Adaptation to vocal expressions reveals multistep perception of auditory emotion
Bestelmeyer, P.E.; Maurage, P.; Rouger, J.; Latinus, M.; Belin, P.

Journal of Neuroscience

DOI:

Published: 11/06/2014

Publisher's PDF, also known as Version of record

Cyswllt i'r cyhoeddiad / Link to publication

Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA):

Hawliau Cyffredinol / General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

• Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying the publication in the public portal

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.
Adaptation to Vocal Expressions Reveals Multistep Perception of Auditory Emotion

Patricia E. G. Bestelmeyer,1 Pierre Maurage,2 Julien Rouger,3 Marianne Latinus,4 and Pascal Belin4,5,6

1School of Psychology, Bangor University, Bangor, Gwynedd LL57 2AS, United Kingdom, 2Cognitive Neurosciences and Clinical Psychology Research Units, Department of Psychology, Catholic University of Louvain, 1348 Louvain-la-Neuve, Belgium, 3Department of Cognitive Neuroscience, University of Maastricht, 6200 MD, Maastricht, The Netherlands, 4Institut de Neurosciences de La Timone, Unité Mixte de Recherche 7289, Aix-Marseille Université, Centre National de la Recherche Scientifique, 13385 Marseille, France, 5Institute of Neuroscience and Psychology, University of Glasgow, Glasgow G12 8QF, United Kingdom, and 6International Laboratory for Brain, Music and Sound Research, University of Montréal/McGill University, Montréal H3C 3J7, Canada

The human voice carries speech as well as important nonlinguistic signals that influence our social interactions. Among these cues that impact our behavior and communication with other people is the perceived emotional state of the speaker. A theoretical framework for the neural processing stages of emotional prosody has suggested that auditory emotion is perceived in multiple steps (Schirmer and Kotz, 2006) involving low-level auditory analysis and integration of the acoustic information followed by higher-level cognition. Empirical evidence for this multistep processing chain, however, is still sparse. We examined this question using functional magnetic resonance imaging and a continuous carry-over design (Aguirre, 2007) to measure brain activity while volunteers listened to non-speech-affective vocalizations morphed on a continuum between anger and fear. Analyses dissociated neuronal adaptation effects induced by similarity in perceived emotional content between consecutive stimuli from those induced by their acoustic similarity. We found that bilateral voice-sensitive auditory regions as well as right amygdala coded the physical difference between consecutive stimuli. In contrast, activity in bilateral anterior insulae, medial superior frontal cortex, precuneus, and subcortical regions such as bilateral hippocampi depended predominantly on the perceptual difference between morphs. Our results suggest that the processing of vocal affect recognition is a multistep process involving largely distinct neural networks. Amygdala and auditory areas predominantly code emotion-related acoustic information while more anterior insular and prefrontal regions respond to the abstract, cognitive representation of vocal affect.

Key words: fMRI; vocal emotion; voice perception

Introduction

Voices convey an array of social information such as the gender, identity, and emotional state of the speaker. Amplitude, fundamental frequency, and voice smoothness are only some of the variables that convey emotion by physiological changes in the vocal production system (Patel et al., 2011). Perceiving these acoustic cues as a complex message that communicates emotions or intentions even in the absence of speech is crucial for everyday social interactions (Wallbott and Scherer, 1986). Research on auditory affect perception has concentrated on the perception of speech prosody with a neuroscientific model proposing the anatomical underpinnings that are involved (Schirmer and Kotz, 2006). The first stage of this model consists of low-level acoustic analyses in bilateral auditory cortices. These areas then project to superior temporal sulci (STSs) and superior temporal gyri (STGs) for more complex processing in which emotionally salient information is synthesized into an emotional “Gestalt” or acoustic object. The STS and STG then feed into frontal areas for higher-order cognition (e.g., evaluative judgments of prosody). This model provides an analytical first step toward understanding the network involved, but more research is necessary to test its predictions and expand it to the perception of other emotional auditory objects.

