Similar Representations of Emotions across Faces and Voices

Lisa Katharina Kuhn[1,2], Taeko Wydell[2], Nadine Lavan[3], Carolyn McGettigan[3], Lúcia Garrido[2]

1. Department of Psychology, Saarland University, Saarbruecken, Germany
2. Department of Life Sciences, Brunel University London, Uxbridge, UK
3. Department of Psychology, Royal Holloway, University of London, Egham, UK

Correspondence to:

Lisa Katharina Kuhn

Experimental Neuropsychology Unit, Department of Psychology, Saarland University,

66123 Saarbrücken, Germany

Tel.: (+49) 1749187903

lisa.kuhn@uni-saarland.de

Revision submission date: 02/12/2016

Acknowledgements

The authors declare no conflict of interest. This research was supported by an ESRC 1+3 PhD studentship to Lisa Kuhn (ES/I90042X/1). We thank Matthew Longo and two anonymous reviewers for comments on an earlier version of the manuscript.

Abstract

Emotions are a vital component of social communication, carried across a range of modalities and via different perceptual signals such as specific muscle contractions in the face and in the upper respiratory system. Previous studies have found that emotion recognition impairments after brain damage depend on the modality of presentation: recognition from faces may be impaired whilst recognition from voices remains preserved, and vice versa. On the other hand, there is also evidence for shared neural activation during emotion processing in both modalities. In a behavioural study, we investigated whether there are shared representations in the recognition of emotions from faces and voices. We used a within-subjects design in which participants rated the intensity of facial expressions and non-verbal vocalisations for each of the six basic emotion labels. For each participant and each modality, we then computed a representation matrix with the intensity ratings of each emotion. These matrices allowed us to examine the patterns of confusions between emotions and to characterise the representations of emotions within each modality. We then compared the representations *across* modalities by computing the correlations of the representation matrices across faces and voices. We found highly correlated matrices across modalities, which suggest similar representations of emotions across faces and voices. We also showed that these results could not be explained by commonalities between low-level visual and acoustic properties of the stimuli. We thus propose that there are similar or shared coding mechanisms for emotions which may act independently of modality, despite their distinct perceptual inputs.

Keywords (max 6)

Emotion recognition – modality – similar representations - faces – voices

Non-verbal expression of emotions has important evolutionary implications for survival as well as for communication (Darwin, 1965/1872; Hampson, Anders, & Mullin, 2006). We need to rapidly classify emotions in order to recognise threat, assess social situations, and behave accordingly — for example, by protecting our offspring from enemies. Previous studies have shown that, although emotions are highly complex, we can perceive and reliably classify basic emotions via different cues and in different modalities, such as faces (Ekman & Friesen, 1971; Smith, Cottrell, Gosselin, & Schyns, 2005), bodies (De Gelder, 2006), and voices (Belin, Fillion-Bilodeau, & Gosselin, 2008; Sauter, Eisner, Calder, & Scott, 2010).

Yovel and Belin (2013) have proposed that, despite their different sensory inputs, faces and voices might be processed using similar coding mechanisms during identity recognition. They reviewed cognitive, developmental, and neural evidence to show that there are many similarities between the representations of person identity from faces and voices, suggesting that there might be unified coding principles across the visual and auditory modalities. These unified coding principles could, for example, explain how ratings of characteristics such as information about height, or masculinity and femininity, can correlate and be matched across independent face and voice stimuli (Smith, Dunn, Baguley, & Stacey, 2016). Behavioural studies of individual differences have also found significant (though not high) correlations between adults' visual and vocal emotion recognition abilities (Borod et al., 2000; Palermo, O'Connor, Davis, Irons, & McKone, 2013). There is little knowledge, however, about whether emotions from faces and emotions from voices are also *represented* using similar coding mechanisms.

Processing Similarities

Similarities in emotion recognition between faces and voices could occur because there are similar neural coding mechanisms across modalities, even if they are implemented in different regions of the brain (comparable to what happens in person identity recognition, as suggested by Yovel & Belin, 2013). Alternatively, it is possible that emotional stimuli from different modalities are, at least in part, processed in the same brain regions and share the same neural mechanisms, irrespective of whether these regions process all emotions or just one emotion (Calder & Young, 2005).

There is some consistent evidence from past studies pointing towards the latter possibility of modality-independent brain areas that show similar processing for both faces and voices. For example, the amygdala is commonly associated with the recognition of a range of emotions in both faces (Adolphs, Tranel, Damasio, & Damasio, 1995; Adolphs, Tranel, & Damasio, 2001; Fitzgerald, Angstadt, Jelsone, Nathan, & Phan, 2006) and voices (Fecteau, Belin, Joanette, & Armony, 2007; Phillips et al., 1998; Scott, Young, Calder, Hellawell, Aggleton, & Johnson, 1997). Furthermore, the right somatosensory cortex has been implicated both in the discrimination of facial emotions (Adolphs, Damasio, Tranel, Cooper, & Damasio, 2000; Pitcher, Garrido, Walsh, & Duchaine, 2008) and vocal emotions (Adolphs, Damasio, & Tranel, 2002; Banissy, Sauter, Ward, Warren, Walsh, & Scott, 2010). The superior temporal sulcus is not only involved in processing facial expressions of emotion (Engell & Haxby, 2007) but also in processing emotional vocalizations (Fecteau et al., 2007). In addition, the medial prefrontal cortex and the left superior temporal sulcus shows highly correlated activity patterns during emotion recognition from faces, voices and bodies in healthy adults (Peelen, Atkinson, & Vuilleumier, 2010).

Studies that tested the same patients in both modalities also support the idea that facial and vocal emotion recognition share the same neural mechanisms. For example, patients with medial temporal lobe epilepsy have impaired recognition of facial as well as vocal expressions of emotion (Bonora et al., 2011). Similarly, patients with ventral frontal lobe damage show impairment in emotion identification across facial and vocal expressions – although not all patients exhibit this association (Hornak, Rolls, & Wade, 1996). Interestingly, a meta-analysis suggested that patients with Parkinson's Disease have difficulties recognising emotions from both voices and faces (Gray & Tickle-Degnen, 2010). Finally, emotion recognition is often impaired for faces and voices in autism (Philip et al., 2010), schizophrenia (Simpson, Pinkham, Kelsven, & Sasson, 2013), and in recently detoxified alcoholics (Kornreich et al., 2012), suggesting the existence of a core emotion network rather than separate modality-specific processes.

There are several reasons why modality-independent mechanisms of emotion recognition would occur. One possibility is that these support the rapid detection of negative emotions or, more generally, all emotional signals, which may aid survival. Indeed, ERP studies have reported fast

responses to emotional faces (Eimer & Holmes, 2007; Kiss & Eimer, 2008; Liddell, Williams, Rathjen, Shevrin, & Gordon, 2004) as well as emotional voices (Sauter & Eimer, 2009), at around 150 to 200 ms post stimulus onset. Alternatively, modality-independent mechanisms could also be related to abstract or higher-level conceptual representations of emotion categories (Scherer, 2009). For example, Skerry and Saxe (2014) have recently suggested that such representations, related to the appreciation of the causes of events, are implemented in the medial prefrontal cortex.

Finally, similarities in emotion recognition across modalities may also originate from perceptual similarities and interdependence of physical features during emotion production. For example, Ohala (1980) suggested that smiling in the animal and human face may have originated as a way of modulating the resonant properties of vocalisations in order to sound more infantile or submissive, and thus avoid attack. Thus, the upward retraction of the lip corners and the accompanying changes in vocal tract resonance may have a common origin.

Processing Dissimilarities

Emotion recognition involves several processes, from perceptual analysis of individual features to the global categorisation of an emotion (Calder & Young, 2005; Haxby, Hoffman, & Gobbini, 2002; Scherer, 2009). It is possible that some of these processes are shared or have similar coding mechanisms across modalities, while others do not. At the sensory level, signals from the voice and from the face are indeed quite different. Within voices, each of the basic emotions has a unique acoustic profile of pitch, amplitude, and spectral cues (Sauter et al., 2010). These properties result from the interaction of laryngeal activity and configurations of the vocal tract. In contrast, facial emotion recognition is based on configural changes within the face initiated by movements of face muscles (Ekman & Friesen, 1976). Different combinations of contracted face muscles presented in specific temporal orders, such as the early wrinkled nose followed by raised upper lip during disgust expressions, allow for an emotion-specific profile which can be reliably discriminated from other emotional facial expressions (Jack, Garrod, & Schyns, 2014). Further, emotions like happiness produce large scale cues in the face, such as an open mouth showing teeth, which aid rapid recognition within the visual domain. The striking perceptual feature of teeth may guide visual search and lead to

fast detection of happiness (Horstmann, Lipp, & Becker, 2012). For expressions of laughter, it has recently been reported that information from auditory cues guided perception of audio-visual stimuli (Lavan & McGettigan, 2016). Therefore, perceptual features in one modality may be more salient than in another modality and hence, emotion processing could be modality-specific.

Supporting the idea of modality-specific emotion recognition, and in contrast to the results reviewed above, some patient studies have shown that the faces and voices can be independently affected. For example, while the amygdala is often involved in the processing of fear in faces (e.g. Adolphs et al., 1995), it has been shown that an intact amygdala is not necessary for the processing of fearful prosody (Bach, Hurlemann, & Dolan, 2013; but see Scott et al. [1997] who also used non-speech emotional vocalisations). Similarly, while Huntington's disease can impair recognition of disgust in both faces and voices (Sprengelmeyer et al., 1996), a study of pre-clinical carriers of Huntington's disease found impairments in recognising disgust in faces but not in voices (Sprengelmeyer, Schroeder, Young, & Epplen, 2006).

Despite providing support for independent neural systems underlying recognition of facial and vocal expressions of emotion, these results do not provide direct evidence regarding whether there are similar coding principles or mechanisms across modalities. Furthermore, the degree of cross-modality similarity may differ across emotions. For example, whereas perceived fear may activate the amygdala irrespective of modality, perceived disgust may activate the insula only with faces, and not with voices (Phillips et al., 1998).

The Present Study

This brief review shows that, despite extensive research on recognition of emotions from faces and voices, it remains unclear whether these processes rely on modality-specific computations or shared, modality-independent mechanisms. A major limitation of previous research is that emotion recognition has mostly been studied separately across faces and voices, in different groups of participants. Consequently, in the present study, we examined whether facial and vocal emotion recognition have similar coding mechanisms using a within-subjects design. This allowed us to directly compare the behavioural emotion recognition profiles across modalities. Participants were

presented with emotional faces and non-verbal affect vocalisations and rated the intensity of each stimulus for each of the six basic emotions (happiness, sadness, anger, fear, surprise, and disgust; Ekman & Keltner, 1970; Ekman & Friesen, 1975). Our analyses then characterised how emotions are represented *within* each modality, and examined how similar these representations are *across* modalities.

As has been done before, we characterised the content of representations *within* each modality using matrices constructed from behavioural rating profiles for all six emotions (e.g. Adolphs et al. 1995; 1999; Belin et al., 2008; Juslin & Laukka, 2001; Sauter et al., 2010). We call these matrices *representation matrices* — they allow us to look at the confusions between all pairs of emotions, and we assume that two emotions which are often confused have more similar representations than two emotions that are never confused. Next, we compared the representations *across* modalities by correlating the representation matrices for faces with the representation matrices for voices. A high correlation would indicate that the representations have similar structure or content across modalities. In other words, if there are similar coding mechanisms for faces and voices, we expect to see that those emotions that are confused in faces are also confused in voices, in line with the idea of a general, modality-independent processor.

Our approach is based on recent analyses of cognitive and perceptual representations in the brain using Representational Similarity Analyses (RSA) (Kriegeskorte, Mur, & Bandettini, 2008a, Kriegeskorte et al., 2008b, Kriegeskorte & Kievit, 2013). Here, we applied similar methods to our behavioural data in order to be able to compare representations across modalities. This approach has the major advantage of allowing us to compare visual and auditory representations without requiring one-to-one correspondence between specific stimuli or emotions in the two modalities (Kriegeskorte et al., 2008a).

Methods

Participants

Participants were 54 British adults recruited through the participant pool at Brunel University, and through social networks. All participants were tested on the paper-based 60-item

Raven's Standard Progressive Matrices (Raven, Raven, & Court, 1998). We excluded nine participants whose scores were below the 10th percentile on this test, according to age specific norms (Raven, Raven, & Court, 2000). On inspection, the low performance seemed to be due to a lack of attention paid to the task, and therefore we decided to exclude all data from these participants before any analyses. The final sample consisted of 45 participants (15 male, 30 female), aged between 18 and 61, (M = 30.8, SD = 16.81) with an average Raven's score of 48.78 (SD = 4.41) out of 60 possible correct answers. Participants presented different educational backgrounds: secondary education and below (N = 15), undergraduate (N = 24) and postgraduate level (N = 6). All participants reported normal or corrected-to-normal vision and hearing. The study was approved by the Ethics Committee of the Department of Psychology, Brunel University, and all participants gave informed consent to participate.

Materials

The Emotion Judgment Task was programmed on the PsychoPy software in Python (Peirce, 2007) running on an Acer ASPIRE 5735 laptop (15.6 inches; resolution 1366 x 768 pixels; refresh rate 60 Hz) and voices were presented via closed-back, on-ear headphones (Sennheiser HD 202). The task contained both face and voice stimuli exhibiting the six basic emotions of happiness, sadness, anger, fear, surprise, and disgust (Ekman & Keltner, 1970).

Faces were displayed by two male (identity JJ and EM) and two female (identity C and SW) white Caucasian actors from the Ekman Pictures of Facial Affect series (Ekman & Friesen, 1975), making a total of 24 pictures (one stimulus per emotion, per actor). The size of the pictures on the screen was 6 x 8 cm. Viewing distance was not formally controlled but was approximately set to 50 cm. From a viewing distance of 50 cm, pictures thus subtended 6.87° x 9.15° of visual angle. The voice stimuli consisted of non-verbal affect bursts and were selected from the Montreal Affective Voices (Belin et al., 2008). Stimuli were produced by two male (identity 42 and 55) and two female (identity 45 and 53) white Caucasian French-Canadian actors, making a total of 24 sounds (one stimulus per emotion, per actor). Sounds consisted of vocal, non-verbal, affect expressions such as laughter or moans based on the vowel /a/ and were presented for the full duration of the stimulus. The

mean durations of the vocal stimuli used in the present study were 1267 ms for happiness, 2039 ms for sadness, 971 ms for anger, 621 ms for fear, 378 ms for surprise, and 1010 ms for disgust. The face and voice stimuli have previously been validated and show high emotion recognition rates (Belin et al., 2008; Ekman & Friesen, 1975).

Design and Procedure

We used a within-subjects design, in which each participant was tested on emotion judgements for both faces and voices. There were four blocks in the Emotion Judgment Task, two with face stimuli and two with voice stimuli. Each block contained stimuli from one male and one female actor portraying each of the six emotions[1]. Blocks were presented in the following order: voices (V1), faces (F1), voices (V2), faces (F2). Faces were presented in the centre of the display and remained until the participant made their response. For each trial, participants had to rate the intensity of the displayed emotion on a 7-point Likert-scale ranging from '*not happy/sad/… at all = 1'*, to '*extremely happy/sad/… = 7'.* Each stimulus was rated with respect to each of the six basic emotions. Thus, within a block, each stimulus was repeated six times, each time with a different label. Overall, there were 72 trials in each block (6 emotions x 2 sex x 6 labels). The 72 trials were presented in random order and no trial was repeated across different blocks. A similar design of rating tasks was used by Adolphs et al. (1995; 2000) to assess emotion recognition in faces.

Participants completed a short practice session, followed by the Emotion Judgment Task and the Raven's Matrices. Each task lasted less than 25 minutes and the order of tasks was counterbalanced. No definitions of emotions were provided. Finally, all participants were debriefed; psychology undergraduate students received credits as part of their course requirement.

Data Analysis

Our first analysis compared the participants' overall task performance for rating faces and voices. Previous research has suggested comparable task-difficulty for non-verbal affect vocalisations

[1] Different actors were presented in all the different blocks in order to make them independent. This was crucial to be able to perform the split-half analysis that we describe below.

and facial expressions (Hawk, van Kleef, Fischer, & van der Schalk, 2009), whilst emotions recognised from speech prosody showed higher error rates (see also Sauter, Panattoni, & Happe, 2013). For the present sample and stimuli, we computed means and standard deviations for three dependent measures, separately for faces and voices: (1) mean reaction times for target emotions (i.e. trials in which the emotion label matched the emotion shown in the stimulus), (2) mean perceived intensity ratings of target emotions, and (3) mean accuracy (we considered a response as 'correct' when the label corresponding to the target emotion received the highest intensity rating compared to all the other labels. For more details on this procedure, see Kornreich et al., 2012). We then used paired t-tests to compare overall performance across modalities for each dependent variable.

Our main analysis, however, aimed to compare the structure of representations of emotional faces and emotional voices. In order to characterise the representations of emotions for faces and voices separately, we computed representation matrices separately for each modality. These matrices of behavioural ratings and other confusion matrices are widely used in research in emotion recognition (e.g. Adolphs et al. 1995; 1999; Banse & Scherer, 1996; Belin et al., 2008; Calder, Burton, Miller, Young, & Akamatsu, 2001; Juslin & Laukka, 2001; Sauter et al., 2010). Our matrices included the responses using all emotion labels for each type of stimulus. Specifically, we analysed the mean intensity ratings for each of six labels given to each type of emotional stimulus, resulting in 36 conditions. In each matrix, each cell shows the mean intensity rating for one emotion label given to one emotion stimulus. We computed these representation matrices for each participant and also the mean across participants.

We then compared the representations across modalities by correlating the representation matrices for faces with the representation matrices for voices. We thus transformed each matrix into a single vector (or rating profile) and investigated the correlations of these using methods commonly applied in the analysis of fMRI neural response patterns (Haxby, Gobbini, Furey, Ishai, Schouten, & Pietrini, 2001; Kriegeskorte et al., 2008a; Kriegeskorte et al., 2008b). In addition to comparing representations across modalities, we also wanted to test the reliability of the responses within each modality. This provided us with a measure of the stability of the representations within each modality, and also allowed us to estimate the maximum correlation that we could expect between the rating

profiles. We thus computed correlations of rating profiles *within*- and *across*-modalities, to respectively examine the reliability of the responses within the same modality, and investigate the information shared across modalities. To be able to conduct these comparisons, we did a split-half analysis of the data. In other words, we divided the data for each participant and each modality in two separate and independent datasets: the first two presentation blocks formed the first half, whilst the third and fourth presentation block formed the second half. Hence, each half contained the same number of stimuli per emotion label and emotion category. This split-half analysis provided us with four datasets for each participant: two datasets with average intensity ratings for each label and each emotion for faces (F1 and F2), and two datasets with average intensity ratings for each label and each emotion for voices (V1 and V2). To examine the similarity of emotional response profiles, we then computed the correlations between the rating profiles within same modality (F1 *versus* F2 and V1 *versus* V2) and across modalities (F1 *versus* V1).

Results

1. Overall task difficulty

Our first analysis compared overall task difficulty for recognizing emotional faces and voices, examining three dependent measures: (1) mean reaction times for target emotions (although it would be difficult to interpret differences in reaction times between the modalities, given the very different nature of the stimuli), (2) mean perceived intensity ratings of target emotions, and (3) mean accuracy. For reaction times, the means were similar across faces (M = 3.94, SD = .97) and voices (M = 4.08, SD = 1.04), and a paired t-test showed no significant difference across modalities ($t(44)$ = -.79, p = .434). For mean intensity ratings of target emotions, the mean was slightly higher for faces (M = 5.80, SD = .56) than for voices (M = 5.52, SD = .64), and a paired t-test showed that this difference was significant ($t(44)$ = 3.34, p = .002). For accuracy, emotional faces (M = .74, SD = .10) were perceived more accurately than emotional voices (M = .67, SD = .12), and this difference was

significant ($t(44) = 4.35$, $p < .001$). Therefore, these results showed that emotional faces received significantly higher intensity ratings than emotional voices, and were also perceived more accurately.

