Social Cognitive and Affective Neuroscience, 2018, 1–13

doi: 10.1093/scan/nsx142
Advance Access Publication Date: 23 November 2017
Original article

Is the voice an auditory face? An ALE meta-analysis comparing vocal and facial emotion processing

Annett Schirmer[1,2,3]

[1]Department of Psychology, [2]Brain and Mind Institute, The Chinese University of Hong Kong, Shatin, Hong Kong, and [3]Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig 04103, Germany

Correspondence should be addressed to Annett Schirmer, Department of Psychology, The Chinese University of Hong Kong, Shatin, Sino Building, NT, Hong Kong. E-mail: schirmer@cuhk.edu.hk

Abstract

This meta-analysis compares the brain structures and mechanisms involved in facial and vocal emotion recognition. Neuroimaging studies contrasting emotional with neutral (face: $N = 76$, voice: $N = 34$) and explicit with implicit emotion processing (face: $N = 27$, voice: $N = 20$) were collected to shed light on stimulus and goal-driven mechanisms, respectively. Activation likelihood estimations were conducted on the full data sets for the separate modalities and on reduced, modality-matched data sets for modality comparison. Stimulus-driven emotion processing engaged large networks with significant modality differences in the superior temporal (voice-specific) and the medial temporal (face-specific) cortex. Goal-driven processing was associated with only a small cluster in the dorsomedial prefrontal cortex for voices but not faces. Neither stimulus- nor goal-driven processing showed significant modality overlap. Together, these findings suggest that stimulus-driven processes shape activity in the social brain more powerfully than goal-driven processes in both the visual and the auditory domains. Yet, whereas faces emphasize subcortical emotional and mnemonic mechanisms, voices emphasize cortical mechanisms associated with perception and effortful stimulus evaluation (e.g. via subvocalization). These differences may be due to sensory stimulus properties and highlight the need for a modality-specific perspective when modeling emotion processing in the brain.

Key words: fMRI; nonverbal; affective; visual; auditory; social; PET

Introduction

> [. . .] an *innate feeling* must have told him that the pretended crying of his nurse expressed grief . . . (Darwin, 1872, concluding remarks, italics added for emphasis)

In his book *The expression of emotions in man and animals*, Darwin recognized emotions as nonprivate experiences and described their characteristic displays. Whereas Darwin could only speculate about the 'innate feeling' by which these displays are perceived and understood, modern science helped characterize relevant mental and neural processes. Efforts to date, however, have overly focused on the processing of facial emotions. Moreover, insights from the face have shaped inquiry into other expressive channels and became to dominate our thinking about social perception more generally (Belin *et al.*, 2011, 2000; Schirmer and Adolphs, 2017). Taking issue with this situation, this article compares and contrasts the brain structures and mechanisms underpinning face perception with those underpinning voice perception.

Why emotion processing may compare for faces and voices

Constraints associated with nonverbal signaling as well as insights into how nonverbal signals are processed support the idea of channel similarities.

Received: 19 September 2017; **Revised:** 16 November 2017; **Accepted:** 19 November 2017

© The Author (2017). Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

For example, facial and vocal expressions are linked physically. If provoked by an emotional event, they emerge in a temporally synchronized manner (Schirmer *et al.*, 2016b). Additionally, their effectors overlap such that changes in one channel necessitate changes in another channel. The mouth forms part of many facial emotions and serves as a resonance body for the voice. Thus, facial expressions changing the shape of the mouth alter vocal acoustics like roughness or loudness. Likewise, the nature of vocal expression inadvertently impacts the look of the face.

Neuroimaging evidence indicates that both the visual and the auditory system are partially specialized for their respective human signal in higher-order association cortex. Specifically, aspects of the bilateral fusiform gyrus, an area along the visual ventral stream, are more important for the processing of faces than the processing of other visual objects. In the context of functional magnetic resonance imaging (fMRI), this so-called fusiform face area (FFA) activates more strongly when participants see human faces as compared to other body parts, non-human faces, cars, or houses (Kanwisher *et al.*, 1997; Halgren *et al.*, 2000; Bernstein and Yovel, 2015). Moreover, damage to the FFA is associated with a deficit in recognizing familiar faces (Gainotti and Marra, 2011). By analogy, vocalizations more than other sounds activate aspects of superior temporal gyrus and sulcus (STG and STS, respectively) along an auditory ventral stream (Belin *et al.*, 2000, 2011; Sammler *et al.*, 2015). These areas are now referred to as temporal voice areas, and damage to them has occasionally been reported to compromise the recognition of familiar voices (Perrodin *et al.*, 2015).

Another point in favor of modality similarities concerns the mental and brain responses to multimodal stimulation. Face perception benefits from the simultaneous presentation of voices and vice versa. Compared to unimodal stimuli, congruent multimodal stimuli yield greater processing speed and accuracy and stronger activation in a number of brain regions involved in auditory, visual or amodal processing (see Klasen *et al.*, 2012, for a review). Moreover, some of the brain regions that respond to multimodal stimulation contain neurons representing input from more than one modality. For example, there are neurons in the medial temporal lobe that respond to both the face and voice of a familiar person (Quiroga *et al.*, 2009). In the superior temporal cortex, especially its posterior part, some neurons respond to both the sound and sight of an action and, more specifically, to the facial and vocal expression of an emotion (Barraclough *et al.*, 2005; Watson *et al.*, 2014).

Last, structural and functional overlap in the processing of faces and voices may result from a shared processing purpose—that is the representation of another's emotion. One may speculate that computations of emotional significance as well as the kind of cognitive operations associated with expression categorization compare for faces and voices. In line with this, the discrimination between personally relevant and irrelevant stimuli has been attributed to the amygdala (Sander, 2012), a collection of gray matter situated in the anterior aspect of the medial temporal lobe. Additionally, frontal regions including the dorsomedial prefrontal cortex (dmPFC) or the inferior frontal gyrus (IFG) have been linked with the judgment of emotions in both faces and voices (Hensel *et al.*, 2015; Dricu and Frühholz, 2016; Molenberghs *et al.*, 2016). Together, this evidence has inspired the notion that the voice is an 'auditory face', recruiting comparable brain processes and mechanisms (Belin *et al.*, 2011).

Why emotion processing may differ for faces and voices

Although there are reasons to assume that findings on face processing generalize to voice processing, caution is warranted (King and Nelken, 2009; Schirmer and Adolphs, 2017). After all, visual stimuli differ substantially from auditory stimuli and one may wager that these differences translate into how images and sounds impress on the brain. Visual impressions arise from light that falls on photoreceptors in the retina. Although vision can be dynamic, it does not have to be. In other words, we can see and recognize a dynamically moving face as well as we can see and recognize a still face. Indeed, vision can be reduced to a mere snapshot (Sternberg, 1966). In contrast, auditory impressions arise from mechanical waves that displace hair cells in the inner ear. Auditory stimulation and perception are inherently dynamic. Sounds are recognized incrementally as their acoustics unfold in time (Schirmer *et al.*, 2016a, 2016c). Thus, it takes more than just a snapshot to tell whether a voice is familiar, female, or happy.

Modality differences in the stimulus are matched by differences in associated sensory systems. The eyes have a sharpest point of vision, which needs to be directed at a stimulus for that stimulus to be properly perceived. Thus, visual orientation and attention are critical. The eyes project information from left and right hemifields to opposite hemispheres via only a few subcortical relays. Moreover, the representation of basic visual features like orientation, color, location or movement depends on cortical computations (King and Nelken, 2009).

The ears are shaped to propagate sound waves to hair cells in the cochlea and head orientation relative to a sound source modulates sound perception. However, there is no sharpest point of hearing and sounds continuously reach us even when we are asleep (Strauss *et al.*, 2015). Information transmission from the ears to the brain is lateralized, but not completely. Thus, both hemispheres receive left and right projections for easy integration. Moreover, a number of nuclei along the auditory pathway deliver highly processed information to primary auditory cortex, which seems more interested in complex stimuli than its visual counterpart (King and Nelken, 2009).

Together these and other differences cast doubt on the popular 'one-modality-fits-all' approach to nonverbal perception. In fact, one may wish to directly compare facial and vocal processing as to ascertain the degree to which both overlap and differ.

