Visual and auditory brain areas share a neural code for perceived emotion

Beau Sievers[1], Carolyn Parkinson[2], Peter J. Kohler[3], James Hughes, Sergey V. Fogelson, Thalia Wheatley[4]

[1]Department of Psychology, Harvard University, Cambridge, MA 02138

[2]Department of Psychology, University of California Los Angeles, Los Angeles, CA 90095

[3]Department of Psychology, York University, Toronto, Ontario

[4]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755

Correspondence to: Thalia Wheatley, 6207 Moore Hall, Dartmouth College, Hanover, NH 03755. thalia.p.wheatley@dartmouth.edu

Abstract

Emotional music and movement are human universals. Further, music and movement are subjectively linked: it is hard to imagine one without the other. One possible reason for the fundamental link between music and movement is that they are represented the same way in the brain, using a shared neural code. To test this, we created emotional music and animation stimuli that were precisely matched on all time-varying structural features. Participants viewed these stimuli while undergoing fMRI of the brain. Using representational similarity analysis (Kriegeskorte & Kievit, 2013), we show that a single model of stimulus features and emotion content fit activity in both auditory and visual brain areas, providing evidence that these regions share a neural code. Further, this code was used in posterior superior temporal cortex during both audition and vision. Across all regions, the shared code represented both prototypical and mixed emotions (e.g., Happy–Sad). Finally, exploratory analysis revealed that stimulus features and emotion content were represented in early visual areas even when stimuli were presented auditorily. This evidence for a shared neural code is consistent with an adaptive signaling account of emotion perception, where perceivers specifically adapted to perceive cross-sensory redundancy accrue an evolutionary advantage.

Introduction

Wherever there is music, there is movement (Kaeppler, 1978; Savage, Brown, Sakai, & Currie, 2015). It is hard to imagine one without the other, and some languages fuse music and dance together, using a single word for both (Baily, 1985). 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 (Phillips-silver & Trainor, 2005). Further, the communication of emotion through music and movement is one of very few human universals, behaviors that are shared even across otherwise dissimilar cultures (Fritz et al., 2009; Sievers, Polansky, Casey, & Wheatley, 2013; Trehub, Becker, Morley, & Trehub, 2015). We suggest that the universality of music and movement is not coincidental, but is due to fundamental similarities in how music and movement are structured and perceived. Supporting this account, previous research has shown that emotional music and movement share structural features: angry music and movement are both fast and move downward, peaceful music and movement are both slow and move upward, and so on (Sievers, Lee, Haslett, Wheatley, & Wheatley, 2019; Sievers et al., 2013). However, these shared features do not explain the subjective, experiential link between music and movement. Here, we test one possible explanation for this link: that the structural features shared by music and movement are also represented using a single, shared neural code that supports direct comparison of auditory and visual percepts. This hypothesis is consistent with an *adaptive signaling* account of emotion perception (Hebets et al., 2016; Johnstone, 1996, 1997). On this account, evolutionary pressure favors signal senders that broadcast important information across multiple sensory channels, as well as signal receivers that exploit this redundancy to improve perceptual speed and accuracy—perhaps by using a shared neural code.

We tested two hypotheses: (1) that both auditory and visual brain areas use a shared neural code for emotion—i.e., that when perceiving emotion, the vector space of patterns of activity in auditory regions is very similar to the vector space of patterns of activity in visual regions (Kriegeskorte & Kievit, 2013)—and (2) that auditory and visual emotions are represented together in one brain area using a supramodal neural code.

A model capturing both dynamic (i.e., time-varying) stimulus features and emotional meaning was found to fit activity in both auditory and visual areas, supporting hypothesis (1). The same model was found to fit activity in posterior superior temporal gyrus (pSTG) for both auditory and visual emotions, supporting hypothesis (2). Additional exploratory analysis showed that auditory and visual areas represent stimulus features and emotion content even when stimuli are presented in each area's non-preferred modality. These results support an adaptive signaling account of emotion perception, where the structure of emotional signals and the brains of receivers have adapted to tightly fit one another, facilitating efficient and reliable signal perception.

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 a range of emotions (Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012; Wager et al., 2015). However, certain aspects of emotion processing are tightly localized. Lesion and neuroimaging studies have demonstrated that some brain areas play a role in the processing of specific emotions; for example, the amygdala for the conscious recognition

of fearful stimuli (Adolphs, Tranel, Damasio, & Damasio, 1994; Tsuchiya, Moradi, Felsen, Yamazaki, & Adolphs, 2009), and the insula for recognizing disgust (Calder, Lawrence, & Young, 2001; Phillips et al., 1997).

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 between emotions 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 (Norman, Polyn, Detre, & Haxby, 2006) or representational similarity analysis (RSA; Kriegeskorte & Kievit, 2013). Below, we summarize previous research taking a multivariate approach.

Peelen et al. (2010) found that patterns of activation in the medial prefrontal cortex (mPFC) and posterior superior temporal sulcus (pSTS) had greater within-emotion similarity than between-emotion similarity across modalities, indicating these areas supramodally represent emotion identity. Chikazoe et al. (2014) found supramodal directional valence (i.e., positive vs. neutral vs. negative) representations in medial and lateral orbitofrontal cortex (OFC), alongside modality-specific directional valence representations for visual scenes in ventral temporal cortex, and for tastes in anterior insular cortex. Skerry & Saxe (2015) presented written stories depicting characters experiencing many different emotions. They found a model fitting 38 appraisal features (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. Kim et al. (2017) presented emotional movie clips and orchestral music, finding 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.

Experimental paradigm

The present study builds on the foundation of previous research in several ways. Our stimuli consisted of short clips of music and animation in which the depicted object—a piano or a bouncing ball—was held constant, and emotion was communicated solely by varying stimulus features. This ensured emotion processing requirements were uniform across the stimulus set. By contrast, collections of images or movies depicting emotionally charged scenes (e.g., the International Affective Picture System; Lang, Bradley, & Cuthbert, 2008) may require a wide variety of processes for emotion evaluation, including moral judgment, memory, and so on.

Stimuli were created by participants in a previously documented experiment (Sievers et al., 2013), who manipulated five stimulus features (speed, irregularity, consonance/spikiness, ratio of big-to-small movements, ratio of upward-to-downward movements) to communicate five emotions (Angry,

Happy, Peaceful, Sad, Scared). This approach did not presuppose in advance what combinations of features would be used for each emotion, but instead allowed participants to explore the entire possibility space. The emotional stimuli created by the participants 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%. This decision was motivated by evidence that many emotions are perceived as mixes of of other emotions (Cowen & Keltner, 2017). Three additional, "neutral" emotions were identified by searching for points in the stimulus feature possibility space that were distant from all emotional feature combinations. Music and animation stimuli were matched, such that for each musical stimulus there was an animation stimulus with analogous 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 fMRI scanning (Figure 1). This approach enabled the use of an exhaustively complete model, including both stimulus features and participants' judgments of emotion content. All inter-stimulus differences were dependent upon parameters explicitly represented in this model. Critically, this model was designed to test only the two hypotheses stated in the introduction. It was not designed to evaluate which of the many possible descriptions of the structure of emotion is ultimately the best fit for brain activity.

