Social Cognitive and Affective Neuroscience, 2016, 821–828

doi: 10.1093/scan/nsw022
Advance Access Publication Date: 16 February 2016
Original article

Neural basis of processing threatening voices in a crowded auditory world

Martin Mothes-Lasch,[1] Michael P. I. Becker,[1] Wolfgang H. R. Miltner,[2] and Thomas Straube[1]

[1]Institute of Medical Psychology and Systems Neuroscience, University Hospital Muenster, Von-Esmarch-Str. 52, D-48149 Münster, Germany and [2]Department of Biological and Clinical Psychology, Friedrich-Schiller-University Jena, Am Steiger 3//1, D-07743 Jena, Germany

Correspondence should be addressed to Thomas Straube, Institute of Medical Psychology and Systems Neuroscience, University Hospital Muenster, Von-Esmarch-Str. 52, D-48149 Münster, Germany. E-mail: thomas.straube@uni-muenster.de.

Abstract

In real world situations, we typically listen to voice prosody against a background crowded with auditory stimuli. Voices and background can both contain behaviorally relevant features and both can be selectively in the focus of attention. Adequate responses to threat-related voices under such conditions require that the brain unmixes reciprocally masked features depending on variable cognitive resources. It is unknown which brain systems instantiate the extraction of behaviorally relevant prosodic features under varying combinations of prosody valence, auditory background complexity and attentional focus. Here, we used event-related functional magnetic resonance imaging to investigate the effects of high background sound complexity and attentional focus on brain activation to angry and neutral prosody in humans. Results show that prosody effects in mid superior temporal cortex were gated by background complexity but not attention, while prosody effects in the amygdala and anterior superior temporal cortex were gated by attention but not background complexity, suggesting distinct emotional prosody processing limitations in different regions. Crucially, if attention was focused on the highly complex background, the differential processing of emotional prosody was prevented in all brain regions, suggesting that in a distracting, complex auditory world even threatening voices may go unnoticed.

Key words: prosody; fMRI; attentional gating; emotion; amygdala; temporal cortex

Introduction

No matter whether you are in New York or Siberia, hearing an angry voice never is an irrelevant signal for human beings. In general, detection and evaluation of emotional prosody are critical for the appreciation of the social and physical environment, since prosodic information indicates safety or potential danger. Accordingly, angry emotional voices represent a signal of social threat. Thus, the fast detection and prioritized processing of threat-related voices is of fundamental importance, and neural responses to angry prosody should reflect this. Indeed, brain areas such as auditory superior temporal cortex regions (STR) and amygdala show increased responses to angry as compared to neutral voices (Grandjean *et al.*, 2005; Sander *et al.*, 2005; Frühholz and Grandjean, 2013) relatively independent from attentional demands (Grandjean *et al.*, 2005; Sander *et al.*, 2005; Quadflieg *et al.*, 2008; Ethofer *et al.*, 2009), at least if presented unimodally (Mothes-Lasch *et al.*, 2011; Mothes-Lasch *et al.*, 2012).

Neuroimaging findings suggest that STR is a key region in the representation of voice prosody (Grandjean *et al.*, 2005; Ethofer *et al.*, 2012; Bestelmeyer *et al.*, 2014). In addition to the encoding of physical voice dimensions such as fundamental frequency (Mesgarani *et al.*, 2014), STR is also implied in coding attentional aspects of voice perception (Mesgarani and Chang, 2012; Leonard and Chang, 2014). Furthermore, a posterior-anterior gradient in STR in voice and prosody processing is discussed, for example in relation to auditory object identification (Schirmer and Kotz, 2006; Leonard and Chang, 2014).

Received: 18 August 2015; **Revised:** 27 November 2015; **Accepted:** 11 February 2016

The amygdala has been identified as a central structure associated with processing of threat-related stimuli (LeDoux, 2000; Ohman and Mineka, 2001; Ohman, 2005; Vuilleumier, 2005). The finding of increased responses to angry voices in the amygdala is in accordance with its proposed role in the detection of threat, in the initiation of defense behavior and in the guidance of attentional processes (LeDoux, 2000; Adolphs *et al.*, 2005; Gamer and Büchel, 2009).

But how robust are findings for the amygdala and STR if angry voices are embedded in auditory backgrounds of varying complexity? In complex real-world scenarios, threatening voices are part of crowded auditory scenes, and voices as well as background may be in the focus of a listener's attention. Under these conditions, neural representation of threatening voices requires complex underlying feature extraction and isolation processes (Darwin, 2008). Especially in the case of complex auditory environments, guidance of selection processes by attentional focus should become even more important for isolation of prosodic voice features. However, it is unknown to what degree neural processing of emotional prosody is modulated by background complexity, attentional focus or the interaction of both factors under real world scenarios.