Comparing facial and vocal emotion processing

A modality comparison may be achieved in two ways. First, one might explore face and voice perception within one study. To date, this option has been rarely pursued (Phillips *et al.*, 1998; Aubé *et al.*, 2015) and results are limit by the specific emotion types, paradigms and other methodological idiosyncrasies. Second, one might conduct a meta-analysis of published findings from separate face and voice perception studies. This approach is advantageous because it incorporates methodological variance across studies and responses from many hundreds of subjects (Eickhoff *et al.*, 2009; Haidich, 2010). As such, findings are fairly robust and generalizable. Adopting a meta-analytic approach, Fusar-Poli and colleagues (2009) explored the brain basis of facial emotion processing. They subjected 105 studies contrasting emotional and neutral faces with a fixation baseline to an activation likelihood estimation (ALE). The ALE method computes brain activation probabilities based on peak coordinates and is popular for the synthesis of neuroimaging data. In the hands of Fusar-Poli and colleagues, it revealed a number of regions including frontal lobe, amygdala, parahippocampal gyrus, fusiform gyrus and middle temporal gyrus, with effects being largely bilateral. These results overlap with a similar approach taken by Sabatinelli and colleagues (2011) who explored contrasts of emotional with neutral faces collected from 100 studies. Additionally, they converge, albeit to a lesser degree, with a report by Del Casale and colleagues (2017) who contrasted empathetic face processing with a range of control conditions across 23 studies.

Looking at the voice, Frühholz and colleagues subjected temporal lobe coordinates from 27 studies to several ALE analyses exploring study–subgroup effects associated with verbal and nonverbal, positive and negative, as well as explicit and implicit processing (2013a). Overall, they identified an effect for emotional voices in primary auditory and surrounding superior temporal cortex. More recently, Dricu and Frühholz examined emotion perception from the voice in conjunction with that from the face and the body. They collated 98 studies employing an explicit emotion task, 32 studies employing a passive paradigm and 46 studies employing an implicit task (e.g. gender discrimination). The authors report a large number of sensory and 'higher-order mind-reading' areas that are more readily engaged during explicit emotion processing as compared with passive and implicit emotion processing. Moreover, they show that the explicit evaluation of faces and voices appears to overlap in the posterior STS (pSTS), the IFG and the dmPFC.

Although the meta-analyses reviewed here suggest overlap and dissociation between facial and vocal processing, insights are limited in a number of ways. First, most published attempts have focused on either face or voice rather than on comparing the two. Moreover, comparing their results in retrospect is complicated by methodological differences including brain coverage, study numbers, the kinds of fMRI contrasts or paradigms and the exact statistical approach (e.g. permutations, thresholding). Second, the one study that included both faces and voices has a number of limitations (Dricu and Frühholz, 2016). For one, it paid limited attention to the actual contrasts that entered statistical analysis. As an example, the analysis of explicit emotion evaluation entailed contrasts of emotional *vs* neutral expressions, emotional expressions *vs* shapes or explicit *vs* implicit tasks and these contrasts differed between faces and voices. As such, it is unclear what mental processes were being isolated and one must worry about modality confounds. Likewise problematic are drastic differences in study numbers for key comparisons. For example, the comparison of explicit emotion perception from faces and voices involved 75 and 18 studies, respectively, making it more likely for significant clusters to emerge for the former relative to the latter condition. Last, although first-level individual analyses were done conservatively, second-order analyses of task and modality effects were done liberally using an uncorrected *P*-value.

Thus, it is worthwhile to pursue the face–voice comparison afresh and to contrast both modalities in a more rigorous manner. Towards this end, two collections of voice studies, of which one included 34 data sets contrasting emotional with neutral stimuli and of which the other included 20 data sets contrasting explicit with implicit or passive processing, were matched with two corresponding collections of face studies. Inspired by existing multi-stage processing models (Haxby *et al.*, 2000; Frühholz and Grandjean, 2013 b; Bernstein and Yovel, 2015; Schirmer and Adolphs, 2017), emotion contrasts (i.e. emotional > neutral) were aimed at identifying basic processing in voice and face regions supporting the representation of emotional significance irrespective of task and attention in a stimulus-driven manner. Task contrasts (i.e. explicit > implicit) were aimed at identifying goal-driven processing in voice and face regions supporting the explicit evaluation or categorization of an expression. Study collections were subjected to a series of ALE analyses that served to statistically compare activation patterns between modalities.

Hypotheses were derived from the evidence reviewed above as well as from extant models of face (Haxby *et al.*, 2000; Atkinson and Adolphs, 2011; Bernstein and Yovel, 2015; Schirmer and Adolphs, 2017) and voice perception (Schirmer and Kotz, 2006; Frühholz and Grandjean, 2013a, 2013b; Schirmer and Adolphs, 2017). Modality-specific effects were expected for the emotion contrast analysis and to a lesser degree for the task contrast analysis. Modality-independent effects were expected for the task contrast analysis and to a lesser degree for the emotion contrast analysis. Modality-specific effects should show in primary and secondary sensory regions, whereas modality-independent effects should show in emotion hot spots like the amygdala and areas presumed to support 'mind reading' like the dmPFC, IFG and pSTS.

Materials and methods

Data set

The studies for this meta-analysis were identified based on previous meta-analyses (Sabatinelli *et al.*, 2011; Frühholz and Grandjean, 2013a; Dricu and Frühholz, 2016) and via two searches on Scopus. The first search was aimed at vocal expression research and used '[fMRI OR PET] AND [emotion* OR affect*] AND [vocal OR voice* OR prosody]' as general search terms. The second search was aimed at facial expression research and used '[fMRI OR PET] AND [emotion* OR affect*] AND [face* OR facial]' as general search terms. Searches were completed by 26 July 2017. Although the focus of this project was unimodal perception, multimodal studies were considered as long as they included relevant unimodal conditions. Specifically, study results were scrutinized for contrasts of emotional with neutral expressions and/or explicit with implicit/passive expression processing. The details of identified studies are presented in the Supplementary Material. Although the initial search revealed a large number of voice studies, many of them had to be excluded because they failed to report the simple contrasts that were of relevance here, because vocal and facial emotions were not examined unimodally but were always combined with other types of non-verbal or verbal emotion signaling (e.g. a negative word spoken in a negative tone), because a whole-brain contrast was unavailable and because only special populations such as patients or children below the age of 10 were tested. In all, we collated data from 34 voice and 76 face studies contrasting emotional with neutral expressions as well as 20 voice and 27 face studies contrasting an explicit with an implicit or a passive task.

Because there were more face than voice studies, the former were carefully selected to match the latter in number and key methodological aspects similar to what is done in the study of special populations such as neurological patients. For the analysis of emotion contrasts, we selected, for each voice study, a matching face study using the same task and emotion(s). Voice studies without a modality match were excluded from the analysis. If a voice study had multiple matching face studies, we considered additional criteria such as the number of participants, the sex ratio and the year of publication and chose the one that minimized differences. The year of publication was considered relevant in connection with changing methodological standards (e.g. number of participants, scanning protocol and statistics) that occurred over the past 2.5 decades. For the analysis of task contrasts, eight face studies were discarded as to reduce task and emotion differences with the voice studies. More specifically, the proportion of studies using a gender recognition task in the implicit condition as well as expressions of fear and disgust was significantly greater in the face as compared with the voice data set. Hence, studies were selected as to minimize these differences. However, some differences remained as further matching would have necessitated the removal of voice studies and compromised the power of the ALE analysis (Eickhoff *et al.*, 2016b). The final matched data sets comprised 26 face and 26 voice studies for the analysis of emotion effects and 20 face and 20 voice studies for the analysis of task effects. These numbers exceeded or adhered with the minimum requirement for an ALE analysis (Eickhoff *et al.*, 2016 b). Relevant study details of the matched data sets are listed in Table 1. The Supplementary Material present an alternative modality comparison. Here, all voice studies were compared with an equal number of randomly drawn face studies across 100 permutations. Results were subjected to a conjunction analysis that identified voxels that were significant in more than 70/100 ALE maps. Findings are fairly similar to the matched approach that is reported below. However, possible methodological confounds as well as issues associated with differences in variance between the smaller/fixed voice and the larger/varying face study sets and with setting the conjunction threshold make the random approach less preferable.