We were interested in examining the neural processing stages of vocal affect perception using a specific type of adaptation paradigm. Behaviorally, adaptation refers to a process during which continued stimulation results in biased perception toward opposite features of the adaptor (Grill-Spector et al., 1999). These aftereffects have been demonstrated behaviorally for vocal affect perception (Bestelmeyer et al., 2010; Skuk and Schweinberger, 2013). Neurally, fMRI adaptation to a specific stimulus feature is typically accompanied by a decrease in the hemodynamic response (also referred to as repetition suppression). Adaptation paradigms are often used to reveal neural populations tuned to the processing of a specific stimulus attribute (i.e., to reveal func-
tional specificity of neural populations). Continuous carry-over designs present an unbroken stream of serially balanced stimuli, thereby allowing the evaluation of adaptation or "carry-over" effects between subsequent stimuli (Aguirre, 2007).

We created morph continua between angry and fearful vocalizations (Bestelmeyer et al., 2010). Carry-over effects between consecutive stimuli were assessed by calculating physical and perceptual differences between morphs in the stimulus sequence and entering these values as regressors in a general linear model. The "physical difference" regressor consisted of the absolute difference in morph step between a given stimulus and the preceding one, directly proportional to the acoustic difference. The "perceptual difference" regressor consisted of the absolute difference in proportion of fear responses between stimuli and was orthogonalized with respect to the "physical regressor" in the design matrix. This design allowed us to examine adaptation effects induced by perceived emotion after variance accounted for by acoustic differences had been explained. We hypothesized that carry-over effects to the physical properties would be evident in voice-sensitive areas with a lateralization toward the right (Belin et al., 2000; Lewis et al., 2009; Bestelmeyer et al., 2011). In contrast, we expected the cognitive representations of emotion to engage separate, higher-level areas involving frontal cortices.

Materials and Methods

Participants. Nineteen healthy volunteers from the under- and postgraduate community of the University of Glasgow took part (13 females, mean age = 22.63, range = 18–32 years, SD = 3.96). All participants were of normal hearing (according to a brief audiometric test) and were reimbursed £12 for their time. Informed consent was obtained from all individuals and the study protocol was approved by the local ethics committee.

Experimental stimuli and fMRI paradigm. Recordings were taken from the Montreal Affective Voices (Belin et al., 2008) in which actors were instructed to produce emotional interjections using the vowel /a/. The voices from six identities (three female) expressed anger and fear. Angry to fearful continua were created separately for each identity in seven steps (fear/anger) 5/95%, 20/80%, 35/65%, 50/50%, 65/35%, 80/20% and 95/5%. The duration of the vocalizations within each such continuum was kept constant and ranged between 0.6 and 0.9 s across each carry-over sequence was defined by eight items: (1) “first voice” of each block (to remove the effect of a stimulus following baseline); (2) “morph step” (i.e., the percentage of fearfulness contained in the voice); (3) physical difference, which was computed as the absolute difference in morph step between two consecutive stimuli; and

Theobald, 2007). Our carry-over sequence was defined by eight items: the seven morph steps of a continuum (always within identity) plus one silent null event. One complete carry-over sequence with eight items comprises 65 stimuli. In this balanced sequence, each stimulus is preceded and followed by every other stimulus an equal number of times. Thus, every item is repeated eight times per sequence (except for the first item, which is repeated nine times; for a sample sequence of eight items, see http://www.bios.ac.uk/people/cm/seqslit.html#08). The end of each carry-over sequence was followed by nine TRs of silences (18 s) before the start of the next carry-over sequence. Each of our six identities was presented once per run, leading to six different carry-over sequences and totaling 390 stimuli per run. Gender was blocked so that an experimental run started with three male carry-over sequences followed by three female carry-over sequences or vice versa. We acquired three experimental fMRI runs.

Sounds were presented binaurally using the electrostatic NNL headphones system (NordicNeuroLab) at an intensity of 80 dB SPL (C). Participants were asked to perform a two-alternative forced choice task in which each affective voice had to be categorized as either angry or fearful by means of two buttons mounted on an MRI-compatible response box (IORP, Current Designs). Participants were asked to react as fast as possible and to keep their eyes closed for the duration of all scans. For each participant, the behavioral responses were used to compute the perceptual difference regressor, which was then fed into the parametric modulation analyses. Following this procedure, we administered a short "voice localizer" scan (for details, see Belin et al., 2000) to localize areas that are particularly sensitive to voices compared with other, environmental sounds.