To answer these questions, the present event-related functional magnetic resonance imaging (fMRI) study aimed at elucidating the respective effects of crowded backgrounds and attentional focus on processing angry voices. We hypothesized that activation in the amygdala and auditory cortex is modulated by attention and background complexity. Specifically, we expected that the combination of external attention and high load prevents amygdala activation. To this end, angry or neutral voices were presented simultaneously with low or high complex auditory scenes that were not related to emotional voices. By task instruction, participants' attention was directed toward voices or it was directed away from voices toward the auditory background during an auditory perceptual decision task.

Materials and methods

Participants

Twenty right-handed healthy subjects (12 women, 8 men, all aged 20–33 years, mean age 25.7 years) with normal or corrected-to-normal vision and normal hearing participated in the study. All participants were right-handed German native speakers and had no history of neurological or psychiatric disorders. Right-handedness was assessed using the Edinburgh Inventory (Oldfield, 1971). Normal hearing was assessed during recruitment by asking participants if their hearing was impaired or if they needed hearing aids. Participants provided written informed consent for the study, which has been approved by the ethics committee of the University of Jena.

Stimuli

Prosodic stimuli consisted of an established set of semantically neutral bi-syllabic nouns spoken in either angry or neutral prosody by two women and two men (Quadflieg *et al.*, 2008; Mothes-Lasch *et al.*, 2011). Stimuli were recorded and digitized with an audio interface of 44.1 kHz sampling rate and 16 bit resolution. Utterances were edited to a common length of 550 ms (Adobe Audition 1.5, San Jose, CA) and evaluated by a naïve sample (Mothes-Lasch *et al.*, 2011). Easily distinguishable animal sounds from the database of the International Affective Digitized Sounds (IADS, Bradley and Lang, 1999) of a cow (No. 113) and a cat (No. 102) served as targets of the auditory perceptual decision task. Complexity was increased by means of four other animal sounds [Bees (No. 115), Rooster (No. 120), Pig (No. 130) and Chickens (No. 132)] that were presented along with the two target sounds. From these sound files, lasting approximately 6 s, representative sequences of 850 ms were selected. The low complexity condition comprised either the cat or cow sound and the high complexity condition comprised either the cat or the cow sound merged with the other animal sound files (Audacity, http://audacity.sourceforge.net). These four animal sound files, as well as all emotional words, were intensity normalized (70 dB) using Praat (www.praat.org).

Experimental design and task

During scanning, auditory stimuli were presented binaurally via headphones that were specifically adapted for the use in fMRI environments (Commander XG MRI audio system, Resonance Technology, Northridge, CA). The instructions were shown via a back-projection screen onto an overhead mirror. In separate runs, participants were instructed either to identify the gender of the speaker (voices attended) or to determine whether the sound of a cow or a cat was presented (background attended) (Figure 1). Responses were made by pressing one of two buttons on a response box (LUMItouch; Photon Control) connected to a PC via fiber optic cable. It was explicitly pointed out to the participants that all four of the animal sound files contained always either a cat or a cow. Prior to scanning, participants were familiarized with the tasks using eight trials with words not presented in the experiment proper. After that, the sound volume was adjusted to a comfortable level. Each word was presented with all four of the animal sound files in each task. Thus, each task comprised 160 trials presented in randomized order in two separate successive runs. Stimulus onset asynchrony ranged between 2360 ms and 5960 ms with a mean of 4150 ms. The order of the tasks and assignment of the buttons to the answer alternatives were counterbalanced across participants. Presentation of the stimuli as well as recording of responses was controlled by Presentation software (version 14, Neurobehavioral Systems, Inc., Albany, CA). Because of equipment malfunction, performance data of one person in the gender task were not recorded. Behavioral data were analyzed by means of 2(prosody) × 2(attentional focus) × 2(complexity) repeated-measures analysis of variance (ANOVA) and post-hoc pairwise *t*-tests using IBM SPSS software (Version 19; SPSS INC., an IBM Company, Chicago, IL). A probability level of $P < 0.05$ was considered statistically significant. All data are reported as mean ± standard deviation.

fMRI data acquisition and analysis

Scanning was performed in a 3 T magnetic resonance scanner (Magnetom TIM Trio; Siemens Medical Systems, Erlangen, Germany). After acquisition of a T1-weighted anatomical scan (TE = 3.03 ms, flip angle = 9°, matrix = 256 × 256, field of view = 256 mm, TR = 2300 ms, 192 slices, thickness = 1 mm), two runs of T2*-weighted echo-planar images consisting of 330 volumes were acquired (TE = 30 ms, flip angle = 90°, matrix = 64 × 64, field of view = 192 mm, TR = 2080 ms). Each volume comprised 35 axial slices (thickness = 3 mm, gap = 0.5 mm, in-plane resolution = 3 × 3 mm). The slices were acquired parallel to the line between anterior and posterior commissure with a tilted orientation to reduce susceptibility artifacts in inferior parts of the anterior brain (Deichmann *et al.*,