2. Comparing the structure of representations of emotional faces and emotional voices

For our main analysis, looking at the structure of representations of emotional faces and emotional voices, we first computed representation matrices separately for faces and voices in order to characterise how emotions are represented within each modality. These representation matrices were based on the mean intensity ratings for each type of emotion label given to each of the six emotions. Figure 1 shows the mean representation matrices (averaged across participants) for faces and voices separately. These matrices reveal interesting similarities between modalities. For example, for both faces and voices, happiness is not usually confused with other emotions, whereas fear and surprise are often confused. More generally, the response profiles for each emotion stimulus are very similar across modalities. In other words, the relationships between emotions recognised from facial expressions appear to be very similar to the relationships between emotions recognised from vocal expressions. Figure 2 shows these similarities more clearly, in which each graph presents one emotion, and the mean rating profile across all six labels for that emotion, separately for faces and voices.

----- *insert Figure 1 here* -----

----- *insert Figure 2 here* -----

We next quantified these similarities across modalities by computing the correlations between the representation matrix (or rating profile) for facial stimuli and the representation matrix for vocal stimuli, separately for each participant. Note that this analysis does not depend on the magnitude of the ratings, but on the *relationship* between the ratings given to all emotion labels. Similar methods have been used previously to analyse behavioural responses to emotional stimuli (Adolphs et al., 2000; 2002; Nummenmaa, Glerean, Hari, & Hietanen, 2014) and to analyse fMRI neural response patterns (Haxby et al., 2001; Kriegeskorte et al, 2008a; Kriegeskorte et al., 2008b). For this analysis, and in

order to perform within-modality and across-modalities comparisons, we divided the data for each participant in four independent datasets: two datasets with average intensity ratings for each label and each emotion for faces (F1 and F2), and two datasets with average intensity ratings for each label and each emotion for voices (V1 and V2). We then computed the correlations between representation matrices (or rating profiles) within same modality (F1 *versus* F2 and V1 *versus* V2) and across modalities (F1 *versus* V1)[2].

We first computed the mean correlations across participants. Since correlations are usually not normally distributed, correlation scores were first Fisher z-transformed. After computing the mean of the z-transformed scores, we computed the inverse transformation of the mean value to obtain more interpretable values between -1 and 1 (for the same procedure, please see Adolphs et al., 1995; 1999). We applied this same procedure to all instances in which we computed the mean of correlation values. The mean correlation across participants of the representation matrices for face stimuli (F1 *versus* F2) was r = .82 (SD = 0.23), which shows that about 67% of the variance in the representation matrices of one half of the face stimuli can be predicted by the representation matrices of the other half of the stimuli (individual and mean correlations are shown in Figure 3). We then aimed to determine whether these correlations were significantly different from zero. For this, we used the non-parametric Wilcoxon signed ranks test, which tests whether the vector of correlations comes from a distribution of values in which the median is zero. Crucially, all rank tests were computed with raw, non-transformed correlations. The signed ranks test showed that the F1 *versus* F2 correlations were significantly different from zero (z = 5.84, p < .001). For voices, the average correlation of the representation matrices (V1 *versus* V2) was r = .74 (SD = 0.26) (*i.e.* about 55% of variance), which was also significantly different from zero (z = 5.84, p < .001[3]). These results show high test-retest reliability of the representations of emotions *within* each modality. Critically, the representation matrices were also highly correlated *across* modalities. The mean correlation between the matrices of face and voice

[2] We also correlated other data-split possibilities, i.e. F2 *vs* V2, F1 *vs* V2, and F2 *vs* V1, and we obtained similar results to the ones presented here (see Appendix 1).

[3] Please note that the results of most of the Wilcoxon signed ranks tests were the same (i.e. z = 5.84). This is because this is a test based on the sum of all positive ranks. Therefore, if all 45 participants had correlation scores above zero, the sum of all possible positive ranks is the same (T+ = 1035), independently of the exact correlation values, and then z = 5.84 (for more details, please see Siegel & Castellan, 1988). We additionally note that all the same comparisons were significant when we used one-sample t-tests.

stimuli (F1 *versus* V1) was $r = .71$ ($SD = 0.24$), which shows that about 50% of the variance in the representation matrices of faces can be predicted by the representation matrices of voices (and vice-versa)[4]. Again these correlations were significantly different from zero ($z = 5.84$, $p < .001$). Hence, the results suggest that the perception of confusions or distinctions between emotions largely overlaps across the visual and auditory modalities.

In order to compare the within- and across-modalities correlations, we conducted a repeated-measures ANOVA with contrasts (3 levels: F1 *versus* F2, V1 *versus* V2, F1 *versus* V1) as the within-subject variable. We used the z-transformed correlations in the ANOVA. The repeated-measures ANOVA revealed a significant main effect for contrast, $F(2, 88) = 38.45$, $p < .001$, $\eta^2 = .47$. Pairwise comparisons revealed that the correlations of the representation matrices of faces (F1 *versus* F2) were significantly higher than the correlations of the representation matrices of voices (V1 *versus* V2) and higher than the correlations of the matrices across modalities (F1 *versus* V1), both $p < .001$. Interestingly, correlations of the representation matrices across the two modalities were not significantly different from the correlations of the representation matrices of voices ($p = .182$). These results show that the representations of emotions from faces are more reliable across different stimuli than the representations of emotions from vocal expressions. Overall, the representations of faces seem to have some unique information that is not shared with voices, given that there are higher correlations for within-modality rather than across-modalities comparisons. However, the structure of representations of vocal expressions of emotion largely overlaps with the representations of facial expressions of emotion because these across-modality correlations are not significantly higher than the within-modality correlations.

----- *insert Figure 3 here* -----

To demonstrate that the high correlations of representation matrices within-modalities (F1 *versus* F2, V1 *versus* V2) and across-modalities (F1 *versus* V1) were not solely driven by the presence

[4] We also used Spearman correlations, and the mean correlations were: F1 *vs* F2 = .77, V1 *vs* V2 = .70, and F1 *vs* V1 = .65. Wilcoxon signed ranks tests showed that all correlations are significantly higher than zero (all $p < .001$).

of matching target-emotions (*i.e*. the diagonals in these matrices), we ran the same correlation analyses again, but this time we excluded the diagonals of all matrices. Again, to compute means across participants, we used z-transformed correlations and here we report the mean values after they had been transformed back to values between -1 and 1. The mean correlation across participants of the representation matrices (without diagonal) for face stimuli (F1 *versus* F2) was r = .68 (SD = 0.28; the Wilcoxon signed ranks test comparing the raw, non-transformed, correlations to zero was z = 5.84, p < .001) and for voice stimuli (V1 *versus* V2) was r = .59 (SD = 0.27; z = 5.84, p < .001). Finally, the mean correlation of the representation matrices across modalities (F1 *versus* V1) was r = .51 (SD = 0.25; z = 5.83, p < .001). This correlation is lower than in the previous analysis, but still suggests that the structure of representations of emotions largely overlaps across the visual and auditory modalities. Individual and mean correlations are shown in Figure 4.

Further, we repeated the repeated-measures ANOVA with contrasts (3 levels: F1 *versus* F2, V1 *versus* V2, F1 *versus* V1) as the within-subject variable, this time without the diagonal target emotions. As before, we used the z-transformed correlations in the ANOVA. The repeated-measures ANOVA revealed a significant main effect for contrast, $F(2, 88) = 17.74$, $p < .001$, $\eta^2 = .29$. Pairwise comparisons revealed that the correlations of the representation matrices of faces (F1 *versus* F2, without diagonal) were significantly higher than the correlations of the representation matrices of voices (V1 *versus* V2, without diagonal) (p = .01) and higher than the correlations of the representation matrices across modalities (F1 *versus* V1, without diagonal) (p < .001). Again, correlations of the representation matrices across the two modalities were not significantly different from the correlations of the representation matrices of voices (p = .086). By removing the target emotions, we found that there was more overlap of emotion rating profiles in the within-modality conditions compared to the between-modalities condition (though this difference was not significant for voices), which suggests that some of the representational content is modality-specific. Yet, the emotion rating profiles across modalities are still moderately to highly correlated, suggesting that a large proportion of the information is shared across faces and voices, even after removing the diagonals in the representation matrices.

----- *insert Figure 4 here* -----

3. Comparing individual representations to mean representations across individuals

Finally, we compared individual representations of emotions with the average representations from the rest of the participants. This allowed us to determine whether the high correlations we observed within and across modalities were related to idiosyncrasies of each participant's representations (or even related to specific ways in which they responded during our task), or whether they were related to representations of emotions that had similar structure across individuals. We therefore correlated each individual's representation matrices to matrices averaged across all the other participants. Here, we report results for the same comparisons as the ones shown in Figure 4, and thus after removing the diagonals from all matrices. The mean correlation (as before, we computed the mean based on z-transformed correlations, and here we report the mean values after being transformed back to values between -1 and 1) of each participant's representation matrix for faces (F1, with half of the stimuli) with the mean matrix for faces across all the other participants (MF2, with the other half of the stimuli) was r = .69 (SD = 0.25; the Wilcoxon signed ranks test comparing non-transformed correlations to zero was z = 5.83, p < .001). Similarly, the mean correlation of each participant's representation matrix for voices (V1) with the mean matrix for voices across all the other participants (MV2) was r = .64 (SD = 0.23; z = 5.84, p < .001). Finally, the mean correlation of each participant's representation matrix for faces (F1) with the mean matrix for voices across all the other participants (MV1) was r = .55 (SD = 0.20; z = 5.83, p < .001), and the mean correlation of each participant's representation matrix for voices (V1) with the mean matrix for faces across all the other participants (MF1) was r = .54 (SD = 0.20; z = 5.84, p < .001). These results show that representations of emotions for facial and vocal expressions have a similar structure across individuals (Figure 5). These similarities explained a substantial portion of the variance in individual profiles (on average, between 30% and 48% of the variance). However, in all the comparisons, there were still large amounts of unexplained variance. This unexplained variance could be related to individual differences in emotion representation. In the future, it will be very interesting to explore potential factors that may contribute to these individual differences.

----- *insert Figure 5 here* -----

4. Comparing behavioural representations to representations of low-level properties of the stimuli

Confusions between different emotions are often attributed to similar perceptual features within one modality, such as image-based properties in faces (Calder et al., 2001), muscle configurations in faces (Jack et al., 2014), or acoustic properties in voices (Banse & Scherer, 1996; Juslin & Laukka, 2001; Sauter et al., 2010). In a similar view, Juslin and Laukka (2003) found several similarities between the patterns of emotion perception across voices and music, and suggested that those could be largely explained by similarities of acoustic cues. It is therefore important to examine whether the high correlations that we observe between the rating profiles across modalities could be explained by the visual and acoustic properties of the stimuli.

It is possible that the visual and acoustic properties of the stimuli are themselves correlated across modalities, as a result of interdependence between the activation of face and vocal tract musculature during emotion expression (e.g. Ohala, 1980). On the other hand, if the behavioural correlations are not due to the acoustic or visual properties of the stimuli, it could be that they result from modality-independent processes, such as abstract representations of emotions (e.g. Scherer, 2009; Skerry & Saxe, 2014). To distinguish between these two possibilities, we carried out analyses of the low-level (visual and acoustic) properties of the emotional stimuli and obtained representation matrices for each visual and acoustic cue. We then computed partial correlations between the behavioural matrices, while removing the variance due to each visual and acoustic cue. We next describe these analyses in detail.

4.1. Visual analysis of faces

First, we carried out an analysis of the visual properties of the faces. For this, we based our methods on Calder et al. (2001), who examined whether principal component analyses (PCA) of images of emotional faces supported facial expression recognition. The authors found that it was possible to categorise the different emotions based on the outputs of the PCA. Critically, the pattern of

miscategorisations was similar to that of human observers. These results showed that linear analysis of the visual information present in images of facial expressions allows the categorisation of emotions in a manner that is consistent to human performance.

Here, we based our analysis on the same visual properties that were used by Calder et al. (2001). Specifically, Calder et al. (2001) included three datasets in which they conducted the PCAs: (1) full images, corresponding to the greyscale pixel values of full face images that had been modified to have the eyes aligned to the same position ; (2) shape-free images, corresponding to the greyscale pixel values of face images that had been modified to have the same average-shape, and (3) shape-only, which corresponded to the x and y coordinates of 35 anatomical feature points on the face. The best results, in terms of accurate categorisation of emotions, were obtained by combining the visual information from shape-free images and shape-only coordinates.

Like Calder et al. (2001), our analyses of visual information considered three separate datasets, each related to a different visual property. The first dataset consisted of pixel values of the face images, after we had aligned the position of the eyes. The second dataset consisted of pixel values of the face images, after we had aligned all images to have all anatomical features in the same position. The third dataset consisted of vectors of the x and y coordinates of 49 features in the face (Calder et al. [2001] did not specify the 35 features that they used, and here we used 49 features that could be clearly identified). We then created several representation matrices, each based on one specific visual property.

In order to prepare all the face stimuli for each dataset, we used Psychomorph (Tiddeman, Stirrat, & Perrett, 2005). As a first step, we removed external features of the faces, such as hair and ears. Then, all stimuli were aligned to have the eyes in the same position (the coordinates of the eyes were based on the average of all the face images). There was no further processing for the images in the first dataset. For the images in the second dataset, we started with the images from the first dataset and created an average of all the images. We then transformed each image to have the same shape (i.e. the same shape coordinates) as the average image. The third dataset consisted of the x and y coordinates of 49 points in the face (using the same images from the first dataset); each point corresponded to a clear anatomical landmark.

We next computed representation matrices based on each of these three datasets. For this, we needed to compute the similarity across pairs of faces within each dataset. Therefore, separately for each dataset, we computed Euclidean distances[5] between each pair of faces. Specifically, for the first two datasets, we computed the Euclidean distances between vectors consisting of the pixel values corresponding to each image. We used the same oval to mask for each face (in order to avoid including the contours of the face), and the vectors only included the greyscale values inside the oval. The similarity between two faces thus consisted of the Euclidean distance between two such vectors. For the third dataset, Euclidean distances were computed using the coordinates of the positions of the 49 features in the face as vectors. We transformed the x and y coordinates for each image in a single vector by concatenating the y coordinates after the x coordinates (Calder et al., 2001), and then computed the Euclidean distance between each pair of vectors corresponding to each pair of faces.

In addition to using Euclidean distances to examine the similarity of the positions of face features (third dataset), we also conducted Procrustes analysis, which specifically allows the comparisons of the shapes of two objects (Bookstein, 1991; Rohlf & Slice, 1990). Procrustes analysis consists of the linear transformation (translation, scaling, and rotation) of the shape of one object to best match the shape of another object. As a measure of similarity, we then used the sum of the squared errors between the transformed (superimposed) shapes. This approach has previously been used to compare the shapes of body parts, such as hands (Longo & Haggard, 2010) and faces (e.g. Fink et al., 2005; Pound, Penton-Voak, & Brown, 2007; Pound et al., 2014). We used Matlab version 8.2.0.701 (Mathworks, Natick, MA, USA) to carry out the Procrustes analysis. The data for each face consisted of the x and y coordinates of the same 49 features that we used in the previous analysis. Then, for each pair of faces, we transformed the shape (i.e. positions of features) of the first face to match the shape of the second face, and computed the Procrustes dissimilarity between the resulting transformed shapes[6].

[5] Given that we compute Euclidean distances, it would be more natural to describe the *dissimilarity* between faces. However, in order to be consistent with the previous sections, we have used the term *similarity* throughout.

[6] As for Euclidean distances, zero means that the two shapes are the same, and higher values indicate more dissimilar shapes. For consistency, we will again use the term ‘similarity’ instead of ‘dissimilarity’.

We computed the similarity of each pair of images by using all four methods described above. We only compared the images of faces with the same identity (i.e. identity of the person shown on the image), and then averaged all the matrices across different identities. We therefore created four representation matrices of visual properties of the stimuli: (1) *Full images*: representation matrix based on the Euclidean distances between vectors consisting of pixel values of the full images which had all eyes aligned, (2) *Shape-Free images*: representation matrix based on the Euclidean distances between vectors consisting of pixel values of shape free images, (3) *Shape-49*: representation matrix based on Euclidean distances between vectors consisting of the coordinates of 49 facial features, (4) *Shape-49-Procrustes*: representation matrix based on Procrustes analysis. These four representation matrices are shown in Appendix 2.

4.2. Acoustic analysis of voices

The analysis of the acoustic properties of the voices was based on methods used by Sauter et al. (2010), who showed that the linear analysis of acoustic properties of non-verbal emotional expressions of emotion could support categorisation of emotions in a psychologically plausible manner. In other words, it is possible to categorise emotions purely based on the analysis of acoustic properties of the stimuli, and the pattern of miscategorisations is consistent with errors made by human observers.

For the present study, we aimed to examine the similarity between vocal expressions of individual emotions based on their acoustic properties alone. For each stimulus, we therefore extracted the same ten acoustic properties used by Sauter et al. (2010), including measures of fundamental frequency (F0), spectral properties, amplitude, and periodicity: (1) total duration (seconds), (2) amplitude: standard deviation (dB), (3) mean intensity (dB), (4) number of amplitude onsets, (5) F0 minimum (Hz), (6) F0 maximum (Hz), (7) F0 mean (Hz), (8) F0 standard deviation (Hz), (9) spectral centre of gravity (Hz), and (10) standard deviation of the spectrum (Hz). We additionally extracted four other acoustic properties to further describe the periodicity of these vocalisations: (11) mean harmonics-to-noise-ratio (dB), (12) jitter, (13), percentage of unvoiced segments, and (14) shimmer.

These acoustic properties are described in more detail in Appendix 3. For each vocal stimulus, we extracted these acoustic properties using PRAAT (Boersma & Weenink, 2015).

We then created representation matrices based on each of these acoustic properties by computing the similarity across pairs of vocal stimuli. For each acoustic property, and for each pair of stimuli, we computed the Euclidean distance between the single values for that property[7]. Similar to the analysis of face stimuli, we only compared pairs of stimuli belonging to the same person identity, and then averaged all the matrices across different identities. We therefore created fourteen representation matrices, each corresponding to an acoustic property. These fourteen representation matrices are shown in Appendix 3.

4.3. Behavioural matrices

In order to be able to compare the visual and acoustic representation matrices with the behavioural representation matrices, all matrices needed to be symmetric. Furthermore, there should be a one-to-one correspondence between the entries in all the matrices. However, whilst the representation matrices of acoustic and visual properties are symmetric (Appendices 2 and 3), the behavioural matrices that we used above (sections 2 and 3) were not symmetric across the diagonal. We therefore needed to change the format of the behavioural representation matrices by computing them in a new manner, comparable to the way in which the representation matrices for the low-level properties were constructed. Briefly, we computed the similarity of each pair of emotional stimuli using the six emotion labels as features. This procedure is also comparable to the way in which representational matrices are computed in fMRI studies (e.g. Kriegeskorte et al., 2008a; 2008b), in which voxels are the features; see also Skerry and Saxe (2015) who recently used a comparable method to compute similarities of behavioural ratings of emotions. More specifically, we computed a representation matrix for each participant, each modality, and each person identity. Each identity was represented by six separate stimuli, each corresponding to one emotion. Each stimulus was rated for six different emotion labels. Therefore, within each modality and for each pair of stimuli of the same

[7] We also built a representation matrix using all the acoustic properties at the same time. In other words, each stimulus was represented by a vector composed of all the values for all acoustic properties. This representation matrix, however, was not correlated with behavior, and therefore we do not show this analysis here.

identity, we computed the Euclidean distance between the two vectors consisting of the ratings for the six emotion labels (i.e. vectors consisting of six rating values)[8]. This analysis resulted in four representation matrices for faces and four representation matrices for voices for each participant. We then averaged all representation matrices within the same modality, which resulted in one representation matrix for emotional faces and one representation matrix for emotional voices for each participant (Appendix 5 also shows results of all the same analyses as described below but using non-averaged matrices for each modality). Crucially, computing the matrices in this manner does not change the previous conclusions, as can be seen in Appendix 4. Appendix 4 shows the mean representation matrices for judgments from faces and voices and also includes the same analyses that had been done for Figures 3 and 4 but now using the new behavioural matrices. These analyses demonstrate that the results are similar using the new matrices. Correlations between the matrix for faces and the matrix for voices (only the lower triangular part of each matrix) for each individual can be seen in Figure 6A (last column). The mean correlation for all participants was r = .60 (*SD* = 0.34; the result of the Wilcoxon signed ranks test comparing non-transformed correlations to zero was z = 5.81, $p < .001$).