Participants listened to and viewed all stimuli during fMRI scanning, and the fitness of the model to activity across the brain during vision and audition was evaluated using searchlight representational similarity analysis (RSA; Kriegeskorte, Goebel, & Bandettini, 2006; Kriegeskorte & Kievit, 2013; Kriegeskorte, Mur, & Bandettini, 2008). RSA tests if differences between stimulus classes in the model track differences between corresponding multi-voxel patterns of BOLD activity. If the model matches brain activity in an area, this supports the hypothesis that the neural code in that area represents the content of the model. Although previous research investigating crossmodal neural representations has used cross-decoding (Kriegeskorte, 2011) or fMRI adaptation (Grill-spector & Malach, 2001), these techniques are not appropriate for testing smooth variation in neural representations or testing across separate populations of neurons,[1] making RSA a better fit for our hypotheses. Importantly,

[1] In cross-decoding, a linear decoder is trained using neural activity recorded during stimulus presentation in one modality (e.g., audition), but tested using data from another modality (e.g., vision). In fMRI adaptation, repeated presentation of stimuli causes neurons representing those stimuli to adapt, creating a decrease in BOLD activity. If this adaptation persists across stimulus classes, then the neural population does not represent the differences between stimuli. RSA is a better fit for our hypotheses for two reasons: (1) RSA can test the hypothesis that neural representations in both visual and auditory brain areas are smoothly distributed across a vector space, where similar emotions are closer to one another. Clear classification boundaries between stimulus categories such as "50% happy and 50% peaceful" and "25% happy and 75% peaceful" are unlikely to exist, and so linear decoding methods that are insensitive to smooth but predictable variation could yield misleading false negative results. (2) RSA can test the hypothesis that *different* areas in the brain use similar neural codes. fMRI adaptation cannot be used to test across different populations of neurons. Further, neural codes in different brain areas may be arbitrarily rotated relative to one another (Goddard et al., 2018; Haxby et al., 2011), such that a classification boundary drawn in one region is unlikely to produce good predictions in other regions, even if those regions represent the same content using the same neural codes. Because

our approach focused exclusively on what emotions participants perceived—no emotions were expressed by participants during fMRI scanning.

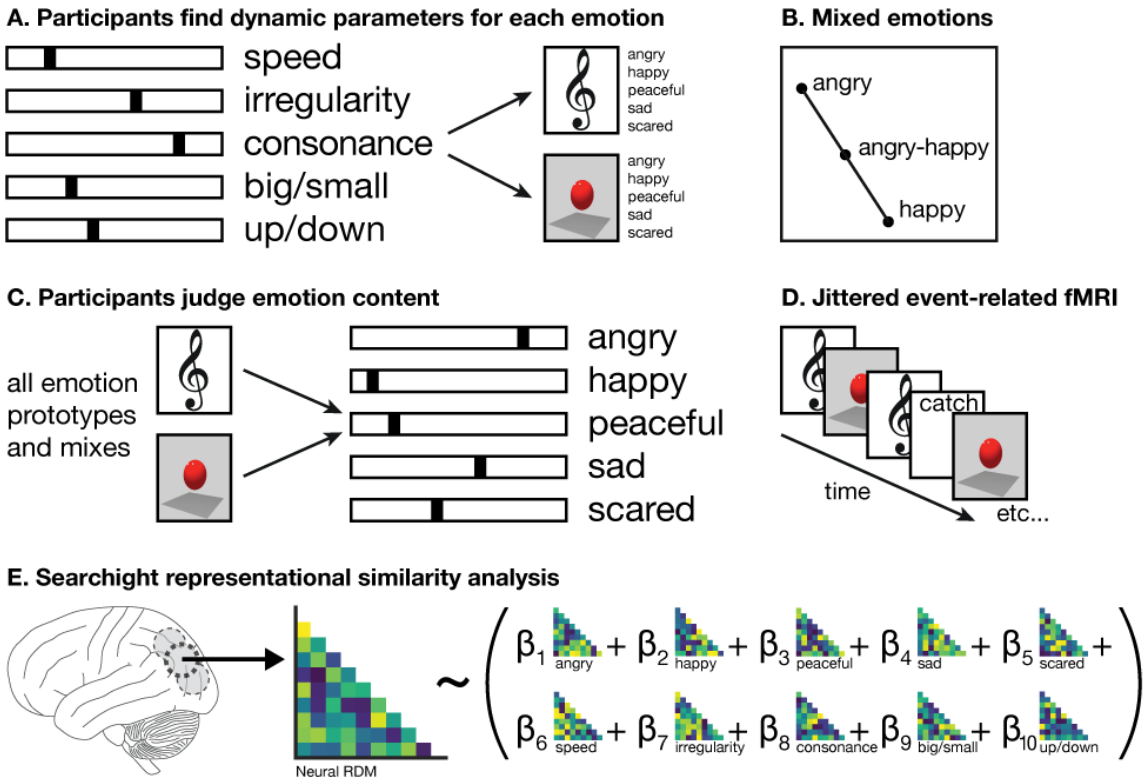

Figure 1: Experimental paradigm. *A.* Participants in Sievers et al. (2013) manipulated stimulus features to generate music and animation communicating five prototypical emotions: Angry, Happy, Peaceful, Sad, and Scared. *B.* Mixed emotions were generated by linear interpolation between the stimulus 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 (Kriegeskorte et al., 2006, 2008; Kriegeskorte & Kievit, 2013). 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.

it uses distances between patterns of activity, RSA is insensitive to arbitrary rotations (i.e., it tests for a second-order isomorphism between the model and the data; Kriegeskorte et al., 2008), making it a more appropriate tool for testing our hypotheses.

Results

Emotion judgments

To test how tightly behavioral participants' emotion judgments were clustered within each stimulus class, we applied a permutation testing procedure. For each emotion, we averaged participants' emotional slider bar ratings, yielding a class mean. We then calculated the Euclidean distance of each judgment to this mean, yielding a distribution of distances to the mean for each stimulus class. A null distribution of distances to the stimulus class means was created by applying this procedure 2000 times, each with a different permutation of the emotion labels over the whole dataset. If participants' judgments were tightly clustered, then the observed distances to the mean should be lower than the corresponding null distances. Welch's independent samples *t*-test was applied to test whether the observed distributions of distances to class means differed from the null. For all stimulus classes except one "neutral" emotion, the observed distances were significantly lower (mean t=-4.37; mean distance=-15.52; mean $p<.001$), indicating participants reached broad agreement about the emotion content of each stimulus class (Supplementary Figure 11).

Representational Similarity Analysis

To test hypotheses (1) and (2), we performed searchlight representational similarity analysis (Kriegeskorte et al., 2006, 2008), testing the fit of a single model to BOLD activity evoked during both music and animation trials. The model used 10 representational dissimilarity matrices (RDMs): five based on the mean parameter settings used to create the stimuli (speed, irregularity/jitter, consonance/spikiness, ratio of big-to-small movements, ratio of upward-to-downward movements), and five based on the mean emotion judgments of the behavioral participants (Angry, Happy, Peaceful, Sad, and Scared). Because participants' responses were clustered together, model RDMs were created using mean slider bar settings, providing a low-noise estimate of the consensus across participants. Each RDM captured the Euclidean distance between every pair of stimuli in terms of a single stimulus feature or emotion judgment parameter (Supplementary Figure 1). If graded differences in patterns of BOLD activity match graded differences in stimulus features or emotion judgments made during the behavioral experiment, this is evidence that the patterns of BOLD activity represent the stimulus features and emotion judgments. Importantly, the same stimulus feature parameter settings were used to create both the music and animation stimuli, and behavioral participants' emotion judgments were averaged across music and animation. These precautions ensured that the modeled distance between any two music stimuli was always equal to the modeled distance between the corresponding animation stimuli, making the model suitable for testing the fundamental similarity of BOLD activity evoked by music and animation.