Activations reported in Talairach space were converted to MNI space using the Brett transform in GingerALE, a Brainmap tool (Eickhoff *et al.*, 2009, 2011, 2012; Turkeltaub *et al.*, 2012; https://www.brainmap.org/ale/). The resulting coordinates included eight outliers that were corrected by changing them to the nearest brain voxel.

Data analysis

The data was subjected to GingerALE. First, each data set (face: emo-neu, voice: emo-neu, face: explicit-implicit, voice: explicit-implicit) was analyzed separately. Face and voice ALE maps were thresholded at the cluster level using 5000 permutations with a *P*-value of 0.05 and a cluster-forming FDRpN-value of 0.01 (conservative, makes no assumptions about how the data are correlated). Then, a modality-combined data set (emo-neu, explicit-implicit) was created and subjected to an identical computation. The ALE maps resulting from these initial steps were subjected to a subtraction/conjunction analysis (Eickhoff *et al.*, 2011) with 5000 permutations and results were thresholded with an FDRpID-value of 0.05 (assumes independence or positive dependence). The minimum cluster size was set to 50 mm^3.

Results

Emotion contrast

An individual ALE analysis of the full voice data set ($N = 34$) revealed four clusters (Table 2, Figure 1). The largest cluster was located in the right STG and from the middle of BA22 extended medially into primary auditory cortex and insula, anteriorly along the superior temporal sulcus and ventrally into the middle temporal gyrus. Second and third clusters were located in the middle and anterior aspect of left STG, respectively, and were fairly circumscribed within BA22. The last and smallest cluster was located in the left IFG in BA45. These results were well matched by those obtained from a reduced data set ($N = 26$) that entered the modality comparison described further below (Table 3, Figure 1).

An individual ALE analysis of the full face data set ($N = 76$) revealed seven clusters (Table 2, Figure 1). The largest cluster centered on the left amygdala and from there extended into BA34 and 28 of the parahippocampal gyrus, the globus pallidus, the putamen and the subcallosal gyrus. A similar but smaller cluster centered on the right globus pallidus. The remaining five clusters peaked in BA37 of the right fusiform gyrus, BA18 of the left and right middle occipital gyrus, the anterior cerebellum and BA45 of the left IFG, respectively. Only the two largest clusters from the full data set survived when analyzing a reduced data set ($N = 26$) matched with the voice data set for modality comparison (Table 3, Figure 1). The larger cluster centered on the left and the smaller on the right parahippocampal gyrus. Their peaks were located in BA34, amygdala and the subcallosal gyrus of the left hemisphere and BA34, BA28 and putamen of the right hemisphere.

A subtraction analysis was performed on the matched data sets to identify differences between vocal and facial emotion processing. This analysis revealed a greater activation likelihood for voices relative to faces in the STG clusters reported above; the IFG cluster failed to reach significance. Faces were more likely than voices to activate the left and right parahippocampal region. A conjunction analysis revealed no significant overlap.

Task contrast

Analysis of the voice data ($N = 20$) revealed one cluster in the dmPFC (Table 4, Figure 1). Its peak centered on BA6. Analysis of the full face data ($N = 27$) and the face data reduced for modality comparison ($N = 20$) was non-significant. A subtraction of face from voice data highlighted the dmPFC in BA6, while the reverse subtraction was non-significant. A conjunction analysis revealed no significant modality overlap (Table 5).

Discussion

This study compared the processing of emotions in voices and faces. Separate meta-analyses conducted on emotion and task contrasts both confirmed and contradicted the hypotheses derived from previous work. As expected, vocal emotion contrasts produced the largest effect in sensory regions in the bilateral temporal cortex. For the face, however, the largest effect occurred subcortically in the medial temporal lobe. Additionally, contrasts of explicit minus implicit processing revealed the dmPFC for voices, but no significant clusters for faces. The following paragraphs discuss these findings in more detail, relate them to current thinking about nonverbal emotion processing and develop an agenda for future research.

Stimulus-driven processing of emotional voices and faces

Voices. The contrast of emotional *vs* neutral expressions may, arguably, reveal largely stimulus-driven processes underpinning the representation of emotion. Its analysis for vocal studies ($N = 34$) revealed clusters in the STG of both right and left hemispheres with cluster sizes being right lateralized. The larger right STG cluster extended medially into primary auditory and multisensory association cortex in Heschl’s gyrus and the posterior insula, respectively (Nieuwenhuys, 2012). Its anterior and ventral extension reached STS and MTG and thus

Table 1. Statistical comparison of key characteristics for the matched voice and face data sets

	Face: Mean (SD)	Voice: Mean (SD)	Statistic
Face vs voice studies included in the emotion contrast			
Number of participants	19.3 (8.88)	17.72 (6.19)	*F*(1, 50)=0.3, *P*=0.583
Sex ratio (F/F+M)	0.51 (0.26)	0.45 (0.23)	*F*(1, 50)=1.34, *P*=0.253
Year of publication	2007.7 (4.82)	2009.24 (5.05)	*F*(1, 50)=1.02, *P*=0.317
Task (count data)[a]			
Explicit & implicit			
Emotional (pos & neg)	1	1	
Explicit			
Angry	4	4	
Emotional (pos & neg)	6	6	
Fearful	1	1	
Happy	1	1	
Implicit			
Angry	7	7	
Disgusted	1	1	
Emotional (pos & neg)	3	3	
Fearful	1	1	
Happy	1	1	
Negative	1	1	
Positive	1	1	
Sad	2	2	
Passive			
Emotional (pos & neg)	1	1	
Pained	1	1	
Positive	1	1	
Face vs voice studies included in the task contrast			
Number of participants	17.75 (7.89)	16.91 (9.97)	*F*(1, 38)=0.00, *P*=.959
Sex ratio (F/F+M)	0.51 (0.17)	0.48 (0.18)	*F*(1, 38)=0.21, *P*=.647
Year of publication	2007.8 (4.5)	2006.68 (5.6)	*F*(1, 38)=0.21, *P*=.649
Baseline task (count data)			
Acoustics	0	2	
Age	3	1	
Gender	9	2	
Identity	4	1	
Linguistic	0	10	
Passive	1	3	
Plausibility	0	1	
Stimulus Type	1	1	
Sensory Motor	1	1	
Shape	1	0	$\chi^2(9) = 22.21$, P = 0.008
Stimulus emotion (count data)			
Angry	12	10	
Doubting	0	1	
Emotional	2	6	
Fearful	6	1	
Happy	14	16	
Ironic	0	1	
Negative	1	2	
Obvious	0	1	
Pleasure	0	1	
Sad	7	12	
Surprised	2	1	$\chi^2(10) = 11.28$, P = 0.336

[a]Please note that some studies examined more than one task and emotion.

traditional voice areas (Belin *et al.*, 2011). Activity in the left STG was more circumscribed and centered more anteriorly in the STS. Together, these findings show that emotional voices excite both early and late auditory processing more strongly than do neutral voices. Effects in primary auditory cortex imply an influence on early acoustic representations, whereas effects in posterior insula and STS/MTG imply an influence on late voice and multimodal representations. Moreover, the overall right hemisphere bias accords with the more general pattern observed for nonverbal signals in the literature (Brauer *et al.*, 2016; Schirmer *et al.*, 2016b; Schirmer and Adolphs, 2017).