4.4. Accounting for low-level properties

For each participant, we first correlated the four representation matrices for visual properties with the behavioural representation matrices for emotional faces. Figure 6A shows the individual correlations, as well as the means across all participants. The results showed that none of the visual properties strongly predicted participants' behavioural responses. Only the two representation matrices with shape information correlated with the behavioural matrices for faces significantly above zero. The mean correlation of the representation matrix using Euclidean distances between shape vectors (*Shape-49*) and the behavioural matrices was r = .11 (*SD* = 0.18), which was significantly above zero (z = 3.43, $p < .001$). The mean correlation of the representation matrix using Procrustes distances between shape vectors (*Shape-49-Procrustes*) and the behavioural matrices was r = .25 (*SD* = 0.19)

[8] Please note that it is also possible to use correlations in order to compute similarities between the two vectors, but we used Euclidean distances to keep it consistent with the analyses of the acoustic properties, in which it would not be possible to compute correlations because there was one single value for each stimulus.

and again, these correlations were significantly above zero ($z = 5.37$, $p < .001$). This latter representation matrix seemed to be the best predictor of the behavioural ratings of emotional faces.

----- *Insert Figure 6 here* -----

In a second step, we computed partial correlations between the individual representations matrices for faces and voices, while controlling for the representation matrices for each visual property. Figure 6B shows that these partial correlations were still very high (all mean $r > 0.61$; Wilcoxon signed ranks tests comparing correlations to zero: all $z > 5.80$, all $p < .001$), and therefore the visual properties that we considered here do not seem to account for the crossmodal behavioural correlations.

We also conducted similar analyses for the acoustic properties, by computing the correlations between each of the fourteen representation matrices for acoustic properties and the behavioural representation matrices for emotional voices for each participant. Figure 7A shows the individual correlations as well as the means across participants. The results showed that several of the acoustic properties describing amplitude, periodicity, and spectral properties of the vocalisations (acoustic cues 1 to 4, and 10 to 14, see Appendix 3) were good predictors of participants' behaviour (all mean $r > 0.22$; Wilcoxon signed ranks tests comparing correlations to zero: all $z > 4.79$, all $p < .001$). However, the correlations between the representation matrices for acoustic cues 5 to 9 and the behavioural matrices for voices were not significantly above zero.

----- *Insert Figure 7 here* -----

Finally, we computed partial correlations between the individual behavioural representations matrices for faces and voices, while controlling for the representation matrices for each acoustic cue. Figure 7B shows that these partial correlations were still quite high (all mean $r > 0.44$; Wilcoxon signed ranks tests comparing correlations to zero: all $z > 5.41$, all $p < .001$), and therefore the acoustic properties that we considered here did not account for most of the crossmodal behavioural correlations. In other words, while the acoustic properties account for significant amounts of variance

of the behavioural representation matrices of emotional voices, they do not seem to account for most of the shared variance between emotional faces and emotional voices.

It is possible, however, that our averaging of the representation matrices of low-level properties across different stimuli may have distorted the results. In particular, the matrices could be quite different for distinct identities and, therefore, the mean matrices that we used (in which we averaged representation matrices of different identities of the stimuli) would be non-interpretable and distort the results. Therefore, we performed separate analyses without averaging the representation matrices across different identities. In these analyses, each entry to a representation matrix was a stimulus, resulting in a 24-by-24 matrix. These analyses are shown in Appendix 5. Appendix 5 shows that the results using these non-averaged matrices were comparable to the results described above, and therefore the averaging procedure did not seem to distort the results.

To this point, our analyses suggested that single low-level visual or acoustic properties do not account for the majority of the shared variance between the representation matrices of emotional faces and emotional voices. However, it is possible that a combination of the low-level properties would be able to better account for this shared variance. We therefore carried out multiple regressions to remove the variance accounted for by multiple low level properties from the behavioural representation matrices. There are, however, a couple of important caveats when conducting these analyses. First, several of the predictors were highly correlated (Appendix 6 has a correlation matrix of all the low-level properties used here). Second, there were many predictors and very few data points per regression. In fact, because of this, it was impossible to conduct multiple regression with all the acoustic predictors using the average representation matrices. Therefore, we conducted multiple regressions using the 24-by-24 non-averaged representation matrices. The results of these analyses are described in Appendix 7 and show that, even when accounting for multiple visual properties of the faces or acoustic properties of the voices, the correlations across faces and voices did not substantially decrease.

We conclude that the visual and acoustic properties of the stimuli do not seem to account for most of the shared variance between the representations of emotional faces and emotional voices. We note, however, that some of the partial correlations decreased when controlling for the low-level

properties, especially when controlling for acoustic properties (see Figure 7 and Appendix 7). This suggests that some of the acoustic properties of the voices may account for some of the shared variance in the ratings of emotions, and it will be interesting to systematically examine the role of these acoustic properties in future studies. Furthermore, the correlation matrix in Appendix 6 also suggests that the representation matrices for some of the visual and acoustic properties are correlated, even if those correlations are low. This could be due to the interdependence between vocal production and change in facial muscles (Ohala, 1980), or could be more generally related to the idea that faces and voices may carry redundant signals, leading to more accurate judgments (Smith et al., 2016). It will be interesting to explore this in future studies, using a greater diversity of stimuli, and an even wider selection of acoustic and visual properties — it would be particularly interesting to conduct these analyses of low-level properties using voices and faces of the same actors, as this would control for extra variability that was introduced by having different people posing the vocal and facial emotions.

Discussion

In this study we aimed to compare the representations of emotions across faces and voices. We used an approach based on Representational Similarity Analysis (Kriegeskorte et al., 2008a, 2008b; Kriegeskorte & Kievit, 2013). Briefly, we examined the structure or geometry of the representations within each modality by computing representation matrices for faces and voices separately, and then we compared the representations *across* modalities by correlating these matrices. Our results showed high correlations between the representation matrices for faces and the representation matrices for voices, which suggest similar representations of the six basic emotions across modalities. In other words, participants associated specific emotion-stimuli with specific emotion-labels, and this pattern was consistent *within*, as well as *across* modalities. We also found that the structure of these representations is quite similar across individuals, though there was also unique

variance for each individual. In the future, it will be very interesting to determine variables that may contribute to individual differences in emotion representation profiles.

We also examined whether the shared variance between representations of facial and vocal expressions of emotion could be explained by the physical properties of the stimuli. Specifically, we computed representation matrices for the faces and the voices based on their physical properties. For faces, we used texture and shape information, based on the analysis of facial expressions by Calder et al. (2001). For the voices, we used acoustic properties related to fundamental frequency, spectral properties, amplitude, and periodicity of the vocalisations, based on the analysis of vocal expressions by Sauter et al. (2010). Correlations of the representation matrices based on physical properties of the stimuli with the behavioural representation matrices showed that they could account for some of the variance within each modality. However, when we removed the variance explained by these low-level properties, there were still moderate (and significant) correlations between behavioural representations of emotional faces and emotional voices.

Our results extend previous studies showing that individual differences in vocal emotion recognition are correlated with individual differences in facial emotion recognition (Borod et al., 2000; Palermo et al., 2013). These past studies, however, had only focused on the overall ability to recognise emotions in each modality. Conversely, here we focused on the full representations of the six basic emotions to demonstrate that their representational structure or geometry is similar across faces and voices, suggesting similar or shared mechanisms across modalities.

Despite these similarities across modalities, we also found some differences between face and voice emotion recognition. These can be seen in the higher correlations of the representation matrices within modalities than the correlations across modalities (though this was only significant for face matrices). The higher within-modality correlations show that there is modality-specific information in these representation matrices that is not shared across modalities. The modality-specific representational content could be related to the physical properties of the stimuli themselves.

Based on the present findings, and in line with previous behavioural studies (Borod et al., 2000; Palermo et al., 2013), we suggest that emotions may be largely categorised by modality-independent mechanisms. Confusions between different emotions are often attributed to similar

perceptual features within one modality, such as image-based properties in faces (Calder et al., 2001), muscle configurations in faces (Jack et al., 2014), or acoustic properties in voices (Banse & Scherer, 1996; Juslin & Laukka, 2001; Sauter et al., 2010). In a similar view, Juslin and Laukka (2003) found several similarities between the patterns of emotion perception across voices and music, and suggested that these could largely be explained by similarities of acoustic cues. However, our findings suggest that explanations based on low-level visual or auditory perceptual features may be incomplete. Instead, emotion recognition may also depend on modality-independent processes.

These modality-independent processes could be rooted in top-down mechanisms. For example, rapid responses to emotional compared to neutral stimuli are seen for both faces (e.g Eimer & Holmes, 2007) and voices (Sauter & Eimer, 2009). There could also be modality-independent representations consisting of abstract representations of emotions, for example linked to the appraisal of situations or events that cause the various emotions (Scherer, 2009; Skerry & Saxe, 2014; 2015). In a similar view, the same semantic representation of emotion categories may be activated across different types of stimulus presentation formats, as has previously been shown for objects depicted in pictures or as written words (Shinkareva, Malave, Mason, Mitchell, & Just, 2011). This explanation could also be related to the specific task we used, which may have relied on the semantic use of emotion categories or labels. Future studies could test whether similarities across modalities still hold up when using tasks that do not rely on labels or categorical contexts, such as perceptual matching tasks.

Our findings are also compatible with studies suggesting that modality-independent mechanisms could be implemented in multimodal brain regions. For example, it seems reasonable that specific subcortical structures such as the amygdala process emotions independently of modality (Phillips et al., 1998). It is also possible that a wider network of structures is active during emotion recognition (Peelen et al., 2010; Skerry & Saxe, 2014) despite the great perceptual differences in sensory inputs from faces and voices. Again, this may be irrespective of whether individual brain regions are emotion-specific or not. However, it is also possible that the two modalities are processed in separate regions that are modality specific, but which have similar coding mechanisms. In either case, we suggest that Yovel and Belin's (2013) proposal of common underlying coding mechanisms

for recognising person-identity from faces and voices may also apply to recognising emotions from faces and voices.

What could be the benefit of modality processing similarities during emotion recognition? In everyday life, emotions in faces or voices may not always be expressed in isolation. In other words, it is very common that emotions are expressed simultaneously across modalities. For the integration of signals across modalities, Hagan, Woods, Johnson, Calder, Green, and Young (2009) found increased activity in posterior regions of the superior temporal sulcus (STS) during the combined processing of fearful static facial expressions and non-verbal emotion vocalizations, compared with unimodal presentations. Further, Kreifelts, Ethofer, Shiozawa, Grodd, and Wildgruber (2009) reported a functional segregation of emotion processing in the STS, with specific parts of the STS being either sensitive to faces or to voices. Interestingly, parts of the STS that spatially overlapped between face and voice-selective regions were active during audio-visual emotion recognition. This shared use of neural structures, such as the STS, as well as the current behavioural processing similarities during emotion recognition across faces and voices may be purposefully linked to facilitate the integration of information from several modalities.

Alternatively, everyday exposure to multimodal emotion expressions may have strengthened our associations of emotion representations from individual modalities so that the recognition from one modality is associated with the recognition from another modality. In line with this, there may be physical interdependence between the activation of face and vocal tract musculature during emotion expression. As Ohala (1980) suggested, the retraction of the corners of the mouth – which typically accompanies a smile – causes an increase in the frequency range of sounds, which typically makes them resemble infantile vocalisations. Similarly, muscular activation of the mouth and tongue, which reflect functionally adaptive behaviours such as vomiting, not only creates the typical face expression of disgust but also modifies configural properties in the upper vocal tract, creating typical vocal expressions of disgust (Scherer, 1994). However, according to this view, we would have expected that the low-level properties of the stimuli would have accounted for a larger proportion of the shared variance across faces and voices. This was not the case here, but future studies using facial and vocal expressions posed by the same actors, and additional visual (for example, looking at muscle activity of

the face over time) and acoustic properties (for example, considering the time-course of the sound, and not just taking a mean value) could further probe this hypothesis. Finally, it could be that confusions between emotions emerge because of shared physiological responses across specific pairs of emotion, and the associated perception of those responses (James, 1884; Nummenmaa et al., 2014). In this view, the modality in which the stimuli are presented does not affect the physiological responses.

In our study, to characterise the structure of representations of emotional stimuli, we initially used the intensity ratings of different emotion labels for each emotional stimulus, as done in previous studies (e.g. Adolphs et al., 1995; 1999; 2000; Calder et al, 2001; Sauter et al., 2010). Furthermore, in our analyses comparing the behavioural representations to representations of low-level properties of the stimuli, we additionally used Euclidean distances between the vectors of intensity ratings (on the six emotion labels) for each stimulus. However, a more standard method of characterising the structure of representations is to use pairwise similarity judgments (i.e. where participants rate the similarity between two stimuli). Kriegeskorte and Mur (2012) have also recently proposed the multiple arrangement method, in which participants arrange multiple stimuli in two-dimensional space according to their perceived similarities. In this latter approach, the similarities are inferred from the distances between stimuli. For visual stimuli, the multiple arrangement method is substantially faster than acquiring pairwise similarity judgments, but may incur additional working memory demands for auditory stimuli. Nonetheless, we think that it would be very interesting in future work to compare our current approach with the outcomes of pairwise similarity judgments and the multiple arrangement method.

Conclusions and Future Studies

Overall, the present behavioural study demonstrates interesting parallel representations for recognising emotions displayed in static faces and vocal affect bursts. Possible (and non-mutually exclusive) explanations for this include modality-independent higher level representations, underlying shared neural networks, or the interdependence of facial and vocal musculature during emotion production. In our study, for the static faces as well as for non-verbal affect bursts, emotions were portrayed in iconic and prototypical ways, which are relatively easy to recognise across both

modalities. For future studies, it would be interesting to compare iconic emotion expressions portrayed by actors in a prototypical manner with more spontaneous and authentic expressions of emotion. Further, it may be interesting to extend this approach to include dynamic face or body expressions from a larger variety of actors in order to investigate whether the present claims of modality-independent emotion processing hold up to a wider range of stimuli. In order to control for sources of variability related to the identity of the people posing the emotions, it would also be informative to compare the representations of emotions from facial and vocal stimuli generated by the same actor. Finally, future studies investigating the origin, development, and neural correlates of these similarities could provide a deeper insight into the common mechanisms between facial and vocal emotion processing.

References

Adolphs, R., Damasio, H., & Tranel, D. (2002). Neural systems for recognition of emotional prosody: A 3-D lesion study. *Emotion*, *2(1)*, 23-51.

Adolphs, R., Damasio, H., Tranel, D., Cooper, G., & Damasio, A. (2000). A Role for somatosensory sortices in the visual recognition of emotion as revealed by three-dimensional lesion mapping. *Journal of Neuroscience*, *20(7)*, 2683–2690.

Adolphs, R., Tranel, D., & Damasio, H. (2001). Emotion recognition from faces and prosody following temporal lobectomy. *Neuropsychology*, *15(3)*, 396-404.

Adolphs, R., Tranel, D., Damasio, H., & Damasio, A. (1995). Fear and the human amygdala. *Journal of Neuroscience*, *15*(9), 5879-5891.

Adolphs, R., Tranel, D., Hamman, S., Young, A., Calder, A., Phelps, E.A.,… & Damasio, A. (1999). Recognition of facial emotion in nine individuals with bilateral amygdala damage. *Neuropsychologia*, *37*, 1111-1117.

Bach, D., Hurlemann, R,. & Dolan, R. (2013). Unimpaired discrimination of fearful prosody after amygdala lesion. *Neuropsychologia*, *51(11)*, 2070–2074.

Banissy, M., Sauter, D., Ward, J., Warren, J., Walsh, V., & Scott. S. (2010). Suppressing sensorimotor activity modulates the discrimination of auditory emotions but not speaker identity. *Journal of Neuroscience, 30(41)*, 13552–13557.

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

Belin, P., Fillion-Bilodeau, S., & Gosselin, F. (2008). The Montreal Affective Voices: A validated set of nonverbal affect bursts for research on auditory affective processing. *Behavior Research Methods*, *40(2)*, 531–539.

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer [Computer program]. Version 5.4.08, retrieved 24 March 2015 from http://www.praat.org/.

Bonora, A., Benuzzi, F., Monti, G., Mirandola, L., Pugnaghi, M., Nichelli, P., & Meletti, S. (2011). Recognition of emotions from faces and voices in medial temporal lobe epilepsy. *Epilepsy & Behavior*, *20(4)*, 648–654.

Bookstein, F.L. (1991). *Morphometric tools for landmark data. Geometry and biology*. Cambridge: Cambridge University Press.

Borod, J. C., Pick, L. H., Hall, S., Sliwinski, M., Madigan, N., Obler, L. K., … Tabert, M. (2000). Relationships among facial, prosodic and lexical channels of emotional perceptual processing. *Cognition and Emotion, 14(2),* 193-211.

Calder, A. J., Burton, A. M., Miller, P., Young, A. W., & Akamatsu, S. (2001). A principal component analysis of facial expressions. *Vision Research, 41(9),* 1179-1208.

Calder, A. J. & Young, A. W. (2005). Understanding the recognition of facial identity and facial expression. *Nature Reviews Neuroscience, 6(8),* 641–651.

Darwin, C. (1965). *The expression of the emotions in man and animals.* Chicago: University of Chicago Press. (Original work published 1872).

De Gelder, B. (2006). Towards the neurobiology of emotional body language. *Nature Reviews Neuroscience, 7(3),* 242-249.

Eimer, M., & Holmes, A. (2007). Event-related brain potential correlates of emotional face processing. *Neuropsychologia, 45(1),* 15-31.

Ekman, P. & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology. 17*(2), 124-129.

Ekman, P. & Friesen, W. V. (1975). *Pictures of facial affect*. Consulting Psychologists Press.

Ekman, P. & Friesen, W. V. (1976). Measuring Facial Movement. *Environmental Psychology and Nonverbal Behavior, 1*(1), 56-75.

Ekman, P., & Keltner, D. (1970). Universal facial expressions of emotion. *California Mental Health Research Digest, 8*(4), 151-158.

Engell, A. D., & Haxby, J. V. (2007). Facial expression and gaze-direction in human superior temporal sulcus. *Neuropsychologia, 45*(14), 3234-3241.

Fecteau, S., Belin, P., Joanette, Y., & Armony, J. (2007). Amygdala responses to nonlinguistic emotional vocalizations. *NeuroImage, 36,* 480–487.

Fink, B., Grammer, K., Mitteroecker, P., Gunz, P., Schaefer, K., Bookstein, F., & Manning, J. (2005). Second to fourth digit ratio and face shape. *Proceedings of the Royal Society B, 271,* 1995-2001.

Fitzgerald, D. A., Angstadt, M., Jelsone, L. M., Nathan, P. J., & Phan, K. L. (2006). Beyond threat: amygdala reactivity across multiple expressions of facial affect. *Neuroimage, 30*(4), 1441-1448.

Gray, H. M., & Tickle-Degnen, L. (2010). A meta-analysis of performance on emotion recognition tasks in Parkinson's disease. *Neuropsychology, 24*(2), 176-191.

Hagan, C. C., Woods, W., Johnson, S., Calder, A. J., Green, G. G. R., & Young, A. W. (2009). MEG demonstrates a supra-additive response to facial and vocal emotion in the right superior temporal sulcus. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(47), 20010–20015.

Hampson, E., Van Anders, S., & Mullin, L. (2006). A female advantage in the recognition of emotional facial expressions: test of an evolutionary hypothesis. *Evolution and Human Behavior*, *27*(6), 401–416.

Hawk, S., van Kleef, G., Fischer, A., & van der Schalk, J. (2009). "Worth a thousand words": absolute and relative decoding of nonlinguistic affect vocalizations. *Emotion, 9*(3), 293–305.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425-2430.

Haxby, J., Hoffman, E., & Gobbini, I. (2002). Human neural systems for face recognition and social communication. *Biological Psychiatry*, *51*(1), 59-67.