Each voxel in the brain was used as the center of a searchlight sphere with a 3-voxel (9mm) radius. Within each searchlight sphere, the ranked correlation distance (1-Spearman's ρ) between each pair of stimulus-dependent patterns of BOLD activity was aggregated to create a neural RDM. 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. Model fit (R^2) was assessed using multiple regression with the ranked model RDMs as predictors and the neural RDM as the target. Because including additional predictors in a multiple regression model increases R^2 even when those predictors are non-informative, R^2 values were adjusted downward using a permutation-testing approach. 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 values (R^2_{adj}.). To evaluate how much variance the model explained relative to the total explainable variance in BOLD activity, we report model fit statistics relative to the upper and lower bounds of the noise ceiling (Nili et al., 2014, see methods). However, β weights from multiple regression reflect only the unique contribution of each predictor, discarding meaningful variance shared among correlated predictors, and the variance of beta weight estimates can be inflated when predictors are correlated with one another (Mason & Perreault, 1991). To address these issues, we assessed the contribution of individual predictors by calculating the ranked correlation (Spearman's ρ) of each predictor RDM to the neural RDM. (See supplementary materials for an assessment of multicolinearity in the model including variance inflation factors for each predictor, as well as β weight maps reflecting the unique contributions of each predictor.) All group-level statistics were corrected for multiple comparisons using a maximum cluster mass sign-flipping permutation test performed with FSL `randomise` (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012; Nichols & Holmes, 2001), with a cluster-determining threshold of p=.01 and a family-wise error rate of .05.

We ran these analyses twice—first, using only music trials to create the neural RDM, then using only animation trials. Each of these two analyses was modality-specific. The critically important feature of this analysis approach is that it does not test separate models for vision and audition, but instead tests the fitness of a single model to brain activity evoked by stimuli in different sensory modalities. If this single model fits different modality-specific brain regions, those brain regions must use a similar neural code.

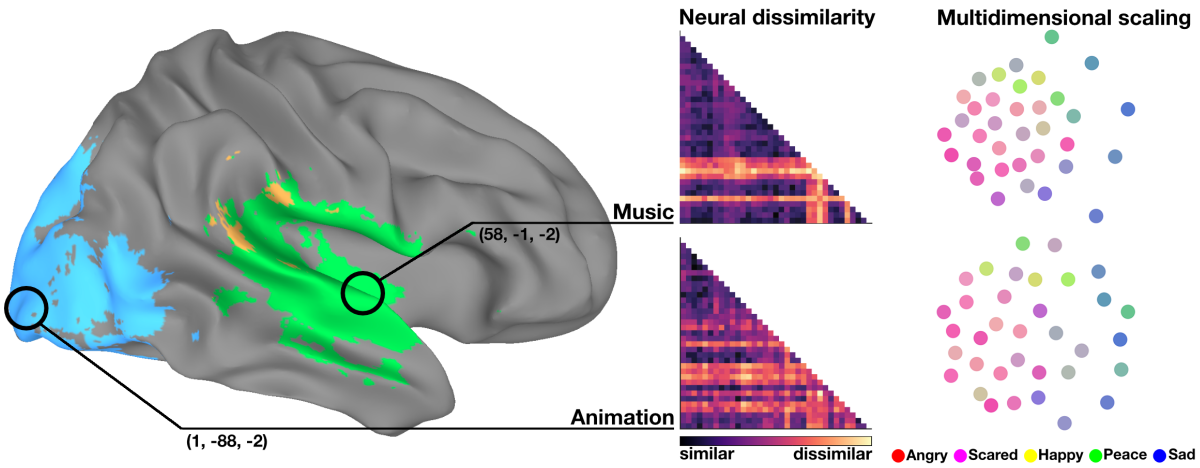

Figure 2: Highlighted brain areas fit a model including stimulus features and emotion judgments as predictors during animation trials (blue) and music trials (green). A significant proportion of participants' model fits overlapped for both trial types (yellow). Neural dissimilarity matrices show pairwise similarity of activity patterns evoked by each stimulus at the locations of best model fit (circled)—medial lingual gyrus (animation) and lateral superior temporal gyrus (music). Fully labled versions of these matrices are shown in Supplementary Figure 7. Multidimensional scaling flattens these dissimilarity 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.

The model explained variance in a range of visual brain regions during animation trials, and auditory brain regions during music trials, providing strong support for hypothesis (1), that these regions share a neural code (Figure 3; Table 1). The peak of the average model fit across participants was in the left medial lingual gyrus for animation trials (mean R^2_{adj}=.15; 95% CI: .08–.21; t(19)=4.68; p=.005 corrected) and in right anterior superior temporal gyrus for music trials (mean R^2_{adj}=.15; 95% CI: .1–.2; t(19)=6.08; p=.01 corrected). 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. Importantly, Spearman's ρ for all stimulus feature and emotion judgment predictors was significant at these peak locations, indicating that model fit was not driven solely by any single stimulus feature or emotion judgment predictor. 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 the lingual gyrus for animation trials (mean individual R^2_{adj}=.31; 95% CI: .24–.38; t(19)=9.2; p<.001 uncorrected) and in bilateral anterior superior temporal gyrus for music trials (mean individual R^2_{adj}=.26; 95% CI: .21–.31; t(19)=10.95; p<.001 uncorrected) (Supplementary Figures 2 and 3).