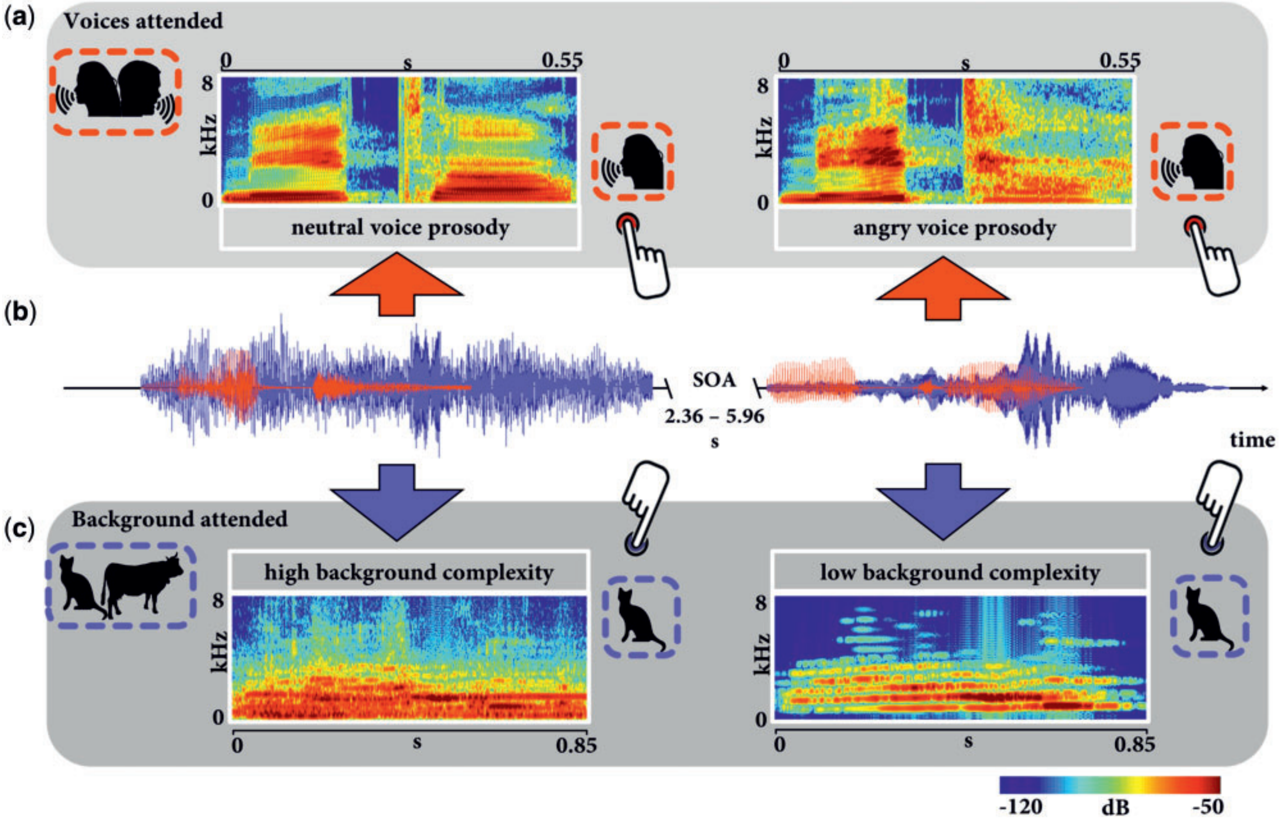

Fig. 1. Manipulation of attentional focus by task and exemplification of sound stimuli. (a) Guiding attention toward voices in gender decision task (see Materials and Methods): spectrograms of the word 'Meter' (['me:te]) phonated in neutral and angry prosody by the same female speaker. (b) Acoustic waveforms of sound stimuli as shown in a (red, voices) and c (purple, backgrounds). (c) Guiding attention toward backgrounds in animal decision task: spectrograms of a background scene in high (left) or low complexity (right).

2003). Furthermore, before imaging, a shimming procedure was performed to improve field homogeneity. The first four volumes of each run were discarded from analysis to ensure steady-state tissue magnetization.