Hornak, J., Rolls, E. T., & Wade, D. (1996). Face and voice expression identification in patients with emotional and behavioural changes following ventral frontal lobe damage. *Neuropsychologia*, *34*(4), 247-261.

Horstmann, G., Lipp, O. V., & Becker, S. I. (2012). Of toothy grins and angry snarls—Open mouth displays contribute to efficiency gains in search for emotional faces. *Journal of Vision*, *12*(5):7, 1-15.

Jack, R., Garrod, O., & Schyns, P. (2014). Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current Biology*, *24*, 187–192,

James, W. (1884). What is an emotion? *Mind*, *9*, 188-205.

Juslin, P. N., & Laukka, P. (2001). Impact of intended emotion intensity on cue utilization and decoding accuracy in vocal expression of emotion. *Emotion*, *1*(4), 381–412.

Juslin, P. N., & Laukka, P. (2003). Communication of emotions in vocal expressions and music performance: Different channels, same code? *Psychological Bulletin*, *129*(5), 770-814.

Kiss, M. & Eimer, M. (2008). ERPs reveal subliminal processing of fearful faces. *Psychophysiology*, *45*(2), 318–326.

Kornreich, C., Brevers, D., Canivet, D., Ermer, E., Naranjo, C., Constant, E., … Noël, X. (2012). Impaired processing of emotion in music, faces and voices supports a generalized emotional decoding deficit in alcoholism. *Addiction*, *108*(1), 80–88.

Kreifelts, B., Ethofer, T., Shiozawa, T., Grodd, W., & Wildgruber, D. (2009). Cerebral representation of non-verbal emotional perception: fMRI reveals audiovisual integration area between voice-and face-sensitive regions in the superior temporal sulcus. *Neuropsychologia, 47*(14), 3059-3066.

Kriegeskorte, N., Mur, M., & Bandettinii, P. (2008a). Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience, 2*(4), 1-28.

Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P.A (2008b)**.** Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron, 60*(6), 1126-1141.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences, 17*(8), 401-412.

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology, 3*, 245, 1-13.

Lavan, N., & McGettigan, C. (2016). Increased discriminability of authenticity from multimodal laughter is driven by auditory information. *The Quarterly Journal of Experimental Psychology*, 1-10.

Liddell, B., Williams, L., Rathjen, J., Shevrin, H., & Gordon, E. (2004). A temporal dissociation of subliminal versus supraliminal fear perception: An event-related potential study. *Journal of Cognitive Neuroscience, 16*(3), 479–486.

Longo, M., & Haggard, P. (2010). An implicit body representation underlying human position sense. *Proceedings of the National Academy of Sciences, 107*, 11727-11732.

Nummenmaa, L., Glerean, E., Hari, R., & Hietanen, J. (2014). Bodily maps of emotions. *Proceedings of the National Academy of Sciences, 111*(2), 646-651.

Ohala, J. (1980). The acoustic origin of the smile. *The Journal of the Acoustical Society of America*, 68, S33. [Abstract].

Palermo, R., O'Connor, K. B., Davis, J. M., Irons, J., & McKone, E. (2013). New tests to measure individual differences in matching and labelling facial expressions of emotion, and their association with ability to recognise vocal emotions and facial identity. *PloS One, 8*(6), e68126.

Peelen, M. V, Atkinson, A. P., & Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *The Journal of Neuroscience, 30*(30), 10127–10134.

Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162*(1-2), 8-13.

Philip, R. C. M., Whalley, H. C., Stanfield, A. C., Sprengelmeyer, R., Santos, I. M., Young, A. W., … Hall, J. (2010). Deficits in facial, body movement and vocal emotional processing in autism spectrum disorders. *Psychological Medicine*, *40*(11), 1919–1929.

Phillips, M. L., Young, A. W., Scott, S. K., Calder, A. J., Andrew, C., Giampietro, V., … Gray, J. A. (1998). Neural responses to facial and vocal expressions of fear and disgust. *Proceedings of the Royal Society B: Biological Sciences*, *265*(1408), 1809–1817.

Pitcher, D., Garrido, L., Walsh, V., & Duchaine, B. C. (2008). Transcranial magnetic stimulation disrupts the perception and embodiment of facial expressions. *Journal of Neuroscience*, *28*(36), 8929–33.

Pound, N., Lawson, D., Toma, A., Richmond, S., Zhurov, A., & Penton-Voak, I. (2014). Facial fluctuating asymmetry is not associated with childhood ill-health in a large British cohort study. *Proceedings of the Royal Society B, 281,* 20141639.

Pound, N., Penton-Voak, I., & Brown, W. (2007). Facial symmetry is positively associated with self-reported extraversion. *Personality and Individual Differences, 43*, 1572-1582.

Raven, J., Raven, J. C., & Court, J. H. (1998). Coloured progressive matrices. Hartcourt Assessments, Texas, USA.

Raven, J., Raven, J. C., & Court, J. H. (2000). Standard Progressive Matrices. NCS Pearson, Texas, USA.

Rohlf, F.J., & Slice, D.E. (1990). Extension of the Proscrustes methods for the optimal superimposition of landmarks. *Systematic Zoology, 39*, 40-59.

Sauter, D. A. & Eimer, M. (2009). Rapid detection of emotion from human vocalizations. *Journal of Cognitive Neuroscience*, *22*(3), 474-481.

Sauter, D. A., Eisner, F., Calder, A., & Scott, S. (2010). Perceptual cues in nonverbal vocal expressions of emotion, *The Quarterly Journal of Experimental Psychology*, *63*(11), 2251-2272.

Sauter, D. A., Panattoni, C., & Happé, F. (2013). Children's recognition of emotions from vocal cues. *The British Journal of Developmental Psychology*, *31(1)*, 97–113.

Sauter, D. A. & Scott, S. (2007). More than one kind of happiness: can we recognize vocal expressions of different positive states? *Motivation and Emotion, 31*, 192-199.

Scherer, K.R., 1994. *Affect bursts*. In: van Goozen, S.H.M., van de Poll, N.E., Sergeant, J.A. (Eds.), Emotions. Lawrence Erlbaum, Hillsdale, NJ.

Scherer, K. R. (2009). The dynamic architecture of emotion: Evidence for the component process model, *Cognition and Emotion*, *23*(7), 1307-1351.

Scott, S., Young, A., Calder, A., Hellawell, D., Aggleton, J., & Johnson, M. (1997). Impaired auditory recognition of fear and anger following bilateral amygdala lesions. *Nature*, 385, 254-257.

Shinkareva, S. V., Malave, V. L., Mason, R. A., Mitchell, T. M., & Just, M. A. (2011). Commonality of neural representations of words and pictures. *Neuroimage*, *54*(3), 2418-2425.

Siegel, S., & Castellan, N.J. (1988). Nonparametric statistics for the behavioral sciences. New York: McGraw-Hill.

Simpson, C., Pinkham, A. E., Kelsven, S., & Sasson, N. J. (2013). Emotion recognition abilities across stimulus modalities in schizophrenia and the role of visual attention. *Schizophrenia Research*, *151*(1-3), 102–106.

Skerry, A. E., & Saxe, R. (2014). A Common Neural Code for Perceived and Inferred Emotion. *The Journal of Neuroscience*, *34*(48), 15997-16008.

Skerry, A. E., & Saxe, R. (2015). Neural representations of emotion are organised around abstract event features. *Current Biology, 25*, 1945-1954.

Smith, H. M., Dunn, A. K., Baguley, T., & Stacey, P. C. (2016). Concordant Cues in Faces and Voices Testing the Backup Signal Hypothesis. *Evolutionary Psychology*, *14*(1), 1474704916630317.

Smith, M. L., Cottrell, G.W., Gosselin, F., & Schyns, P.G. (2005). Transmitting and decoding facial expressions. *Psychological Science*, *16*(3), 184-189.

Sprengelmeyer. R., Young, A., Calder, A., Karnat, A., Lange, H., Homberg, V., Perrett,, D., & Rowland, D. (2006). Loss of disgust: Perception of faces and emotions in Huntington's disease. *Brain, 119*, 1647-1665.

Sprengelmeyer, R., Schroeder, U., Young, A W., & Epplen, J. T. (2006). Disgust in pre-clinical Huntington's disease: a longitudinal study. *Neuropsychologia, 44*(4), 518–33.

Tiddeman, B., Stirrat, M., & Perrett, D. (2005). Towards realism in facial transformation: results of a wavelet MRF method. *Computer Graphics Forum, 24*, 449-456.

Yovel, G. & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, *17*(6). 263-271.

Figure 1.

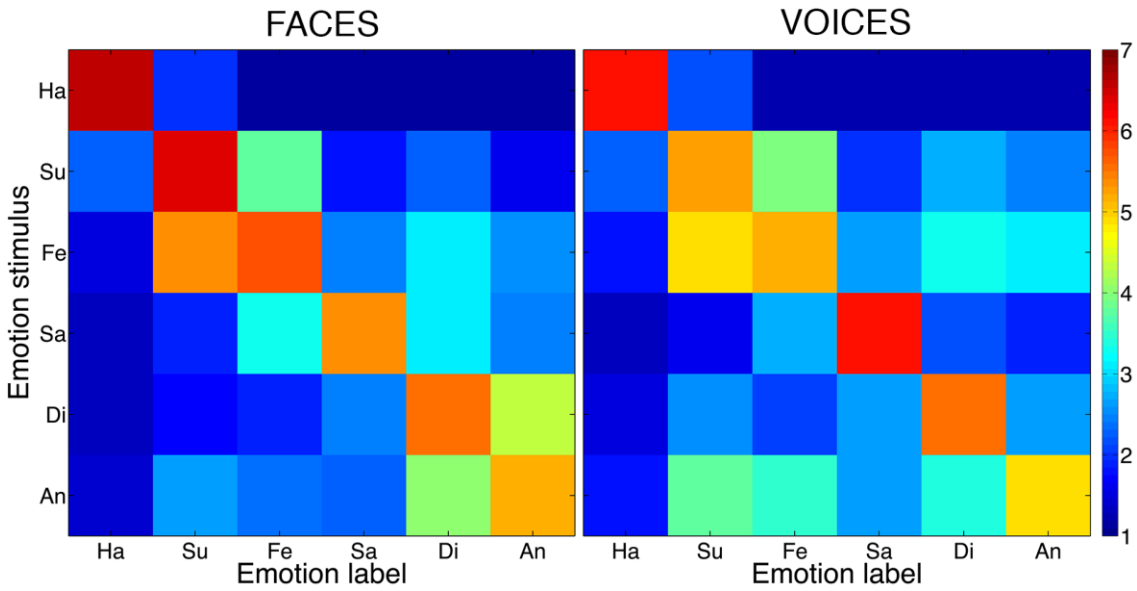


Figure 1. **Representation matrices for faces and voices with mean intensity ratings across participants.** Each cell shows the mean intensity rating for one emotion label (x-axis) given to one type of emotion stimulus (y-axis). Ha = happiness, Su = surprise, Fe = fear, Sa = sadness, Di = disgust, An = anger.

Figure 2.

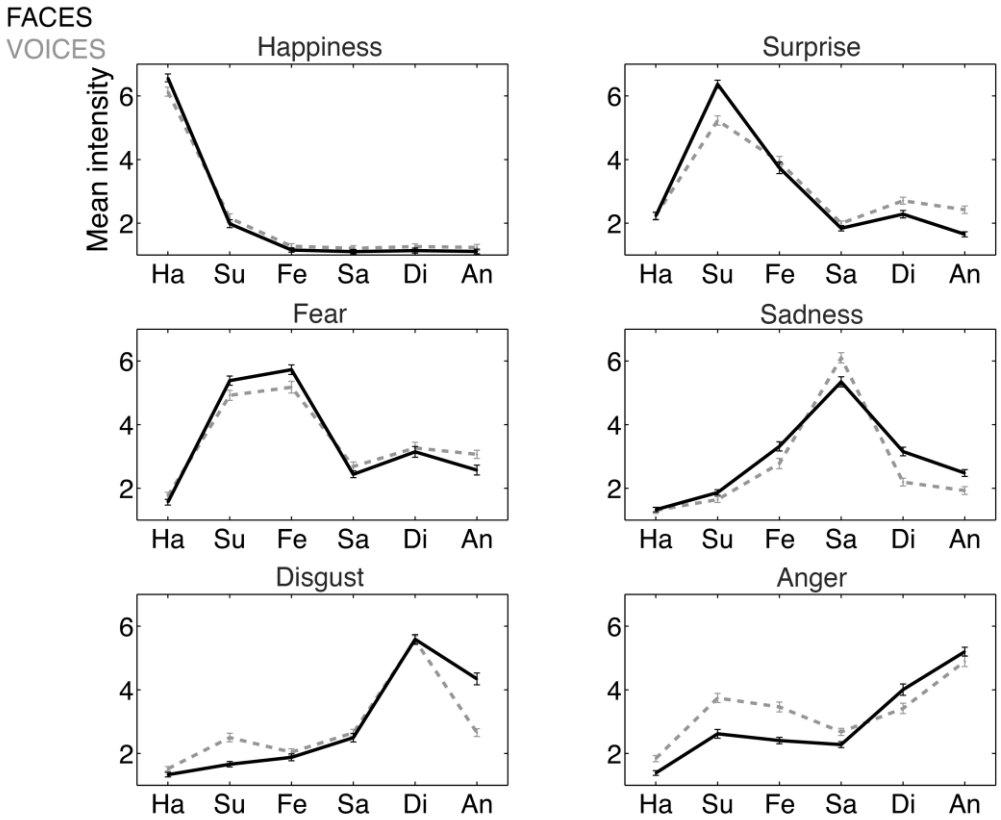


***Figure 2*. Mean emotion rating judgements and standard errors (SE) for each of the six emotions presented, split by modality.** The word at the top of each plot shows the emotion of the stimulus presented, and the x-axis shows the six emotion labels. Ha = happiness, Su = surprise, Fe = fear, Sa = sadness, Di = disgust, An = anger.

Figure 3.

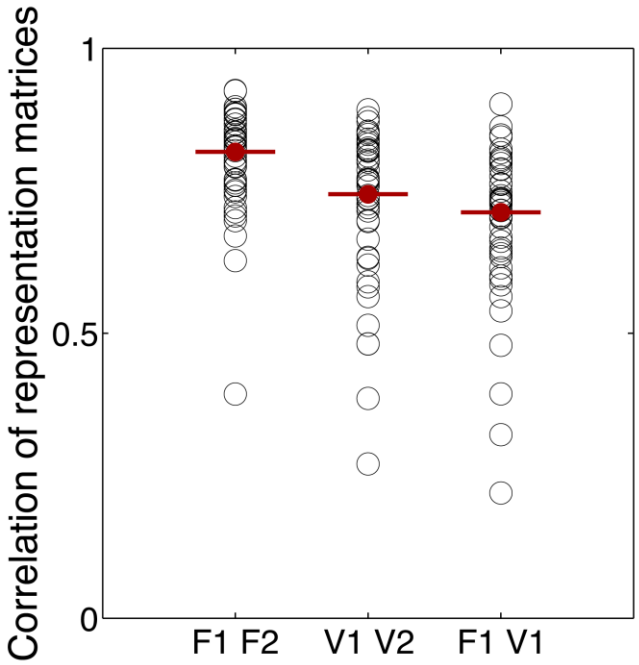


Figure 3. **Correlations of the representation matrices (or rating profiles) of emotional stimuli either *within* the same modality (F1 *vs* F2 and V1 *vs* V2) or *across* different modalities (F1 *vs* V1).** Each black empty circle represents one participant, and the red filled circles with a line indicate the mean correlations across participants.

Figure 4.

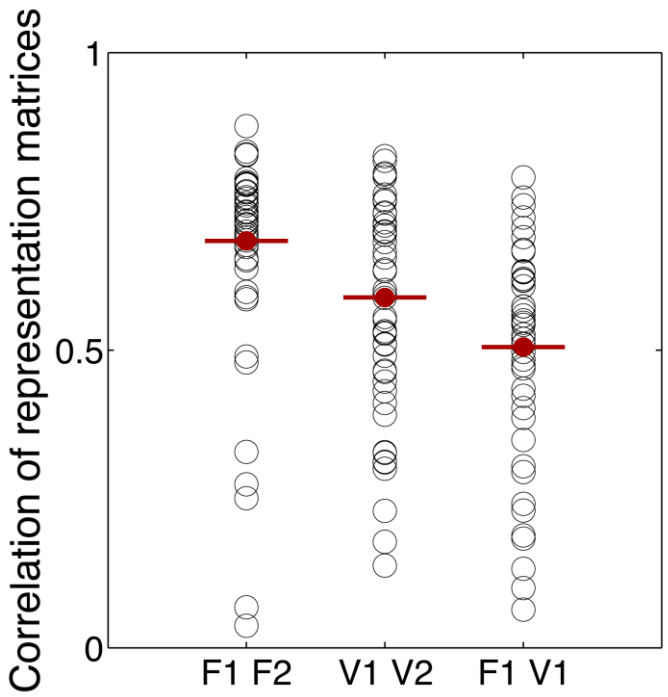


Figure 4. **Correlations of the representation matrices (or rating profiles) of emotional stimuli either *within* the same modality (F1 *vs* F2 and V1 *vs* V2, both without diagonal) or *across* different modalities (F1 *vs* V1, without diagonal).** Each black empty circle represents a participant, and the red filled circles with a line indicate the mean correlations across participants.

Figure 5.

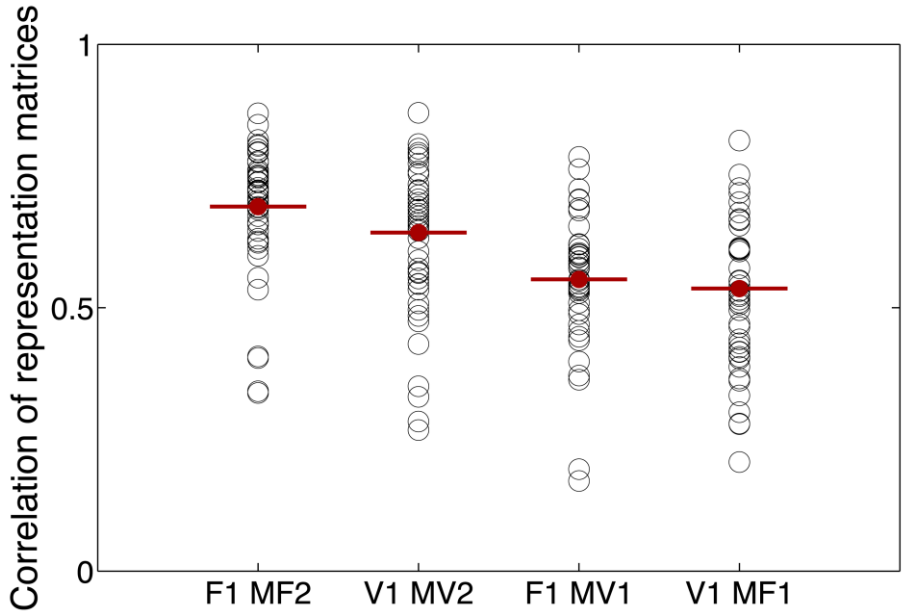


Figure 5. **Correlations of the representation matrices (or rating profiles) of emotional stimuli either within the same modality (F1 *vs* MF2 and V1 *vs* MV2, both without diagonal) or across different modalities (F1 *vs* MV1 and V1 *vs* MF1, both without diagonal).** These correlations were computed between an individual representation matrix and the mean representation matrix of all other participants. M indicates mean across participants. Each black empty circle represents a participant, and the red filled circles with a line indicate the mean correlations across participants.

Figure 6.