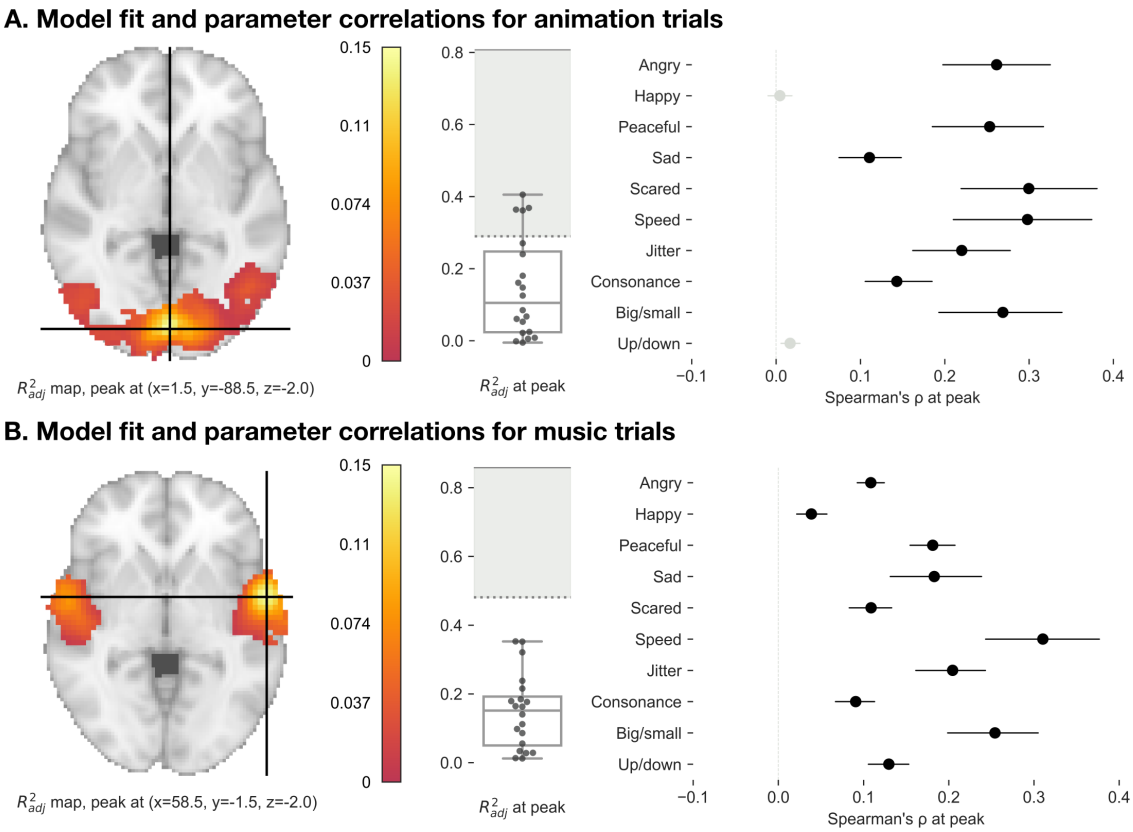

Figure 3: Maps of the mean coefficient of determination (R^2_{adj}) across participants. The model included 5 stimulus feature predictors and 5 emotion judgment predictors, and was separately fit to animation and music trials. Maps thresholded at FWER=.05. Box plots show the median, quartiles, and range of the per-participant R^2_{adj} values at the location of best model fit at the group level. The light grey dotted line indicates the lower bound of the noise ceiling, and the dark grey solid line indicates the upper bound. Points within the shaded grey area are within the bounds of the noise ceiling. For per-parameter Spearman's ρ and β weight maps, see Supplementary Figures N and N.

Table 1: Brain regions fitting the stimulus feature and emotion judgment model during animation trials.

x	y	z	Nearest atlas label (Destrieux, 2009)	R^2_{adj}	95% CI	p
2.0	-88.0	-2.0	L Lingual gyrus, lingual part of the medial occipito-temporal gyrus, (O5)	.15	.08–.21	.005
46.0	-68.0	1.0	R Inferior occipital gyrus (O3) and sulcus	.04	.01–.07	.005

x	y	z	Nearest atlas label (Destrieux, 2009)	R^2_{adj}	95% CI	p
22.0	-82.0	31.0	R Superior occipital gyrus (O1)	.03	.01–.06	.005

Table 2: Brain regions fitting the stimulus feature and emotion judgment model during music trials.

x	y	z	Nearest atlas label (Destrieux, 2009)	R^2_{adj}	95% CI	p
58.0	-2.0	-2.0	R Lateral aspect of the superior temporal gyrus	.15	.10–.20	.011
-62.0	-16.0	7.0	L Lateral aspect of the superior temporal gyrus	.09	.05–.12	.011

To rule out the possibility that these brain regions were a good fit for our model, but not truly a shared neural code (i.e., 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 ρ=-.007, whereas the inter-region similarity at the locations of peak model fit was ρ=.68 (p<.001), more similar than any pair of coordinates in the null distribution.

Overlapping auditory and visual model fit

To locate brain regions where emotional music and animation were both represented using the same supramodal code, we created binary overlap masks, selecting voxels where both music and animation model fits were significant at the individual level (permutation p<.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. Supramodal representations were found in bilateral posterior superior temporal gyrus (pSTG) in 60% of participants (95% CI 36%–84%, p<.001 corrected), providing

support for hypothesis (2) (Fig 4). 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}=.03, 95% CI: .02–.05, t(19)=5.78, p=.01 corrected; animation mean R^2_{adj}=.03, 95% CI: .01–.05, t(19)=2.94, p=.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 emotion representations. Manual inspection of the overlap masks showed supramodal emotion representations in pSTG were consistent across participants (Supplementary Figure 4).

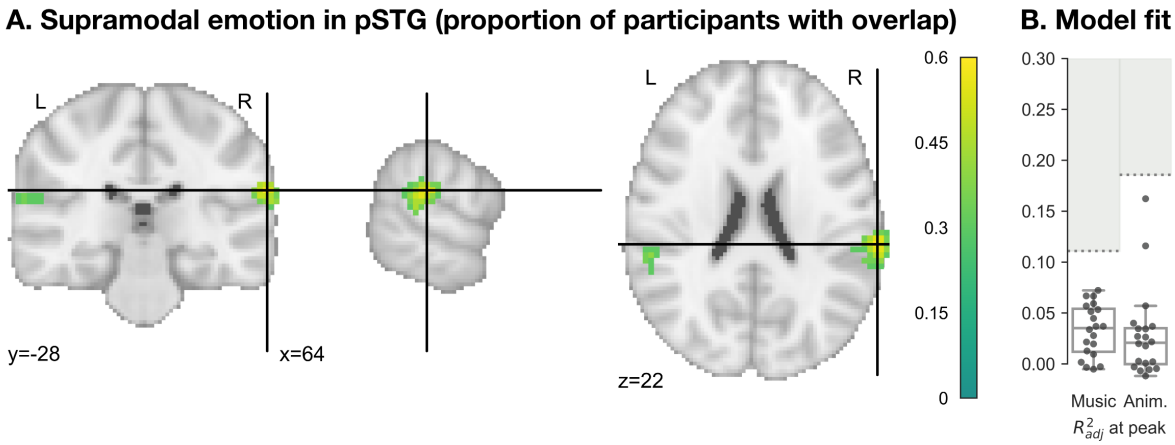

Figure 4: A. Binary overlap masks were created per participant, selecting voxels significant at the individual participant level for both music and animation trials. Maps show the voxelwise average of these overlap masks; i.e., the proportion of participants representing emotion in music and animation in the same brain areas. Maps thresholded at voxelwise FWER=.05. B. Box plots show the median, quartiles, and range of R^2_{adj} for music and animation trials at the location where most participants exhibited supramodal emotion representations.

Table 3: Brain regions fitting the stimulus feature and emotion judgment model during both music and animation trials.

x	y	z	Nearest atlas label (Destrieux, 2009)	Proportion of participants	95% CI	p
64.0	-28.0	22.0	R Supramarginal gyrus	.6	.36–.84	<.001
-56.0	-32.0	19.0	L Supramarginal gyrus	.4	.16–.64	.008

Exploratory intermodal RSA

To find brain areas that represented emotional stimuli even when those stimuli were presented in the area's non-preferred modality, we performed an exploratory intermodal RSA. In this analysis, both the neural target RDM and the predictor RDMs contained only between-modality distances, corresponding to the lower-left square region of the larger triangular RDM created using stimuli from both modalities (Figure 5A). 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 considers between-modality distances, it could not in principle identify any modality-specific activity.

Intermodal RSA revealed a bilateral set of areas across occipital, superior parietal, temporal, cingulate, and frontal cortex that each represented emotions presented in their non-preferred modality (Figure 5B; Table 4). 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}=.28; 95% CI: .20–.37; t(19)=6.9; p<.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 indicates that the model captures the majority of explainable variance in intermodal activity, suggesting intermodal activity in left lingual gyrus may be dominated by representations related to emotion content. However, the lower bound of the intermodal noise ceiling was relatively low (.07), suggesting that most reliable neural activity in this region was modality-specific.