Preprocessing and analyses of the functional data were performed with Brain Voyager QX software (Version 2.4; Brain Innovation, Maastricht, The Netherlands). First, all volumes were realigned to the first volume in order to minimize artifacts resulting from head movements. Subsequently, a slice-timing correction was conducted. Further data preprocessing comprised spatial (8-mm full-width half-maximum isotropic Gaussian kernel) as well as temporal smoothing (high pass filter: 10 cycles per run; linear trend removal). The anatomical and functional images were coregistered and normalized to Talairach space (Talairach and Tournoux, 1988).

Statistical analyses were performed by multiple linear regression of the signal time course at each voxel. The expected BOLD signal change for each event type (= predictor) was modeled by a hemodynamic response function. Within-group statistical comparisons were conducted using a mixed effect analysis, which considers inter-subject variance and permits population-level inferences. In the first step, predictor estimates were generated for each individual. In the second step, predictor estimates were analyzed across subjects. The analysis was conducted in two regions of interest (ROI), the STR, including the superior temporal gyrus and transverse temporal gyrus, and the amygdala. ROIs were defined by the Talairach daemon parameters (Lancaster *et al.*, 2000) available in the WFU PickAtlas (Maldjian *et al.*, 2003) without 3D dilation. Statistical parametric maps resulting from the voxel-wise analysis were considered statistically significant for values that survived a cluster-based correction for multiple comparisons as implemented in Brain Voyager (Goebel *et al.*, 2006), which is based on a 3D extension of the randomization procedure described by Forman *et al.* (1995). First, voxel-level threshold was set at $P < 0.001$ (uncorrected). Thresholded maps were then submitted to a ROI-based correction criterion based on the estimate of the map's spatial smoothness and on an iterative procedure (Monte Carlo simulation) for estimating cluster-level false-positive rates. After 1000 iterations, the minimum cluster size threshold that yielded a cluster-level false-positive rate of 5% was applied to the statistical maps. The splitclustercoords function of Neuroelf (v0.9c, http://neuroelf.net/) was used to assess local maxima of clusters.

Results

Performance data

A 2(Attentional focus) × 2(Prosody) × 2(Complexity) ANOVA of reaction times (RT) and error rates (ER) revealed main effects for Prosody (RT: $F_{(1,18)} = 11.3$, $P < 0.05$; ER: $F_{(1,18)} = 5.3$, $P < 0.05$), Complexity (RT: $F_{(1,18)} = 38.2$, $P < 0.05$; ER: $F_{(1,18)} = 91.4$, $P < 0.05$) and Attentional focus (RT: $F_{(1,18)} = 61.8$, $P < 0.05$; ER: $F_{(1,18)} = 45.3$, $P < 0.05$). RT were slower and ER higher for angry compared to neutral prosody, for high compared to low background complexity and for the animal compared to the gender decision task (see Table 1 for descriptive data). Furthermore, both RT and ER

Table 1. Behavioral data

Performance measures	Attentional focus			
	Gender		Animal	
	Anger	Neutral	Anger	Neutral
RT (in ms)				
High complexity	870 (s.d. 255)	837 (s.d. 248)	1282 (s.d. 288)	1292 (s.d. 284)
Low complexity	856 (s.d. 216)	804 (s.d. 206)	1017 (s.d. 164)	1003 (s.d. 167)
ER (in %)				
High complexity	6.1 (s.d. 7.3)	2.1 (s.d. 3.0)	30.1 (s.d. 12.1)	30.4 (s.d. 14.8)
Low complexity	5.0 (s.d. 5.4)	1.9 (s.d. 3.6)	2.1 (s.d. 2.5)	2.5 (s.d. 2.8)

Note. RT and ER to angry and neutral prosody presented for backgrounds with high or low complexity in both tasks.

showed significant interactions of Attentional focus by Prosody (RT: $F_{(1,18)} = 11.0$, $P < 0.05$; ER: $F_{(1,18)} = 5.2$, $P < 0.05$) and of Attentional focus by Complexity (RT: $F_{(1,18)} = 28.0$, $P < 0.05$; ER: $F_{(1,18)} = 93.2$, $P < 0.05$). Post hoc analysis showed that differential effects for RT and ER depending on prosody were significant in the gender decision task (RT: $t_{(18)} = 6.00$, $P < 0.05$, ER: $t_{(18)} = -3.52$, $P < 0.05$) but not in the animal decision task (RT: $t_{(19)} = 0.17$, $P = 0.87$, ER: $t_{(19)} = 0.28$, $P = 0.78$, see Table 1). Differential effects for RT and ER depending on background were only significant in the animal decision task (RT: $t_{(19)} = 6.77$, $P > 0.05$, ER: $t_{(19)} = -10.27$, $P > 0.05$; gender decision task: RT: $t_{(18)} = 1.16$, $P = 0.26$, ER: $t_{(19)} = -1.10$, $P = 0.29$, see Table 1). Neither the interaction of Complexity by Prosody (RT: $F_{(1,18)} = 1.0$, $P > 0.05$; ER: $F_{(1,18)} = 0.28$, $P > 0.05$) nor the interaction of Attentional focus by Complexity by Prosody (RT: $F_{(1,18)} = 0.004$, $P > 0.05$; ER: $F_{(1,18)} = 0.002$, $P > 0.05$) was statistically significant.