Image acquisition and analysis. All scans were acquired in a 3.0 tesla Siemens Tim Trio scanner using a 12-channel head coil. Whole-brain T1-weighted anatomical scans were performed using fast gradient echo known as T1 magnetization-prepared rapid acquisition gradient echo, consisting of 192 axial slices of 1 mm thickness with an in-plane resolution of 1 × 1 × 1 [field of view (FOV) = 256°] and a matrix of 256 × 256 performed at the end of the experimental session. T2*-weighted functional scans were acquired using an interleaved ascending sequence consisting of 32 slices of 3 mm thickness (0.3 mm gap) with an in-plane resolution of 3 × 3 × 3 (FOV = 126°), an acquisition matrix of 70 × 70, and a flip angle of 77°. The three runs of the experimental scan (TR = 2 s, TE = 30 ms) consisted of 446 volumes each, while the voice localizer (TR = 2 s, TE = 30 ms) scan consisted of one run of 310 volumes and allows reliable identification of the temporal voice areas by contrasting vocal with nonvocal sounds.

All MRI data were analyzed using SPM8 (Wellcome Department of Cognitive Neurology, University College London, http://www.fil.ion.ucl.ac.uk/spm/). Preprocessing of the data consisted of anterior commissure–posterior commissure alignment of the anatomical images (and application of the orientation change to all functional images acquired in the same session). Functional scans were corrected for head motion (trilinear interpolation) by aligning all scans to the first scan of the last run, and a mean image was created. The anatomical scan was coregistered to the mean image. Functional and anatomical data were transformed to Montreal Neurological Institute space after segmentation of the anatomical scan. Normalized data were spatially smoothed by applying a Gaussian kernel of 8 mm full-width at half-maximum.

We used parametric modulations to analyze our data ( Büchel et al., 1998). We linearly rescaled the parametric variables so that all measurements were on a comparable scale lying between 0 and 1 using the “Min-Max” normalization. This transformation preserves the relationships among the original data values and ensures that the regressors have equivalent weights. For each auditory stimulus (except the first sound in each carry-over sequence), we calculated the physical and perceptual differences between it and the preceding stimulus (Fig. 1). The design matrix of first-level analyses contained all voice onset events as the first regressor which was followed by four parametric regressors, as follows: (1) "first voice" of each block (to remove the effect of a stimulus following baseline); (2) "morph step" (i.e., the percentage of fearfulness contained in the voice); (3) physical difference, which was computed as the absolute difference in morph step between two consecutive stimuli; and
(4) perceptual difference, which was calculated in the same way as regressor 2 but was based on the participant’s behavioral responses (Fig. 2; Aguirre, 2007; Charest et al., 2013). Thus, for the perceptual difference regressor of each run, we calculated the mean behavioral response to each morph step and then calculated the absolute difference in mean responses between two consecutive stimuli. The perceptual difference regressor was therefore based on the participant’s own subjective judgments of the affective morphs.

Note that each of the four parametric regressors was orthogonalized (i.e., rendering collinear regressors linearly independent) with respect to the previous ones to account only for variance not explained by previous regressors. We also included the silent null events presented within the carry-over sequences and six movement parameters in our model. Second-level analysis of the morph step and physical difference regressors consisted of one-sample t tests for each regressor, allowing for the evaluation of positive and negative correlations. The analysis of the perceptual difference regressor consisted of an F test across linear and quadratic contrasts in a full-factorial ANOVA with two levels (level 1: T-contrast images from the linear expansion; level 2: T-contrast images from the quadratic expansion). To illustrate the parametric modulations in Figures 3 and 4, we used the built-in function of SPM8 (spm_graph.m) to extract the β estimates at the peak maxima within a sphere of 6 mm.

Figure 1. Illustration of the calculation of the physical (red) and perceptual (green) difference regressors. The physical difference regressor is calculated based on the absolute difference in morph step between two consecutively presented stimuli. Similarly, the perceptual difference regressor is calculated based on the absolute difference in the averaged participant’s behavioral responses (two alternative forced choice responses) to each morph step between two consecutively presented stimuli.

Figure 2. A, Results of the emotion categorization task. B, Reaction time as a function of morph step. Error bars represent the SEM.
Unlike the physical difference, behavioral responses did not average to discrete steps but, rather, resulted in a range of numbers. Therefore, to account for the interindividual variability of behavioral responses and to allow the averaging across participants, we grouped averaged behavioral responses into five bins of 20%. Results are illustrated on an average anatomical scan using MRIcon (Rorden et al., 2007). Anatomical cluster location was assessed with xjview version 8.1 (http://www.alivelearn.net/xjview) and cross-checked with the brain atlas of Duvernoy (1999) to ensure accuracy.

