A multi-sensory code for emotional arousal

Beau Sievers[a], Caitlyn Lee[b], William Haslett[c] & Thalia Wheatley[b]

[a]Department of Psychology, Harvard University, Cambridge, MA 02138
[b]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH 03755
[c]Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, NH 03755

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

Abstract

People express emotion using their voice, face and movement, as well as through abstract forms as in art, architecture and music. The structure of these expressions often seems intuitively linked to its meaning: e.g., romantic poetry is written in flowery curlicues, while the logos of death metal bands use spiky script. Here we show that these associations are universally understood because they are signaled using a multi-sensory code for emotional arousal. Specifically, variation in the central tendency of the frequency spectrum of a stimulus—its spectral centroid—is used by signal senders to express emotional arousal, and by signal receivers to make emotional arousal judgments. We show that this code is used across sounds, shapes, speech, and human body movements, providing a strong multi-sensory signal that can be used to efficiently estimate an agent's level of emotional arousal.

Introduction

Arousal is a fundamental dimension of emotional experience that shapes our behavior. We bark in anger and slink away in despair; we sigh in contentment and jump for joy. The studies presented here are motivated by the observation that these and other expressions of emotional arousal seem intuitively to match each other, even across the senses: Angry shouting is frequently accompanied by jagged, unpredictable flailing, and peaceful reassurance with deliberate, measured movement. These correspondences abound in the art and artifacts of everyday life. Lullabies and Zen gardens both embody a tranquil simplicity; anthems and monuments a kind of inspiring grandeur. And, remarkably, expressions of emotional arousal are interpretable not only across cultures (Russell, Lewicka, & Niit, 1989; Sauter, Eisner, Ekman, & Scott, 2010), but across species as well (Faragó et al., 2014; Filippi et al., 2017). What accounts for the multi-sensory consistency and universal understanding of emotional arousal?

Previous research has shown that low-level physical features of shapes, speech, colors, and movements influence emotion judgments (Holland & Wertheimer, 1964; Lim & Okuno, 2012; Lundholm, 1921; Palmer, Schloss, Xu, & Prado-León, 2013; Poffenberger & Barrows, 1924; Sievers, Polansky, Casey, & Wheatley, 2013). These physical features include the arrangement and number of lines, corners, and rounded edges in shapes, the pitch, volume, and timbre of speech, the hue, saturation, and brightness of colors, and the speed, jitter, and direction of movements. In particular, spectral features have been shown to predict emotional arousal judgments of human speech (Banse & Scherer, 1996), nonverbal vocalizations (Sauter, Eisner, Calder, & Scott, 2010), and music (Gingras, Marin, & Fitch, 2014), as well as of dog barks (Faragó et al., 2014; Pongrácz, Molnár, & Miklósi, 2006; Pongrácz, Molnár, Miklósi, & Csányi, 2005), and of vocalizations across all classes of terrestrial vertebrates (Filippi et al., 2017). Accordingly, determining exactly how information about internal states (including emotions) is transmitted using low-level stimulus properties is a necessary step toward understanding inter-species communication (Marler, 1961; Seyfarth & Cheney, 2017).

Based on these findings, we propose that expressions of emotional arousal are universally understood across cultures and species because they are signaled (Otte, 1974; Owren, Rendall, & Ryan, 2010) using a multi-sensory code. Here, we test this hypothesis, showing over five studies that signal senders and receivers both encode and decode variation in emotional arousal using variation in the spectral centroid, a low-level stimulus feature that is shared across sensory modalities. We further show that the spectral centroid predicts arousal across a range of emotions, regardless of whether those emotions are positively or negatively valenced. (N.B.: Our hypothesis is limited to emotional arousal and the spectral centroid. We do not intend to suggest, for example, that the

complete form of a shape with a given level of emotional arousal is recoverable from a similarly expressive sound—only that, all else being equal, shapes and sounds with the same level of emotional arousal will have similar spectral centroids.)

Study 1 tests this hypothesis using a classic sound–shape correspondence paradigm (Köhler, 1929; Ramachandran & Hubbard, 2003), comparing judgments of emotional arousal for shapes and sounds with either high or low spectral centroids. Study 2 tests that participants in fact use the spectral centroid to express emotion in the visual domain by asking them to draw many emotionally expressive shapes and comparing their spectral centroids. Study 3 tests that the spectral centroid predicts emotional arousal better than other candidate features across a wide range of shapes and sounds. Study 4 tests that the spectral centroid predicts judgments of emotional arousal in natural stimuli. Finally, Study 5 tests that the spectral centroid predicts continuously varying emotional arousal across a wide range of emotion categories. Findings from all studies are summarized in Table 1.

	Task	Stimuli	Emotions	Result
Study 1	Emotion judgment (categorical)	Shapes and sounds based on Köhler (1929)	Angry, Sad, Excited, Peaceful	77–89% accuracy
Study 2	Emotion expression (categorical)	Shapes drawn by participants	Angry, Sad, Excited, Peaceful	71–78% accuracy
Study 3	Emotion judgment (categorical)	Procedurally generated shapes (n=390) and sounds (n=390)	Angry, Sad, Excited, Peaceful	78% accuracy
Study 4	Emotion judgment (categorical)	Natural speech and body movements	Angry, Sad	86–88% accuracy
Study 5	Emotional arousal judgment (continuous)	Natural speech and body movements	Anger, Boredom, Disgust, Anxiety/Fear, Happiness, Sadness, Neutral	SC β=.76

Table 1. Summary of findings. Each study tested how well the spectral centroid predicted emotional arousal. Studies 1–4 used categorical emotion labels, while Study 5 used continuous judgments of emotional arousal. The Result column indicates the accuracy of a logistic regression classifier that uses the spectral centroid to predict emotional arousal, except for Study 5, where we report the beta weight associated with the spectral centroid for continuous prediction of emotional arousal.

Data, code, stimuli, and materials

Data, code, stimuli, and materials for all studies and analyses can be downloaded at: https://github.com/beausievers/supramodal_arousal

Study 1: The spectral centroid predicts emotional arousal in shapes and sounds

Methods

Participants were 262 students, faculty, and staff of Dartmouth College. Because these experiments were brief (1–2 minutes), participants viewed an information sheet in lieu of a full informed consent procedure and were compensated with their choice of bite-sized candies. Sample sizes for all tasks reported in this manuscript were determined as described in the Supplementary Information. Ethical approval was obtained from the Dartmouth College institutional review board. For all studies and tasks, participants were excluded from analysis if they reported knowledge of Köhler's (1929) experiment. Eight participants were excluded from Study 1 for this reason.

For all reported studies, we compared auditory and visual stimuli in the same terms: their observable frequency spectra, obtained via the Fourier transform. In time-varying stimuli such as human speech and movement we quantify the central tendency of the frequency spectrum as the spectral centroid (SC). SC is defined in Equation 1:

$$SC(M_i) = \frac{\sum f_i M_i}{\sum M_i} \quad (1)$$

Where M is a magnitude spectrum possessing frequency components f, both indexed by i (Schubert & Wolfe, 2006). The SC can be understood as the "center of mass" of a spectrum—the frequency around which the majority of the energy is concentrated.

We estimated the SC of two-dimensional drawings using a different method. When a two-dimensional contour is closed and non-overlapping, it can be represented by a Fourier shape descriptor (Zahn & Roskies, 1972). Shapes with more sharp corners require higher magnitude high-frequency components, pushing up their SC (see Figure 1 for a visual explanation). Here, we analyzed drawings containing open, overlapping contours, and therefore did not use Fourier descriptors. Instead, we took advantage of the link between corners and high-frequency energy components, estimating the SC of shapes using Harris corner detection (Harris & Stephens, 1988).