In addition to the temporal clusters, emotional voices activated a smaller cluster in the left IFG overlapping with Broca's area. A white matter tract called arcuate fasciculus connects posterior STG with Broca's area and is believed to support the translation of acoustical into articulatory representations

Table 2. Results of the emotion contrast analysis for the full data sets

Cluster ID	Size	Anatomical structure	BA
Voice (*N* = 34)			
1	1864	R superior temporal gyrus (centered at 52.4, −19.7, 2.8; 16 studies)	
1.1	760	R superior temporal gyrus	22
1.2	368	R superior temporal gyrus	41
1.3	264	R superior temporal gyrus	13
1.4	184	R insula	13
1.5	112	R transverse temporal gyrus	41
1.6	72	R superior temporal gyrus	*
2	312	L superior temporal gyrus (centered at − 50.8, −11.4, −3; 4 studies)	
2.1	224	L superior temporal gyrus	22
2.2	56	L superior temporal gyrus	*
3	144	L superior temporal gyrus (centered at − 58, −23.2, 1.1; 1 study)	
3.1	64	L superior temporal gyrus	22
4	80	L inferior frontal gyrus (centered at − 47.8, 24.6, 3; 2 studies)	
4.1	56	L inferior frontal gyrus	45
Face (*N* = 76)			
1	2808	L amygdala (centered at − 21.5, −3.2, −15.7; 31 studies)	
1.1	920	L amygdala	
1.2	472	L parahippocampal gyrus	34
1.3	368	L putamen	
1.4	312	L globus pallidus	
1.5	264	L parahippocampal gyrus	28
1.6	184	L globus pallidus	
1.7	136	L subcallosal gyrus	34
1.8	136	L amygdala	
2	1592	R globus pallidus (centered at 22.3, −4.2, −15.6; 13 studies)	
2.1	544	R amygdala	
2.2	304	R parahippocampal gyrus	34
2.3	224	R globus pallidus	
2.4	200	R globus pallidus	
2.5	120	R parahippocampal gyrus	28
2.6	104	R putamen	
2.7	80	R amygdala	
3	456	R fusiform gyrus (centered at 42.6, −48.3, −19.2; 7 studies)	
3.1	256	R fusiform gyrs	37
3.2	160	R cerebellum, anterior lobe, culmen	
4	424	L middle occipital gyrus (centered at − 28.7, −91.8, 2.9; 7 studies)	
4.1	248	L middle occipital gyrus	18
4.2	120	L inferior occiptal gyrus	18
4.3	56	L middle occipital gyrus	*
5	280	R middle occipital gyrus (centered at 33.5, -89.9, 5.3; 6 studies)	
5.1	216	R middle occipital gyrus	18
6	128	L cerebellum (centered at − 42.7, −51.8, −22.5; 2 studies)	
6.1	112	L Cerebellum, anterior lobe, culmen	
7	112	L inferior frontal gyrus (centered at − 45.4, 21.7, −2.1; 1 study)	
7.1	56	L inferior frontal gyrus	45

(Hickok and Poeppel, 2007). Thus, one may speculate that the effects observed here relate to the articulatory demands of representing vocal and in particular speech sounds. Yet, separate post-hoc analyses of verbal (*N* = 23) and nonverbal (*N* = 11) studies revealed no IFG effect. Moreover, the IFG was also conspicuously silent when looking at explicit (*N* = 10) and implicit/passive studies (*N* = 19). Because this may be due to the small sample and because left IFG emerged for explicit processing previously from a larger mixed-modality study (Dricu and Frühholz, 2016), vocal and facial emotion contrasts derived from explicit paradigms were combined (*N* = 35) for another exploratory analysis. Again, no IFG activation emerged suggesting that IFG function cannot be neatly categorized as previously proposed (Schirmer and Kotz, 2006; Wildgruber *et al.*, 2006; Frühholz and Grandjean, 2013b). Moreover, the potential involvement of this anatomically heterogeneous region in articulatory and/or explicit processing may be multiconditional depending, for example, on stimulus difficulty (e.g. subtlety, ambiguity; Kuhlmann *et al.*, 2016; Valk *et al.*, 2017) and the difficulty arising from stimulation context (e.g. emotionally incongruous *vs* congruous; Schirmer *et al.*, 2004).

Perhaps surprisingly, the amygdala did not emerge as a significant contributor to the perception of vocal emotions. Ten out of 34 articles mentioned the amygdala as being more active for emotional as compared with neutral voices. However, their activation site was inconsistent (two left, five bilateral, three right), the statistical threshold was typically reduced relative to a whole-brain analysis and the voxel numbers were typically small (< 15). Moreover, the stimuli consistently conveyed high-arousal states like anger (Sander *et al.*, 2005; Quadflieg *et al.*,

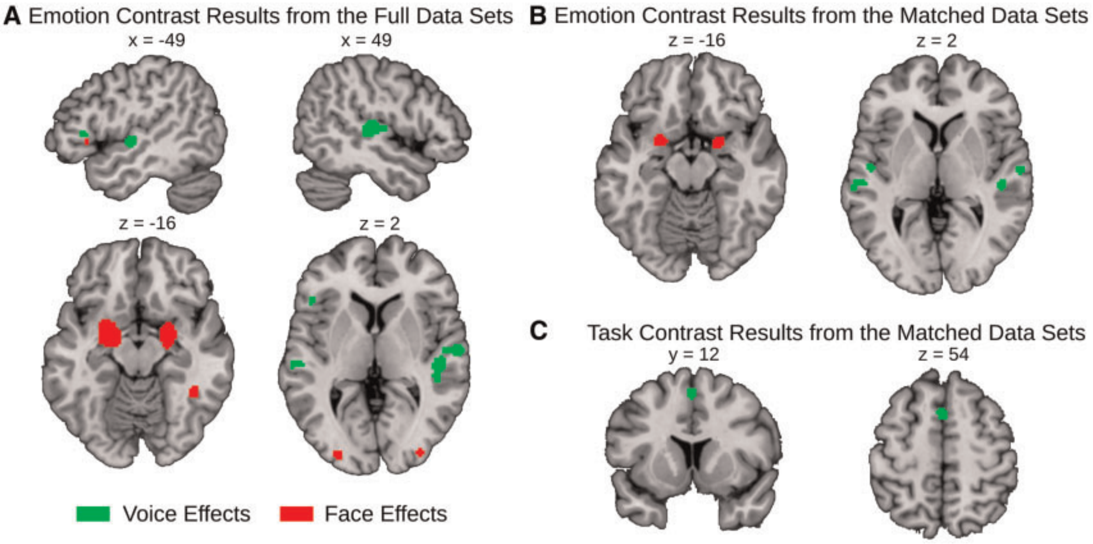
A Emotion Contrast Results from the Full Data Sets
x = -49
x = 49
z = -16
z = 2
B Emotion Contrast Results from the Matched Data Sets
z = -16
z = 2
C Task Contrast Results from the Matched Data Sets
y = 12
z = 54
🟩 Voice Effects
🟥 Face Effects

Fig. 1. ALE results. A, Emotion contrast results from the full face (N = 76) and voice (N = 34) data sets. Significant face clusters are indicated in red and significant voice clusters are indicated in green. B, Emotion contrast results from the modality comparison (N = 26). Areas of greater activation likelihood for faces than for voices are indicated in red. The reversed subtraction is presented in green. C, Task contrast results from the full face (N = 27) and voice (N = 20) data sets. There was only one significant cluster for voices and no significant cluster for faces. The modality subtraction analysis revealed the same voice and no face cluster.

2008; Ethofer *et al.*, 2009; Mothes-Lasch *et al.*, 2011; Frühholz *et al.*, 2012; Goerlich-Dobre *et al.*, 2014), fear (Phillips *et al.*, 1998; Morris *et al.*, 1999; Fecteau *et al.*, 2007) or sexual desire (Fecteau *et al.*, 2007). In line with this, patient data concerning the amygdala's role in vocal emotions are inconsistent. Whereas some authors find preserved emotion recognition following amygdala damage (Anderson and Phelps, 1998), others suggest impairments (Scott *et al.*, 1997; Frühholz *et al.*, 2015). Together, this indicates that amygdala contributions to voice perception may be relatively small, unreliable and state-specific.

Faces. Analysis of emotion contrasts for the full face data set (N = 76) revealed seven clusters. The two largest clusters were situated in the medial temporal lobe and comprised a range of structures including amygdala, entorhinal cortex of the parahippocampal gyrus (BA28 and 34), globus pallidus, putamen and subcallosal gyrus. Notably, the left hemisphere cluster was almost twice as large as the right hemisphere cluster. Five small clusters were identified in the right fusiform gyrus, bilateral visual association cortex in the occipital lobe, the left cerebellum and BA45 in the left IFG.