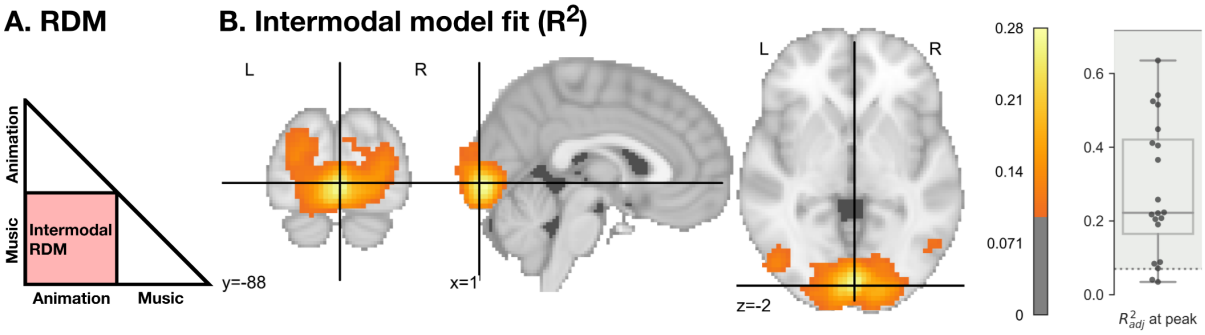

Figure 5: Intermodal model fit. We created behavioral and neural intermodal RDMs to locate brain areas that represented emotional stimuli even when those stimuli were presented in the area's non-preferred modality. A. Intermodal RDMs capture the stimulus feature, emotion judgment, or neural pattern distances between emotional music and animation. The intermodal RDM is the lower-left square of the the full RDM created using both music and animation. B. Intermodal model fit, thresholded at FWER=.05. R^2_{adj} values below .1 hidden for visual clarity. Box plot shows the median, quartiles, and range of per-participant R^2_{adj} values at the location of best model fit at the group level. The light grey dotted lines indicate the lower bound of the noise ceiling for that trial type, and the dark grey solid lines indicate the upper bound. Points within the shaded grey areas are within the bounds of the noise ceiling.

Table 4: Brain regions fitting intermodal model; i.e., regions which fit the stimulus feature and emotion judgment model even when the stimulus is presented in the non-preferred modality.

x	y	z	Nearest atlas label (Destrieux, 2009)	R^2_{adj}	95% CI	p
2.0	-88.0	-2.0	L Lingual gyrus, lingual part of the medial occipito-temporal gyrus, (O5)	.28	.20–.37	<.001
64.0	-28.0	22.0	R Supramarginal gyrus	.09	.06–.12	<.001
-56.0	-4 .0	22.0	L Planum temporale or temporal plane of the superior temporal gyrus	.08	.05–.10	<.001
32.0	-56.0	61.0	R Superior parietal lobule (lateral part of P1)	.07	.05–.09	<.001
-32.0	-56.0	64.0	L Superior parietal lobule (lateral part of P1)	.06	.04–.08	<.001
-16.0	-22.0	4 .0	L Marginal branch (or part) of the cingulate sulcus	.05	.03–.07	<.001

x	y	z	Nearest atlas label (Destrieux, 2009)	R^2_{adj}	95% CI	p
-28.0	-58.0	-53.0	L Lateral occipito-temporal gyrus (fusiform gyrus, O4-T4)	.04	.03–.06	<.001
-46.0	44.0	22.0	L Middle frontal gyrus (F2)	.03	.02–.04	<.001
-4.0	64.0	22.0	L Superior frontal gyrus (F1)	.03	.02–.04	<.001

Discussion

On adaptive signaling accounts of emotion perception (Dezecache, Mercier, & Scott-Phillips, 2013; Hebets et al., 2016; Huron, 2012; Lorenz, 1970), the human brain should show adaptations specific to the crossmodally redundant structure of emotion. To investigate this, we tested two hypotheses: (1) that auditory and visual brain areas use a shared neural code for emotion that represents the same underlying parameters, and (2) that in some brain areas, both auditory and visual emotions are represented using a single, supramodal neural code. We tested these hypotheses by fitting a single model that captured stimulus features and participants' emotion judgments to brain activity evoked by both auditory and visual stimuli. Critically, this single model fit brain activity in both auditory and visual sensory brain areas, indicating these regions use a shared neural code for emotion. Further, a direct test of inter-region representational similarity showed that auditory and visual representations were significantly more similar to each other than representations in randomly selected brain regions. Both of these findings support hypothesis (1). The model described above also fit activity in pSTG during both animation and music trials, indicating this region uses a supramodal emotion representation, supporting hypothesis (2).

Previous studies have shown multimodal processing in unimodal areas (for reviews, see Bulkin & Groh, 2006; Ghazanfar & Schroeder, 2006; Kayser & Logothetis, 2007), which may depend on direct inter-areal projections between unimodal regions (Cappe & Barone, 2005; Falchier, Clavagnier, Barone, & Kennedy, 2002; Rockland & Ojima, 2003). Critically, our 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 neural code that is shared across modalities.

Exploratory intermodal representational similarity analysis further showed that low-level visual areas represented stimulus features and emotion content even when presented in the non-preferred modality. However, the intermodal noise ceiling was relatively low, indicating that most reliable neural activity is modality-specific. Accordingly, the reported results should not be interpreted as showing that visual or auditory areas are modality-general.

Although the stimulus features captured by the reported model comprehensively accounted for all of the low-level differences between emotions, participants' emotion judgments could not have been reconstructed from a linear combination of the stimulus features. This implies that a non-linear transformation of features to judgments must have taken place. The simultaneous presence of significant stimulus feature and emotion judgment betas in early auditory and visual regions suggests that this non-linear transformation occurs early in the perceptual processing hierarchy. This tuning of low-level sensory representations to the structure of emotions shows that the need to identify such signals has exerted a profound shaping force on low-level perceptual processes. We do not see or hear the actions of others as raw sense impressions first, and later encode them as communicating emotion after a chain of intermediary processing steps occurring in encapsulated cognitive modules (Fodor, 1985). Rather, we begin accumulating evidence for an emotional interpretation from the lowest levels of sensory processing.

Supramodal representation in pSTG/pSTS

Our findings in pSTG overlap with previously reported pSTS activation during action understanding (Beauchamp, Lee, Argall, & Martin, 2004; Wyk, Hudac, Carter, Sobel, & Pelphrey, 2009) and emotion perception tasks (Kreifelts, Ethofer, Grodd, Erb, & Wildgruber, 2007; Robins, Hunyadi, & Schultz, 2009; Watson et al., 2014). The pSTS also exhibits greater activation for combined audio–visual presentation than for either modality alone (Beauchamp et al., 2004; Wright, Pelphrey, Allison, McKeown, & McCarthy, 2003), and the amplitude of these responses predicts object categorization performance (Werner & Noppeney, 2010). Damage to the pSTS does not impair voice recognition (Jiahui et al., 2017), suggesting its representations are downstream from low-level feature detectors. Alongside these results, our findings are consistent with the hypothesis that the pSTG/pSTS acts as a general-purpose hub for transforming unimodal inputs into a common supramodal representation (Schirmer & Adolphs, 2017). Interestingly, visual and auditory selectivity in pSTS are linked, with areas sensitive to moving mouths responding strongly to voices, but not non-vocal sounds (Zhu & Beauchamp, 2017). This suggests crossmodal selectivity in pSTS may be shaped by co-occurrence statistics in the environment.

Adaptive signaling of emotion

The findings reported here are consistent with an *adaptive signaling* account of emotion perception (Hebets et al., 2016; Huron, 2012). As a radically social species, our survival depends on quickly and accurately understanding others' thoughts and feelings (Allport, 1924; Tooby & Cosmides, 1990). This is no easy task, as communication transpires across a *noisy channel*—imprecise gestures, sounds, and speech, must pierce through a chaotic environment to maximize their chances of perception

by distracted and inattentive observers. Effective communication requires signals that can survive the noisy channel, and brains adapted to perceive them (Dezecache et al., 2013; Huron, 2012; Lorenz, 1970). Cross-sensory redundancy is one strategy for ensuring a signal can travel across a noisy channel (Evans & Treisman, 2011; Huron, 2012). For example, rattlesnakes hold their tail aloft when rattling, ensuring that it is both seen and heard. This combined audio–visual display is harder to miss or misinterpret than either movement or rattling alone. Perceivers' brains must be sensitive to such crossmodal redundancies in order to take advantage of them, and the shared neural code identified here may be an evolutionary adaptation specific to this purpose.

