. Saarimäki, H., Ejtehadian, L.F., Glerean, E., Jääskeläinen, I.P., Vuilleumier, P., Sams, M., and Nummenmaa, L. (2018). Distributed affective space represents multiple emotion categories across the human brain. Soc. Cogn. Affect. Neurosci. *13*, 471–482.

64. Wheeler, M.E., Petersen, S.E., and Buckner, R.L. (2000). Memory's echo: vivid remembering reactivates sensory-specific cortex. Proc. Natl. Acad. Sci. USA *97*, 11125–11129.

65. Barrett, L.F., Lindquist, K.A., and Gendron, M. (2007). Language as context for the perception of emotion. Trends Cogn. Sci. *11*, 327–332.

66. Hoemann, K., Xu, F., and Barrett, L.F. (2019). Emotion words, emotion concepts, and emotional development in children: a constructionist hypothesis. Dev. Psychol. *55*, 1830–1849.

67. Spector, F., and Maurer, D. (2009). Synesthesia: a new approach to understanding the development of perception. Dev. Psychol. *45*, 175–189.
68. Sievers, B., and Wheatley, T. (2021). Rapid dissonant grunting, or, but why does music sound the way it does? PsyArXiv. https://doi.org/10.31234/osf.io/89d2h.
69. Filippi, P., Congdon, J.V., Hoang, J., Bowling, D.L., Reber, S.A., Pašukonis, A., Hoeschele, M., Ocklenburg, S., de Boer, B., Sturdy, C.B., et al. (2017). Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: evidence for acoustic universals. Proc. Biol. Sci. *284*, 1–9.
70. Parkinson, C., Walker, T.T., Memmi, S., and Wheatley, T. (2017). Emotions are understood from biological motion across remote cultures. Emotion *17*, 459–477.
71. Margulis, E.H. (2017). An exploratory study of narrative experiences of music. Music Percept. *35*, 235–248.
72. Barrett, L.F., Mesquita, B., and Gendron, M. (2011). Context in emotion perception. Curr. Dir. Psychol. Sci. *20*, 286–290.
73. Deneux, T., Harrell, E.R., Kempf, A., Ceballo, S., Filipchuk, A., and Bathellier, B. (2019). Context-dependent signaling of coincident auditory and visual events in primary visual cortex. eLife *8*, 1–23.
74. Murphy, D.L.K., King, C.D., Schlebusch, S.N., Shera, C.A., and Groh, J.M. (2020). Evidence for a system in the auditory periphery that may contribute to linking sounds and images in space. bioRxiv. https://doi.org/10.1101/2020.07.19.210864.
75. Hebets, E.A., Barron, A.B., Balakrishnan, C.N., Hauber, M.E., Mason, P.H., and Hoke, K.L. (2016). A systems approach to animal communication. Proc. Biol. Sci. *283*, 20152889.
76. Johnstone, R.A. (1996). Multiple displays in animal communication: 'backup signals' and 'multiple messages'. Philos. Trans. R. Soc. B: Biol. Sci. *351*, 329–338.
77. Johnstone, R.A. (1997). The evolution of animal signals. In Behavioral Ecology, J.R. Krebs, and N.B. Davies, eds. (Oxford University), pp. 155–178.
78. Huron, D. (2012). Understanding music-related emotion: lessons from ethology. In Proceedings of the 12th International Conference on Music Perception and Cognition and the 8th Triennial Conference of the European Society for the Cognitive Sciences of Music, E. Cambouropoulos, C. Tsougras, P. Mavromatis, and K. Pastiadis, eds., pp. 473–481.
79. Saffran, J.R., Aslin, R.N., and Newport, E.L. (1996). Statistical learning by 8-month-old infants. Science *274*, 1926–1928.
80. Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav. Brain Sci. *36*, 181–204.
81. Kok, P., Brouwer, G.J., van Gerven, M.A.J., and de Lange, F.P. (2013). Prior expectations bias sensory representations in visual cortex. J. Neurosci. *33*, 16275–16284.
82. de Lange, F.P., Heilbron, M., and Kok, P. (2018). How do expectations shape perception? Trends Cogn. Sci. *22*, 764–779.
83. Laland, K.N., Odling-Smee, J., and Feldman, M.W. (2000). Niche construction, biological evolution, and cultural change. Behav. Brain Sci. *23*, 131–146, discussion 146–175.
84. Sievers, B., Parkinson, C., Kohler, P.J., Hughes, J., Fogelson, S.V., and Wheatley, T. (2021). Visual and auditory brain areas share a representational structure that supports emotion perception: fMRI data. https://openneuro.org/datasets/ds003715/versions/1.0.0.
85. Sievers, B. (2021). Visual and auditory brain areas share a representational structure that supports emotion perception: code and materials. https://osf.io/kvbqm/.
86. Zicarelli, D.D. (1998). An extensible real-time signal processing environment for Max. In Proceedings of the 1998 International Computer Music Conference, pp. 463–466.
87. Peirce, J.W. (2007). PsychoPy–psychophysics software in Python. J. Neurosci. Methods *162*, 8–13.
88. Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav. Res. Methods *39*, 175–191.
89. Hanke, M., Halchenko, Y.O., Sederberg, P.B., Hanson, S.J., Haxby, J.V., and Pollmann, S. (2009). PyMVPA: a python toolbox for multivariate pattern analysis of fMRI data. Neuroinformatics *7*, 37–53.
90. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Louppe, G., Prettenhofer, P., Weiss, R., et al. (2011). Scikit-learn: machine learning in Python. J. Mach. Learn. Res. *12*, 2825–2830.
91. Oliphant, T.E. (2006). A Guide to NumPy (Trelgol).
92. Jones, E., Oliphant, E., and Peterson, P. (2001). SciPy: open source scientific tools for Python.
93. Peres-Neto, P.R., Legendre, P., Dray, S., and Borcard, D. (2006). Variation partitioning of species data matrices: estimation and comparison of fractions. Ecology *87*, 2614–2625.
94. Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., and Smith, S.M. (2012). FSL. Neuroimage *62*, 782–790.
95. Nichols, T.E., and Holmes, A.P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. *15*, 1–25.
96. Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A toolbox for representational similarity analysis. PLoS Comput. Biol. *10*, e1003553.
97. Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., Gramfort, A., Thirion, B., and Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. Front. Neuroinform. *8*, 14.
98. Saad, Z.S., Reynolds, R.C., Argall, B., Japee, S., and Cox, R.W. (2004). SUMA: an interface for surface-based intra- and inter-subject analysis with AFNI. In 2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE), pp. 1510–1513.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Neuroimaging data	Sievers et al.[84]	https://doi.org/10.18112/openneuro.ds003715.v1.0.0
Software and algorithms		
Preprocessing and analysis code	Sievers[85]	https://doi.org/10.17605/OSF.IO/KVBQM
Other		
Auditory and visual stimuli	Sievers[85]	https://doi.org/10.17605/OSF.IO/KVBQM