fMRI data

A three-way ANOVA with factors complexity, attentional focus and prosody revealed that, over and above of bilateral main effects of prosody in mid STR (mSTR; left: $F_{(1,19)} = 32.6$, $P < 0.05$, corr.; right: $F_{(1,19)} = 29.5$, $P < 0.05$, corr.), a pronounced interaction of prosody and complexity proved significant in this region (left peak x, y, z: −60,−16,10, $F_{(1,19)} = 72.6$, cluster size 9504 mm^3, $P < 0.05$, corr.; right peak x, y, z: 60,−16,16, $F_{(1,19)} = 97.6$, cluster size 11 529 mm^3, $P < 0.05$, corr.). This interaction clearly showed that differential prosody encoding in mSTR (increased responses to angry as compared to neutral prosody) is only detectable under low (left: $t_{(19)} = 6.0$, $P < 0.05$, corr.; right: $t_{(19)} = 5.5$, $P < 0.05$, corr.) but not under high background complexity (left: $t_{(19)} = -0.5$, $P = 0.91$; right: $t_{(19)} = -1.4$, $P = 0.96$). While the locus of maximal activation of this interaction cluster ($[\text{angry-neutral}]_{\text{low complexity}} > [\text{angry-neutral}]_{\text{high complexity}}$) lay in mSTR (Figure 2), the cluster also comprised more posterior aspects of STR (Table 2). Interactions of prosody and attentional focus were found in anterior STR (aSTR; peak x, y, z: −45,−2,−14, $F_{(1,19)} = 32.8$, cluster size 702 mm^3, $P < 0.05$, corr.; Figure 3) and amygdala (peak x, y, z:−21,−10,−14, $F_{(1,19)} = 31.8$, cluster size 162 mm^3, $P < 0.05$, corr.; Figure 3). Increased responses to angry as compared to neutral prosody in these regions were only observed if participants focused their attention on voices (aSTR: $t_{(19)} = 4.7$, $P < 0.05$, corr.; amygdala: $t_{(19)} = 4.3$, $P < 0.05$, corr.) but not if they had to attend the background (aSTR: $t_{(19)} = -0.9$, $P > 0.05$; amygdala: $t_{(19)} = -0.6$, $P > 0.05$). These interaction clusters ($[\text{angry-neutral}]_{\text{voices attended}} > [\text{angry-neutral}]_{\text{background attended}}$, Figure 3) did not show additional local maxima. If complexity was high and voices were unattended, neither amygdala nor STR showed differential prosody processing even at very liberal thresholds ($P = 0.05$, uncorrected).

Discussion

This study was designed to investigate the processing of threat-related voices in crowded auditory environments. Irrespective of background complexity, we found stronger activation of the amygdala and aSTR to angry *vs* neutral voices if the voices were attended but not if voices were unattended. Conversely, we observed stronger activation of the mSTR to angry *vs* neutral voices under low but not under high auditory complexity, independently of attentional focus. If the highly complex auditory background was in the focus of attention, no brain region responded differentially to angry *vs* neutral prosody. These findings suggest that emotional prosody processing in mSTR on the one hand and amygdala and aSTR on the other hand are subject to different sets of capacity limitations, respectively.

Differential encoding of emotional prosody in amygdala and aSTR required focused attention to the voice-related auditory stream regardless of complexity of the competing auditory input. Thus, particularly amygdalar gating of angry voices may allow to alert the listener even in crowded auditory environments. Yet, without sufficient attentional ressources, threatening voices are not necessarily salient enough to activate the amygdala. These results show that the activation of the amygdala is not automatically triggered by auditory threat-related stimuli. Rather, attention must be directed to the auditory stream containing the relevant acoustical features during the presence of distracting input. This conclusion is further supported by recent cross-modal studies (Mothes-Lasch *et al.*, 2011, 2012) in which emotional voices and visual symbols were presented simultaneously, while the attentional focus was either directed to voices or to visual symbols. Although stronger activation of the amygdala to emotional voices was found if the voices were attended, this effect was absent if voices were not attended (Mothes-Lasch *et al.*, 2011, 2012).