Limitations and future directions: toward comprehensive experiments and models

The present study was designed only to test a shared neural code for emotion across vision and audition, and was not designed to map all of the brain regions involved in the emotion perception process. Indeed, the auditory and visual representations identified here likely provide input to a range of complimentary processes that also contribute to emotion perception (Wager et al., 2015), and this may explain why they use the same neural code. Further, the reported model was not designed to explain all of the meaningful variance in brain activity during emotion perception, and was not designed to take individual differences into account. Instead, we focused on a small collection of parameters that were perceived similarly across participants, and that were shared between vision and audition, excluding all modality-specific information. We do not claim that the neural code in auditory and visual brain regions is the same in every respect; rather, our claim is that these neural codes are the same for the included parameters. Further, we do not claim that the included model parameters are the only features involved in emotion perception, or the only features represented in auditory and visual brain regions. For example, Russell's (1980) circumplex model organizes emotions in terms of their valence (positive–negative) and degree of arousal. These and other possible parameters, such as warmth and dominance (Kiesler, 1983) or value (Levy & Glimcher, 2012; Shuster & Levy, 2018) certainly covary with the stimulus features and emotion judgments used in the model described above. An advantage of a model based on plain language emotion judgments is that, insofar as other parameters can be mapped to emotion labels, the reported results remain interpretable. However, the omission of parameters that might ultimately be shown to better describe the underlying structure of emotion may have biased the reported model fit downward.

Critically, the reported model succeeded despite these limitations, explaining variance in separate brain areas that are normally understood to perform completely different sensory functions. This shows that vision and audition have more in common than previously appreciated, especially with respect to emotion perception. This said, we do believe that developing a comprehensive map of emotion processing across the brain that explains all emotion-related variance in neural activity is a laudable goal. To achieve this, researchers will need to run experiments that include a wide range of

emotion perception tasks, leverage multiple analysis techniques including cross-decoding and fMRI adaptation, test models that take individual differences in emotion perception into account, and compare models that test different theories of emotion and include task-specific information.

While participants in our study perceived emotions based on configurations of stimulus features, emotions may also be read from semantic content (Chikazoe et al., 2014; Kim et al., 2017; Skerry & Saxe, 2015). The emotional meanings of stimuli used in semantic emotion perception studies (e.g., detailed written stories; images from the International Affective Picture System, Lang et al., 2008) depend on participants recognizing *what* is depicted and *why* it is emotionally relevant. This is fundamentally different from reading emotional meaning from stimulus features, as in movement or prosody. Our experiment used music and animation in which the depicted object was held constant, and relatively low-level stimulus features were manipulated, capturing a wide range of emotions. Studies that seek to explain the maximum amount of variance in brain activity will need comprehensive experiments that impose both semantic and perceptual processing demands.

We anticipate that advances in automatic feature extraction (McNamara, Vega, & Yarkoni, 2017) will enable the use of naturalistic stimuli and complex models spanning not only the stimulus feature and emotion spaces examined here, but also additional dimensions of semantic meaning, context dependence, self- and other-relevance, appraisal features, and so on. Such future experiments will be the best of both (or many) worlds, allowing researchers to disentangle the underlying mechanisms supporting emotion perception.

Adaptive signaling vs. "peg fits hole"

One possible reading of these results is that humans have evolved neural detectors specific to emotional stimulus features, and that these are present from birth. On this "peg fits hole" interpretation, any sensory input with the right structure should be perceived as emotion. While this may be true in some basic cases, such as infants' reactions to shouting or motherese, cross-cultural variation in emotion perception places a limit on the "peg fits hole" interpretation. Although many of the same features are implicated in emotion perception across cultures (Ekman, 1992; Jack, Sun, Delis, Garrod, & Schyns, 2016; Sievers et al., 2013), there are also substantial cross-cultural differences (Jack, Caldara, & Schyns, 2012; Jack et al., 2016; Yuki, Maddux, & Masuda, 2007). The neural mechanisms supporting emotion perception must therefore flexibly accommodate culture-specific emotion dialects and display rules. These mechanisms need not be present from birth, and need not be specific to emotion. Rather, emotion perception may exploit statistical learning and predictive coding processes (Clark, 2013; Saffran, Aslin, & Newport, 1996), or may arise later in development as a cognitive strategy for coping with a complex social world (Blakemore, 2008). On this account, the stimulus features used to communicate emotion, the brains of emotion perceivers, and their cultural–environmental niche are interlinked and evolve together. The cross-cultural intelligibility of emotion can be explained by

globally shared contextual factors, including the evolutionary inheritance of the human body, the challenge of cooperating with others in a dangerous, unpredictable, resource-limited world, and the related need to estimate others' internal states. On this account, cross-cultural differences can be understood as path-dependent adaptations specific to a regional cultural–environmental niche.

Conclusion

Emotional stimulus features are shared across music and movement, and this is reflected in the organization of the brain: the same neural code is used to represent emotion in auditory, visual, and supramodal areas. Surprisingly, unimodal auditory and visual areas represent stimuli shown in their non-preferred modality. Such efficient organization is consistent with the adaptive signaling account of emotion perception. This theory predicts both that emotion signals be crossmodally redundant in order to survive communication across a noisy channel, and that receivers be specifically adapted to the crossmodal nature of the signal's structure. In other words, human emotion perception is optimized "end-to-end"—all levels of the processing hierarchy are tuned to support the social goal of understanding the emotional states that predict others' behavior.

Acknowledgements

We thank Sam Nasatase and Matteo Visconti di Oleggio Castello 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 John Templeton Foundation and the Neukom Institute for Computational Science.

Materials and Methods

Participants

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.

Stimuli

Emotion stimuli were generated using a model developed for a prior study (Sievers et al., 2013) 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., 2013) (music *N*=25; movement *N*=25; total *N*=50) used this model to communicate five prototype emotions: Angry, Happy, Peaceful, Sad, and Scared. The main finding of (Sievers et al., 2013) was that participants chose similar music and movement parameter settings for each emotion, showing that music and movement share an underlying structure. The median parameter settings across music and movement from Sievers (2013) 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 (Zicarelli, 1998) 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 use the slider bars to evaluate what emotion or mix of emotions they perceived.

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 3s in duration and followed by fixation periods of randomly varying duration (range: 0.5s–20s). 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 (Peirce, 2007). 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; .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 .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.

Representational similarity analysis

Representational similarity analysis (RSA) (Kriegeskorte et al., 2006, 2008) was conducted using PyMVPA (Hanke et al., 2009), Scikit-Learn (Pedregosa et al., 2012), NumPy (Oliphant, 2006), and SciPy (Jones, Oliphant, & Peterson, 2001). Stimulus feature representational distance matrices (RDMs) for each of the parameters described in (Sievers et al., 2013) (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 5A).

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. Intermodal neural RDMs were created as described above, using neural data instead of stimulus features or emotion judgments (Figure 5A). 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: 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.

Noise ceiling

The upper and lower bounds of the noise ceiling were calculated using an approach based on Nili et al. (2014), but adapted for use with multiple regression. The approach described by Nili et al. (2014) 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<.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=.01; FWER p=.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 `randomise` (Jenkinson et al., 2012; Nichols & Holmes, 2001) (cluster-determining threshold p=.01; family-wise error rate p=.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 (Abraham et al., 2014) and AFNI SUMA (Saad, Reynolds, Argall, Japee, & Cox, 2004).

Data and code availability

The datasets generated and analyzed during the current study, and the code used the perform the analyses, are available from the corresponding author on reasonable request.