Our two regressors of interest (physical and perceptual difference) are not naturally orthogonal. Given our design, the physical regressor, entered before the perceptual regressor, may therefore contain variance that is shared by both regressors (i.e., variance that is neither purely physical nor purely perceptual). We therefore computed two further “single” models for each participant, as follows: (1) a model with the physical regressor but without the perceptual regressor; and (2) a model with the perceptual regressor but without the physical regressor. These models were otherwise identical to the one reported above.

Results

Behavioral results

Results of the behavioral data are summarized in Figure 2. The response data were averaged as a function of the seven morph steps, and a psychophysical curve (based on the hyperbolic tangent function) was fitted to the mean data (Fig. 2A; see also Bestelmeyer et al. (2010) for an identical pattern obtained without scanner noise). Participants more frequently categorized the first three morph steps as anger and the last three morph steps as fear, while the 50% morph was perceived as the most emotionally ambiguous. Reaction times were lowest for the least ambiguous stimuli and highest for the 50% morph with >150 ms difference between extreme versus central points of the continuum (Fig. 2B).

fMRI results

We used parametric modulation analyses to investigate areas that respond to a modulation of the morph step, physical difference, and then perceptual difference from a combined general linear model (see (A) and (B), below). All results are illustrated at a threshold of \( p < 0.001 \) (uncorrected) and an extent threshold of 30 voxels. Unless stated otherwise, statistical significance was assessed at the cluster level with a threshold of \( p < 0.05 \) and was FWE corrected for multiple comparisons across the whole brain. No significant modulations with morph step were observed (i.e., no brain area reacted more to anger than to fear and vice versa).

(A) Carry-over effect of the physical difference between consecutive morphs

Parametric modulation analysis of the physical difference regressors showed significant positive correlations between physical difference and BOLD signal in bilateral superior/mid-temporal gyri (right: \( F_{1(18)} = 28.78, k = 169 \); left: \( F_{1(18)} = 38.80, k = 153 \)), bilateral precuneus (right: \( F_{1(236)} = 2.58, k = 240 \); left: \( F_{1(236)} = 37.75, k = 336 \)), bilateral middle temporal gyrus (MTG), superior temporal gyrus (STG; right STG: \( F_{1(236)} = 21.52, k = 75 \); left STG: \( F_{1(236)} = 25.34, k = 131 \); left MTG: \( F_{1(236)} = 23.73, k = 241 \)), supplementary motor area/mid-cingulum (\( F_{1(236)} = 32.90, k = 205 \)), left superior frontal region (\( F_{1(236)} = 25.20, k = 66 \)), left medial superior frontal (\( F_{1(236)} = 19.40; k = 98 \)), right lingual (\( F_{1(236)} = 20.21, k = 83 \)), and a positive quadratic relationship in a cluster covering bilateral precuneus (\( F_{1(236)} = 32.44; k = 516 \)).

At a more lenient threshold of \( p < 0.001 \) (corrected), using small-volume corrections (6 mm), a number of subcortical structures survived FWE correction at the peak level (i.e., clusters covering bilateral hippocampus and amygdala; right: \( F_{1(236)} = 20.51, k = 19 \), MNI coordinates = \(-4 -4 -23\); left: \( F_{1(236)} = 14.09, k = 20 \), MNI coordinates = \(-24 -7 20\) ) and thalamus (right: \( F_{1(236)} = 16.38, k = 29 \), MNI coordinates = \(-12 -19 4\); left: \( F_{1(236)} = 17.02, k = 28 \), MNI coordinates = \(-12 -16 4\) ). Figure 3 illustrates significant clusters and illustrations of the parameter estimates in bilateral anterior insulae, mid-cingulum, and precuneus. Significant clusters are summarized in Table 1.

We then computed parametric modulation analyses of two single general linear models (reported in (C), below) to address whether regions dealing with physical and perceptual differences of vocal expressions involve partially overlapping regions.