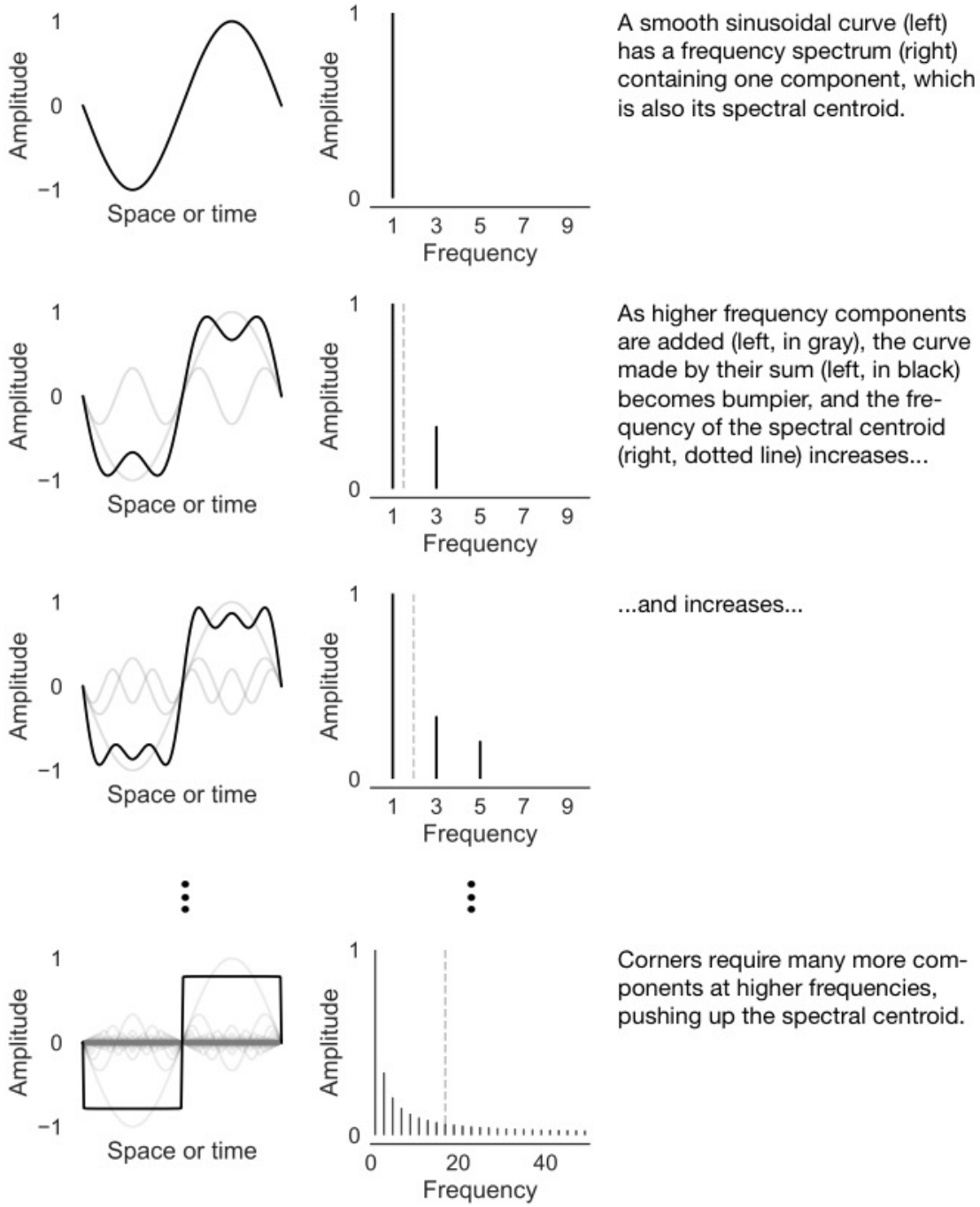

Figure 1. Visual explanation of the link between spectral centroid and corners.

We hypothesized that sounds and shapes with higher SC relative to similar stimuli would be associated with high arousal emotions. Two shape pairs with differing SCs were created based on Köhler's (1929) shapes, where the general outline, positions of line crossings, and size were matched for each pair (Figure 2). We used two non-phonetic sounds, 1.5s in duration, with differing mean SCs

(Figure 3). One sound, designed to match the spiky shape, consisted of four 54ms bursts of white noise in an asymmetrical rhythm. The second sound, designed to match the rounded shape, consisted of an amplitude-modulated sine wave at 250 Hz with volume peaks creating an asymmetrical rhythm similar to the noise burst pattern. To account for the possibility that the effect of SC differed across positive and negative emotions, we used two emotion pairs (Angry/Sad and Excited/Peaceful) differing in valence (Russell, 1980). In all tasks, each participant completed a single trial, and the order of shapes, sounds, and emotion words was counterbalanced across participants.

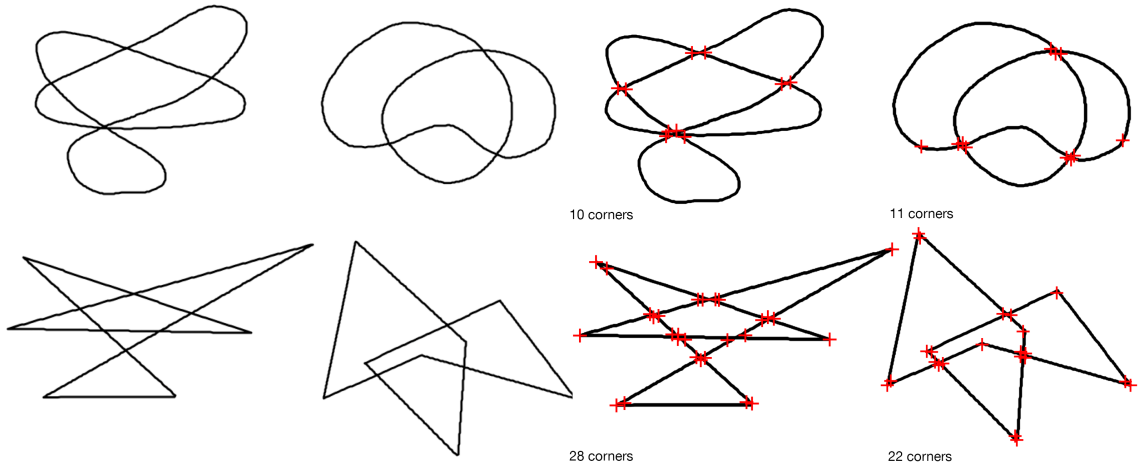

Figure 2. Shape pairs used in tasks 1 and 2, based on Köhler's (1929) shapes, before and after Harris corner detection. *Top row:* Low SC, low corner count shapes. *Bottom row:* High SC, high corner count shapes.

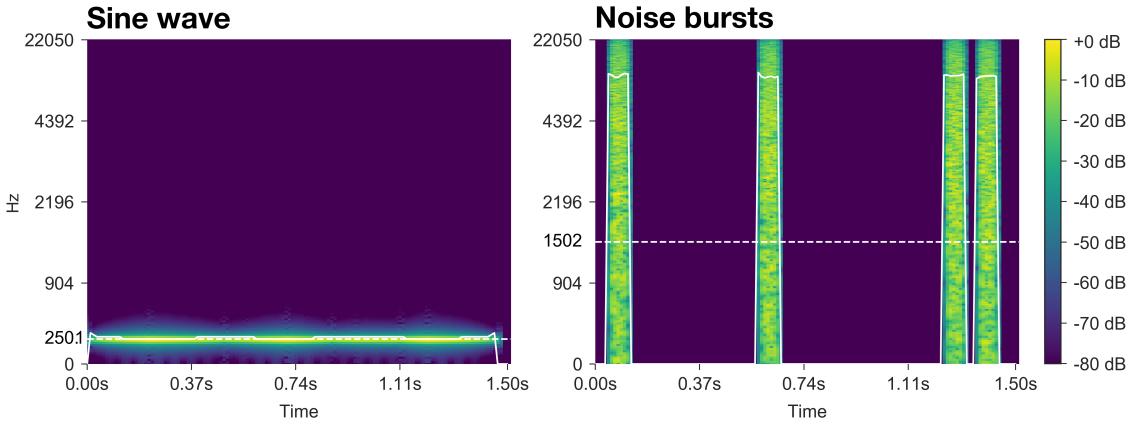

Figure 3. Spectrograms of sound stimuli. Solid white line shows instantaneous spectral centroid; dotted white line shows mean spectral centroid.