Bibliography

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Muller, A., Kossaifi, J., … Varoquaux, G. (2014). Machine Learning for Neuroimaging with Scikit-Learn, *8*(February), 1–10. https://doi.org/10.3389/fninf.2014.00014

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1994). Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. https://doi.org/10.1038/372669a0

Allport, F. H. (1924). *Social Psychology*. New York, NY: Houghton Mifflin.

Baily, J. (1985). Music Structure and Human Movement. In P. Howell, I. Cross, & R. West (Eds.), *Musical structure and cognition* (pp. 237–258). London: Academic Press.

Beauchamp, M. S., Lee, K., Argall, B., & Martin, A. (2004). Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, *41*, 809–823. https://doi.org/10.1016/S0896-6273(04)00070-4

Blakemore, S.-J. (2008). The social brain in adolescence. *Nature Reviews Neuroscience*, *9*(4), 267–277. https://doi.org/10.1038/nrn2353

Bulkin, D. A., & Groh, J. M. (2006). Seeing sounds: visual and auditory interactions in the brain. *Current Opinion in Neurobiology*, *16*(4), 415–419. https://doi.org/10.1016/j.conb.2006.06.008

Calder, A. J., Lawrence, A. D., & Young, A. W. (2001). Neuropsychology of Fear and Loathing. *Nature Reviews Neuroscience*, *2*(5), 352–363. https://doi.org/10.1038/35072584

Cappe, C., & Barone, P. (2005). Heteromodal connections supporting multisensory integration at low levels of cortical processing in the monkey. *European Journal of Neuroscience*, *22*(11), 2886–2902. https://doi.org/10.1111/j.1460-9568.2005.04462.x

Chikazoe, J., Lee, D. H., Kriegeskorte, N., & Anderson, A. K. (2014). Population coding of affect across stimuli, modalities and individuals. *Nature Neuroscience*, *17*(8), 1114–1122. https://doi.org/10.1038/nn.3749

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Cowen, A. S., & Keltner, D. (2017). Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, *114*(38), E7900–E7909. https://doi.org/10.1073/pnas.1702247114

Dezecache, G., Mercier, H., & Scott-Phillips, T. C. (2013). An evolutionary approach to emotional communication. *Journal of Pragmatics*, *59*, 221–233. https://doi.org/10.1016/j.pragma.2013.06.007

Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, *6*(3), 169–200. https://doi.org/10.1080/02699939208411068

Evans, K. K., & Treisman, A. (2011). Natural cross-modal mappings between visual and auditory features. *Journal of Vision*, *10*(1), 6–6. https://doi.org/10.1167/10.1.6

Falchier, A., Clavagnier, S., Barone, P., & Kennedy, H. (2002). Anatomical evidence of multimodal integration in primate striate cortex. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *22*(13), 5749–59. https://doi.org/20026562

Fodor, J. A. (1985). Précis of The Modularity of Mind. *Behavioral and Brain Sciences*, *8*, 1–42.

Fritz, T., Jentschke, S., Gosselin, N., Sammler, D., Peretz, I., Turner, R., … Koelsch, S. (2009). Report Universal Recognition of Three Basic Emotions in Music. *Current Biology*, *19*(7), 573–576. https://doi.org/10.1016/j.cub.2009.02.058

Ghazanfar, A. A., & Schroeder, C. E. (2006). Is neocortex essentially multisensory? *Trends in Cognitive Sciences*, *10*(6), 278–285. https://doi.org/10.1016/j.tics.2006.04.008

Goddard, E., Klein, C., Solomon, S. G., Hogendoorn, H., Thomas, A., & Klein, C. (2018). Interpreting the dimensions of neural feature representations revealed by dimensionality reduction. *NeuroImage*, *180*(Part A), 41–67. https://doi.org/10.1016/j.neuroimage.2017.06.068

Grill-spector, K., & Malach, R. (2001). fMR-adaptation : a tool for studying the functional properties of human cortical neurons. *Acta Psychologica*, *107*, 293–321.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, *7*(1), 37–53. https://doi.org/10.1007/s12021-008-9041-y

Haxby, J. V., Guntupalli, J. S., Connolly, A. C., Halchenko, Y. O., Conroy, B. R., Gobbini, M. I., … Ramadge, P. J. (2011). A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, *72*(2), 404–16. https://doi.org/10.1016/j.neuron.2011.08.026

Hebets, E. A., Barron, A. B., Balakrishnan, C. N., Hauber, M. E., Mason, P. H., & Hoke, K. L. (2016). A systems approach to animal communication. *Proceedings of the Royal Society B: Biological Sciences*, *283*(1826), 20152889. https://doi.org/10.1098/rspb.2015.2889

Huron, D. (2012). Understanding Music-Related Emotion: Lessons from Ethology. *Proceedings of the 12th International Conference on Music Perception and Cognition*, 473–481. Retrieved from http://icmpc-escom2012.web.auth.gr/sites/default/files/papers/473{_}Proc.pdf

Jack, R. E., Caldara, R., & Schyns, P. G. (2012). Internal representations reveal cultural diversity in expectations of facial expressions of emotion. *Journal of Experimental Psychology: General*, *141*(1), 19–25. https://doi.org/10.1037/a0023463

Jack, R. E., Sun, W., Delis, I., Garrod, O. G. B., & Schyns, P. G. (2016). Four not six: Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology: General*, *145*(6), 708–730. https://doi.org/10.1037/xge0000162

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., & Smith, S. M. (2012). FSL. *NeuroImage*, *62*, 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015

Jiahui, G., Garrido, L., Liu, R. R., Susilo, T., Barton, J. J., & Duchaine, B. (2017). Normal voice processing after posterior superior temporal sulcus lesion. *Neuropsychologia*, *105*(September 2016), 215–222. https://doi.org/10.1016/j.neuropsychologia.2017.03.008

Johnstone, R. A. (1996). Multiple displays in animal communication: 'backup signals' and 'multiple messages'. *Proceedings of the Royal Society B: Biological Sciences*, *351*, 329–338.

Johnstone, R. A. (1997). The evolution of animal signals. In J. Krebs & N. Davies (Eds.), *Behavioral ecology* (pp. 155–178). Oxford: Oxford University Press.

Jones, E., Oliphant, E., & Peterson, P. (2001). SciPy: Open Source Scientific Tools for Python.

Kaeppler, A. L. (1978). Dance in anthropological perspective. *Annual Review of Anthropology*, *7*, 31–49.

Kayser, C., & Logothetis, N. K. (2007). Do early sensory cortices integrate cross-modal information? *Brain Structure and Function*, *212*(2), 121–132. https://doi.org/10.1007/s00429-007-0154-0

Kiesler, D. J. (1983). The 1982 interpersonal circle: a taxonomy for complementarity in human transactions. *Psychological Review*, *90*(3), 185–214.

Kim, J., Shinkareva, S. V., & Wedell, D. H. (2017). Representations of modality-general valence for videos and music derived from fMRI data. *NeuroImage*, *148*(January), 42–54. https://doi.org/10.1016/j.neuroimage.2017.01.002

Kreifelts, B., Ethofer, T., Grodd, W., Erb, M., & Wildgruber, D. (2007). Audiovisual integration of emotional signals in voice and face: An event-related fMRI study. *NeuroImage*, *37*(4), 1445–1456. https://doi.org/10.1016/j.neuroimage.2007.06.020

Kriegeskorte, N. (2011). NeuroImage Pattern-information analysis : From stimulus decoding to computational-model testing. *NeuroImage*, *56*(2), 411–421. https://doi.org/10.1016/j.neuroimage.2011.01.061

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–8. https://doi.org/10.1073/pnas.0600244103

Kriegeskorte, N., & Kievit, R. a. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–12. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(November), 4. https://doi.org/10.3389/neuro.06.004.2008

Lang, P., Bradley, M., & Cuthbert, B. (2008). *International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8.* Gainesville, FL: University of Florida.