(C) Overlap between the carry-over effect for the physical and perceptual difference from the two single models

Parametric modulation analyses of the two single models (either with the physical but not the perceptual regressor or vice versa) showed very similar patterns of activations to our combined model reported above. The analyses confirm that while the two steps necessary for the acoustic and perceptual processing of vocal emotion are largely distinct, there is some overlap between the two processes in right midtemporal and left superior temporal gyrus and left putamen (Fig. 5). Significant clusters of both single models are summarized in Table 2.

Discussion

We were interested in disentangling the different regions of the vocal emotion network that process acoustic differences between emotional vocalizations from regions that deal with the cognition and evaluation of vocally expressed emotions by using a new adaptation design (Aguirre, 2007). We morphed nonverbal expressions on a continuum between anger and fear, and presented the stimuli in a continuous and balanced sequence (Nonyane and Theobald, 2007). We found linear neuronal adaptation effects in response to the physical difference between two consecutively presented voice morphs in bilateral voice-sensitive cortices (Belin and Zatorre, 2000; Lewis et al., 2009; Bestelmeyer et al., 2011) with larger activations in the right hemisphere, and right amygdala. In response to the perceived difference between stimuli, we found repetition effects in more widespread regions in-
including bilateral anterior insulae, precentral gyri, and medial superior frontal areas, as well as amygdalae, hippocampi, and thalami at a more lenient threshold. Our study provides support for the key concept described in Schirmer and Kotz (2006) of multistep processes involved in the coding of auditory emotion but also highlights the unpredicted complexity of the network necessary for the evaluative judgments of affective vocalizations, even in the absence of speech.

Schirmer and Kotz (2006) propose a hierarchical model for the processing of emotional prosody (for a multimodal version, see also Wildgruber et al. (2009); Brück et al. (2011)). This model consists of three steps starting with a low-level acoustic analysis in bilateral auditory cortices followed by mid-level analyses in bilateral STS/STG, which synthesize emotionally salient information into an acoustic object. Finally, STS/STG then feed into frontal areas for higher-order cognition and evaluation. Most of the neuroimaging studies cited in support of the model used subtraction designs, which make a demonstration of the existence of separate steps and differential functions of STS/STG and IFGs difficult. Importantly, also, these studies do not typically differentiate activity patterns that process general physical differences between stimuli from those that process more abstract emotional information independent of lower-level acoustic changes. To our knowledge, only one study (Wiethoff et al., 2008) has used multiple regression to demonstrate that the passive-listening response to emotionally spoken words in the right mid-STG could be explained only by a combination of various acoustic parameters, but the study did not elucidate the areas responsible for the higher-level representation of emotion.

The continuous carry-over design allowed us to isolate the neural processes engaged in lower-level stages dealing with physical differences between sounds from higher-level stages involved in vocal affect cognition. Thus, we first computed correlations of neural activity patterns that process general physical differences between stimuli from those that process more abstract emotional information independent of lower-level acoustic changes. To our knowledge, only one study (Wiethoff et al., 2008) has used multiple regression to demonstrate that the passive-listening response to emotionally spoken words in the right mid-STG could be explained only by a combination of various acoustic parameters, but the study did not elucidate the areas responsible for the higher-level representation of emotion.

The continuous carry-over design allowed us to isolate the neural processes engaged in lower-level stages dealing with physical differences between sounds from higher-level stages involved in vocal affect cognition. Thus, we first computed correlations of neural activity patterns that process general physical differences between stimuli from those that process more abstract emotional information independent of lower-level acoustic changes. To our knowledge, only one study (Wiethoff et al., 2008) has used multiple regression to demonstrate that the passive-listening response to emotionally spoken words in the right mid-STG could be explained only by a combination of various acoustic parameters, but the study did not elucidate the areas responsible for the higher-level representation of emotion.

The continuous carry-over design allowed us to isolate the neural processes engaged in lower-level stages dealing with physical differences between sounds from higher-level stages involved in vocal affect cognition. Thus, we first computed correlations of neural activity patterns that process general physical differences between stimuli from those that process more abstract emotional information independent of lower-level acoustic changes. To our knowledge, only one study (Wiethoff et al., 2008) has used multiple regression to demonstrate that the passive-listening response to emotionally spoken words in the right mid-STG could be explained only by a combination of various acoustic parameters, but the study did not elucidate the areas responsible for the higher-level representation of emotion.