Task 1: Shape–emotion matching. In the negative valence shape–emotion task, participants (*N*=60) viewed a single shape and chose one of two emotion labels. The instructions took the form: “Press 1 if this shape is angry. Press 2 if this shape is sad.” Participants in the positive valence shape–emotion task (*N*=71) viewed a single shape and answered a pen-and-paper survey reading “Is the above shape Peaceful or Excited? Circle one emotion.” Data in the negative shape–emotion matching task were originally collected and reported by Sievers et al. (2017).

Task 2: Sound–emotion matching. In the negative valence sound–emotion task, participants (*N*=59) heard one sound and chose one of two emotion labels (Angry and Sad). The instructions took the form: “Click on the emotion that goes with *[noise bursts].*” Participants in the positive valence sound–emotion task (*N*=71) heard one of the sounds and answered a pen-and-paper survey reading “Is the above shape Angry or Sad? Circle one emotion.”

Results

Matching. Across all matching tasks, participants associated high mean SC sounds (noise burst) and shapes (spiky) with high arousal emotions (Angry, Excited), and low mean SC sounds (sine wave) and shapes (rounded) with low arousal emotions (Sad, Peaceful).

The sound–emotion estimated success rate was 89% (95% CI: .80–.96, *N*=59) for negative emotions and 85% (95% CI: .75–.92, *N*=68) for positive emotions, and the shape–emotion estimated success rate was 83% (95% CI: .73–.91, *N*=60) for negative emotions and 77% (95% CI: .67–.87, *N*=63) for positive emotions (Figure 4).

These results are consistent with the hypothesis that expressions of emotional arousal are universally understood because they share observable, low-level features, even across sensory domains. Specifically, the spectral centroid of both auditory and visual stimuli predicted perceived emotional arousal across sounds and shapes.

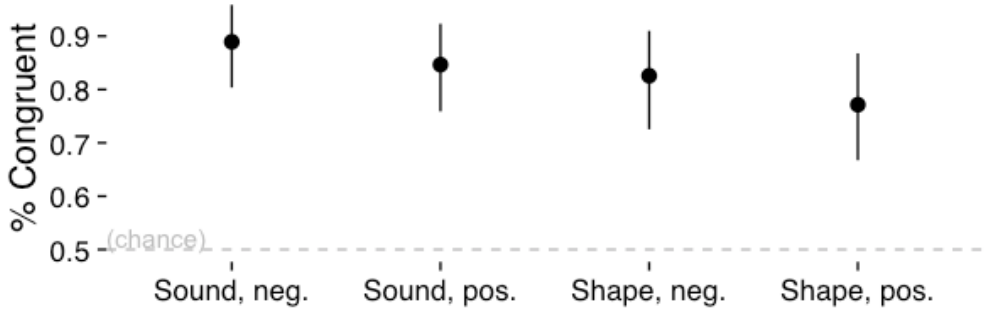

Figure 4. Study 1 results. Angry and Excited (high arousal) were associated with high spectral centroid stimuli (noise bursts and spiky shapes), while Sad and Peaceful (low arousal) were associated with low spectral centroid stimuli (sine wave and rounded shapes).

Study 2: Drawing emotional shapes

Methods

To directly test the hypothesis that people use SC variation to express emotional arousal, we asked participants to draw abstract shapes that expressed emotions. Participants (*N*=68) completed a pen-and-paper survey. The first page of the survey showed a single shape with the text: **"**Is the above shape Angry or Sad? Circle one emotion." Below this question, participants were instructed to draw a shape that expressed the emotion they did *not* circle. On the second page, participants answered two free-response questions: "*What is different between the shape we provided and the shape you drew?"* and "*Why did you draw your shape the way you did?"* Participants were given no information about our hypothesis, and their responses were not constrained to be either spiky or rounded. We replicated this study using a second survey (*N*=152) that did not show an example shape, and included the positively valenced emotions Excited and Peaceful. This survey had two questions: "Below, draw an abstract shape that is [emotion]" and "Why did you draw your shape the way you did?" Surveys are available at https://github.com/beausievers/supramodal_arousal.

Bayesian logistic regression classification was used to determine whether the SC, estimated using Harris corner detection (Harris & Stephens, 1988), was sufficient to predict emotional arousal. All classification analyses used 5-fold stratified cross-validation. All Receiver Operating Characteristic (ROC) analyses

were conducted using the same cross-validation training and test sets as their corresponding classification analysis.

Results

In responses to our first survey, including a prompt image and using only negatively valenced emotions, drawings of Angry shapes had a mean of 23.3 corners, while Sad shapes had 6.6 corners. Bayesian logistic regression classified Angry and Sad shapes with 78% accuracy and 85% area under the curve (AUC) based on the number of corners alone. Responses to our second survey, with no prompt image and including both positive and negative emotions, showed Angry shapes had a mean of 24.2 corners, Sad shapes had a mean of 8.8 corners, Excited shapes had a mean of 17.1 corners, and Peaceful shapes had a mean of 6.8 corners (Figure 5). Bayesian logistic regression classified Angry and Sad shapes with 75% accuracy and 84% AUC, and Excited and Peaceful shapes with 71% accuracy and 77% AUC (Figure 6).

Most participants were aware of their strategy, using terms such as “jagged,” and “sharp,” noting that Angry emotion implied “pointed” images with many crossing lines while Sad emotion implied “round” or “smooth” images with fewer or no crossing lines. Participants sometimes used crossmodal and affective language to describe their drawings, describing Angry as having “fast movement,” being “frantic,” etc., while Sad was “soft.”

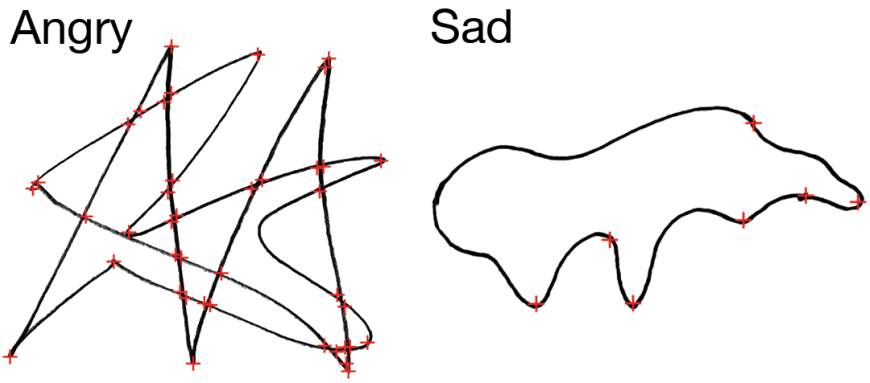

Figure 5. Characteristic Angry and Sad drawings from Study 3, after smoothing and corner detection. Corners are marked with red ‘+’ signs. Among our participants, Angry drawings had a mean of 23.3 corners, while Sad drawings had a mean of 6.6 corners.