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Thalia Wheatley (thalia.p.wheatley@dartmouth.edu).

Materials availability

Stimuli have been deposited to osf.io.[85] DOIs and accession numbers are listed in the Key resources table.

Data and code availability

De-identified fMRI data have been deposited to OpenNeuro.[84] All original code has been deposited to osf.io.[85] DOIs are listed in the Key resources table. Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

79 participants (47 female) were recruited from the Dartmouth College student community to participate in the emotion evaluation task (experiment 1). 20 of these participants (11 female) also participated in the fMRI of emotion viewing task (experiment 2). All fMRI participants were right-handed and had normal or corrected-to-normal vision. All participants provided written informed consent, and the study was approved by the Dartmouth College Committee for the Protection of Human Subjects.

METHOD DETAILS

Stimuli

Emotion stimuli were generated using a model developed for a prior study[7] that used movement across a number line to create both music (simple piano melodies) and animated movement (a bouncing ball). The model had five stimulus feature parameters: speed, irregularity/jitter, consonance/spikiness, ratio of big-to-small movements, and ratio of upward-to-downward movements. Each time the model was run, it probabilistically generated a new stimulus based on its current parameter settings. Participants in Sievers et al.[7] (music N = 25; movement N = 25; total N = 50) used this model to communicate five prototype emotions: Angry, Happy, Peaceful, Sad, and Scared. Critically, participants were split into separate music and movement groups, each of which had no knowledge of the other. Participants chose similar music and movement parameter settings for each emotion across modalities, showing that music and movement share an underlying structure. The median parameter settings across music and movement from the United States participants in Sievers et al.[7] were used to generate the stimuli used in the present studies. All stimuli are available at https://osf.io/kvbqm/.