The continuous carry-over design allowed us to isolate the neural processes engaged in lower-level stages dealing with physical differences between sounds from higher-level stages involved in vocal affect cognition. Thus, we first computed correlations of neural activity patterns that process general physical differences between stimuli from those that process more abstract emotional information independent of lower-level acoustic changes. To our knowledge, only one study (Wiethoff et al., 2008) has used multiple regression to demonstrate that the passive-listening response to emotionally spoken words in the right mid-STG could be explained only by a combination of various acoustic parameters, but the study did not elucidate the areas responsible for the higher-level representation of emotion.

The continuous carry-over design allowed us to isolate the neural processes engaged in lower-level stages dealing with physical differences between sounds from higher-level stages involved in vocal affect cognition. Thus, we first computed correlations of neural activity patterns that process general physical differences between stimuli from those that process more abstract emotional information independent of lower-level acoustic changes. To our knowledge, only one study (Wiethoff et al., 2008) has used multiple regression to demonstrate that the passive-listening response to emotionally spoken words in the right mid-STG could be explained only by a combination of various acoustic parameters, but the study did not elucidate the areas responsible for the higher-level representation of emotion.

The continuous carry-over design allowed us to isolate the neural processes engaged in lower-level stages dealing with physical differences between sounds from higher-level stages involved in vocal affect cognition. Thus, we first computed correlations of neural activity patterns that process general physical differences between stimuli from those that process more abstract emotional information independent of lower-level acoustic changes. To our knowledge, only one study (Wiethoff et al., 2008) has used multiple regression to demonstrate that the passive-listening response to emotionally spoken words in the right mid-STG could be explained only by a combination of various acoustic parameters, but the study did not elucidate the areas responsible for the higher-level representation of emotion.

The continuous carry-over design allowed us to isolate the neural processes engaged in lower-level stages dealing with physical differences between sounds from higher-level stages involved in vocal affect cognition. Thus, we first computed correlations of neural activity patterns that process general physical differences between stimuli from those that process more abstract emotional information independent of lower-level acoustic changes. To our knowledge, only one study (Wiethoff et al., 2008) has used multiple regression to demonstrate that the passive-listening response to emotionally spoken words in the right mid-STG could be explained only by a combination of various acoustic parameters, but the study did not elucidate the areas responsible for the higher-level representation of emotion.

The continuous carry-over design allowed us to isolate the neural processes engaged in lower-level stages dealing with physical differences between sounds from higher-level stages involved in vocal affect cognition. Thus, we first computed correlations of neural activity patterns that process general physical differences between stimuli from those that process more abstract emotional information independent of lower-level acoustic changes. To our knowledge, only one study (Wiethoff et al., 2008) has used multiple regression to demonstrate that the passive-listening response to emotionally spoken words in the right mid-STG could be explained only by a combination of various acoustic parameters, but the study did not elucidate the areas responsible for the higher-level representation of emotion.
overlapped with voice-sensitive regions, correlated positively with an increased difference in physical properties between consecutive stimuli (for similar results with different task and design, see also Frühholz et al. (2012)). The bilateral correlation in voice-sensitive cortex was lateralized to the right, confirming previous neuropsychological studies (Van Lancker and Sidtis, 1992; Hornak et al., 1996; Shamay-Tsoory et al., 2004; Ross and Monnot, 2008) and neuroimaging studies (Imaizumi et al., 1997; Morris et al., 1999; Buchanan et al., 2000; Rama et al., 2001; Wildgruber et al., 2002; Mitchell et al., 2003; Grandjean et al., 2005; Wildgruber et al., 2005; Etohofer et al., 2006, 2009a,b; Leitman et al., 2010; Etohofer et al., 2012; for a recent meta-analysis, see Wittenman et al., 2012) that provide support for the importance of these regions in the perception of emotional prosody. While STS/STG are classically highlighted in studies on affective prosody perception, they do not seem to be specific to the processing of auditory emotion, as activations in these areas have been reported for the perception of identity (Andics et al., 2010; Latinus et al., 2011, 2013), gender (Charest et al., 2013), and attractiveness (Bestelmeyer et al., 2012). Research on auditory object perception suggests that while subregions of the STG/STS may be specific to certain sound categories, these activations may not as such be indicative of representations of the semantic category but instead of a shared acoustic profile common to a particular group of auditory objects (e.g., vocalizations, musical instruments; Lewis et al. (2009); Leaver and Rauschecker, 2010). Thus, bilateral STS/STG may be related to acoustic feature extraction of paralinguistic information more generally.