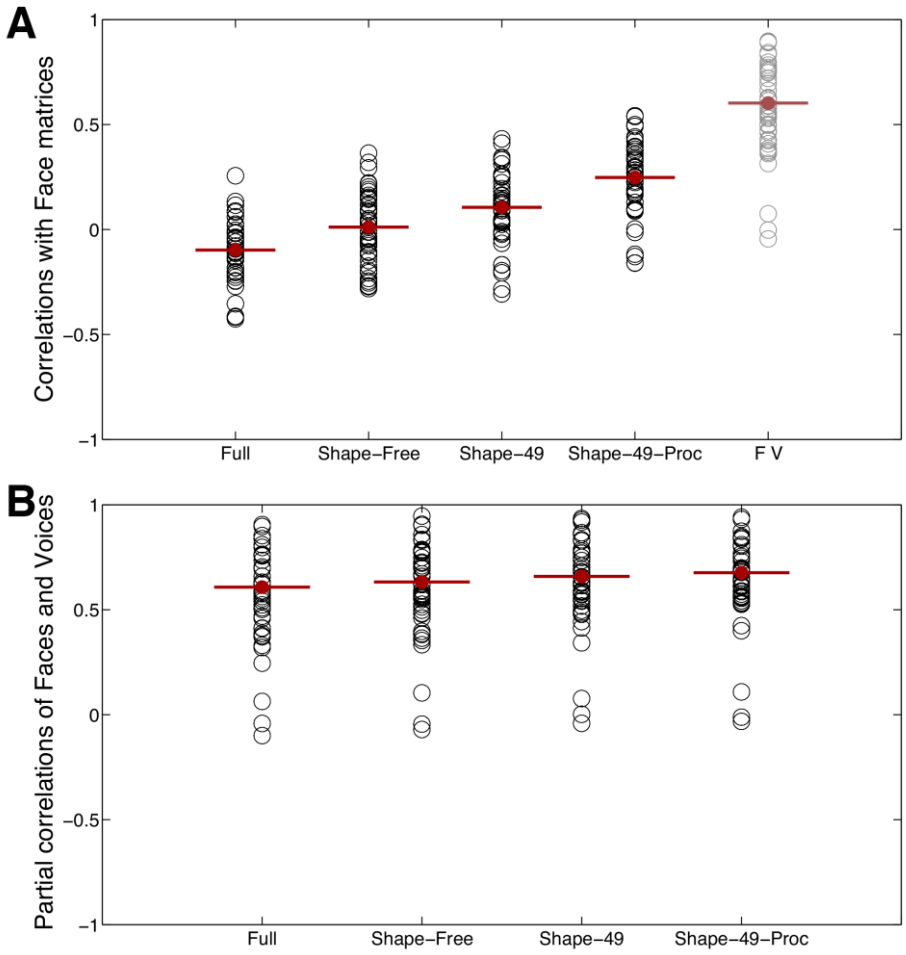


Figure 6. **Analysis of low-level visual properties.** Panel A shows correlations of representation matrices using visual properties of the images (1: Full images, 2: Shape-Free images, 3: Shape-49, 4: Shape-49-Procrustes) and the behavioural matrices for faces. Each circle shows the corrrelation for one participant, and the red full circle shows the mean across participants. The matrices using shape information seem to be the best predictors of behaviour. The last column in Panel A shows the correlations of the representation matrices for emotional faces and emotional voices. Panel B shows the partial correlations between the representation matrices for faces and the representation matrices for voices, while controlling for each of the visual properties. Each circle shows the partial corrrelation for one participant, and the red full circle shows the mean across participants. All partial correlations are still high, even after controlling for the variance of the visual properties of the images.

Figure 7.

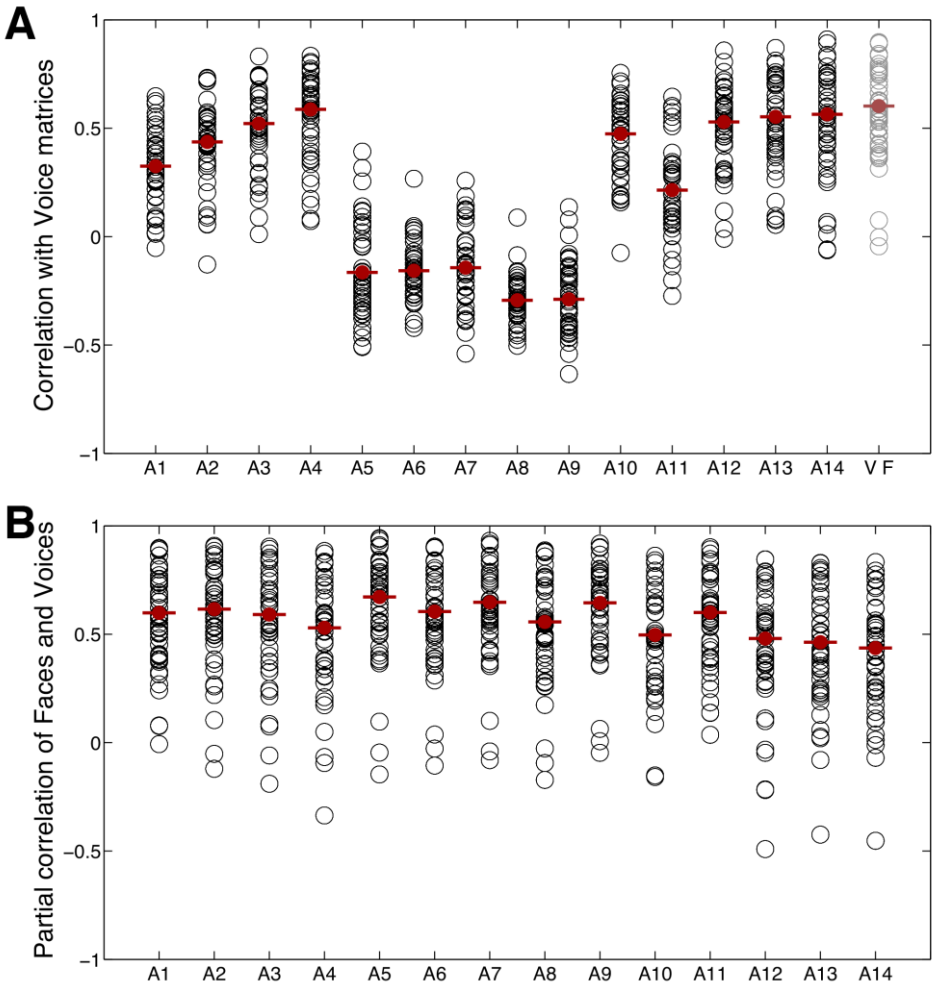


Figure 7. **Analysis of low-level acoustic properties.** Panel A shows correlations of representation matrices using acoustic properties of the images (A1: Total duration, A2: Amplitude SD, A3: Mean intensity, A4: Numbe of amplitude onsets, A5: F0 minimum, A6: F0 maximum, A7: F0 mean, A8: F0 SD, A9: Spectral centre of gravity, A10: Spectral SD, A11: Mean HNR, A12: Jitter, A13: Percentage of unvoiced segments, A14: Shimmer — see Appendix 2 for description of each of these properties) and the behavioural matrices for voices. Each circle shows the corrrelation for one participant, and the red full circle shows the mean across participants. The last column in Panel A shows the correlations of the representation matrices for emotional faces and emotional voices. Panel B shows the partial correlations between the representation matrices for voices and the representation matrices for faces, while controlling for each of the acoustic properties. Each circle shows the partial corrrelation for one participant, and the red full circle shows the mean across participants. All partial correlations are still high, even after controlling for the variance of the acoustic properties of the sounds.

Figure 1

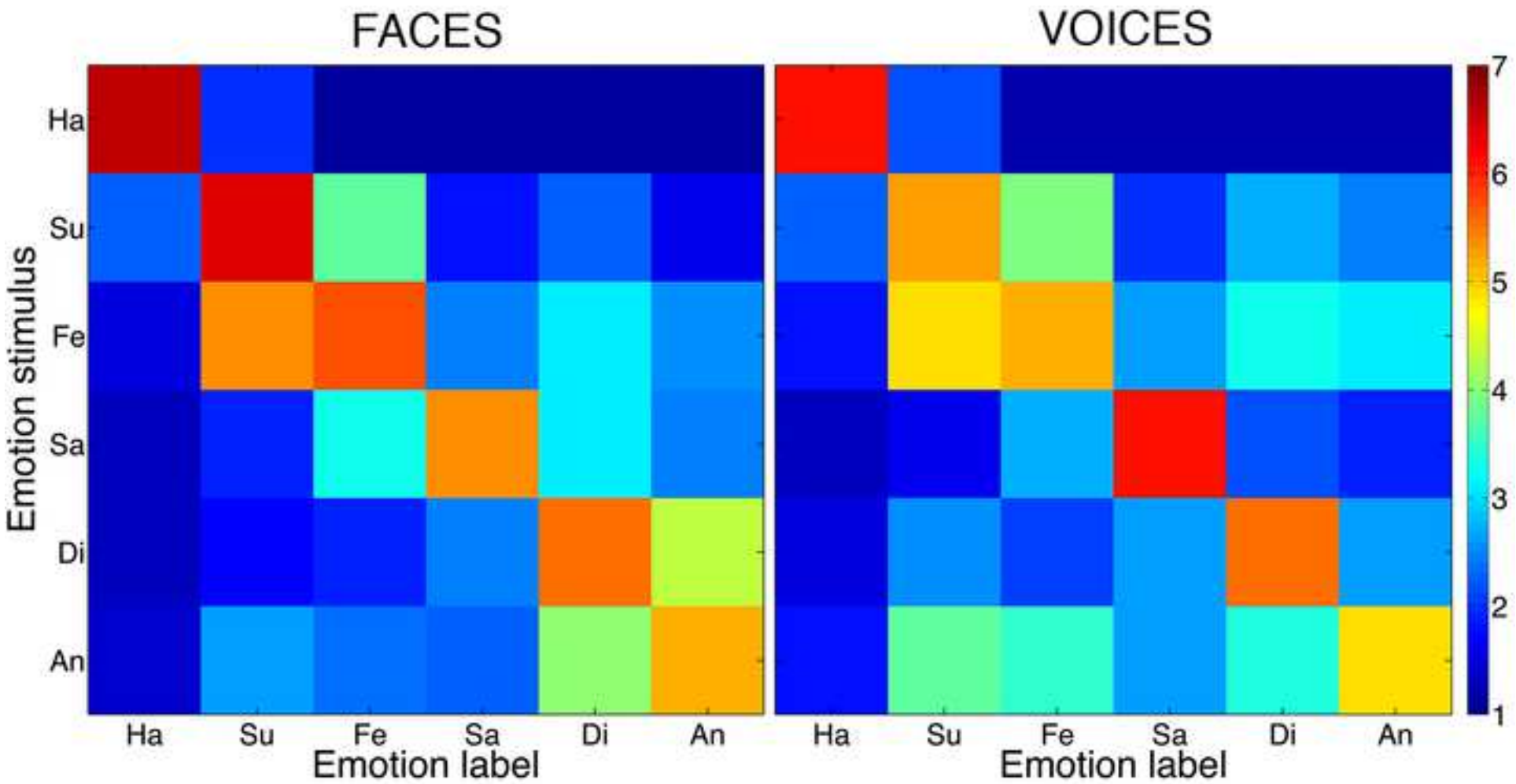

FACES
VOICES
Emotion stimulus
Ha
Su
Fe
Sa
Di
An
Emotion label
Ha
Su
Fe
Sa
Di
An
Emotion label
7
6
5
4
3
2
1

Figure 2

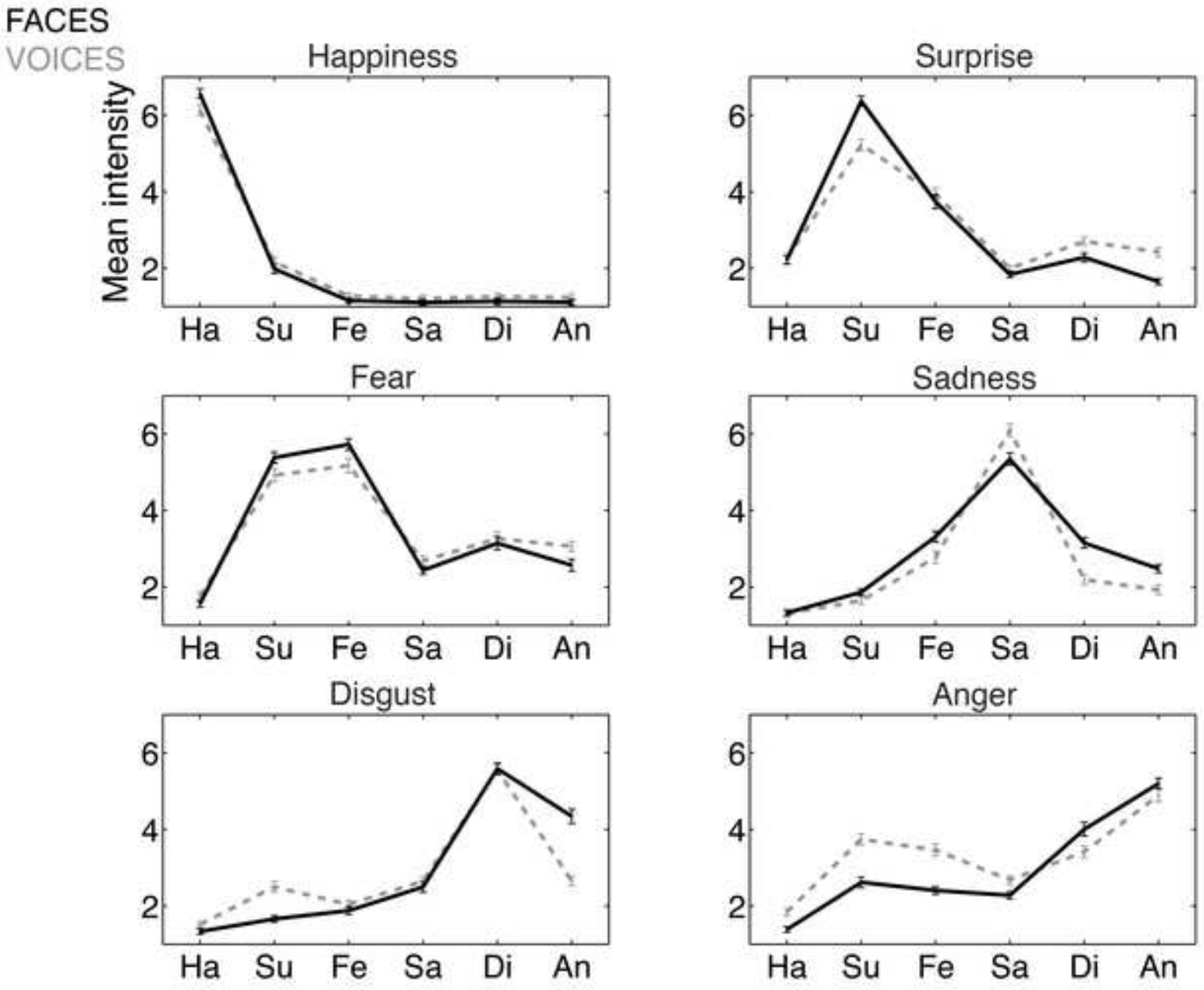

FACES
VOICES
Happiness
Surprise
Fear
Sadness
Disgust
Anger
Mean intensity
2
4
6
Ha
Su
Fe
Sa
Di
An