In addition to the prototype emotions, mixed emotion stimuli were created by interpolating linearly between the parameter settings for each prototype emotion pair; 25%, 50%, and 75% mixes were used. We also added three putatively "neutral" or "non-emotional" parameter settings that were selected to be distant from all other stimuli. "Search One" and "Search Four" were selected by a Monte Carlo search algorithm, and consisted of extreme values for all five parameters. "Biggest Gap" was created by selecting the midpoint of the largest gap between the five prototype emotions and the stimulus feature parameter endpoints.

For each prototype, mixed, and "non-emotional" parameter setting in each modality, we probabilistically generated 20 exemplars, for a total of 1,520 stimuli (38 emotions x 2 modalities x 20 exemplars). To eliminate the possibility of generating unusual outlier stimuli,

each candidate exemplar was compared to a larger, separate sample of 5000 same-emotion exemplars, and was re-generated if found to be further than one standard deviation from the emotion mean along any parameter.

Experiment 1 (emotion evaluation)

Participants (N = 79, 47 female) evaluated the emotion content of the stimuli. Stimuli were presented using a computer program developed using Max/MSP version 5[86] that displayed five slider bars, one for each emotion prototype (Angry, Happy, Peaceful, Sad, and Scared). The on-screen order of slider bars and emotion stimuli were randomized across participants. Participants viewed or listened to each stimulus at least three times, and were asked "to evaluate the amount or intensity of emotion expressed by the music or animation by positioning the slider bars."

Experiment 2 (fMRI of emotion viewing)

During each fMRI run, participants (N = 20, 11 female) viewed 18 randomly selected exemplars from each of the 76 stimulus classes described above. Each stimulus class was shown once per run, and participants completed 18 runs across 3 separate scanning sessions (~3 hours of scan time, 1,368 stimulus impressions). Each scan session was scheduled for approximately the same time of day, and no more than one week elapsed between scan sessions.

Stimuli were truncated to 3 s in duration and followed by fixation periods of randomly varying duration (range: 0.5 s–20 s). The ratio of stimulus presentation to fixation was 1:1. A Monte Carlo procedure was used to select separate, optimized stimulus presentation orderings and timings for each participant. This procedure used AFNI make_random_timing.py to generate thousands of possible stimulus timings, and AFNI 3dDeconvolve to select the timings that best supported deconvolving unique patterns of brain activity for each stimulus. Stimuli were presented using PsychoPy version 1.84.2.[87] Participants were instructed to attend to the emotion content of the stimuli. During randomly interspersed catch trials (10 per run), participants used a button box to rate on a four-point scale whether the most recently presented stimulus had emotion content that was "more mixed" or "more pure." To ensure familiarity with the stimuli, all fMRI participants had previously completed the emotion evaluation task.

fMRI acquisition

Participants were scanned at the Dartmouth Brain Imaging Center using a 3T Phillips Achieva Intera scanner with a 32-channel head coil. Functional images were acquired using an echo-planar sequence (35ms TE, 3000ms TR; 90° flip angle; 3x3x3mm resolution) with 192 dynamic scans per run. A high resolution T1-weighted anatomical scan (3.7 ms TE; 8200ms TR; 0.938x.938x1mm resolution) was acquired at the end of each scanning session. Sound was delivered using an over-ear headphone system. Foam padding was placed around participants' heads to minimize motion.

fMRI preprocessing

Anatomical images were skull-stripped and aligned to the last TR of the last EPI image using AFNI align_epi_anat.py. EPI images were motion corrected and aligned to the last TR of the last EPI image using AFNI 3dvolreg. Rigid body transformations for aligning participants' anatomical and EPI images to the AFNI version of the MNI 152 ICBM template were calculated using AFNI @auto_tlrc. Alignment transformations were concatenated and applied in a single step using AFNI 3dAllineate. EPI images were scaled to show percent signal change and concatenated. EPI images were not smoothed. TRs where inter-TR motion exceeded a euclidean norm threshold of 0.3 were censored, along with the immediately preceding TR. The general linear model was used to estimate BOLD-responses evoked by each of the 76 emotional stimulus classes using AFNI 3dREMLfit. All six demeaned motion parameters as well as polynomial trends were included as regressors of no interest.