These results overlap with those reported in previous meta-analyses (Fusar-Poli *et al.*, 2009; Sabatinelli *et al.*, 2011; Del Casale *et al.*, 2017) and suggest a broad emotional enhancement of visual processing. This enhancement occurs most consistently within the amygdala, which presumably serves as a relevance detector (Sander *et al.*, 2003), supports identity processing (Rutishauser *et al.*, 2015) and facilitates emotional learning especially in the context of fear (Anderson and Phelps, 2001; Fox *et al.*, 2015). Structures adjacent to the amygdala are also strongly implicated but received less attention to date and do not feature in current models of facial emotion perception. Nevertheless, they may be relevant as suggested here and elsewhere. Specifically, the entorhinal cortex has been implicated in the memory for an event's context (Suh *et al.*, 2011; Aminoff *et al.*, 2013). Its role in emotion may be explained by the fact that emotion stimuli can be ambiguous and require context for disambiguation or that emotions serve as context for the processing of other stimuli (e.g. faces). Globus pallidus and putamen form part of the basal ganglia and are traditionally thought to support movement. However, a recent meta-analysis of neuroimaging studies implicated them in a range of processes that include, apart from movement, working memory and executive function, pain, reward and emotional judgments (Arsalidou *et al.*, 2013). Moreover, activations associated with emotional judgments overlapped most closely with those observed here indicating a convergence of results. Possibly the ventromedial basal ganglia support emotional judgments by linking an emotional expression with an emotional experience. Last, the subcallosal gyrus lies below the frontal aspect of the corpus callosum and shows high local connectivity with a range of structures including amygdala and entorhinal cortex. Individuals with depression show subcallosal volume reduction and benefit from the electrical stimulation of this region (Hamani *et al.*, 2011). Possibly then its integrity may be relevant for healthy emotion functioning including the perception of emotion signals.

Whereas emotions are traditionally conceived of as a function of the right hemisphere (Schirmer and Kotz, 2006), the present face results were left-laterlized. This pattern is difficult to compare with previous meta-analyses of non-clinical studies (Fusar-Poli *et al.*, 2009; Sabatinelli *et al.*, 2011; Del Casale *et al.*, 2017), which contrasted emotional expressions against a non-face baseline, mixed emotion with task contrasts and/or failed to report cluster sizes. However, a meta-analysis exploring emotional face perception in socially anxious and healthy individuals sheds light on lateralization (Binelli *et al.*, 2014). Increased emotional responding in the former relative to the latter group produced a left-lateralized parahippocampal activity similar to the one observed here. Thus, facial emotion processing may indeed be left-lateralized for a range of reasons including the verbalization of perceived emotions or an association with a right-lateralized movement.

Faces vs voices. In order to compare voice with face perception, we created two data sets that matched each other in size (N = 26), task and emotion. Individual analyses replicated the results from the full data set for voices but not for faces.

Table 3. Results of the modality comparison for the emotion contrast analysis (matched data set, $N = 26$)

Cluster ID	Size	Anatomical structure	BA
Voice			
1	456	R superior temporal gyrus (centered at 49.1, −22, 5.2; 4 studies)	
1.1	184	R superior temporal gyrus	13
1.2	88	R superior temporal gyrus	41
1.3	72	R superior temporal gyrus	22
2	312	L superior temporal gyrus (centered at −50.2,−11.2,−3; 4 studies)	
2.1	208	L superior temporal gyrus	22
3	144	L superior temporal gyrus (centered at −59.7,−24.1, 1.5; 1 study)	
3.1	80	L superior temporal gyrus	22
4	144	R superior temporal gyrus (centered at 62.1,−13.2, 1.9; 3 studies)	
4.1	120	R superior temporal gyrus	22
5	136	L Inferior frontal gyrus (centered at −47.9, 24.5, 3.3; 2 studies)	
5.1	96	L Inferior frontal gyrus	45
Face			
1	688	L Parahippocampal gyrus (centered at −22.2, 0.4,−17.3; 6 studies)	
1.1	336	L Parahippocampal gyrus	34
1.2	216	L Amygdala	
1.3	80	L Subcallosal gyrus	34
2	384	R Parahippocampal gyrus (centered at 20.8, 0.3, −16.7; 5 studies)	
2.1	136	R Parahippocampal gyrus	34
2.2	72	R Putamen	
2.3	64	R Parahippocampal gyrus	28
Voice > face			
1	488	R superior temporal gyrus (centered at 49.1, −22, 5.2; 3 studies)	
1.1	184	R superior temporal gyrus	13
1.2	88	R superior temporal gyrus	41
1.3	72	R superior temporal gyrus	22
2	280	L superior temporal gyrus (centered at −50.3, −11.7, −3.1; 4 studies)	
2.1	184	L superior temporal gyrus	22
3	144	L superior temporal gyrus (centered at −59.7, −24.4, 1.4; 1 study)	
3.1	80	L superior temporal gyrus	22
4	144	R superior temporal gyrus (centered at 62, −13.3, 1.8; 3 studies)	
4.1	120	R superior temporal gyrus	22
Face > voice			
1	400	L subcallosal gyrus (centered at −22.8, 2.4, −17.2; 5 studies)	
1.1	232	L parahippocampal gyrus	34
1.2	80	L subcallosal gyrus	34
2	336	R parahippocampal gyrus (centered at 20.8, 1, −16.7; 4 studies)	
2.1	136	R parahippocampal gyrus	34
2.2	72	R putamen	
Voice & face			
NS			

Specifically, in the case of faces, only the two mediotemporal clusters survived. Given the consistent voice effects and the fact that our data set size exceeded the minimum recommendation of 20 studies (Eickhoff *et al.*, 2016b), one may argue that the five unreplicated face clusters are unreliable and/or that they play a subordinate role in representing facial emotion. Indeed, existing models of face perception attribute non-emotional processes to occipital, fusiform and inferior frontal cortices. Occipital and fusiform cortices are associated with the early and later processing of face form, respectively, and inferior frontal cortex with that of dynamic facial features (Haxby *et al.*, 2000; Bernstein and Yovel, 2015). Their occasional emotion sensitivity may presuppose high emotional intensity and reflect secondary effects of attention or resource allocation.

The matched modality comparison revealed four bilateral superior temporal clusters that were more strongly activated by emotional voice as compared to face processing. Two of these clusters were located in the right and two in the left hemisphere each overlapping with the clusters obtained in the individual voice analysis. Two mediotemporal clusters located in left and right hemispheres were more strongly associated with face than voice perception and likewise overlapped with those obtained in the individual face analysis.

These results point to a clear dissociation between the auditory and the visual modality. Compared to visual, auditory emotion processing more readily engages primary and higher-order sensory regions. The reason for this is most likely that auditory signals arriving at the cortex are more processed and complex than their visual counterparts (King and Nelken, 2009). As such, they may readily represent basic aspects of emotion. For example, primary auditory cortex may be sensitive to affective cues of valence and/or arousal (Frühholz *et al.*, 2016). Arising affective representations (e.g. positively aroused) may then feed into more sophisticated emotion processing in secondary auditory and association cortex enabling expression categorization (e.g. happy).

Table 4. Results of the task contrast analysis for the full data sets

Cluster ID	Size	Anatomical structure	BA
Voice (N = 20)			
1	256	L superior frontal gyrus (centered at – 0.3, 11.1, 53.8; 3 studies)	
1.1	168	L superior frontal gyrus	6
Face (N = 27)			
NS			

Table 5. Results of the task contrast analysis (matched data set, N = 20)

Cluster ID	Size	Anatomical structure	BA
Voice			
1	256	L superior frontal gyrus (centered at – 0.3, 11.1, 53.8; 3 studies)	
1.1	168	L superior frontal gyrus	6
Face			
NS			
Voice > face			
1	208	L superior frontal gyrus (centered at – 0.1, 10.5, 53.9; 2 studies)	
1.1	120	L superior frontal gyrus	6
Face > voice			
NS			
Voice & face			
NS			

Compared to auditory emotion processing, visual emotion processing more readily engages a mediotemporal region with the amygdala at its center. Initially implicated in fear conditioning and fear responding, the amygdala was later shown to support emotion and social perception more broadly. Moreover, it is thought to discriminate mundane from personally relevant stimuli either very crudely based on low-level thalamic input (e.g. low spatial frequency) (Garvert *et al.*, 2014; Méndez-Bértolo *et al.*, 2016) or in a more sophisticated manner based on input from sensory cortex (for a review see Schirmer and Adolphs, 2017). That the amygdala and its local emotion network are more relevant for facial than for vocal expressions may be because primary and secondary visual areas focus on basic structural representations outsourcing emotion representations to dedicated emotion centers. Moreover, visual compared to auditory emotion signals may be computationally less complex and thus less dependent on higher-order cortical contributions. Although visual more than auditory stimuli excite the amygdala, the latter may happen via the thalamic route for simple sounds such as fear-conditioned tones and, occasionally, via the cortical route for more complex sounds for which emotional meaning derives from multiple, dynamically changing acoustic signals (e.g. pitch, intensity, harmonic-to-noise ratio, tempo) that denote a heightened state of arousal.