In agreement with previous studies, aSTR showed a similar activation profile as the amygdala, underlining this region's involvement in processing of emotional prosody as part of an auditory 'what'-pathway (Schirmer and Kotz, 2006; Wiethoff *et al.*, 2008; Frühholz *et al.*, 2012). In particular, constancy of prosodic features across individual voice exemplars, which usually differ in characteristics such as fundamental frequencies and formant variability (Latinus *et al.*, 2013), is thought to be represented here (Frühholz and Grandjean, 2013). Hence, in the present task activation of aSTR could represent decoding of

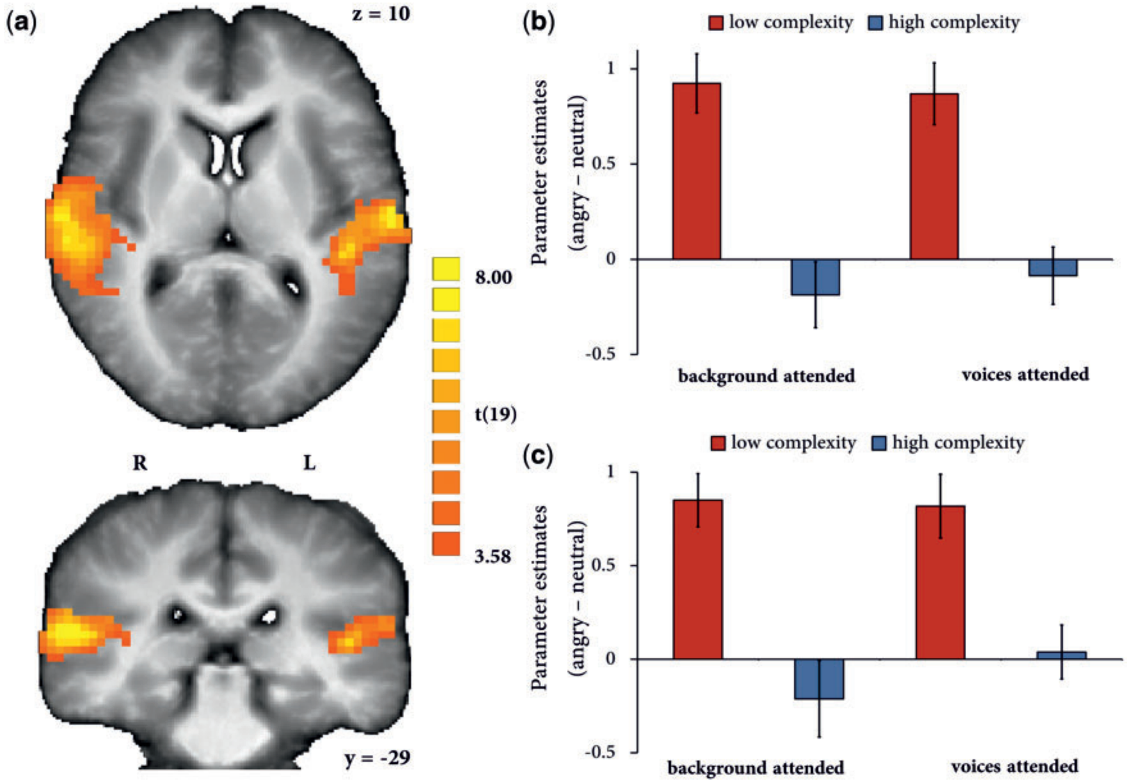

Fig. 2 Encoding of voice prosody in mSTR is prevented by highly complex auditory background environments and restricted to low complexity conditions. (a) Bilaterally in mSTR, *t*-maps (thresholded at $P < 0.001$) indicate that activation of this region is gated by background complexity and not by attentional focus. Bar graphs of the mean difference of parameter estimates between angry and neutral prosody (±SEM) of the activated clusters in right (b) and left mSTR (d) show that differential encoding of voice prosody only takes place if background complexity is low, independent of attentional focus.

Table 2. Coordinates of global and local maxima of the interaction cluster (([angry-neutral]$_{\text{low complexity}}$ > [angry-neutral]$_{\text{high complexity}}$) in mSTR depicted in Figure 2

Global and local maxima in mSTR clusters	Laterality	Talairach coordinates			mm^3	$t_{(19)}$
		x	y	z		
Right mSTR						
Transverse temporal gyrus (global maximum)	R	60	−16	16	17 091	9.9
Superior temporal gyrus (local maximum)	R	60	−28	13	1053	9.1
Left mSTR						
Transverse temporal gyrus (global maximum)	L	−60	−16	10	12 420	8.5
Superior temporal gyrus (local maximum)	L	−45	−31	13	810	7.8
Superior temporal gyrus (local maximum)	L	−57	−31	19	405	6.8
Superior temporal gyrus (local maximum)	L	−51	−4	1	1134	6.2
Superior temporal gyrus (local maximum)	L	−42	−43	16	162	4.6

prosodic voice-features that are independent of the physical differences in male and female speaker's voices. It is possible that aSTR differs functionally from the amygdala, in that the former extracts general prosodic features from voice exemplars, while the latter detects aspects of biological significance independently of specific voices (Pessoa and Adolphs, 2010).