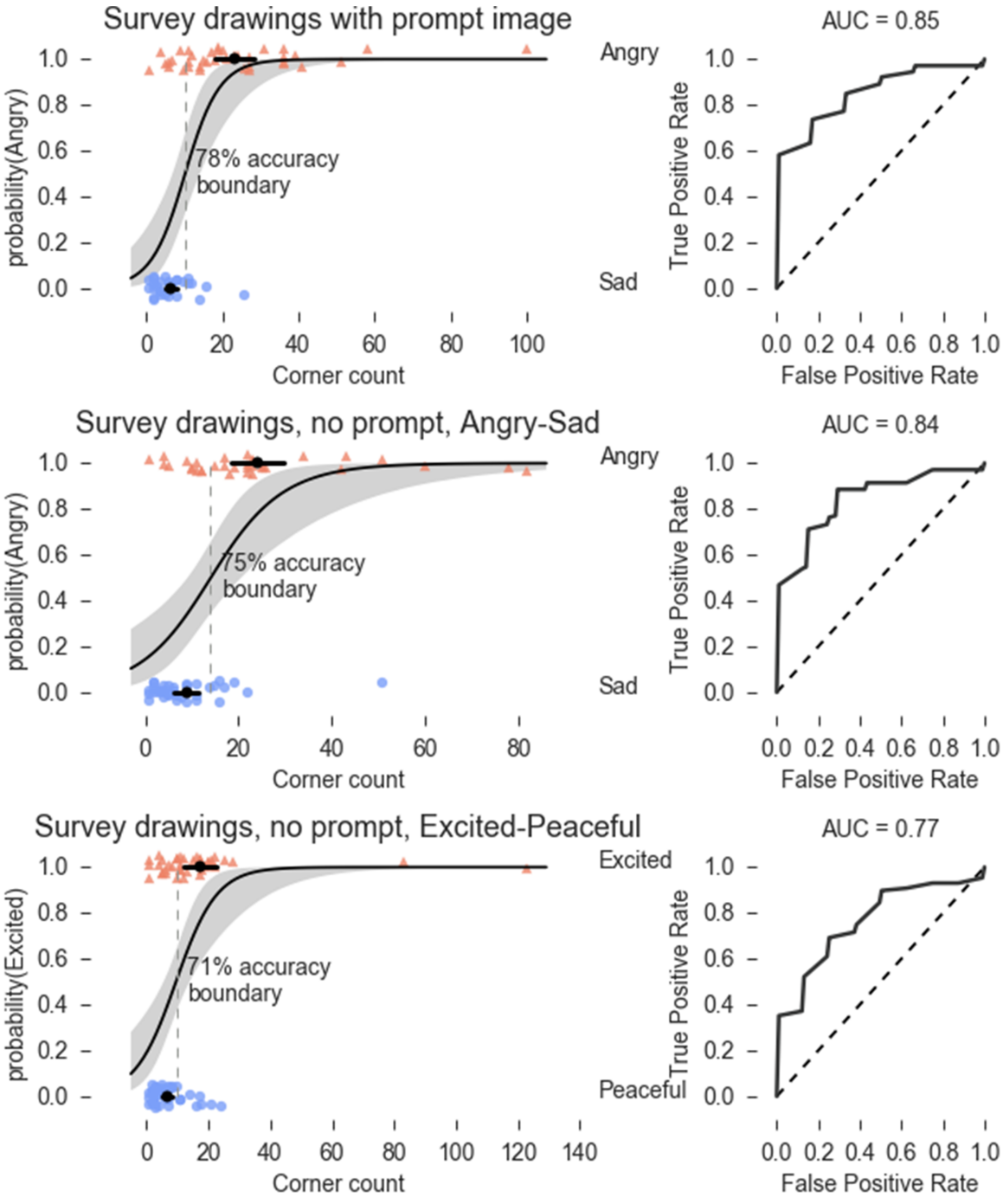

Figure 6. Bayesian logistic regression classification of Angry, Sad, Excited, and Peaceful drawings from our survey participants. For all logistic regression plots, black dots and bars show means and 95% confidence intervals of the mean, and the dotted line shows the 50% probability boundary. *Top:* Participants were shown a prompt image and asked to draw a shape conveying the opposite arousal negative emotion. *Middle & Bottom:* Participants were shown no prompt image and asked to draw a shape that was Angry, Sad, Excited, or Peaceful.

Study 3: The spectral centroid predicts emotional arousal across many shapes and sounds

Methods

In Study 3, we procedurally generated 390 shapes and 390 sounds that varied broadly in their observable features. These features included convexity, area, and corner count for shapes, and onset strength, the number of onsets, duration, and SC for sounds (see Supplementary Information for details). Stimuli were separated into 13 equally spaced bins of 30 stimuli each based on Harris corner count (range: 1–14) or SC (range: 0–1950Hz). Stimuli were generated such that SC and corner count were not strongly correlated with other features that could contribute to emotion judgment (Figure 7). This enabled accurate estimation of effect sizes for each feature.

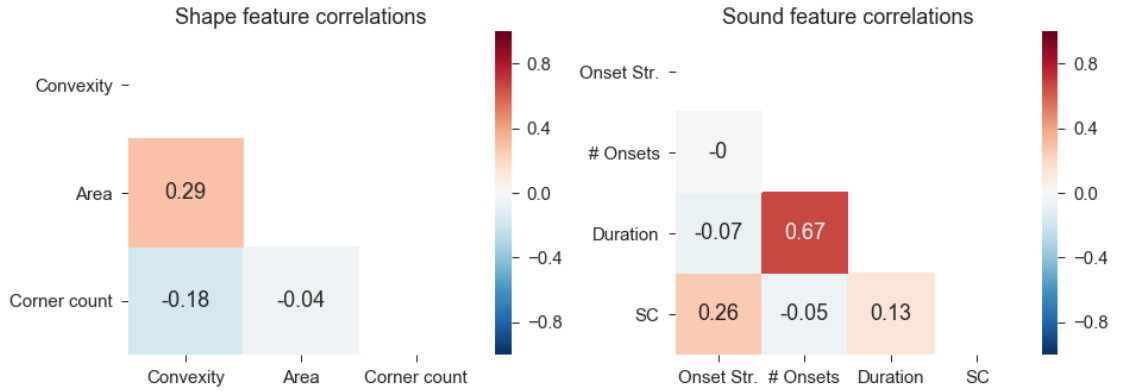

Figure 7. Correlation matrices for features of procedurally generated shape (left) and sound stimuli (right). Corner count and spectral centroid were not strongly correlated with other features, enabling accurate estimation of effect sizes for each feature.

Participants in the positive valence condition (N=20) judged whether stimuli were Excited or Peaceful, and a separate group of participants in the negative valence condition (N=20) judged whether stimuli were Angry or Sad. A judgment trial consisted of attending to a randomly selected stimulus (presented on screen or in headphones) then clicking on the best fitting emotion label. Stimulus order was randomized, and the position of emotion label buttons was counterbalanced across participants. Participants performed 40 trials for each stimulus bin, totaling 520 trials. 4 participants performed fewer trials (156, 156, 286, and 416 trials) due to a technical error, and 104 trials were excluded from analysis because participants indicated they could not hear the sound, giving 9,862 total shape trials and 9,768 total sound trials.