It is noteworthy that the left IFG identified in the voice analysis was non-significant in both the modality subtraction and the conjunction analysis, suggesting that its role, although not clearly modality-specific is also not clearly modality-unspecific. Left IFG involvement in emotion evaluation may be slightly stronger for vocal than facial expressions due to the relevance of articulation and speech. Although covert mimicry and other IFG functions may also be triggered by faces, the small cluster obtained only in the full face data set suggests a minor role.

Goal-driven processing of emotional voices and faces

The contrast of explicit *vs* implicit/passive emotion processing may, arguably, reveal largely top-down mechanisms that support effortful stimulus categorization. In this meta-analysis, this contrast produced a significant effect for the voice data set (N = 20) in dmPFC centered in BA6. Interestingly, an exploratory analysis of vocal emotion contrasts with an explicit task design (N = 10) revealed a comparable effect corroborating the association between BA6 and explicit vocal processing.

BA6 comprises lateral and medial cortex that is classified as premotor cortex or supplementary motor area and as such is crucial for the planning, initiation and monitoring of movement. Relevant for this purpose is that medial BA6 supports vocalizing as demonstrated when speaking is compared with whispering (Schulz *et al.*, 2005) or when regional cerebral blood flow is correlated with muscle changes indicative of spontaneous laughter (Iwase *et al.*, 2002). Moreover, vocal production effects in medial BA6 overlap with voice perception pointing to a shared underlying mechanism (Belyk *et al.*, 2016; Lima *et al.*, 2016). Thus, together these and the present results suggest that listeners internally mimic a sound or vocal expression when trying to infer its emotional meaning.

Importantly, it is probable that medial BA6 serves multiple and/or a more general, modality-independent function that is facilitated when mental state inferences are made explicitly. Evidence for this comes from a recent meta-analysis comparing different mind-reading tasks (Molenberghs *et al.*, 2016). Medial BA6 was most reliably activated when another's mental state had to be explicitly inferred and when that state was an emotional one. In line with this, a meta-analysis looking at the anatomical and functional parcellation of dmPFC linked this brain region with a range of both emotional and cognitive processes implying a role in effortful, top-down control (Eickhoff *et al.*, 2016a).

Unlike voice studies, face studies from the full and matched data sets (N = 27/20) revealed no differences between explicit and implicit processing. Moreover, a subtraction analysis indicated that medial BA6 was more active for voices than faces, while there were no common regions and no regions more

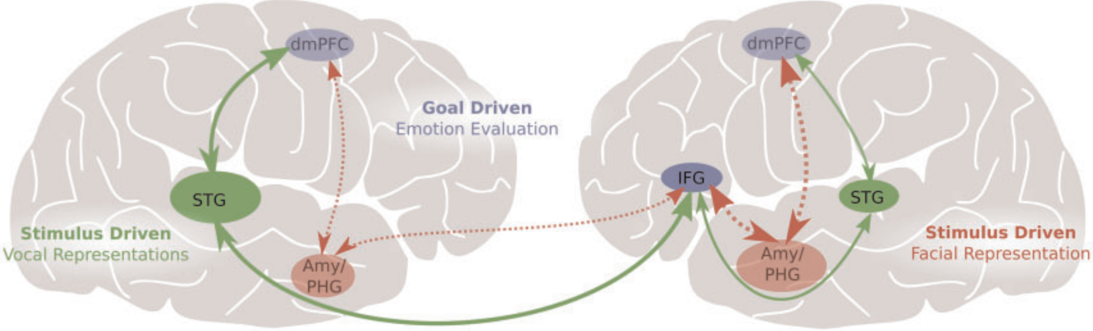

Fig. 2. Summary of brain regions highlighted in this meta-analysis. Lateral and medial areas are marked in nontransparent and transparent color, respectively. Early modality specific processing is indicated for faces in red and for voices in green. Later, potential modality convergence is indicated in violet. Arrows illustrate hypothesized up- and downstream modulations (not tested in the present study). Modality effects with strong evidence are marked by solid lines and those with weak evidence or with evidence from previous studies are marked by dashed lines. Although dmPFC failed to show for faces in this meta-analysis, there is other work implicating this region when the analysis of facial expressions is challenging (e.g. reading the mind in the eyes test). Left IFG activity was found in the emotion contrast of the full voice and face data sets. However, its exact functionality and activation conditions in the context of emotion perception remain to be determined. Amy, amygdala; dmPFC, dorso-medial prefrontal cortex; PHG, parahippogampal gyrus; IFG, inferior frontal gyrus; STG, superior temporal gyrus.

active for faces than for voices. This seems in apparent conflict with the idea that medial BA6 supports emotion perception cross-modally by producing an internal model of another's vocal or facial movement and by facilitating intentional representations of another's mental state. Perhaps, the absent face effect here may be due to the kind of expressions that were used in the face study set. Those were typically posed and exaggerated like the Ekman expressions. Moreover, studies suggest that such stimuli are more easily categorized when compared with vocal expressions and may hence be processed fairly effortlessly (Schirmer and Adolphs, 2017). By contrast, more challenging displays (e.g. when expressions are reduced to the eye region) appear to reliably activate medial BA6 (Molenberghs *et al.*, 2016).

Of note is that neither voices nor faces more strongly activated IFG and pSTS during explicit as compared with implicit processing. This is at odds with existing social perception models in which these regions play a central role and are thought to support higher-order processes characterized by modality convergence (Schirmer and Kotz, 2006; Frühholz and Grandjean, 2013a, 2013b; Bernstein and Yovel, 2015; Dricu and Frühholz, 2016; Schirmer and Adolphs, 2017). As discussed above, the recruitment of IFG and, possibly, pSTS may depend on both stimulus and task characteristics that make social inferences challenging or particularly relevant. What these characteristics are, however, and whether and how the modalities converge in lateral cortex has to be tackled by future research. Figure 2 provides an overview of the key points discussed here.

Directions for future research

The meta-analytic approach is powerful and overcomes many problems associated with individual studies. Moreover, given the many individual studies appearing on a weekly basis, it helps synthesize relevant findings. Yet, conducting a meta-analysis successfully presupposes that individual studies adhere to basic principles. For example, the current effort depended on the reporting of two specific contrast types. Although many studies explored vocal or facial emotions with emotional and neutral conditions as well as explicit and implicit tasks, many failed to report basic contrasts in favor of a more sophisticated analysis. As this greatly limited the present data pool and hampers meta-analyses in general, it would be useful if future studies report simple contrast results, even if only in a supplementary section. Second, prospective research should place greater emphasis on individual differences and report results from different subject populations. For example, there is much evidence that women are more sensitive than men to nonverbal emotions that are task irrelevant (Proverbio *et al.*, 2008; Schirmer *et al.*, 2013, 2016c) and that they differ in the brain correlates of emotion (Fusar-Poli *et al.*, 2009). Yet, very few studies used a sex-balanced sample and even fewer explored males and females separately. Last, although visual perception needs only a snapshot, it should not be reduced to that. Many more studies are needed that use real (instead of morphed) dynamic facial expressions. This would enable a fairer comparison with the auditory modality, which strictly depends on dynamic stimuli. Moreover, it would show whether some of the modality differences reported here are a mere artifact of experimental choices.

Although the meta-analytic approach has many advantages, it cannot replace carefully designed individual studies in addressing a particular question. Moreover, the benefit of looking at many data points comes at the cost of potential confounds which, despite best efforts at matching irrelevant experimental conditions, creep into the analysis. This is a particular concern for this investigation of goal-driven processes based on the explicit *vs* implicit/passive task contrasts. The latter task category differed widely between voice and face studies (Table 1), and this may be another factor contributing to the lack of modality convergence. Thus, the results obtained here should be probed further in a proper experiment designed to avoid confounds.