Remarkably, while amygdala and aSTR are able to resolve complex auditory scenes and to encode angry prosody under conditions of voice-focused attention, the activation of mSTR to angry *vs* neutral voices depends on stimulus complexity but not on attentional focus. Thus, the auditory areas in STR differentiate between angry and neutral prosody only in auditory environments of low complexity, comprising few figurative features. Stronger brain activation to angry *vs* neutral voices of the mSTR under the low complexity condition is assumed to reflect the activation of representations of behaviorally relevant auditory objects (Nelken *et al.*, 2003; King and Nelken, 2009). The increased representation of angry prosody, however, seems to be impaired by simultaneous, competing complex auditory stimuli regardless of the task relevance of the competing auditory stimuli. Results from human electrophysiology (Mesgarani and Chang, 2012) and neuroimaging (Mothes-Lasch *et al.*, 2011) converge with our finding that voice representations within mSTR are subject to capacity limitations, which reflect the

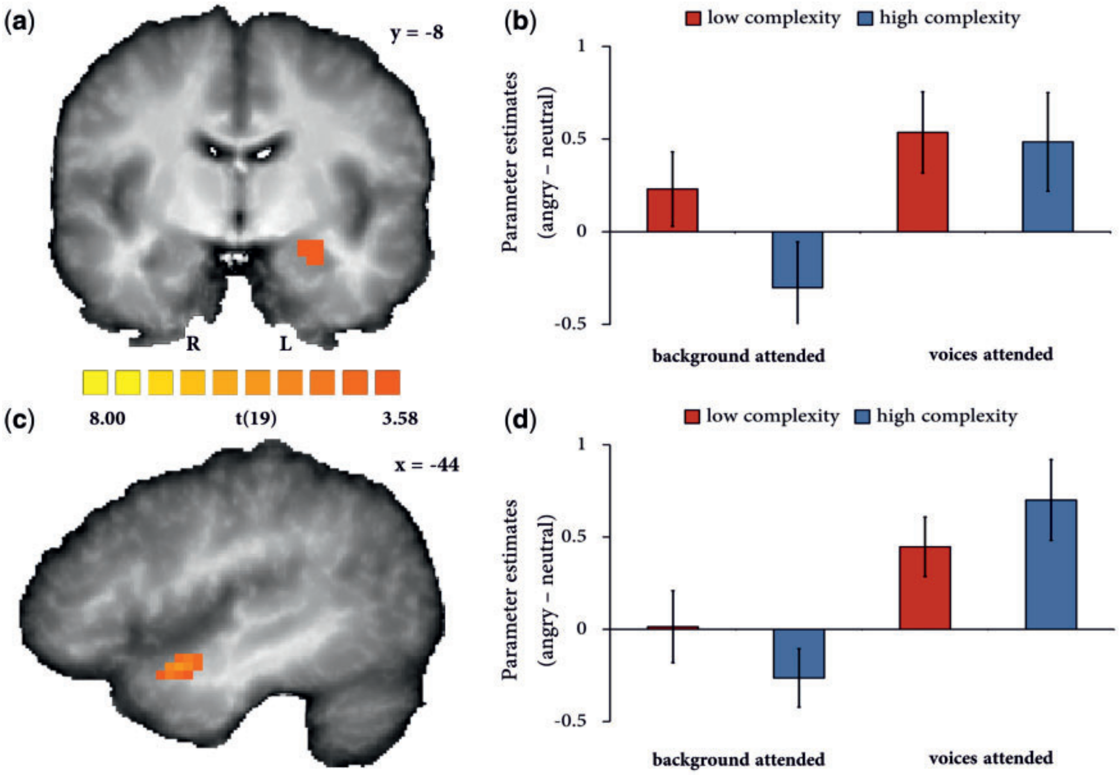

Fig. 3 Attentional gating of encoding of voice prosody in amygdala (a,b) and aSTR (c,d). In both regions, voice prosody encoding occurs irrespective of the complexity of auditory background if voices are attended to, but not if background is attended to. All *t*-maps of significant clusters are overlaid on averaged brains (a,c) and thresholded at $P < 0.001$. Bar graphs show mean difference of parameter estimates between angry and neutral prosody (±SEM) of the activated clusters in left amygdala (b) and left aSTR (d) search regions for high (red) and low (blue) complexity while either background or voices were attended (see Materials and Methods).