Figure 3

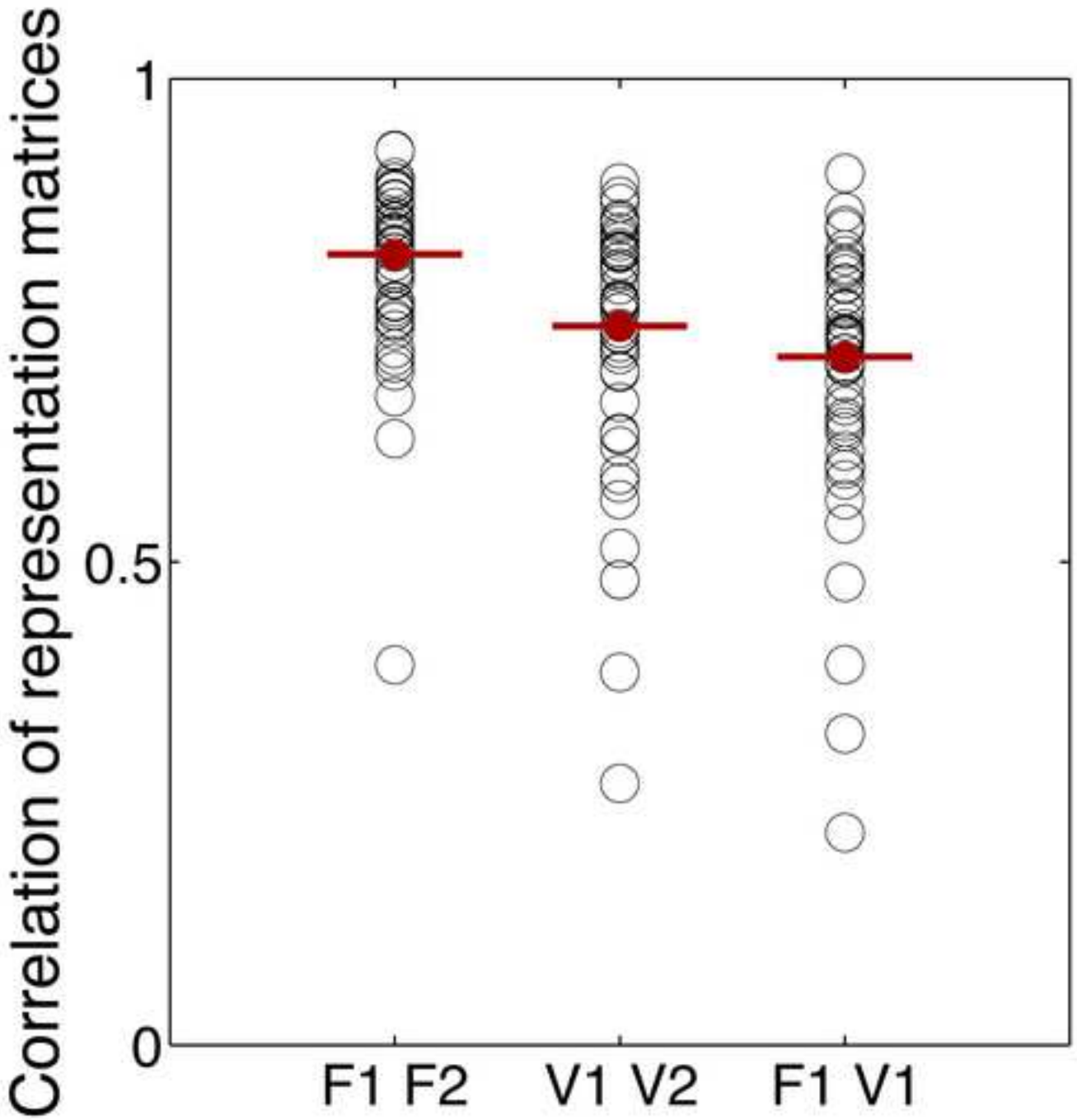

Correlation of representation matrices
1
0.5
0
F1 F2
V1 V2
F1 V1

Figure 4

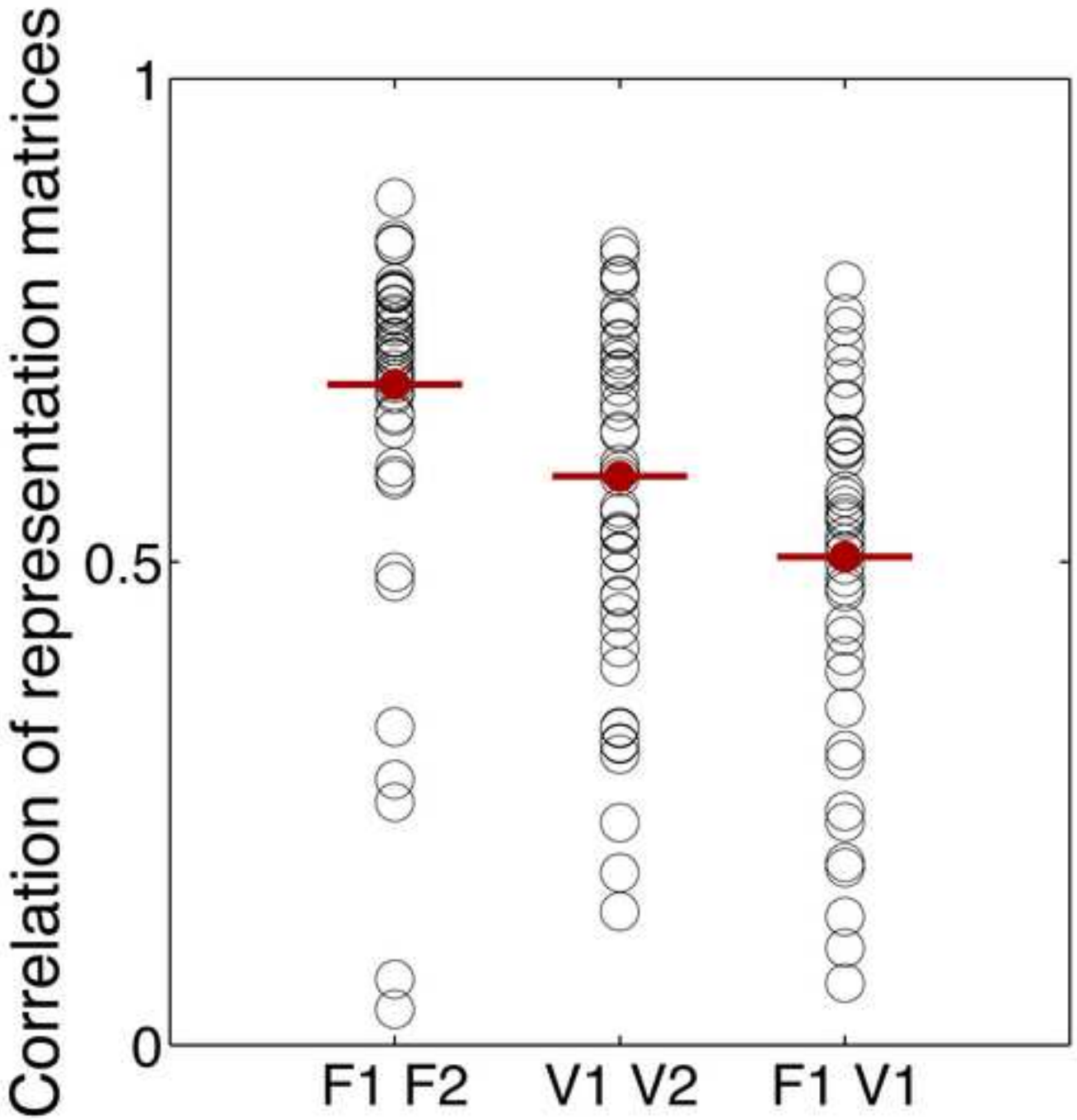

Correlation of representation matrices
1
0.5
0
F1 F2
V1 V2
F1 V1

Figure 5

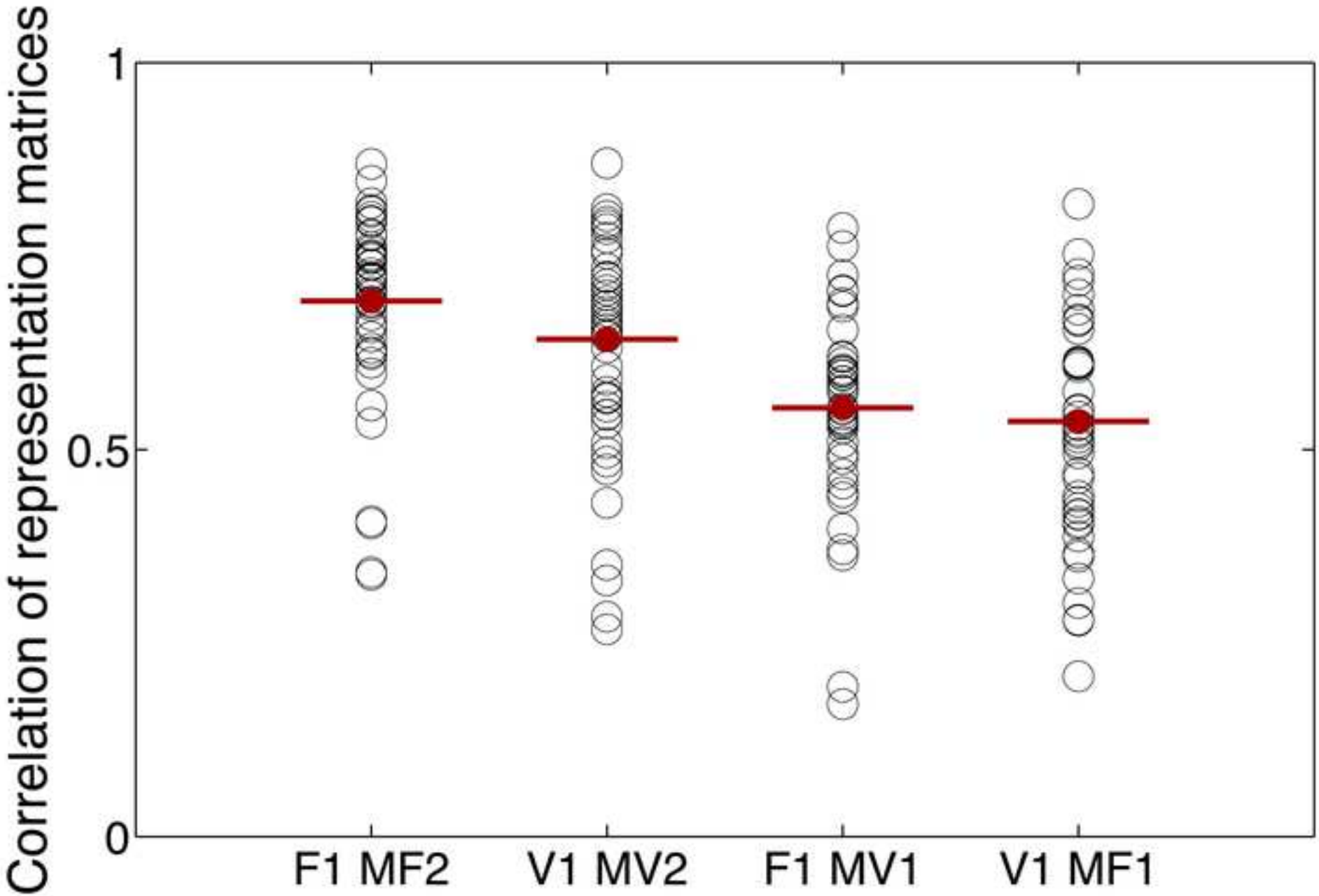

Correlation of representation matrices
1
0.5
0
F1 MF2
V1 MV2
F1 MV1
V1 MF1

Figure 6

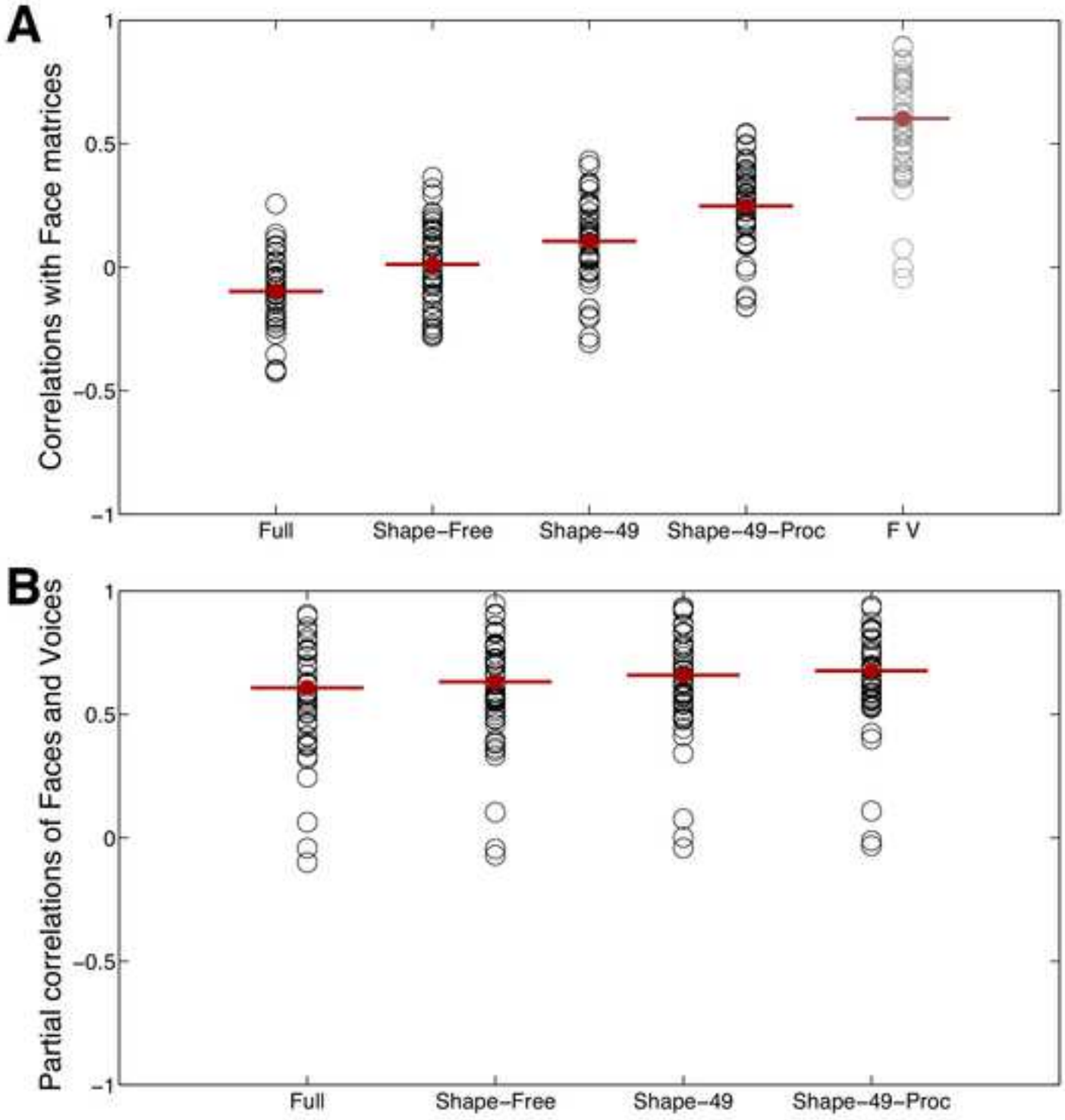
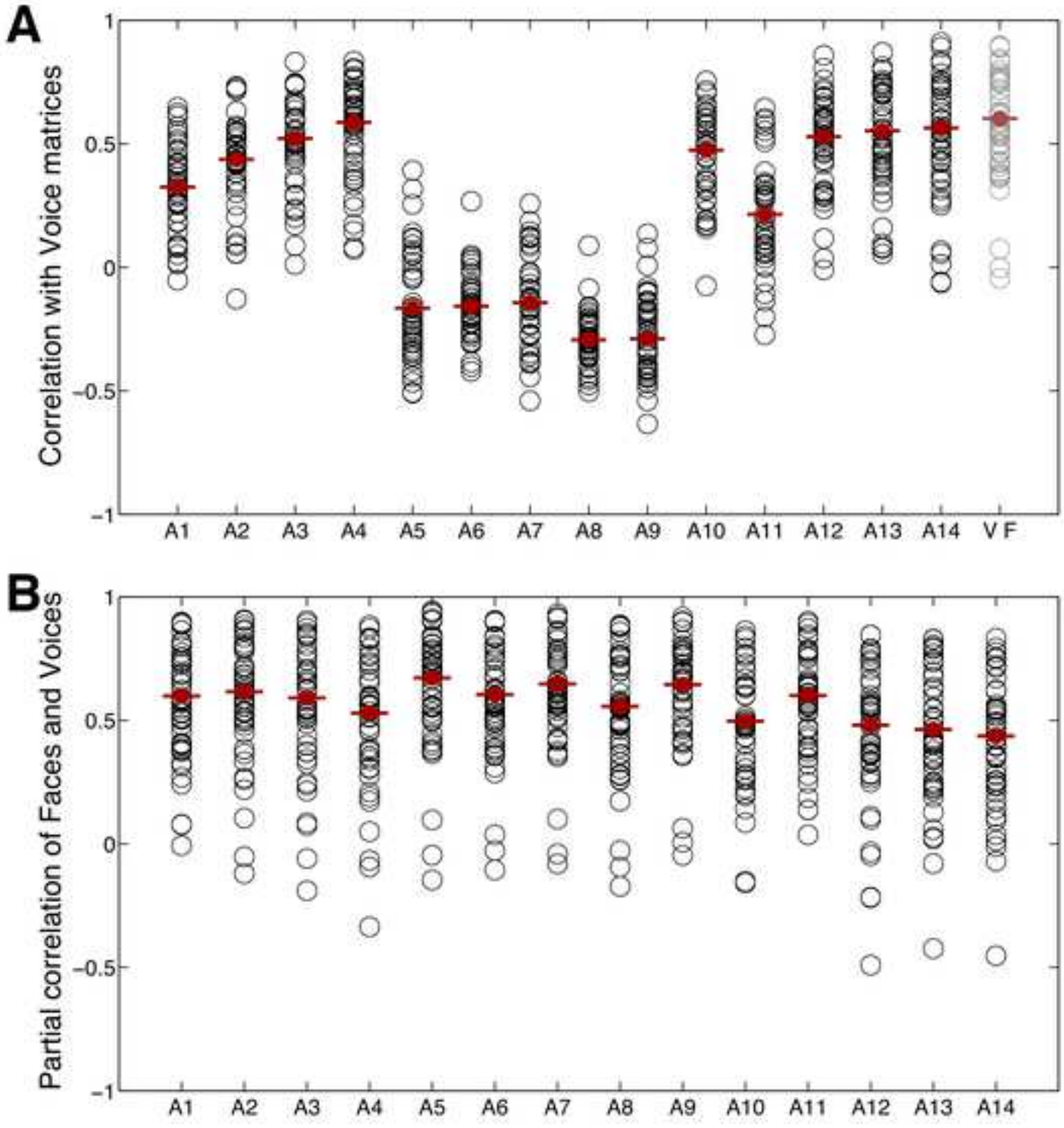

A
Correlations with Face matrices
1
0.5
0
−0.5
−1
Full
Shape−Free
Shape−49
Shape−49−Proc
F V
B
Partial correlations of Faces and Voices
1
0.5
0
−0.5
−1
Full
Shape−Free
Shape−49
Shape−49−Proc

Figure 7


A
Correlation with Voice matrices
1
0.5
0
−0.5
−1
A1
A2
A3
A4
A5
A6
A7
A8
A9
A10
A11
A12
A13
A14
V F
B
Partial correlation of Faces and Voices
1
0.5
0
−0.5
−1
A1
A2
A3
A4
A5
A6
A7
A8
A9
A10
A11
A12
A13
A14

APPENDICES: Similar representations of emotions across faces and voices

APPENDIX 1: Correlations of response profiles across modalities

In the main text, we presented results for the correlation of F1 *versus* V1. The mean correlation between ratings of F1 versus V1 was r =.82 (*SD* = 0.24). Correlations of the other splits of the data led to very similar results. The mean correlation between ratings of F1 versus V2 was r = .67 (*SD* = 0.23), the mean correlation between ratings of F2 versus V1 was r = .76 (*SD* = 0.31), and the mean correlation between ratings of F2 versus V2 was r = .73 (*SD* = 0.27). All these correlations were significantly different from zero (all $p < .001$).

Similar results were observed when we removed the diagonal. The mean correlation between ratings of F1 versus V1, without diagonal, was r = .51 (*SD* = 0.25), and correlations of the other splits of the data led to very similar results. The mean correlation between ratings of F1 versus V2 was r = .45 (*SD* = 0.24), the mean correlation between ratings of F2 versus V1 was r = .61 (*SD* = 0.34), and the mean correlation between ratings of F2 versus V2 was r = .54 (*SD* = 0.28). All these correlations were significantly different from zero (all $p < .001$).

We also performed analyses in which we did not do any averaging of ratings per block. We did this by splitting all the emotion ratings by identity of the stimulus. There were four identities for the faces and four identities for the voices, so we obtained a representation matrix or rating profile for each of these identities (four for faces and four for voices). We then performed correlations across all possible splits of the data, both for within-modality correlations, and across-modality correlations. For the within-modality correlations, there were six possible combinations of splits of the data. For the across-modalities correlations, there were 16 possible combinations of splits of the data. Figures A1-1 and A1-2 show the results of these analyses.

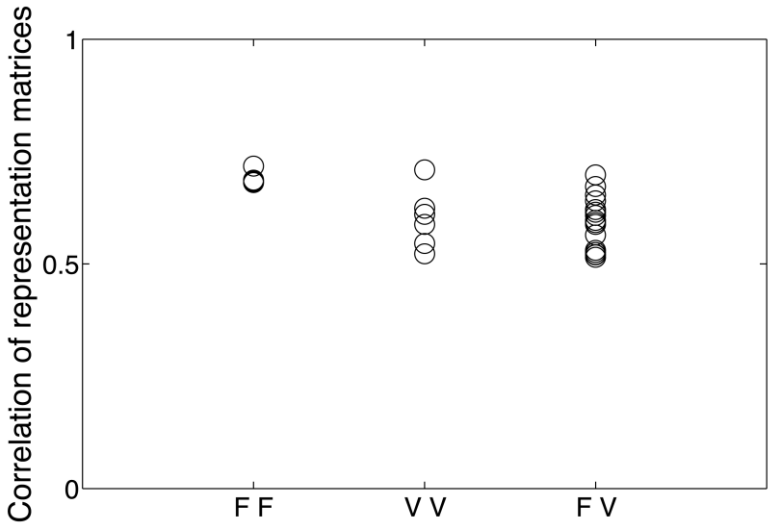


***Figure A1-1*: Correlations of representation matrices for faces and voices, using separate matrices for each identity**. There were four face matrices and four voice matrices for each participant. Each circle shows the mean correlation across all participants for one combination of two of those matrices. There were six possible combinations of splits of the data for the within-modality correlations (FF are the within-modality correlations for faces and VV are the within-modality correlations for voices) and there were 16 possible combinations for the across-modalities correlations (FV). FF correlations ranged between .68 and .72 (all $z = 5.84$, all $p < .001$). VV correlations ranged between .52 and .71 (all $z > 5.78$, all $p < .001$). FV correlations ranged between .52 and .67 (all $z > 5.78$, all $p < .001$). As expected, these correlations are lower that the ones presented in the main text and above, as we do not perform any averaging across trials, and therefore the representation matrices are more stimuli-specific. Nevertheless, all mean correlations were of medium size (and showing substantial amount of shared variance) and all significanly different from zero.

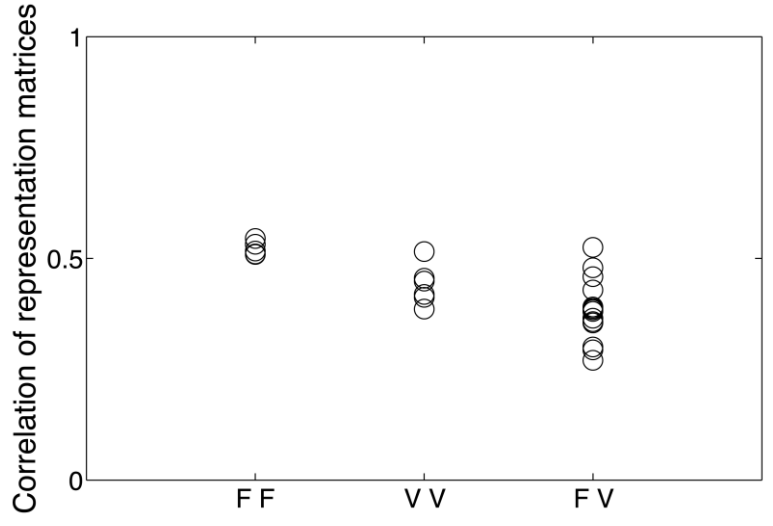


***Figure A1-2*: Correlations of representation matrices for faces and voices (without using the diagonals of the matrices), using separate matrices for each identity.** This is the same analysis as done for Figure A1-1, but here the diagonals were removed. FF correlations ranged between .51 and .55 (all $z > 5.76$, all $p < .001$). VV correlations ranged between .39 and .52 (all $z > 5.69$, all $p < .001$). FV correlations ranged between .27 and .52 (all $z > 5.32$, all $p < .001$). As expected, these correlations are lower that the ones presented in the main text and above, as we do not perform any averaging across trials, and therefore the representation matrices are more stimuli-specific. Nevertheless, all mean correlations were of small to medium size (and showing substantial amount of shared variance) and all significanly different from zero.