Conclusions

This study compared stimulus- and goal-driven emotion processing between faces and voices by subjecting published brain activations to a meta-analysis. Looking at stimulus-driven processes, emotional *vs* neutral expressions were found to depend largely on the bilateral amygdala for faces and on superior temporal cortex for voices. Looking at goal-driven processes, the contrast of explicit *vs* implicit tasks revealed a dmPFC cluster for voices and non-significant effects for faces. Together, these findings point to modality similarities and differences. Across modalities, emotion effects appear to modulate brain activity more extensively than task effects suggesting that both vocal and facial emotions shape mental functioning powerfully in a bottom-up manner. However, whereas voices emphasize cortical mechanisms, faces emphasize subcortical mechanisms. These differences imply that modality-specific sensory features shape processes beyond basic stimulus perception. In the case of voices, they promote higher-order perceptual and evaluative processing (e.g. via mimicry), whereas in the case of faces, they promote core emotional and mnemonic processes. Although future research is needed to replicate these results, they underline the importance of considering the modalities as unique and to study them both in isolation and combination.

Acknowledgements

The author thanks Lukas Hensel for sharing the results of an unpublished contrast and Ralph Adolphs for helpful comments on an earlier version of this manuscript.

Supplementary data

Supplementary data are available at *SCAN* online.

Conflict of interest. None declared.

References

Aminoff, E.M., Kveraga, K., Bar, M. (2013). The role of the parahippocampal cortex in cognition. *Trends in Cognitive Sciences*, **17**(8), 379–90.

Anderson, A.K., Phelps, E.A. (2001). Lesions of the human amygdala impair enhanced perception of emotionally salient events. *Nature*, **411**, 305–9.

Anderson, A.K., Phelps, E.A. (1998). Intact recognition of vocal expressions of fear following bilateral lesions of the human amygdala. *NeuroReport*, **9**, 3607–13.

Arsalidou, M., Duerden, E.G., Taylor, M.J. (2013). The centre of the brain: Topographical model of motor, cognitive, affective, and somatosensory functions of the basal ganglia. *Human Brain Mapping*, **34**(11), 3031–54.

Atkinson, A.P., Adolphs, R. (2011). The neuropsychology of face perception: beyond simple dissociations and functional selectivity. *Philosophical Transactions of the Royal Society*, **366**(1571), 1726–38.

Aubé, W., Angulo-Perkins, A., Peretz, I., Concha, L., Armony, J.L. (2015). Fear across the senses: brain responses to music, vocalizations and facial expressions. *Social Cognitive and Affective Neuroscience*, **10**(3), 399–407.

Barraclough, N.E., Xiao, D., Baker, C.I., Oram, M.W., Perrett, D.I. (2005). Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Journal of Cognitive Neuroscience*, **17**(3), 377–91.

Belin, P., Bestelmeyer, P.E.G., Latinus, M., Watson, R. (2011). Understanding voice perception. *British Journal of Psychology*, **102**(4), 711–25.

Belin, P., Zatorre, R.J., Lafaille, P., Ahad, P., Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, **403**(6767), 309–12.

Belyk, M., Pfordresher, P.Q., Liotti, M., Brown, S. (2016). The neural basis of vocal pitch imitation in humans. *Journal of Cognitive Neuroscience*, **28**(4), 621–35.

Bernstein, M., Yovel, G. (2015). Two neural pathways of face processing: a critical evaluation of current models. *Neuroscience & Biobehavioral Reviews*, **55**, 536–46.

Binelli, C., Subirà, S., Batalla, A., *et al*. (2014). Common and distinct neural correlates of facial emotion processing in social anxiety disorder and Williams syndrome: a systematic review and voxel-based meta-analysis of functional resonance imaging studies. *Neuropsychologia*, **64**, 205–17.

Brauer, J., Xiao, Y., Poulain, T., Friederici, A.D., Schirmer, A. (2016). Frequency of maternal touch predicts resting activity and connectivity of the developing social brain. *Cerebral Cortex*, **26**(8), 3544–52.

Darwin, C. 1872. *The Expression of the Emotions in Man and Animals*, John Murray, London.

Del Casale, A., Kotzalidis, G.D., Rapinesi, C., *et al*. (2017). Neural functional correlates of empathic face processing. *Neuroscience Letters*, **655**, 68–75.

Dricu, M., Frühholz, S. (2016). Perceiving emotional expressions in others: activation likelihood estimation meta-analyses of explicit evaluation, passive perception and incidental perception of emotions. *Neuroscience & Biobehavioral Reviews*, **71**, 810–28.

Eickhoff, S.B., Bzdok, D., Laird, A.R., Kurth, F., Fox, P.T. (2012). Activation likelihood estimation meta-analysis revisited. *NeuroImage*, **59**(3), 2349–61.

Eickhoff, S.B., Bzdok, D., Laird, A.R., *et al*. (2011). Co-activation patterns distinguish cortical modules, their connectivity and functional differentiation. *NeuroImage*, **57**(3), 938–49.

Eickhoff, S.B., Laird, A.R., Fox, P.T., Bzdok, D., Hensel, L. (2016). Functional segregation of the human dorsomedial prefrontal cortex. *Cerebral Cortex*, **26**(1), 304.

Eickhoff, S.B., Laird, A.R., Grefkes, C., Wang, L.E., Zilles, K., Fox, P.T. (2009). Coordinate-based activation likelihood estimation meta-analysis of neuroimaging data: a random-effects approach based on empirical estimates of spatial uncertainty. *Human Brain Mapping*, **30**, 2907–26.

Eickhoff, S.B., Nichols, T.E., Laird, A.R., *et al*. (2016). Behavior, sensitivity, and power of activation likelihood estimation characterized by massive empirical simulation. *NeuroImage*, **137**, 70–85.

Ethofer, T., Kreifelts, B., Wiethoff, S., *et al*. (2009). Differential influences of emotion, task, and novelty on brain regions underlying the processing of speech melody. *Journal of Cognitive Neuroscience*, **21**(7), 1255–68.

Fecteau, S., Belin, P., Joanette, Y., Armony, J.L. (2007). Amygdala responses to nonlinguistic emotional vocalizations. *NeuroImage*, **36**(2), 480–7.

Fox, A.S., Oler, J.A., Tromp, D.P.M., Fudge, J.L., Kalin, N.H. (2015). Extending the amygdala in theories of threat processing. *Trends in Neuroscience*, **38**(5), 319–29.

Frühholz, S., Ceravolo, L., Grandjean, D. (2012). Specific brain networks during explicit and implicit decoding of emotional prosody. *Cerebral Cortex*, **22**(5), 1107–17.

Frühholz, S., Grandjean, D. (2013). Multiple subregions in superior temporal cortex are differentially sensitive to vocal expressions: a quantitative meta-analysis. *Neuroscience & Biobehavioral Reviews*, **37**(1), 24–35.

Frühholz, S., Grandjean, D. (2013). Processing of emotional vocalizations in bilateral inferior frontal cortex. *Neuroscience & Biobehavioral Reviews*, **37**(10), 2847–55.

Frühholz, S., Hofstetter, C., Cristinzio, C., *et al*. (2015). Asymmetrical effects of unilateral right or left amygdala damage on auditory cortical processing of vocal emotions. *Proceedings of the National Academy of Sciences of the United States of America*, **112**(5), 1583–8.

Frühholz, S., van der Zwaag, W., Saenz, M., *et al*. (2016). Neural decoding of discriminative auditory object features depends on their socio-affective valence. *Social Cognitive and Affective Neuroscience*, **11**(10), 1638–49.

Fusar-Poli, P., Placentino, A., Carletti, F., *et al.* (2009). Functional atlas of emotional faces processing: a voxel-based meta-analysis of 105 functional magnetic resonance imaging studies. *Journal of Psychiatry & Neuroscience JPN*, **34**, 418–32.

Gainotti, G., Marra, C. (2011). Differential contribution of right and left temporo-occipital and anterior temporal lesions to face recognition disorders. *Frontiers in Human Neuroscience*, **5**, 55. https://doi.org/10.3389/fnhum.2011.00055.

Garvert, M.M., Friston, K.J., Dolan, R.J., Garrido, M.I. (2014). Subcortical amygdala pathways enable rapid face processing. *NeuroImage*, **102 Pt 2**, 309–16.

Goerlich-Dobre, K.S., Witteman, J., Schiller, N.O., van Heuven, V.J.P., Aleman, A., Martens, S. (2014). Blunted feelings: Alexithymia is associated with a diminished neural response to speech prosody. *Social Cognitive and Affective Neuroscience*, **9**(8), 1108–17.

Haidich, A.B. (2010). Meta-analysis in medical research. *Hippokratia*, **14(Suppl 1)**, 29–37.