To test our hypothesis that expressions of emotional arousal are universally understood because they share observable, low-level features across sensory domains, judgments of emotional arousal were predicted using Bayesian hierarchical logistic regression. The contribution of SC to emotional arousal judgments in shapes and sounds was estimated by training a full model on all trials in both modalities. Predictors included SC/corner count bin, modality, and valence. Additionally, to facilitate comparison of the predictors, separate models were trained for each single predictor. To assess the contribution of modality-specific features to emotion judgments, this process was repeated for each modality. For sounds and shapes, we constructed a full model including all modality-specific predictors, as well as additional models for each single predictor. For shapes, predictors included corner count bin, rectangular bounding box area, and convexity (the ratio of shape area to bounding box area). For sounds, predictors included SC bin, duration, number of onsets, and mean onset strength. The number of onsets is sensitive to peaks in amplitude, and onset strength to changes in spectral flux. Onset features were calculated using LibROSA's onset_detect and onset_strength functions (McFee et al., 2015).

All models used random slopes and intercepts per participant, as well as weakly informative priors that shrank parameter estimates toward zero (see Supplementary Information). All scalar predictors were z-scored before model fitting. Model accuracy was determined using 5-fold cross-validation, with folds stratified by participant ID to ensure each fold contained trials from all participants. ROC analyses were conducted using the same cross-validation training and test sets used to determine model accuracy. Chance thresholds were determined by the base rates of participant emotion judgments. Note that because these models capture inter-participant uncertainty, cross-validated accuracy and AUC are conservative estimates of goodness of fit.

Results

Results are summarized in Figure 8. The full multimodal model predicted high arousal emotion judgments of shapes and sounds with 78% accuracy (SD < .001, 86% AUC, N=40). The multimodal SC/corner count bin-only model was comparably accurate (77% acc., SD=.006, 84% AUC, N=40), and the modality and valence models did not exceed the chance threshold, providing strong evidence that SC drives judgments of emotional arousal across modalities. Estimates of interaction weights for SC bin with valence (M=.72, SD=.17) and modality (M=.69, SD=.15) were positive, indicating sounds were perceived as having slightly higher arousal than shapes, and that participants made more high arousal judgments when the choice of emotions was positively valenced. Random effects terms indicated moderate variability in slopes and intercepts across participants, suggesting individual differences in the tendency to identify stimuli as having high emotional arousal (SD=.36), and in the effect of SC on

emotional arousal judgments (SD=.42). See Supplementary Information for tables summarizing all estimated model weights.

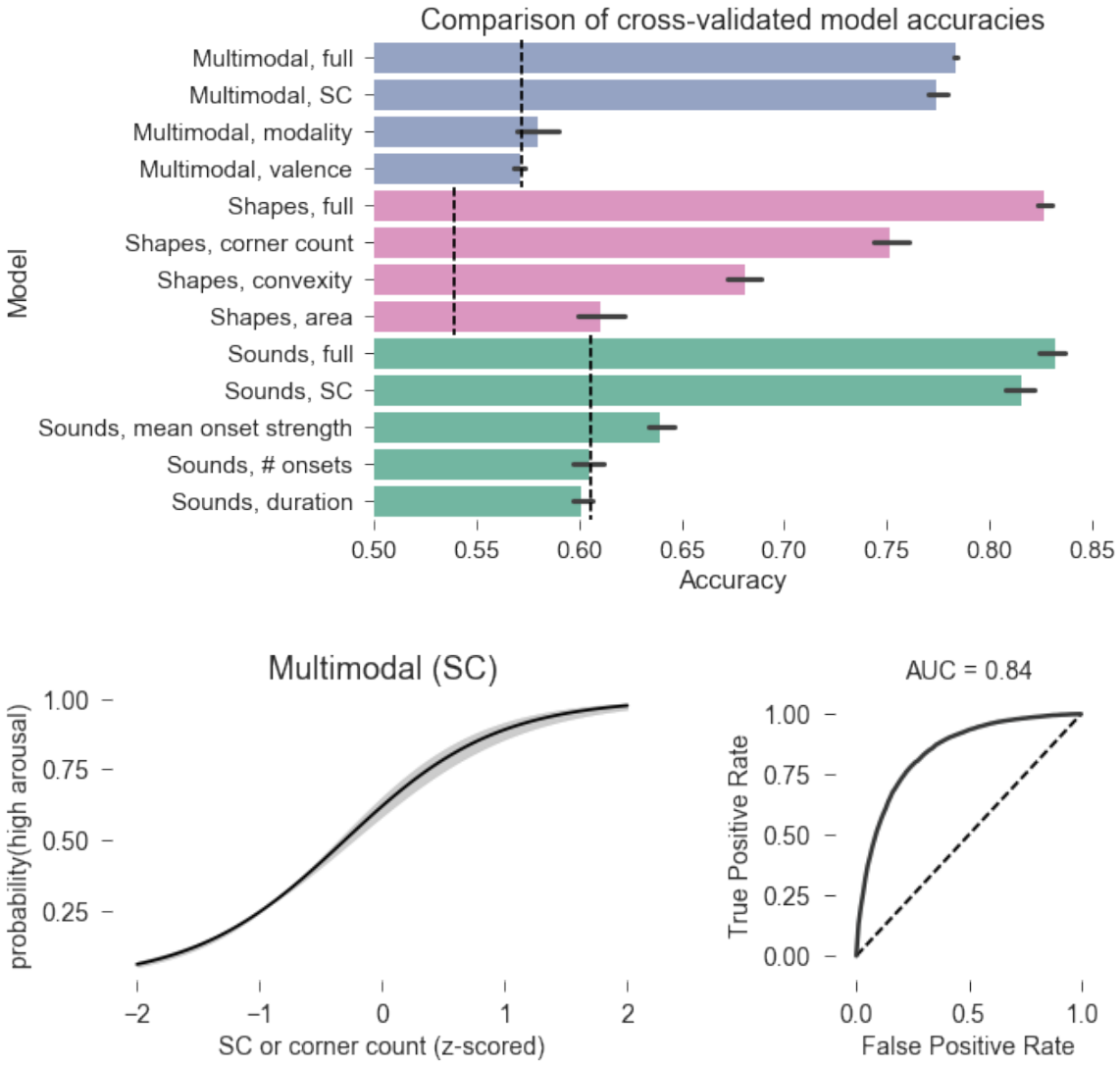

Figure 8. Bayesian logistic regression classification of participant emotion judgments of procedurally generated shapes and sounds. *Top:* Comparison of model accuracies. Dashed lines indicate chance performance. “Full” models included all predictors and interactions, others were limited to the single labeled predictor. All models included random slopes and intercepts per participant. *Bottom:* Fixed effect of SC in the multimodal, single-predictor model. Lines and shaded areas show mean and 95% credible interval. See Supplementary Information for all model fit plots.

Modality-specific models revealed SC was the best single predictor of high arousal emotion judgments for shapes (75% acc., SD=.011, 83% AUC, N=20) and sounds (82% acc., SD=.008, 89% AUC, N=20). For shapes, convexity (68% acc., SD=.01, 73% AUC, N=20) and bounding box area (61% ACC., SD=.01, 64% AUC, N=20) also predicted high arousal emotion better than chance. For sounds, mean onset strength weakly predicted high arousal emotion (63% acc., SD=.007, 69% AUC, N=20), while single-predictor models using the number of onsets and duration did not perform better than chance. The inclusion of modality-specific predictors may explain the slightly higher accuracy of modality-specific models (shapes: 83% acc., SD=.004, 91% AUC, N=20; sounds: 83% acc., SD=.008, 91% AUC, N=20).