QUANTIFICATION AND STATISTICAL ANALYSIS

Post hoc power analysis

Because the present study is the first to use the reported paradigm, we did not conduct an *a priori*/prospective power analysis. Because accurate assessment of effect size is impossible without stable patterns, we prioritized having a large number of fMRI trials per participant. The number of trials per stimulus class per participant was determined by consulting studies that used similar analysis methods (MVPA/RSA). E.g., Peelen et al.[17] used 12 trials per class per participant for 18 participants, Kim et al.[32] used 10 trials per class per participant for 20 participants, and the present study used 18 trials per class per participant for 20 participants. A post hoc/retrospective power analysis using G*Power 3.1[88] showed that this provided 85% power for music trials and 99% power for animation trials for the effects reported in Figure 3.

Representational similarity analysis

Representational similarity analysis (RSA)[21,22] was conducted using PyMVPA,[89] Scikit-Learn,[90] NumPy,[91] and SciPy.[92] Stimulus feature representational distance matrices (RDMs) for each parameter (speed, irregularity/jitter, consonance/spikiness, ratio of big-to-small movements, ratio of upward-to-downward movements) were created by aggregating the Euclidean distances between the mean slider bar settings for each pair of emotions, including mixed emotions. Both music and animation stimuli were created using the same slider bar settings for each emotion, making it unnecessary to create modality-specific feature RDMs. Emotion RDMs were created by aggregating the Euclidean distances between the mean of each emotion judgment parameter in experiment

1 (Angry, Happy, Peaceful, Sad, and Scared) for each pair of emotions, including mixed emotions. Emotion judgments were averaged across music and animation, making it unnecessary to create modality-specific emotion judgment RDMs. Intermodal RDMs were built by calculating the full multi-modality RDM including both music and animation stimuli and selecting its lower-left square region (Figure 6B).

Representational similarity analysis was conducted separately for music trials, animation trials, and (for the intermodal analysis) music and animation trials together. Each analysis used a spherical searchlight with a 3-voxel (9mm) radius. In each searchlight sphere, music and animation neural RDMs were created by aggregating the ranked correlation distances (1-Spearman's ρ) between the estimated stimulus-evoked pattern of BOLD activation for each emotion. The use of correlation distance ensured that the analysis would not mistake differences in the mean level of BOLD activity across music and animation trials for differences in representational similarity. Intermodal neural RDMs were created as described above, using neural data instead of stimulus features or emotion judgments (Figure 6B). The fit of the model to stimulus-evoked patterns of BOLD activation was assessed using multiple regression, with the ranked model RDMs as predictors and the neural RDM as the target. This produced coefficient of determination (R^2) and β weight maps for each participant and each analysis. R^2 values were adjusted using a permutation approach:[93] Multiple regression was performed an additional 1000 times with randomly selected permutations of each predictor, and the mean R^2 from this null distribution was subtracted from the reported R^2 values (R^2_{adj}). Multiple regression β weights reflect only the unique contribution of each predictor, resulting in β weight maps that do not reflect the shared contributions of correlated predictors. To assess the contribution of individual predictors we calculated the ranked correlation (Spearman's ρ) of each predictor to the neural RDM.

All group-level statistics (including R^2_{adj}, β weights, Spearman's ρ, p values, and any other values reported unless otherwise noted) were corrected for multiple comparisons using a maximum cluster mass sign-flipping permutation test performed with FSL randomize,[94,95] with a cluster-determining threshold of $p = 0.01$ and a family-wise error rate of 0.05.