APPENDIX 2: Representation matrices for visual properties of the emotional faces

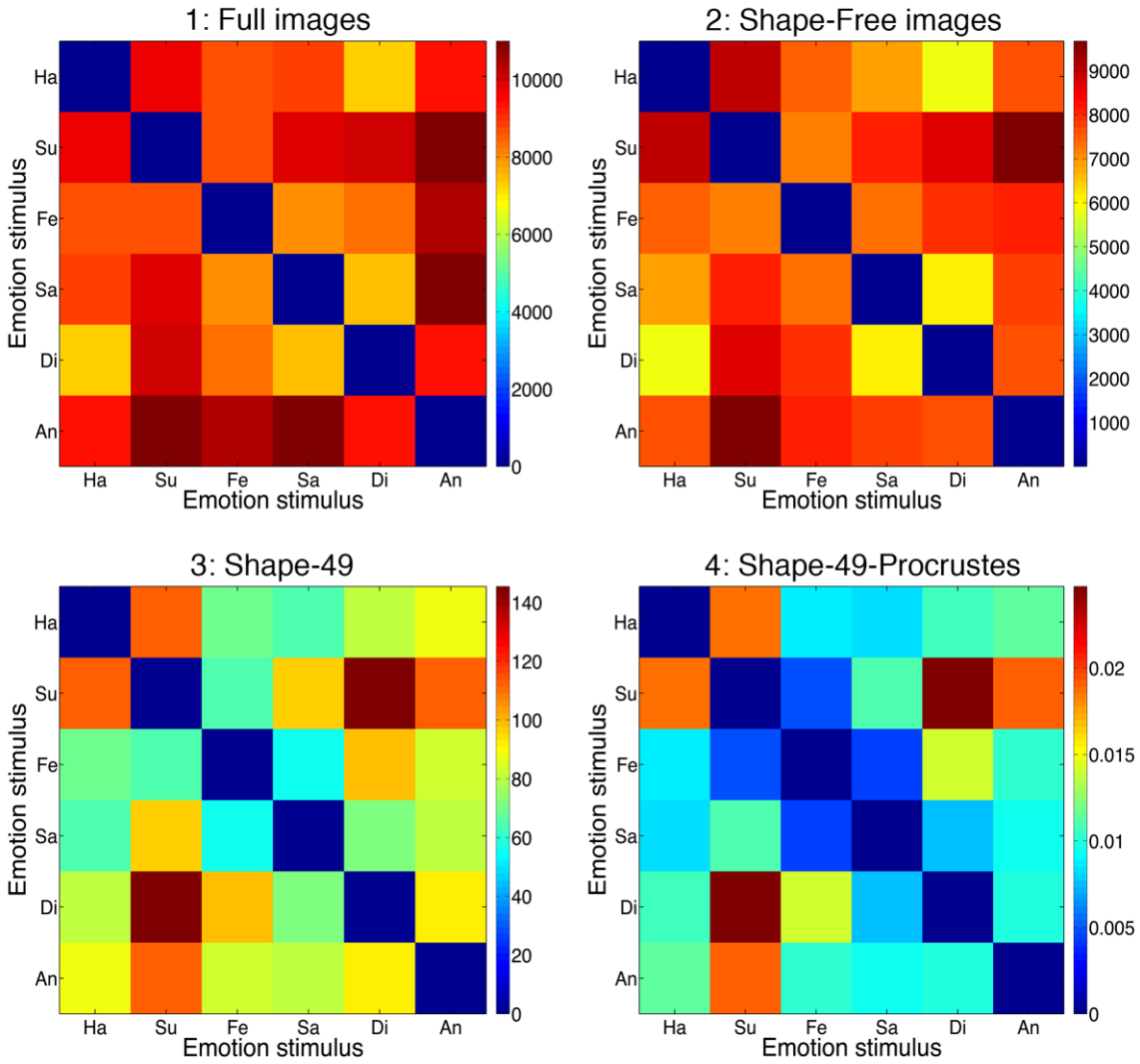


***Figure A2-1*: Mean representation matrices for each of the visual properties**. 1: Full images, 2: Shape-Free images, 3: Shape-49, and 4: Shape-49-Procrustes. See main text for more details about how these matrices were computed. Ha = happiness, Su = surprise, Fe = fear, Sa = sadness, Di = disgust, An = anger.

APPENDIX 3: Acoustic properties of the emotional voices

We analysed 14 acoustic properties and here we provide a brief description of each: (1) total duration (in seconds, defined as the interval between the first zero-crossing of the onset to the final zero crossing after the offset of the vocalisation), (2) amplitude: standard deviation (in pascal, defined as the variability in the amplitude profile over the duration of the sound), (3) mean intensity (in dB, defined as the average intensity of the vocalisation relative to the auditory threshold), (4) number of amplitude onsets (the amplitude onsets were manually labelled to describe the structure of the sounds' threshold (e.g. laughter has multiple onsets whereas a scream typically only has one amplitude onset)), (5) F0 minimum (in Hz, defined as the lowest F0 measurement within a vocalisation, which was manually labelled to reduce the impact of doubling/halving error on these measures), (6) F0 maximum (in Hz, defined as the highest F0 measurement within a vocalisation, which was manually labelled to reduce the impact of doubling/halving error on these measures), (7) F0 mean (in Hz, computed using the auto-correlation method in PRAAT. F0 floor was set at 75 Hz and the F0 ceiling at 1000 Hz to include potentially high-pitched vocalisations such as screams and laughter), (8) F0 standard deviation (in Hz, defined as the standard deviation of the F0 mean), (9) spectral centre of gravity (in Hz, measure for the mean height of the frequencies for each vocalisation, which captures the weighting of energy in the sound across the frequency range), (10) standard deviation of the spectrum (in Hz, measure describing the dispersion of spectral energy across the frequency range), (11) mean harmonics-to-noise-ratio (in dB, defined as the mean ratio of quasi periodic to non-period signals across time segments), (12) jitter (in dB, defined as the average absolute difference between consecutive periods, divided by the average period, i.e., microfluctuations in the duration of each period), (13), percentage of unvoiced segments (percentage of frames lacking harmonic structure), and (14) shimmer (in dB, defined as the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude).

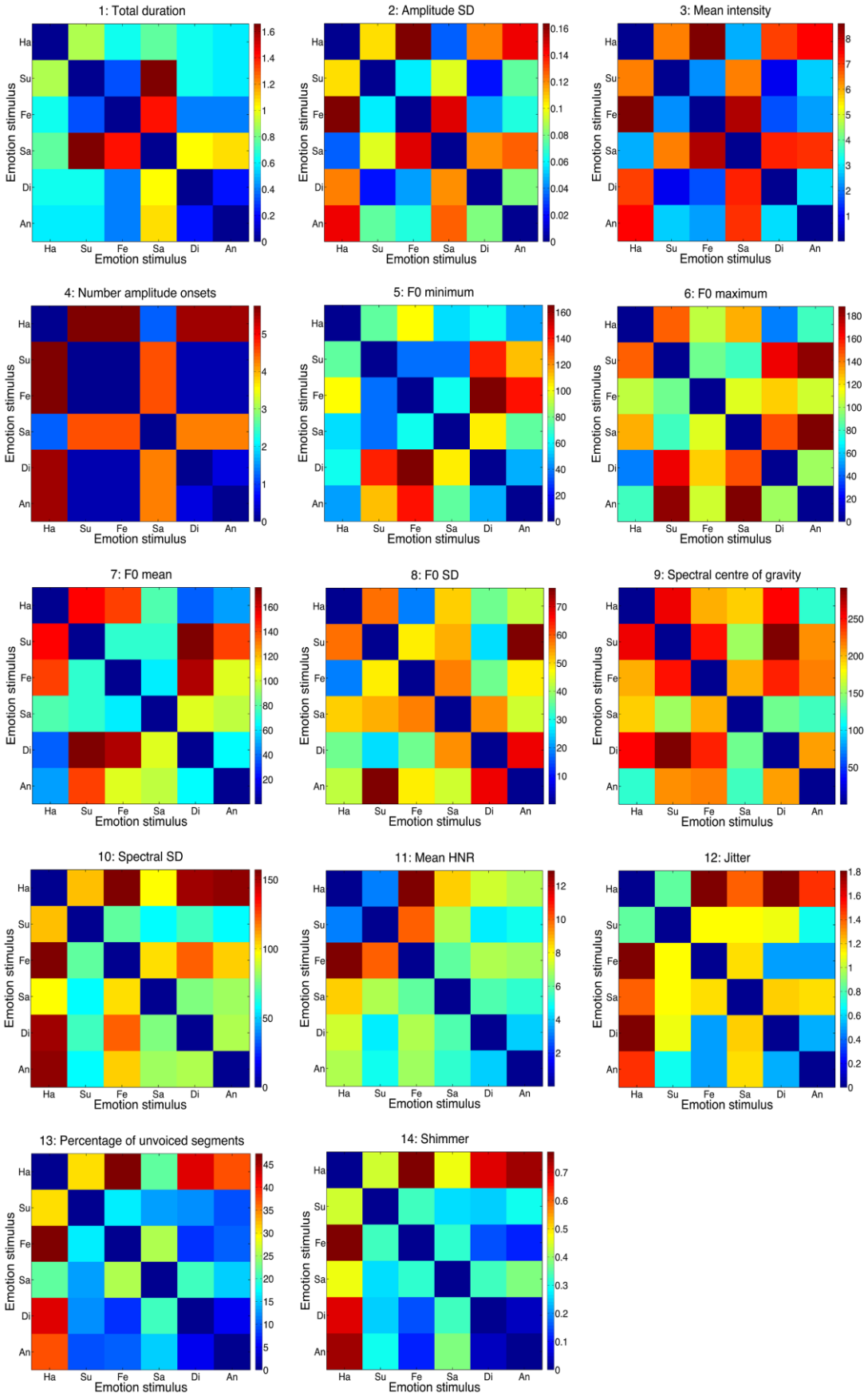


***Figure A3-1*: Mean representation matrices for each of the acoustic properties.** See main text for more details about how these matrices were computed. Ha = happiness, Su = surprise, Fe = fear, Sa = sadness, Di = disgust, An = anger.

APPENDIX 4: New behavioural representation matrices

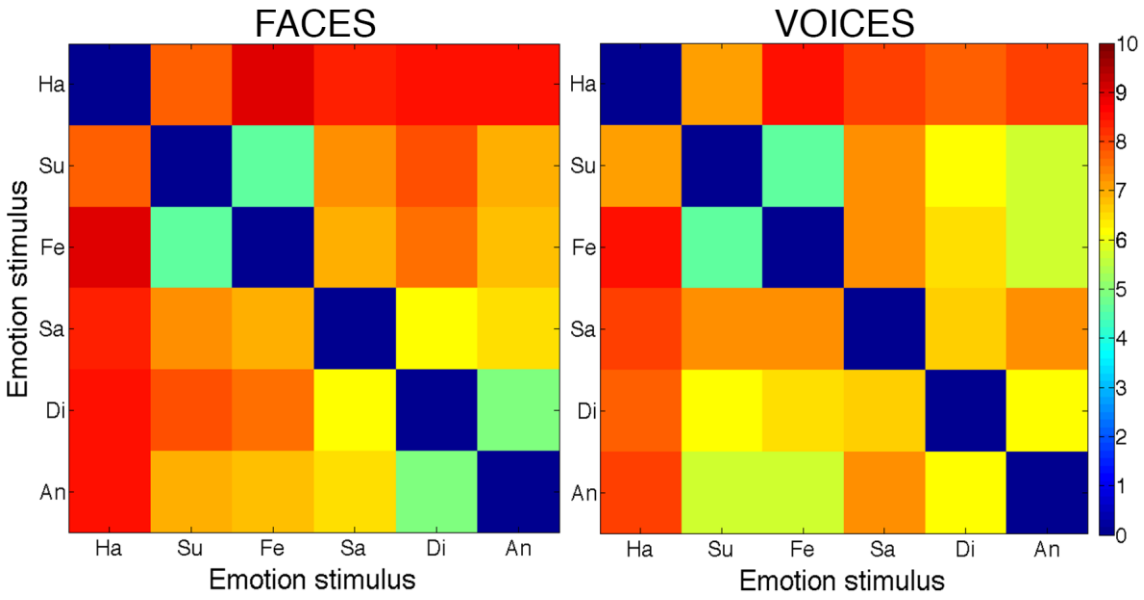


***Figure A4-1*: Mean representation matrices for the behavioural ratings of emotional faces and voices.** For each pair of stimuli, we computed the Euclidean distance between the ratings on the six emotion labels that were given to each stimulus (i.e., we computed the Euclidean distance of two vectors, each with ratings on six emotion labels). For each participant, we averaged the matrices of different identities of the stimuli, separately for faces and voices. For this figure only, we averaged the matrices across all the participants. Ha = happiness, Su = surprise, Fe = fear, Sa = sadness, Di = disgust, An = anger.

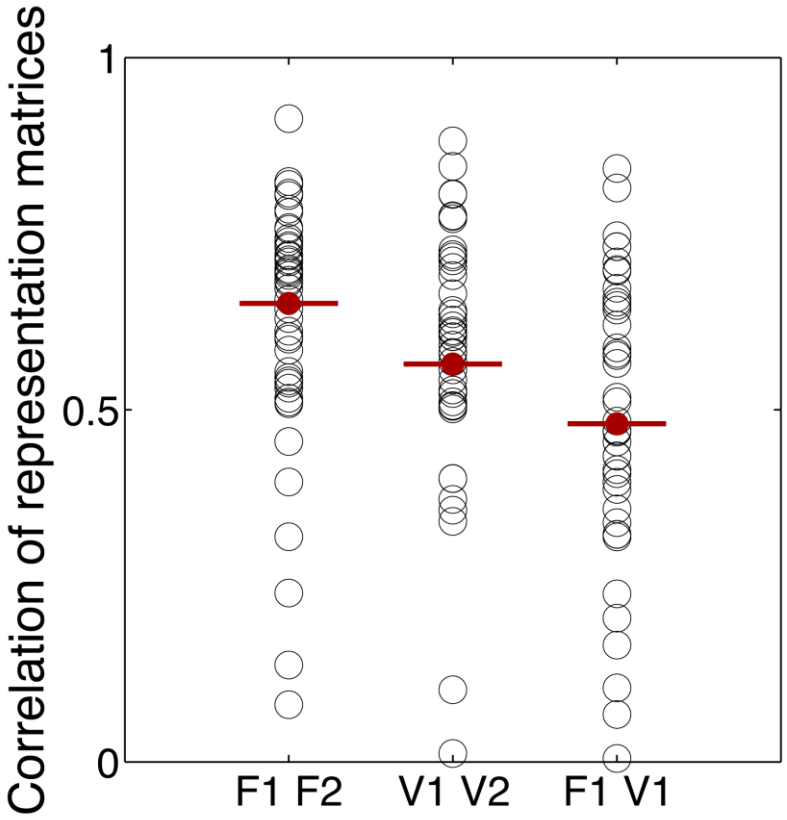


***Figure A4-2*: Individual correlations of representation matrices within- (F1 *versus* F2, V1 *versus* V2) and across-modalities (F1 *versus* V1) for the new behavioural matrices.** For this analysis, which was conducted in the same manner as the main analysis in the manuscript (section 2 of the Results) we averaged matrices within each block of stimuli, resulting in four representation matrices per individual: F1, F2, V1, and V2. This is the same analysis as the ones shown in Figures 3 and 4 of the main text, but using the new matrices. All correlations were only computed with the lower triangular part of the matrices. Means across participants are shown in red (to compute these means, we first z-trasnformed all individual correlations, then averaged the trasnformed values, and finally reverse trasnsformed the mean to a value between -1 and 1). The mean correlation of the representation matrices for face stimuli (F1 *versus* F2) was $r = .65$ ($SD = .27$; the result of the Wilcoxon signed ranks test comparing non-transformed correlations to zero was $z = 5.84$, $p < .001$). The mean correlation of the representation matrices for voice stimuli (V1 *versus* V2) was $r = .56$ ($SD = .34$; $z = 5.65$, $p < .001$). The mean correlation of the representation matrices across modalities (F1 *versus* V1) was $r = .48$ ($SD = .32$; $z = 5.68$, $p < .001$). These results are comparable to the ones shown in Figure 4, and suggest that this new method of computing similarities across behavioural responses yields very similar results to the previous method. In the main text, Figures 6 and 7 also present correlations of representation matrices across modalities, but those matrices were averaged across all stimuli for each modality. Appendix 5 shows correlations across modalities without any averaging.

APPENDIX 5: Analysis of low-level properties using non-averaged matrices

For each of the visual properties that we considered, and each pair of stimuli, we computed the Euclidean distance between the respective feature vectors (described in section 4.1 of the main text), resulting in four 24-by-24 representation matrices. We then computed the behavioural representation matrices for faces in the same manner: for each participant, and each pair of stimuli, we computed the Euclidean distance between the respective feature vectors (each vector consisting of ratings on six emotion labels), resulting in a 24-by-24 matrix per participant. We note that each entry in the visual representation matrices corresponds to one stimulus, which matches exactly the same stimulus on the face behavioural representation matrix. Figure A5-1 shows correlations and partial correlations using these large, non-averaged matrices. Briefly, these results show that not averaging across representation matrices of faces produced results largely comparable to the ones that we described in the main text, using averaged matrices (compare Figure A5-1 with Figure 6).

Similarly, for each of the 14 acoustic properties that we considered, and each pair of stimuli, we computed the Euclidean distance between the respective feature vectors (described in section 4.2 of the main text and Appendix 3, and each corresponding to a single value on an acoustic property), resulting in fourteen 24-by-24 representation matrices. We then computed behavioural representation matrices for voices, again comparing each pair of stimuli (using the same procedure as above for faces). Figure A5-2 shows correlations and partial correlations using these large, non-averaged matrices. In a similar manner as it was found for faces, the results show that not averaging across representation matrices of voices produced results largely comparable to the ones we described before, using averaged matrices (compare Figure A5-2 with Figure 7).