Levy, D. J., & Glimcher, P. W. (2012). The root of all value: a neural common currency for choice. *Current Opinion in Neurobiology*, *22*(6), 1027–1038. https://doi.org/10.1016/j.conb.2012.06.001

Lindquist, K. A., Wager, T. D., Kober, H., Bliss-Moreau, E., & Barrett, L. F. (2012). The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences*, *35*(03), 121–143. https://doi.org/10.1017/S0140525X11000446

Lorenz, K. (1970). *Studies in Animal and Human Behavior, Volume 1*. London: Methuen.

Mason, C. H., & Perreault, W. D. (1991). Collinearity, Power, and Interpretation of Multiple Regression Analysis. *Journal of Marketing Research*, *28*(3), 268–280.

McNamara, Q., Vega, A. de la, & Yarkoni, T. (2017). Developing a comprehensive framework for multimodal feature extraction. Retrieved from http://arxiv.org/abs/1702.06151

Nichols, T. E., & Holmes, A. P. (2001). Nonparametric Permutation Tests For Functional Neuroimaging : A Primer with Examples. *Human Brain Mapping*, *25*(1), 1–25.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., & Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, *10*(4). https://doi.org/10.1371/journal.pcbi.1003553

Norman, K. a, Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–30. https://doi.org/10.1016/j.tics.2006.07.005

Oliphant, T. E. (2006). *A guide to NumPy*. USA: Trelgol Publishing.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python, *12*, 2825–2830. https://doi.org/10.1007/s13398-014-0173-7.2

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

Peirce, J. W. (2007). PsychoPy-Psychophysics software in Python. *Journal of Neuroscience Methods*, *162*(1-2), 8–13. https://doi.org/10.1016/j.jneumeth.2006.11.017

Phillips, M. L., Young, A. W., Senior, C., Brammer, M., Andrew, C., Calder, A. J., … David, A. S. (1997). A specific neural substrate for percieving facial expressions of disgust. *Nature*, *389*(2 October), 495–498.

Phillips-silver, J., & Trainor, L. J. (2005). Feeling the Beat : Movement Influences Infant Rhythm Perception. *Science*, *308*, 1430.

Robins, D. L., Hunyadi, E., & Schultz, R. T. (2009). Superior temporal activation in response to dynamic audio-visual emotional cues. *Brain and Cognition*, *69*(2), 269–278. https://doi.org/10.1016/j.bandc.2008.08.007

Rockland, K. S., & Ojima, H. (2003). Multisensory convergence in calcarine visual areas in macaque monkey. *International Journal of Psychophysiology*, *50*(1-2), 19–26. https://doi.org/10.1016/S0167-8760(03)00121-1

Russell, J. a. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161–1178. https://doi.org/10.1037/h0077714

Saad, Z. S., Reynolds, R. C., Argall, B., Japee, S., & Cox, R. W. (2004). SUMA: an interface for surface-based intra- and inter-subject analysis with AFNI. *Biomedical Imaging: Nano to Macro, 2004. IEEE International Symposium on*, (October 2015), 1510–1513 Vol. 2. https://doi.org/10.1109/ISBI.2004.1398837

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science (New York, N.Y.)*, *274*(5294), 1926–1928. https://doi.org/10.1126/science.274.5294.1926

Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, *112*(29), 8987–8992. https://doi.org/10.1073/pnas.1414495112

Schirmer, A., & Adolphs, R. (2017). Emotion Perception from Face, Voice, and Touch: Comparisons and Convergence. *Trends in Cognitive Sciences*, *21*(3), 216–228. https://doi.org/10.1016/j.tics.2017.01.001

Shuster, A., & Levy, D. J. (2018). Common Sense in Choice: The Effect of Sensory Modality on Neural Value Representations. *eNeuro*, *5*(2), 1–14.

Sievers, B., Lee, C., Haslett, W., Wheatley, T., & Wheatley, T. (2019). A multi-sensory code for emotional arousal. *Proceedings of the Royal Society B: Biological Sciences*, *286*(20190513), 1–10.

Sievers, B., Polansky, L., Casey, M., & Wheatley, T. (2013). Music and movement share a dynamic structure that supports universal expressions of emotion. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(1), 70–5. https://doi.org/10.1073/pnas.1209023110

Skerry, A. E., & Saxe, R. (2015). Neural Representations of Emotion Are Organized around Abstract Event Features. *Current Biology : CB*, *25*(15), 1945–54. https://doi.org/10.1016/j.cub.2015.06.009

Tooby, J., & Cosmides, L. (1990). The past explains the present: Emotional adaptations and the structure of ancestral environments. *Ethology and Sociobiology*, *11*(4), 375–424. https://doi.org/10.1016/0162-3095(90)90017-Z

Trehub, S. E., Becker, J., Morley, I., & Trehub, S. E. (2015). Cross-cultural perspectives on music and musicality. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(20140096), 1–9.

Tsuchiya, N., Moradi, F., Felsen, C., Yamazaki, M., & Adolphs, R. (2009). Intact rapid detection of fearful faces in the absence of the amygdala. *Nature Neuroscience*, *12*(10), 1224–1225. https://doi.org/10.1038/nn.2380

Wager, T. D., Kang, J., Johnson, T. D., Nichols, T. E., Satpute, A. B., & Barrett, L. F. (2015). A Bayesian Model of Category-Specific Emotional Brain Responses. *PLoS Computational Biology*, *11*(4), 1–27. https://doi.org/10.1371/journal.pcbi.1004066

Watson, R., Latinus, M., Noguchi, T., Garrod, O., Crabbe, F., & Belin, P. (2014). Crossmodal Adaptation in Right Posterior Superior Temporal Sulcus during Face-Voice Emotional Integration. *Journal of Neuroscience*, *34*(20), 6813–6821. https://doi.org/10.1523/JNEUROSCI.4478-13.2014

Werner, S., & Noppeney, U. (2010). Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cerebral Cortex*, *20*(8), 1829–1842. https://doi.org/10.1093/cercor/bhp248

Wright, T. M., Pelphrey, K. A., Allison, T., McKeown, M. J., & McCarthy, G. (2003). Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral Cortex*, *13*(10), 1034–1043. https://doi.org/10.1093/cercor/13.10.1034

Wyk, B. C. V., Hudac, C. M., Carter, E. J., Sobel, D. M., & Pelphrey, K. a. (2009). Action understanding in the superior temporal sulcus region. *Psychological Science : A Journal of the American Psychological Society / APS*, *20*(6), 771–7. https://doi.org/10.1111/j.1467-9280.2009.02359.x

Yuki, M., Maddux, W. W., & 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. *Journal of Experimental Social Psychology*, *43*(2), 303–311. https://doi.org/10.1016/j.jesp.2006.02.004

Zhu, L. L., & Beauchamp, M. S. (2017). Mouth and Voice: A Relationship between Visual and Auditory Preference in the Human Superior Temporal Sulcus. *The Journal of Neuroscience*, *37*(10), 2697–2708. https://doi.org/10.1523/JNEUROSCI.2914-16.2017

Zicarelli, D. D. (1998). An extensible real-time signal processing environment for Max. *Proceedings of the 1998 International Computer Music Conference*, 463–466. Retrieved from http://www.citeulike.org/group/12573/article/6483609