Intermodal RSA

Intermodal RSA differed from the RSA analysis described above in that both the neural target RDM and the predictor RDMs used only between-modality distances, corresponding to the lower-left square region of the larger triangular RDM created using stimuli from both modalities (Figure 6B). If activity in a brain area was unrelated to stimuli presented in its non-preferred modality, then the intermodal neural RDM should be uncorrelated with the intermodal model RDMs. However, if a brain area was even weakly representing emotion content across modalities, then the intermodal neural RDM should be correlated with the intermodal model RDMs. Note that because this analysis only considered between-modality distances, it could not in principle have identified any modality-specific activity.

Model-free similarity analysis

To rule out the possibility that the identified brain regions were a good fit for the stimulus features and emotion judgments in the reported model, but did not truly share a representational geometry (i.e., were not directly similar to each other), we performed a permutation test of inter-region representational similarity. This test assessed whether the representations at the locations of peak model fit were more similar than representations at randomly selected locations. Analogously, the claim "San Francisco and Oakland are close to each other," is weaker than the claim "San Francisco and Oakland are closer to each other than 95% of all pairs of American cities." To build a null distribution, we randomly selected 2000 pairs of coordinates in the right hemisphere of the brain. For each coordinate pair, we measured the ranked correlation (Spearman's ρ) of the mean neural RDM for music trials at the first coordinate and the mean neural RDM for animation trials at the second coordinate. The mean inter-region similarity in the null distribution was $\rho = -.007$, whereas the inter-region similarity at the locations of peak model fit was $\rho = 0.68$ ($p < 0.001$), more similar than any pair of coordinates in the null distribution.

Noise ceiling

The upper and lower bounds of the noise ceiling were calculated using an approach based on Nili et al.,[96] but adapted for use with multiple regression. The approach described by Nili et al.[96] depends on a simple principle: for any dataset, the model that accounts for the most variance in the data will always be derived from the data itself. For correlation, this best-fitting model is the mean of the data. Given measurement error and individual differences across the dataset, no model could possibly outperform the mean, and so the model fit of the mean establishes a ceiling against which other models can be usefully compared. Analogously, for a multiple regression model with n predictors, the best-fitting model is the mean of the data along with the top n predictors identified using principal component analysis (PCA). No multiple regression model with the same number of predictors could possibly outperform this mean-and-PCA model. In the present study, the upper bound of the noise ceiling was calculated at each searchlight center by performing a multiple regression analysis that used the mean neural RDM and the top 10 principal components of the neural RDM as predictors. The lower bound of the noise ceiling was calculated using a leave-one-subject-out cross-validation approach: For each subject, the same multiple regression procedure was applied, but the mean neural RDM and the top 10 principal components were calculated with that subject left out.

Overlap maps

Overlap maps were created for each participant by identifying voxels where both music and animation model fits were significant at the individual level (permutation $p < 0.05$, uncorrected). Overlap maps were set to 1 if both model fits were significant, and

0 otherwise. Multiple comparisons correction of the overlap maps was performed at the group level (CDT = 0.01; FWER p = 0.05; see below), testing the proportion of individuals that showed overlap in a region against the null hypothesis that no participants showed overlap in that region.

Multiple comparisons correction

Group level maps were calculated and corrected for multiple comparisons using a maximum cluster mass sign-flipping permutation test FSL randomize[94,95] (cluster-determining threshold p = 0.01; family-wise error rate p = 0.05). Tests for R^2_{adj} were 1-sided. Tests for β weights and Spearman's ρ were two-sided. Maps were visualized using Nilearn[97] and AFNI SUMA.[98]

Emotion judgments permutation procedure

For each emotion, we averaged participants' emotion judgment ratings, yielding a class mean. We then calculated the Euclidean distance of each individual judgment to this mean, scaled by the maximum possible distance (determined by the limits of each slider), yielding a distribution of scaled distances to the mean for each stimulus class. A null distribution of scaled distances to the class means was created by applying this procedure 2000 times, each with a different permutation of the emotion labels over the whole dataset. Welch's independent samples t test was applied to test whether the observed distributions of scaled distances to class means differed from the null. This approach was chosen because it accounts for the simultaneous use of five rating scales and conservatively respects the dependency structure of the data.