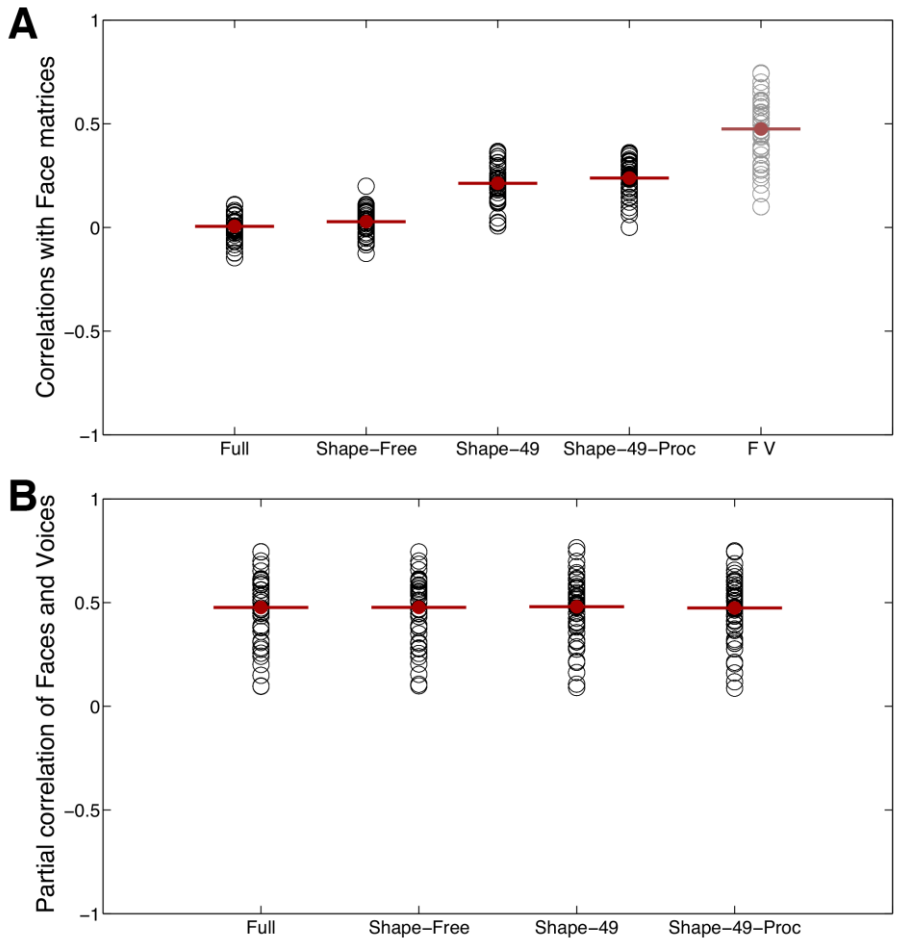


***Figure A5-1*: Analysis of low-level visual properties using 24-by-24, non-averaged matrices.** Panel A shows correlations of representation matrices using visual properties of the images (1: Full images, 2: Shape-Free images, 3: Shape-49, 4: Shape-49-Procrustes — see main text for description of each of these properties) and the behavioural matrices for faces. Each circle shows the corrrelation for one participant, and the red full circle shows the mean across participants. Each representation matrix consisted of a 24-by-24 matrix in which each entry was a stimulus. Three of the matrices describing low-level visual properties correlated with the behavioural matrices for faces significantly above zero: Shape-Free (mean r = .03; z = 2.92, p < .0035), Shape-49 (mean r = .21; z = 5.84, p < .001) and Shape-49-Procrustes (mean r = .24; z = 5.84, p < .001). Despite the significant correlation between the Shape-Free images and the behavioural matrices for faces, the effect size was very small. Conversely, like for the analysis presented in Figure 6, the matrices using shape information seem to be better predictors of behaviour. The last column in grey shows the correlation of the behavioural matrix for faces and the behavioural matrix for voices for each participant. The mean correlation across participants was r = .48 (z = 5.84, p < .001). Please note that these correlations were computed still using 24-by-24 matrices, but in the case of these correlation across modalities, there is not a perfect correspondence between the entries on face and voice matrices, given that the identity of the faces are not the same as the identities of the voices. Therefore, we arbitrarily matched the identities of faces with the identities of voices. Nevertheless, we presented these correlations here for completeness, and because it is important to compare these correlations to the partial correlations in panel B.

Panel B shows the partial correlations between the representation matrices for faces and the representation matrices for voices, while controlling for each of the visual properties. Each circle shows the partial corrrelation for one participant, and the red full circle shows the mean across participants. All partial correlations were still high, even after controlling for the variance of the visual properties of the images (all mean r > .47; all z = 5.84, all p < .001). Please note again that, in this

case, we were using partial correlations of two 24-by-24 behavioural matrices in which the entries did not match entirely, i.e., they matched on the emotion of the stimulus but not on their identity.

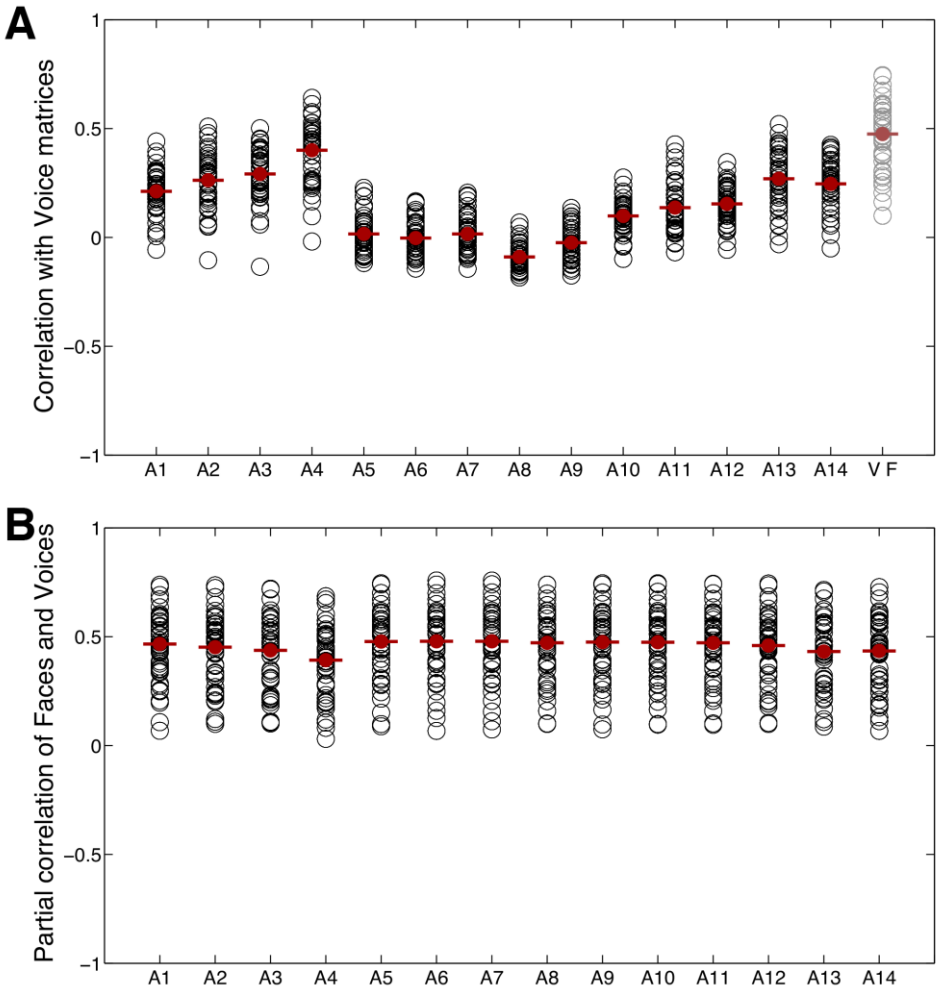


***Figure A5-2*: Analysis of low-level acoustic properties using 24-by-24, non-averaged matrices.** Panel A shows correlations of representation matrices using acoustic properties of the sounds (A1: Total duration, A2: Amplitude SD, A3: Mean intensity, A4: Number of amplitude onsets, A5: F0 minimum, A6: F0 maximum, A7: F0 mean, A8: F0 SD, A9: Spectral centre of gravity, A10: Spectral SD, A11: Mean HNR, A12: Jitter, A13: Percentage of unvoiced segments, A14: Shimmer — see Appendix 2 for description of each of these properties) and the behavioural matrices for voices. Each circle shows the corrrelation for one participant, and the red full circle shows the mean across participants. Each representation matrix consisted of a 24-by-24 matrix in which each entry was a stimulus. Matrices describing low-level acoustic properties A1, A2, A3, A4, A10, A11, A12, A13, and A14 correlated with the behavioural matrices for voices significantly above zero (all mean $r > .10$; all $z > 5.37$, all $p < .001$). The last column in grey shows the correlation of the behavioural matrix for faces and the behavioural matrix for voices for each participant. The mean correlation across participants was $r = .48$ ($z = 5.84$, $p < .001$). Please note that these correlations were computed still using 24-by-24 matrices, but in the case of these correlation across modalities, there is not a perfect correspondence between the entries on face and voice matrices, given that the identity of the faces are not the same as the identities of the voices. Therefore, we arbitrarily matched the identities of faces with the identities of voices. Nevertheless, we presented these correlations here for completeness, and because it is important to compare these correlations to the partial correlations in panel B.

Panel B shows the partial correlations between the representation matrices for faces and the representation matrices for voices, while controlling for each of the acoustic properties. Each circle shows the partial corrrelation for one participant, and the red full circle shows the mean across participants. All partial correlations were still high, even after controlling for the variance of the

acoustic properties of the images (all mean $r > .39$; all $z = 5.84$, all $p < .001$). Please note again that, in this case, we were using partial correlations of two 24-by-24 behavioural matrices in which the entries did not match entirely, i.e., they matched on the emotion of the stimulus but not on their identity.

APPENDIX 6: Correlation of all low-level representation matrices

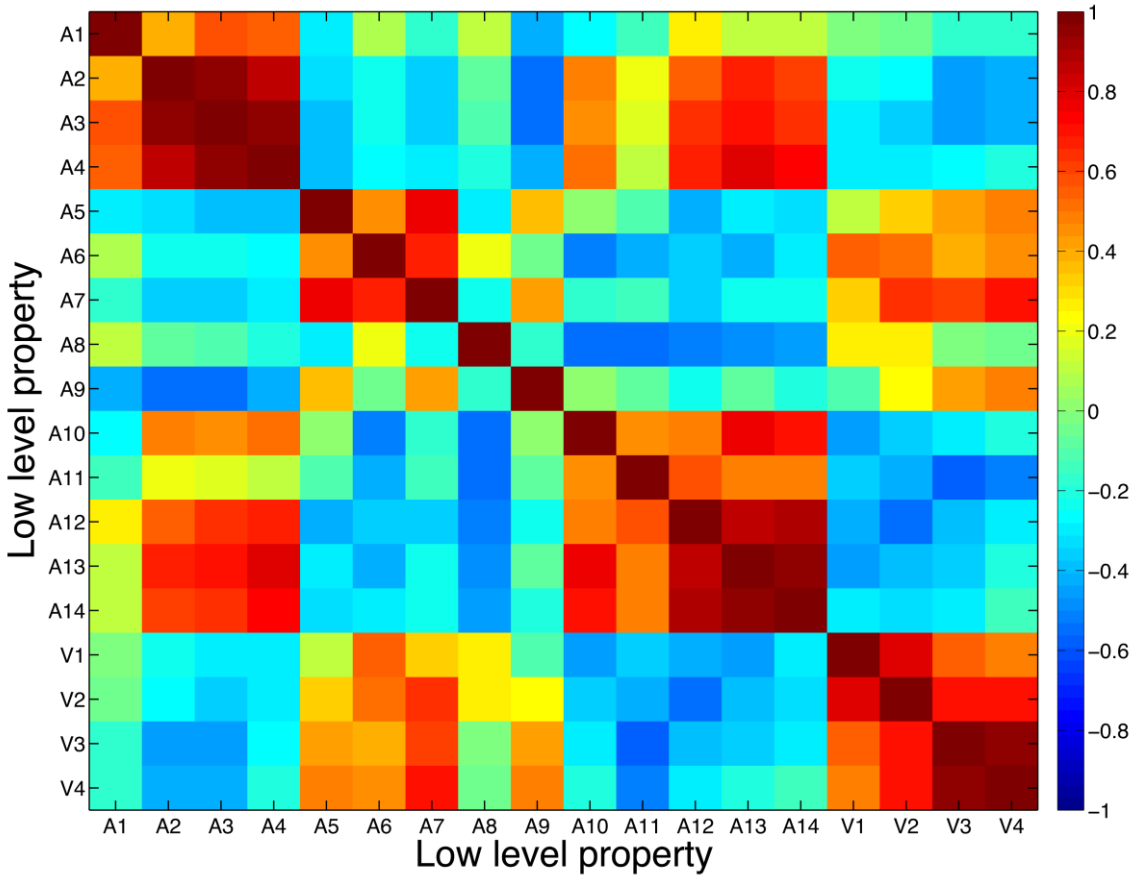


***Figure A6-1*: Correlations between representation matrices of low-level properties.** Each entry to this matrix is a representation matrix of an acoustic (A) property of the emotional voices, or of a visual (V) property of the emotional faces. All representation matrices for each property are the mean of the representation matrices for all identities of the stimuli. A1: Total duration, A2: Amplitude SD, A3: Mean intensity, A4: Number of amplitude onsets, A5: F0 minimum, A6: F0 maximum, A7: F0 mean, A8: F0 SD, A9: Spectral centre of gravity, A10: Spectral SD, A11: Mean HNR, A12: Jitter, A13: Percentage of unvoiced segments, A14: Shimmer, V1: Full images, V2: Shape-Free images, V3: Shape-49, V4: Shape-49-Procrustes. See main text for details of how these representation matrices were computed.

APPENDIX 7: Controlling for multiple low-level properties of the stimuli

For the emotional faces, we conducted multiple regression for each participant, in which the outcome was the behavioural representation matrix for faces, and the predictors were the four representation matrices of visual properties of the faces (we also conducted separate analyses combining just some of the predictors). After removing the variance from the visual properties, we correlated the remaining residuals with the behavioural representation matrix for voices. The results of these analyses are in Figure A7-1, which shows that even when accounting for multiple visual properties of the stimuli, the correlations across faces and voices did not substantially decrease.

Similarly, for emotional voices, we conducted multiple regression for each participant, in which the outcome was the behavioural representation matrix for voices, and the predictors were the 14 representation matrices of acoustic properties of the faces (we also conducted separate analyses combining just some of the predictors). After removing the variance from the acoustic properties, we correlated the remaining residuals with the behavioural representation matrix for faces. The results of these analyses are in Figure A7-2, which shows that even when accounting for multiple acoustic properties of the stimuli, the correlations across faces and voices did not substantially decrease.

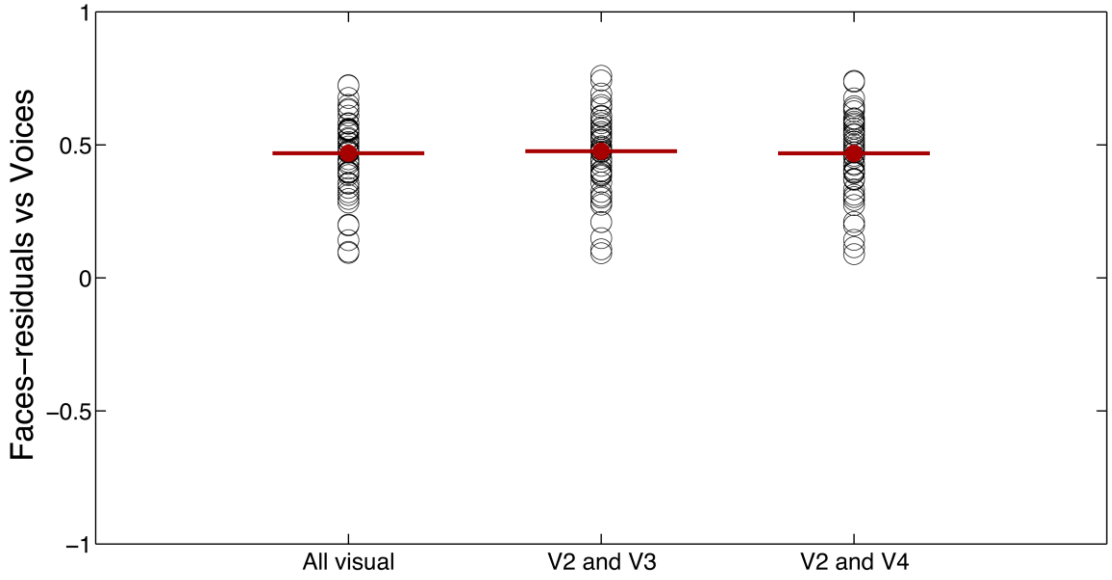


***Figure A7-1*: Controlling for multiple low-level visual properties of the faces using multiple regression.** The multiple regressions were done for each participant, and we conducted three separate analyses: the first analysis (All visual) regressed out all four representation matrices of visual properties, the second analysis (V2 and V3) regressed out the representation matrix based on shape-free information (Shape-Free images) and the one based on vectors with coordinates (Shape-49), and the third analysis (V2 and V4) regressed out the representation matrix based on shape-free information (Shape-Free images) and the one based on Procrustes distances (Shape-49-Procrustes). All these analyses were performed on non-averaged 24-by-24 matrices. After removing the variance accounted for by these visual properties, we correlated the residuals with the behavioural matrices for voices. Each circle shows the individual correlations, and the red filled circles show the mean correlations across participants. It is clear that even after removing the variance of multiple visual properties, there was not a substantial decrease of the correlations between representation matrices for emotional faces and emotional voices (all mean $r > .47$; all $z = 5.84$, all $p < .001$). For comparison, the mean correlation between the matrices for faces and the matrices for voices (without any regressions) was $r = .48$ ($z = 5.84$, $p < .001$). Please note that in this case we correlated two 24-by-24 behavioural matrices (Faces-residuals *versus* Voices) in which the entries did not match entirely, i.e., they matched on the emotion of the stimulus but not on their identity.

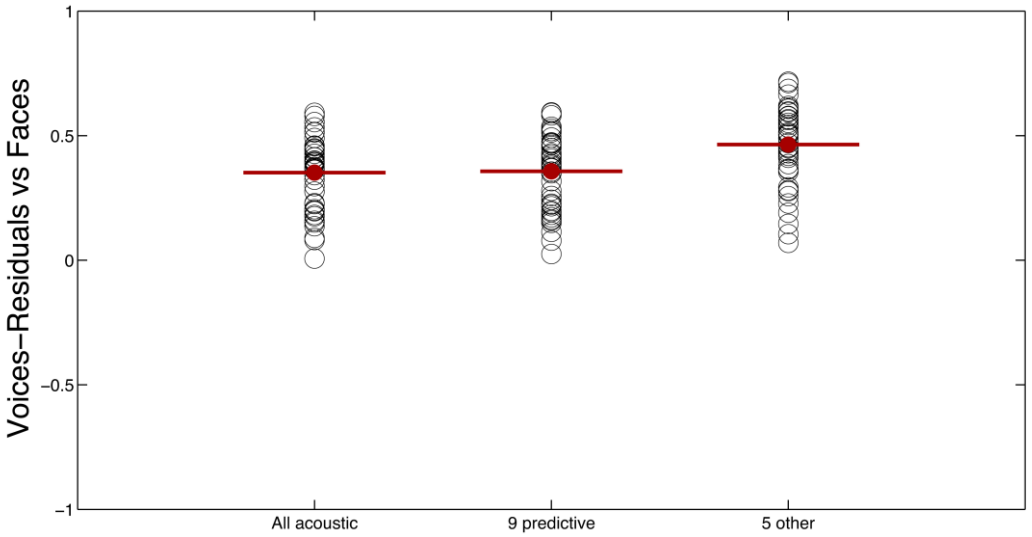


***Figure A7-2*: Controlling for multiple low-level acoustic properties of the voices using multiple regression.** The multiple regressions were done for each participant, and we conducted three separate analyses: the first analysis (All acoustic) regressed out all 14 representation matrices of acoustic properties, the second analysis (9 predictive) regressed out the representation matrices that had been shown to significantly predict behaviour (A1, A2, A3, A4, A10, A11, A12, A13), and the third analysis (5 other) regressed out the representation matrices that had been shown to not predict behaviour (A5, A6, A7, A8, A9). All these analyses were performed on non-averaged 24-by-24 matrices. After removing the variance accounted for by these acoustic properties, we correlated the residuals with the behavioural matrices for faces. Each circle shows the individual correlations, and the red filled circles show the mean correlations across participants. After regressing out all 14 acoustic cues (All acoustic), the mean correlation between the residuals of the voice matrices and the face matrices was $r = .35$; $z = 5.84$, $p < .001$. For the second analysis (9 predictive), the mean correlation was $r = .36$; $z = 5.84$, $p < .001$. For the third analysis (5 other), the mean correlation was $r = .46$; $z = 5.84$, $p < .001$. For comparison, the mean correlation between the matrices for faces and the matrices for voices (without any regressions) was $r = .48$ ($z = 5.84$, $p < .001$). These results show that the correlations for the first two analyses decreased slightly after removing the variance of multiple acoustic properties. However, most of the variance shared between the representation matrices for emotional faces and emotional voices was not accounted for by these acoustic properties. Please note that in this case we correlated two 24-by-24 behavioural matrices (Voices-residuals *versus* Faces) in which the entries did not match entirely, i.e., they matched on the emotion of the stimulus but not on their identity.