Halgren, E., Raij, T., Marinkovic, K., Jousmäki, V., Hari, R. (2000). Cognitive response profile of the human fusiform face area as determined by MEG. *Cerebral Cortex*, **10**(1), 69–81.

Hamani, C., Mayberg, H., Stone, S., Laxton, A., Haber, S., Lozano, A.M. (2011). The subcallosal cingulate gyrus in the context of major depression. *Biological Psychiatry*, **69**(4), 301–8.

Haxby, J.V., Hoffman, E.A., Gobbini, M.I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, **4**(6), 223–33.

Hensel, L., Bzdok, D., Müller, V.I., Zilles, K., Eickhoff, S.B. (2015). Neural correlates of explicit social judgments on vocal stimuli. *Cerebral Cortex*, **25**(5), 1152–62.

Hickok, G., Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, **8**(5), 393–402.

Iwase, M., Ouchi, Y., Okada, H., *et al.* (2002). Neural substrates of human facial expression of pleasant emotion induced by comic films: a PET study. *NeuroImage*, **17**(2), 758–68.

Kanwisher, N., McDermott, J., Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, **17**, 4302–11.

King, A.J., Nelken, I. (2009). Unraveling the principles of auditory cortical processing: can we learn from the visual system?. *Nature Neuroscience*, **12**(6), 698–701.

Klasen, M., Chen, Y.-H., Mathiak, K. (2012). Multisensory emotions: perception, combination and underlying neural processes. *Reviews in the Neurosciences*, **23**, 381–92.

Kuhlmann, M., Hofmann, M.J., Briesemeister, B.B., Jacobs, A.M. (2016). Mixing positive and negative valence: affective-semantic integration of bivalent words. *Scientific Reports*, **6**(1), 30718.

Lima, C.F., Krishnan, S., Scott, S.K. (2016). Roles of supplementary motor areas in auditory processing and auditory imagery. *Trends in Neuroscience*, **39**(8), 527–42.

Méndez-Bértolo, C., Moratti, S., Toledano, R., *et al.* (2016). A fast pathway for fear in human amygdala. *Nature Neuroscience*, **19**(8), 1041–9.

Molenberghs, P., Johnson, H., Henry, J.D., Mattingley, J.B. (2016). Understanding the minds of others: a neuroimaging meta-analysis. *Neuroscience & Biobehavioral Reviews*, **65**, 276–91.

Morris, J.S., Scott, S.K., Dolan, R.J. (1999). Saying it with feeling: neural responses to emotional vocalizations. *Neuropsychologia*, **37**(10), 1155–63.

Mothes-Lasch, M., Mentzel, H.-J., Miltner, W.H.R., Straube, T. (2011). Visual attention modulates brain activation to angry voices. *Journal of Neuroscience*, **31**(26), 9594–8.

Nieuwenhuys, R. (2012). The insular cortex: a review. *Progress in Brain Research*, **195**, 123–63.

Perrodin, C., Kayser, C., Abel, T.J., Logothetis, N.K., Petkov, C.I. (2015). Who is that? Brain networks and mechanisms for identifying individuals. *Trends in Cognitive Sciences*, **19**(12), 783–96.

Phillips, M.L., Young, A.W., Scott, S.K., *et al.* (1998). Neural responses to facial and vocal expressions of fear and disgust. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, **265**(1408), 1809–17.

Proverbio, A.M., Zani, A., Adorni, R. (2008). Neural markers of a greater female responsiveness to social stimuli. *BMC Neuroscience*, **9**, 56.

Quadflieg, S., Mohr, A., Mentzel, H.-J., Miltner, W.H.R., Straube, T. (2008). Modulation of the neural network involved in the processing of anger prosody: the role of task-relevance and social phobia. *Biological Psychology*, **78**(2), 129–37.

Quiroga, R.Q., Kraskov, A., Koch, C., Fried, I. (2009). Explicit encoding of multimodal percepts by single neurons in the human brain. *Current Biology*, **19**(15), 1308–13.

Rutishauser, U., Mamelak, A.N., Adolphs, R. (2015). The primate amygdala in social perception—insights from electrophysiological recordings and stimulation. *Trends in Neuroscience*, **38**(5), 295–306.

Sabatinelli, D., Fortune, E.E., Li, Q., *et al.* (2011). Emotional perception: meta-analyses of face and natural scene processing. *NeuroImage* **54**(3), 2524–33.

Sammler, D., Grosbras, M.-H., Anwander, A., Bestelmeyer, P.E.G., Belin, P. (2015). Dorsal and ventral pathways for prosody. *Current Biology*, **25**(23), 3079–85.

Sander, D. (2012). The role of the amygdala in the appraising brain. *Behavioral and Brain Sciences*, **35**(3), 161.

Sander, D., Grafman, J., Zalla, T. (2003). The human amygdala: an evolved system for relevance detection. *Reviews in the Neurosciences*, **14**, 303–16.

Sander, D., Grandjean, D., Pourtois, G., *et al.* (2005). Emotion and attention interactions in social cognition: brain regions involved in processing anger prosody. *NeuroImage*, **28**(4), 848–58.

Schirmer, A., Adolphs, R. (2017). Emotion perception from face, voice, and touch: comparisons and convergence. *Trends in Cognitive Sciences*, **21**(3), 216–28.

Schirmer, A., Escoffier, N., Cheng, X., Feng, Y., Penney, T.B. (2016). Detecting temporal change in dynamic sounds: on the role of stimulus duration, speed, and emotion. *Frontiers in Psychology*, **6**, 2055.

Schirmer, A., Kotz, S.A. (2006). Beyond the right hemisphere: brain mechanisms mediating vocal emotional processing. *Trends in Cognitive Sciences*, **10**(1), 24–30.

Schirmer, A., Meck, W.H., Penney, T.B. (2016). The socio-temporal brain: connecting people in time. *Trends in Cognitive Sciences*, **20**(10), 760–72.

Schirmer, A., Ng, T., Escoffier, N., Penney, T.B. (2016). Emotional voices distort time: behavioral and neural correlates. *Timing & Time Perception*, **4**(1), 79–98.

Schirmer, A., Seow, C.S., Penney, T.B. (2013). Humans process dog and human facial affect in similar ways. *PLoS One*, **8**(9), e74591.

Schirmer, A., Zysset, S., Kotz, S.A., Yves von Cramon, D. (2004). Gender differences in the activation of inferior frontal cortex during emotional speech perception. *NeuroImage*, **21**(3), 1114–23.

Schulz, G.M., Varga, M., Jeffires, K., Ludlow, C.L., Braun, A.R. (2005). Functional neuroanatomy of human vocalization: an H215O PET study. *Cerebral Cortex*, **15**(12), 1835–47.

Scott, S.K., Young, A.W., Calder, A.J., Hellawell, D.J., Aggleton, J.P., Johnsons, M. (1997). Impaired auditory recognition of fear and anger following bilateral amygdala lesions. *Nature*, **385**(6613), 254–7.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, **153**(3736), 652–4.

Strauss, M., Sitt, J.D., King, J.-R., *et al.* (2015). Disruption of hierarchical predictive coding during sleep. *Proceedings of the National Academy of Sciences of the United States of America*, **112**(11), E1353–62.

Suh, J., Rivest, A.J., Nakashiba, T., Tominaga, T., Tonegawa, S. (2011). Entorhinal cortex layer III input to the hippocampus is crucial for temporal association memory. *Science*, **334**(6061), 1415–20.

Turkeltaub, P.E., Eickhoff, S.B., Laird, A.R., Fox, M., Wiener, M., Fox, P. (2012). Minimizing within-experiment and within-group effects in activation likelihood estimation meta-analyses. *Human Brain Mapping*, **33**(1), 1–13.

Valk, S.L., Bernhardt, B.C., Trautwein, F.-M., *et al.* (2017). Structural plasticity of the social brain: Differential change after socio-affective and cognitive mental training. *Science Advances*, **3**(10), e1700489.

Watson, R., Latinus, M., Charest, I., Crabbe, F., Belin, P. (2014). People-selectivity, audiovisual integration and heteromodality in the superior temporal sulcus. *Cortex*, **50**, 125–36.

Wildgruber, D., Ackermann, H., Kreifelts, B., Ethofer, T. (2006). Cerebral processing of linguistic and emotional prosody: fMRI studies. *Progress in Brain Research*, **156**, 249–68.