Second, to reveal the higher-level processes involved in vocal affect cognition, we computed correlations between the perceptual difference regressor (using an individual’s behavioral data) and BOLD signal after the variance explained by the physical difference regressor had been explained. We found negative quadratic correlations in bilateral insulae (with IFG), mid-cingulum, superior frontal gyrus, and precentral gyrus, and a positive correlation in bilateral precuneus. The involvement of anterior insulae has frequently been reported in the perception of a variety of emotional, multimodal stimuli and experimental tasks (Duerden et al., 2013), emphasizing its importance in the processing and possible integration of emotional information from various senses. A recent review (Craig, 2009) reveals that the response of the anterior insulae is not specific to emotion because an astonishing number of studies covering a range of topics (e.g., emotional awareness, error awareness, and attention to pain)

### Table 2. Areas revealing significant modulations with physical and perceptual difference from the single models

<table>
<thead>
<tr>
<th>Modulator</th>
<th>Anatomical definition</th>
<th>MNI coordinates (peak voxel)</th>
<th>t/F value</th>
<th>Cluster size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single physical difference</td>
<td>Right superior temporal gyrus</td>
<td>63</td>
<td>25</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Right superior temporal gyrus</td>
<td>57</td>
<td>24</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Left superior temporal gyrus</td>
<td>60</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Left superior temporal gyrus</td>
<td>66</td>
<td>22</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Left superior temporal gyrus</td>
<td>51</td>
<td>43</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Left putamen</td>
<td>21</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Precuneus</td>
<td>9</td>
<td>49</td>
<td>37</td>
</tr>
<tr>
<td>Single perceptual difference</td>
<td>Right amygdala</td>
<td>24</td>
<td>7</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Left supplementary motor area</td>
<td>-3</td>
<td>11</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>Left anterior cingulum</td>
<td>23</td>
<td>25</td>
<td>13.21</td>
</tr>
<tr>
<td></td>
<td>Right insula</td>
<td>33</td>
<td>26</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Left insula</td>
<td>30</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Left putamen</td>
<td>21</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Left putamen</td>
<td>18</td>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Left thalamus</td>
<td>12</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Left mid-occipital</td>
<td>52</td>
<td>16</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Left angular gyrus</td>
<td>57</td>
<td>64</td>
<td>25</td>
</tr>
<tr>
<td></td>
<td>Left angular gyrus</td>
<td>54</td>
<td>61</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Right precentral gyrus</td>
<td>45</td>
<td>5</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>Right inferior frontal gyrus</td>
<td>54</td>
<td>8</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Right precentral gyrus</td>
<td>48</td>
<td>43</td>
<td>13.42</td>
</tr>
<tr>
<td></td>
<td>Left precentral gyrus</td>
<td>21</td>
<td>29</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>Left mid-frontal gyrus</td>
<td>30</td>
<td>17</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Left precentral gyrus</td>
<td>42</td>
<td>2</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Left rolandic operculum</td>
<td>48</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Left precentral gyrus</td>
<td>54</td>
<td>5</td>
<td>37</td>
</tr>
<tr>
<td></td>
<td>Left supramarginal gyrus</td>
<td>66</td>
<td>22</td>
<td>16</td>
</tr>
<tr>
<td></td>
<td>Left superior temporal gyrus</td>
<td>57</td>
<td>19</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Left post-central gyrus</td>
<td>60</td>
<td>16</td>
<td>22</td>
</tr>
<tr>
<td></td>
<td>Left inferior parietal area</td>
<td>45</td>
<td>-28</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Left inferior parietal area</td>
<td>33</td>
<td>-37</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Left inferior parietal area</td>
<td>39</td>
<td>-40</td>
<td>52</td>
</tr>
<tr>
<td></td>
<td>Left precuneus</td>
<td>9</td>
<td>-58</td>
<td>28</td>
</tr>
<tr>
<td></td>
<td>Left precuneus</td>
<td>6</td>
<td>-46</td>
<td>43</td>
</tr>
<tr>
<td></td>
<td>Left mid-cingulum</td>
<td>-3</td>
<td>28</td>
<td>40</td>
</tr>
<tr>
<td></td>
<td>Left precentral gyrus</td>
<td>30</td>
<td>-7</td>
<td>46</td>
</tr>
<tr>
<td></td>
<td>Left mid-temporal gyrus</td>
<td>-51</td>
<td>40</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>Left superior temporal gyrus</td>
<td>-51</td>
<td>43</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td>Left superior temporal gyrus</td>
<td>-63</td>
<td>46</td>
<td>13</td>
</tr>
</tbody>
</table>