Study 4: Spectral centroid analysis of naturalistic emotion movement and speech databases

Methods

To test whether spectral centroid predicts emotional arousal judgments in naturalistic stimuli, we analyzed the Berlin Database of Emotional Speech (BDES; Burkhardt, Paeschke, Rolfes, Sendlmeier, & Weiss, 2005) and the PACO Body Movement Library (PBML; Ma, Paterson, & Pollick, 2006). The PBML consists of point-light movement recordings of male and female actors expressing several emotions. The BDES consists of audio recordings of male and female actors reading the same emotionally neutral German sentences while expressing several emotions. We hypothesized that SC would predict the emotional arousal of stimuli in both databases. For the BDES, we used the mean SC of each emotional expression. For the PBML, a collection of point-light movements, we used the mean SC for each joint axis, reducing each expression to 45 features. The BDES includes 127 Angry and 62 Sad expressions, while the PBML has 60 expressions per emotion.

Results

Mean SC was consistently higher in Angry vs. Sad speech. This pattern was robust across genders and held for eight of ten speakers. To compensate for inter-speaker differences, SC was standardized per speaker. Using SC alone, Bayesian logistic regression classified Angry and Sad expressions with 86% accuracy and 90% AUC (Figure 9).

To assess whether the high dimensional structure of emotional movement data in the PBML could be reduced to a single arousal dimension, we performed a Bayesian linear discriminant analysis using stratified 10-fold cross-validation. The mean SCs of 45 movement dimensions (15 joints by 3 axes per joint) were projected on to a single linear discriminant, resulting in 88% classification

accuracy and 88% AUC (Figure 10). See Supplementary Information for additional results.

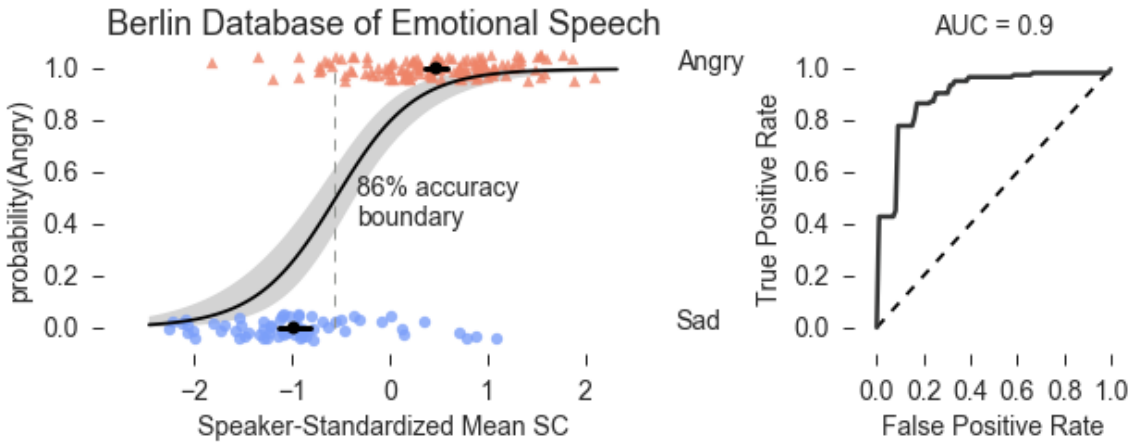

Figure 9. Bayesian logistic regression classification of Angry and Sad examples from the Berlin Database of Emotional Speech.

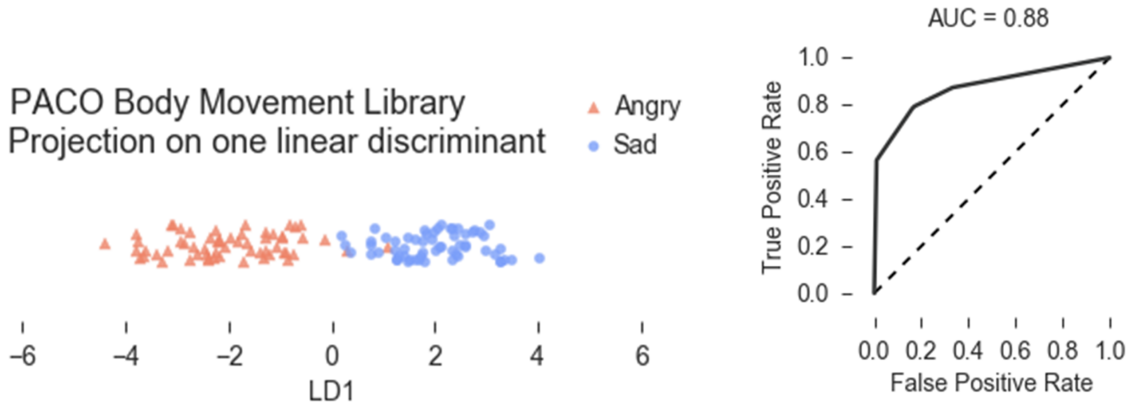

Figure 10. *Left:* Projection of Angry and Sad movements from the PACO Body Movement Library onto a single linear discriminant using Bayes rule. *Right:* ROC analysis of LDA-based classification.

Study 5: Continuous arousal across emotions

The previous studies tested the emotions Angry, Sad, Excited, and Peaceful, but did not directly test the hypothesis that the spectral centroid predicts continuously varying arousal levels across a wide range of emotion categories. To test this, we used the full BDES and PACO stimulus databases introduced in Study 4, including 775 unique stimuli drawn from all emotion categories (BDES: Anger, Boredom, Disgust, Anxiety/Fear, Happiness, Sadness, Neutral; PACO: Angry, Happy, Sad, Neutral). Participants (N=50) that were fluent in English and did not understand German completed 220 arousal judgment trials each, where they attended to a single randomly selected speech or movement stimulus and

used slider bars to rate its valence and arousal on continuous scales from 0 to 100. Seven participants completed fewer than 220 trials, giving a final total of 10,934 trials.

Results were analyzed using Bayesian hierarchical linear regression, with valence, arousal, sensory modality (sound vs. vision) and all interactions as predictors, and with spectral centroid as the dependent variable. Note that predicting the continuous spectral centroid is more challenging than predicting discrete emotion category labels. All models used random slopes and intercepts per participant and all values were z-scored before model fitting. All parameters, including the coefficient of determination (R^2), were fit using 5-fold cross-validation, with folds stratified by participant ID.

Results

The model was able to perform the challenging task of predicting the spectral centroid of a stimulus from its valence and arousal ratings (R^2=.3), across emotion categories and modalities. Stimuli rated higher in arousal had higher spectral centroids (β=.76, 95% CI: .71–.81), while stimuli rated as having more positive valence had slightly lower spectral centroids (β=-.09, 95% CI: -.13–-.05). These main effects were qualified by arousal-modality (β=-.22, 95% CI: -.26–-.18) and valence-modality (β=-.06, 95% CI: -.11–-.02) interactions, indicating that arousal and valence judgments both had slightly stronger effects on the predicted spectral centroids of movements than speech. We also observed a valence-arousal interaction (β=-.1, 95% CI: -.13–.07), indicating stimuli that were both highly arousing and positively valenced had slightly lower spectral centroids and a valence-arousal-modality interaction (β=.15, 95% CI: .11–.18) indicating that this effect was slightly weaker for speech than for movement. The relationships between valence, arousal, and spectral centroid are illustrated in Figure 11.