Bestelmeyer et al. • Multistep Perception of Vocal Affect
J. Neurosci., June 11, 2014 • 34(24):8098 – 8105 • 8103
highlight activation of the anterior insulae with no consistent coactivation of other areas. The anterior insulae are therefore now thought of as a correlate of awareness or consciousness (Craig, 2009).

The precuneus and its functional significance have only recently received attention. Its anatomical location makes it hard to study, and it is therefore a comparatively less accurately mapped area of cortex (Cavanna and Trimble, 2006). The precuneus has widespread reciprocal corticocortical and subcortical connections. The principal connections outside parietal cortex are with the frontal lobes (including supplementary motor areas) and insulae, and it is noteworthy that the precuneus is not directly connected to any primary sensory area (Margulies et al., 2009). This pattern of connections implies that the precuneus is a major association area that may underpin a variety of different behaviors (Cavanna and Trimble, 2006). An inverse relationship between activity in anterior insulae and activity in precuneus has been associated with task-related attention (Dosenbach et al., 2007). We see a similar inverse pattern of activation in precuneus compared with anterior insulae and mid-cingulum, suggesting that these areas could be related to the attentional demands of the task rather than emotion perception per se. Importantly, albeit at a more lenient threshold, we observed activation in bilateral subcortical structures (amygdalae/hippocampi and thalami), all of which have been traditionally implicated in emotion perception.

The quadratic response to the perceptual difference was unexpected. In bilateral insulae, identical and maximally different morphs of precentral gyri and mid-cingulum produced lower responses than when it was less clear whether the morph belonged to the same category as the preceding one. Such complex response forms to continuous changes in stimulus features may be a property of higher-order association cortex and are an important topic for future investigation. While unexpected, these responses have been reported previously in visual psychophysics studies (Aguiire, 2007; Drucker et al., 2009) as well as a study on voice gender (Charest et al., 2013). Intuitively, these quadratic effects could be related to the experimental task, although these have been very different across studies (orthogonal vs active tasks).

We used nonverbal stimuli and a relatively simple task. Nevertheless our data suggest a dissociation between the processing of acoustic and cognitive aspects. It is plausible that the areas responsive to perceptual differences in affect are supramodal. In fact, a recent behavioral adaptation study suggests that there might be neurons which code expressions independently of modality (Skuk and Schweinberger, 2013). There are additional aspects which deserve future attention. First, the perceptual difference regressor was based on a simple categorization task and could be fine-tuned by incorporating more dimensions of emotion, e.g., intensity. We opted, in the first instance, to use data that can be obtained from the participant during scanning. Second, due to the nonverbal nature of our stimuli an interesting extension of this study could investigate the complex interactions between speech content and vocal affect. Furthermore, while our analytical approach was appropriate for teasing apart the processes dealing with acoustic and evaluative aspects of vocal expressions, it is limited in that it assumes our regressors of interest are orthogonal (when in fact they are to some extent correlated). As such, some of the variance attributed to the physical difference regressor is neither purely physical nor purely perceptual but shared. Therefore a strict dichotomy between cerebral effects driven by physical or perceptual features of the sounds is oversimplified.

Our results support the core idea described in Schirmer and Kozt (2006) that vocal affect recognition is a multistep process. Our findings extend that idea by identifying largely distinct regions of the emotion network for the coding of acoustic features of vocalization and for the processing of the cognitive representation of emotion. While the model originally proposes right IFG and orbitofrontal cortex as the hubs for evaluative judgments and left IFG for the processing of semantic content of emotional prosody, our findings highlight the involvement of a more complex, largely bilateral set of regions necessary to perform evaluative judgments of emotional vocalizations even in the absence of speech.

References

Etheoer T, Kreifelts B, Wietthoff S, Wolf J, Grodd W, Vuilleumier P, Wildgru-