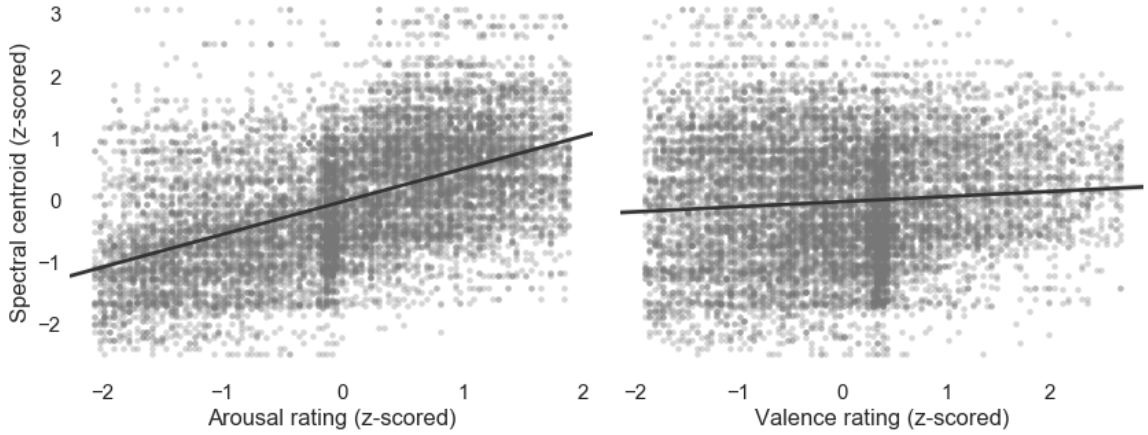

Figure 11. Visualizations of the relationship between arousal, valence, and the spectral centroid. See above for complete hierarchical Bayesian linear regression results. *Left:* Stimuli rated as having higher arousal also have higher spectral centroids. *Right:* The relationship between valence and spectral centroid is relatively weak.

Discussion

Across a series of empirical studies and analyses of extant databases, the spectral centroid predicted emotional arousal in abstract shapes and sounds, body movements, and human speech. Study 1 showed that participants map the SCs of shapes and sounds (similar to those used in Köhler's classic demonstration of the *Bouba-Kiki* effect) to the arousal level of emotions. Study 2 asked many participants to draw shapes expressing high or low emotional arousal. With no additional instruction, participants matched the SCs of their drawings to the level of emotional arousal, showing that SC matching is a consistent feature of emotion expression. Study 3 additionally tested the generalizability of this matching principle to 390 different shapes and sounds, showing that the SC (over and above other stimulus features) predicted emotional arousal judgments. Study 4 further generalized this finding to previously-existing databases of body movements and emotional speech. Finally, Study 5 demonstrated that the spectral centroid predicts continuous emotional arousal judgments across a wide range of emotion categories, showing that the predictive power of the SC does not depend on forced-choice experimental paradigms or discrete emotion labels. Taken together, these studies provide strong evidence that expressions of emotional arousal are universally understood because they are signaled using a multi-sensory code, where signal senders and receivers both encode and decode variation in emotional arousal using variation in the central tendency of the frequency spectrum.

Across all of these studies, we estimated the central tendency of the frequency spectrum using SC and Harris corner detection. We do not suggest that SC and Harris corner detection are the only tools appropriate for this task, nor that matching in the frequency domain is the sole cause of emotional arousal judgments. Study 3 identified several other features that predict arousal (e.g., convexity in shapes, mean onset strength in sounds), although none were as accurate as the SC. Additionally, other measures that closely track the central tendency of the frequency spectrum (e.g., local entropy, pitch) should also predict emotional arousal.

We limited our studies to perceptual modalities where the frequency spectrum is observable, and predict similar results to obtain whenever this is the case (e.g., in touch and vibration). However, our results do not speak to crossmodal mappings when the frequency spectrum cannot be observed. This is the case for taste and smell, which are sensitive to variation in molecular shape. In such cases, we predict crossmodal correspondences depend on statistical learning or deliberative processes. A difference in the underlying mechanism may create measurably different behavior, and may explain why some crossmodal correspondences involving taste vary across cultures (Bremner et al., 2013; Knoeferle, Woods, Käppler, & Spence, 2015).

The findings reported here are consistent with the “common code for magnitude” theory of Spector and Maurer (2009), where “more” in one modality corresponds with “more” in other modalities. Following this theory, is possible that people use a single supramodal spectral centroid representation to assess whether there is more emotional arousal. Alternatively, it is possible that perception of emotional arousal depends on higher-level abilities such as language use and reasoning (Martino & Marks, 1999). Although these two accounts are not mutually exclusive, developmental studies show that perception of relevant crossmodal correspondences occurs before language and reasoning competence: pre-linguistic infants show the *Bouba-Kiki* effect (Ozturk, Krehm, & Vouloumanos, 2013) and associate auditory pitch with visual height and sharpness (Walker et al., 2010). Further, words learned early in language acquisition are more iconic than those learned later (Monaghan, Shillcock, Christiansen, & Kirby, 2014; Perry, Perlman, Winter, Massaro, & Lupyan, 2018), suggesting that crossmodal correspondences come prior to, and are important for the development of, language competence. Although these findings support a “common code” or supramodal representation account, additional research is required to characterize the specific internal representations employed and to chart the course of their development. In particular, it would be valuable to assess to what extent the internal representations used are innate “core knowledge” endowed by evolution (Spelke & Kinzler, 2007), and to what extent they develop over the lifetime due to the operation of modality-general statistical learning and predictive coding mechanisms (Lupyan & Clark, 2015; Saffran & Kirkham, 2018).

A supramodal spectral centroid representation may arise from natural selection. To wit, one function of emotional states is facilitating or constraining action (Damasio & Carvalho, 2013). For example, high emotional arousal facilitates big, spiky vocalizations and movements that necessarily have high spectral centroids. These crossmodal gestures function as crossmodal emotion signals. Organisms that can send and receive these signals have obvious fitness advantages: Senders share their emotional states so receivers know when to approach or avoid them, making both senders and receivers more likely to receive care and avoid harm. All else being equal, organisms with more efficient systems for detecting and representing emotion signals should have better reproductive fitness than those with less efficient systems. We should therefore expect a wide range of species to have an efficient means of detecting crossmodal signals of emotional arousal, and a supramodal spectral centroid representation is an excellent fit for this purpose. (Note that we should expect this even if there exist more accurate means of detecting emotion, such as language and reasoning, as long as those means tend to be less efficient.) Accordingly, preliminary evidence suggests the spectral centroid predicts emotional arousal across cultures (Sievers et al., 2017) and across species (Faragó et al., 2014; Filippi et al., 2017). By understanding how the brain extracts low-level, crossmodal features to determine meaning, we can build a deeper understanding of how communication can transcend immense geographic, cultural, and genetic variation.

Acknowledgements

We thank Professor Robert Caldwell at Dartmouth for helping us clarify the relationship between the Fourier transform and Harris corner detection, as well as Rebecca Drapkin, Evan Griffith, Erica Westenberg, Jasmine Xu, and Emmanuel Kim for assistance collecting data. This research was supported in part by a McNulty Grant from the Nelson A. Rockefeller Center (TW), and a Neukom Institute for Computational Science Graduate Fellowship (BS).

Author Contributions

B. Sievers and T. Wheatley contributed equally to study design. W. Haslett, B. Sievers, and C. Lee wrote the software. B. Sievers and C. Lee collected data. B. Sievers performed data analysis. B. Sievers, T. Wheatley, and C. Lee wrote the paper.

References

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636.

Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8851745

Bremner, A. J., Caparos, S., Davidoff, J., de Fockert, J., Linnell, K. J., & Spence, C. (2013). “Bouba” and “Kiki” in Namibia? A remote culture make similar shape-sound matches, but different shape-taste matches to Westerners. *Cognition*, *126*(2), 165–172. https://doi.org/10.1016/j.cognition.2012.09.007

Burkhardt, F., Paeschke, a, Rolfes, M., Sendlmeier, W., & Weiss, B. (2005). A Database of German Emotional Speech. *Ninth European Conference on Speech Communication and Technology*, *2005*, 3–6. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.130.8506&rep=rep1&type=pdf

Damasio, A., & Carvalho, G. B. (2013). The nature of feelings: Evolutionary and neurobiological origins. *Nature Reviews Neuroscience*, *14*(2), 143–152. https://doi.org/10.1038/nrn3403

Faragó, T., Andics, A., Devecseri, V., Kis, A., Gácsi, M., & Miklósi, Á. (2014). Humans rely on the same rules to assess emotional valence and intensity in conspecific and dog vocalizations. *Biology Letters*, *10*(1). https://doi.org/10.1098/rsbl.2013.0926

Filippi, P., Congdon, J. V., Hoang, J., Bowling, D. L., Reber, S. A., Pašukonis, A., … Güntürkün, O. (2017). Humans recognize emotional arousal in vocalizations across all classes of terrestrial vertebrates: evidence for acoustic universals. *Proceedings of the Royal Society of London B: Biological Sciences*, *284*(1859), 1–9. Retrieved from http://rspb.royalsocietypublishing.org/content/284/1859/20170990?etoc

Gingras, B., Marin, M. M., & Fitch, W. T. (2014). Beyond intensity: Spectral features effectively predict music-induced subjective arousal. *Quarterly Journal of Experimental Psychology (2006)*, *67*(7), 1428–1446. https://doi.org/10.1080/17470218.2013.863954

Harris, C., & Stephens, M. (1988). A Combined Corner and Edge Detector. *Procedings of the Alvey Vision Conference 1988*, 147–151. https://doi.org/10.5244/C.2.23

Holland, M., & Wertheimer, M. (1964). Some Physiognomic Aspects of Naming, or, Maluma and Takete Revisited. *Perceptual and Motor Skills*, *19*, 111–117. Retrieved from http://www.amsciepub.com/doi/pdf/10.2466/pms.1964.19.1.111

Knoeferle, K. M., Woods, A., Käppler, F., & Spence, C. (2015). That sounds sweet: using cross-modal correspondences to communicate gustatory attributes. *Psychology & Marketing*, *32*(1), 107–120.

Köhler, W. (1929). *Gestalt Psychology*. Oxford, England: Liveright.

Lim, A., & Okuno, H. G. (2012). Using Speech Data to Recognize Emotion in

Human Gait. In A. Salah, J. Ruiz-del-Solar, Ç. Meriçli, & P.-Y. Oudeyer (Eds.), *Human Behavior Understanding* (pp. 52–64). Berlin Heidelberg: Springer.

Lundholm, H. (1921). The affective tone of lines: experimental researches. *Psychological Review*, *28*(1), 43–60.

Lupyan, G., & Clark, A. (2015). Words and the world: Predictive coding and the language-perception-cognition interface. *Current Directions in Psychology*, 1–10. https://doi.org/10.1177/0963721415570732

Ma, Y., Paterson, H. M., & Pollick, F. E. (2006). A motion capture library for the study of identity, gender, and emotion perception from biological motion. *Behavior Research Methods*, *38*(1), 134–141. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/16817522

Marler, P. (1961). The logical analysis of animal communication. *Journal of Theoretical Biology*, *1*(1961), 295–317. https://doi.org/10.1016/0022-5193(61)90032-7

Martino, G., & Marks, L. E. (1999). Perceptual and linguistic interactions in speeded classification: Tests of the semantic coding hypothesis. *Perception*, *28*(7), 903–923. https://doi.org/10.1068/p2866

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and Music Signal Analysis in Python. *Proceedings of the 14th Python in Science Conference (SCIPY 2015)*, (Scipy), 18–25.

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *369*(1651), 20130299-. https://doi.org/10.1098/rstb.2013.0299

Otte, D. (1974). Effects and Functions in the Evolution of Signaling Systems. *Annual Review of Ecology and Systematics*, *5*(1), 385–417. https://doi.org/10.1146/annurev.es.05.110174.002125

Owren, M. J., Rendall, D., & Ryan, M. J. (2010). Redefining animal signaling: Influence versus information in communication. *Biology and Philosophy*, *25*(5), 755–780. https://doi.org/10.1007/s10539-010-9224-4

Ozturk, O., Krehm, M., & Vouloumanos, A. (2013). Sound symbolism in infancy: evidence for sound-shape cross-modal correspondences in 4-month-olds. *Journal of Experimental Child Psychology*, *114*(2), 173–186. https://doi.org/10.1016/j.jecp.2012.05.004

Palmer, S. E., Schloss, K. B., Xu, Z., & Prado-León, L. R. (2013). Music-color associations are mediated by emotion. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(22), 8836–8841.

https://doi.org/10.1073/pnas.1212562110

Perry, L. K., Perlman, M., Winter, B., Massaro, D. W., & Lupyan, G. (2018). Iconicity in the speech of children and adults. *Developmental Science*, *21*(3), 1–8. https://doi.org/10.1111/desc.12572

Poffenberger, A., & Barrows, B. (1924). The feeling value of lines. *Journal of Applied Psychology*, *8*(2), 187–205.

Pongrácz, P., Molnár, C., & Miklósi, Á. (2006). Acoustic parameters of dog barks carry emotional information for humans. *Applied Animal Behaviour Science*, *100*(3–4), 228–240. https://doi.org/10.1016/j.applanim.2005.12.004

Pongrácz, P., Molnár, C., Miklósi, Á., & Csányi, V. (2005). Human listeners are able to classify dog (Canis familiaris) barks recorded in different situations. *Journal of Comparative Psychology*, *119*(2), 136–144. https://doi.org/10.1037/0735-7036.119.2.136

Ramachandran, S., & Hubbard, E. M. (2003). Hearing colors, tasting shapes. *Scientific American*, *288*, 43–49.

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

Russell, J. A., Lewicka, M., & Niit, T. (1989). A cross-cultural study of a circumplex model of affect. *Journal of Personality and Social Psychology*, *57*(5), 848–856. https://doi.org/10.1037/0022-3514.57.5.848

Saffran, J. R., & Kirkham, N. Z. (2018). Infant Statistical Learning. *Annual Review of Psychology*, *69*, 181–203. https://doi.org/10.1146/annurev-psych-122216-011805

Sauter, D. A., Eisner, F., Calder, A. J., & Scott, S. K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology*, *63*(11), 2251–2272. https://doi.org/10.1080/17470211003721642

Sauter, D. A., Eisner, F., Ekman, P., & Scott, S. K. (2010). Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences*, *107*(6), 2408–2412. https://doi.org/10.1073/pnas.1508604112

Schubert, E., & Wolfe, J. (2006). Does timbral brightness scale with frequency and spectral centroid? *Acta Acustica United with Acustica*, *92*(5), 820–825.

Seyfarth, R. M., & Cheney, D. L. (2017). The origin of meaning in animal signals. *Animal Behaviour*, *124*, 339–346. https://doi.org/10.1016/j.anbehav.2016.05.020

Sievers, B., Parkinson, C., Walker, T., Haslett, W., & Wheatley, T. (2017). Low-level percepts predict emotion concepts across modalities and cultures.

Retrieved from https://psyarxiv.com/myg3b/

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), 70–75. https://doi.org/10.1073/pnas.1209023110

Spector, F., & Maurer, D. (2009). Synesthesia: a new approach to understanding the development of perception. *Developmental Psychology*, *45*(1), 175–189. https://doi.org/10.1037/a0014171

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Walker, P., Bremner, J. G., Mason, U., Spring, J., Mattock, K., Slater, A., & Johnson, S. P. (2010). Preverbal infants' sensitivity to synaesthetic cross-modality correspondences. *Psychological Science : A Journal of the American Psychological Society / APS*, *21*(1), 21–25. https://doi.org/10.1177/0956797609354734

Zahn, C. T., & Roskies, R. Z. (1972). Fourier descriptors for plane closed curves. *IEEE Transactions on Computers*, *C–21*(3), 269–281. https://doi.org/10.1109/TC.1972.5008949