content- and goal-specific allocation of limited processing ressources. Our study, however, implies that prosody processing in mSTR is altered not only during focused attention to auditory background. Even if attention is focused on voices, the complexity of auditory background strongly modulates responses to emotional prosody in the auditory cortex. Thus, the capacity limitations in mSTR appear to be subject to the number of interacting sound features reciprocally masking themselves, possibly reflecting highly overlapping energy profiles of most single acoustic features within the sound mixtures across the spectral and temporal domains (Mesgarani and Chang, 2012). While not in the focus of analysis, the cluster comprising mSTR activation also comprises voxels in more posterior regions of STR, particularly in the left hemisphere. These posterior regions have previously been implied in target localization during presentation of complex auditory scenes (Zündorf *et al.*, 2014). Activation of these posterior regions to voice prosody during low complexity might reflect unimpaired segregation of auditory scenes.

In addition to the reported dissociation between different brain regions, we found also that the experimental combination of complex background and attentional focus on background prevented differential processing of angry *vs* neutral voices in all brain regions. This finding clearly supports accounts that stipulate capacity limitations for automatic processes even if the organism is confronted with threat. In particular, activation of the amygdala is not independent from awareness and attention and might arbitrate between processing of saliency and non-saliency features rather than reflect automatic processing, per se (Pessoa, 2014).

We would like to address some limitations of our study. First, we have used a limited set of background stimuli, which only reflects excerpts of the range of possible auditory environmental stimuli. While the set consisted of ecologically meaningful stimuli, it still might have captured the attentional focus to a higher degree than other environmental sounds. Furthermore, we cannot fully exclude the possibility that in some instances specific mixtures of voice stimuli and background stimuli might have been perceived as incongruent by the listeners. Thus, angry voices in the context of animal sounds might produce stronger incongruity effects than neutral prosody. However, the pattern of results shows a modulation of prosody by background complexity in STR, an effect which is inconsistent with a simple explanation of the above mentioned incongruity effect. Second, while it would be interesting to pinpoint the dynamics of the precise mechanisms at play during this attention-based attenuation of STR activation, the temporal resolution of our data does not allow for reliable differentiation of early from late attentional subprocesses. It might be worthwhile in future studies to concurrently record EEG during fMRI acquisition to use the more fine-grained temporal resolution of EEG to supplement analyses of fMRI data (Becker *et al.*, 2014). By EEG-informed modeling of BOLD responses, future studies might in particular investigate differential association of early and late potentials with amygdala and STR activation. Third, the continuous acquisition protocol used here did not include a 'silent' gap with gradient fields switched off. Despite the high-level of acoustical shielding realized by the headphones used in this study, we cannot fully exclude the possibility that scanner noise contributed to the sound signal mixtures under investigation.

However, given the randomization of the design and the jittering of stimulus onsets relative to slice acquisition times, it is unlikely that this factor confounded any of the effects reported here in a systematical manner.

Conclusions

This study investigated the roles of attentional focus and auditory scene complexity on brain responses to angry voices. We found a dissociation of brain responses with stronger activation of the amygdala to attended emotional voices but not to unattended emotional voices, irrespective of auditory background. Conversely, we found stronger activation to angry voices in the mSTR under low but not under high auditory complexity, irrespective of the attentional focus. Thus, the combination of crowding and attentional distraction prevented the differential processing of angry and neutral prosody in both regions. These findings imply that under real-world search conditions at least two distinct brain regions mediate the representation of threatening voices and that a combination of different sources of capacity limitations may effectively preclude the processing of angry voices in the human brain.

Acknowledgements

We thank Kathrin Blume, Christiane Fritz, Claudia Heller, Flora Mehrabi, Volkan Yildirim and Holger Hecht for help during acquisition and analysis of the data.

Funding

This research was supported by grants from the German Research Foundation (DFG project numbers: STR 987/3-1 & SFB/TRR-58, C07 & FOR 1097: Person Perception).

Conflicts of interest. None declared.

References

Adolphs, R., Gosselin, F., Buchanan, T.W., Tranel, D., Schyns, P., Damasio, A.R. (2005). A mechanism for impaired fear recognition after amygdala damage. *Nature*, **433**, 68–72.

Becker, M.P., Nitsch, A.M., Miltner, W.H., Straube, T. (2014). A single-trial estimation of the feedback-related negativity and its relation to BOLD responses in a time-estimation task. *Journal of Neuroscience*, **34**, 3005–12.

Bestelmeyer, P.E., Maurage, P., Rouger, J., Latinus, M., Belin, P. (2014). Adaptation to vocal expressions reveals multistep perception of auditory emotion. *Journal of Neuroscience*, **34**, 8098–105.

Bradley, M.M., Lang, P.J. (1999). International affective digitized sounds (IADS): Stimuli, instruction manual and affective ratings (Tech. Rep. No. B-2). University of Florida, Gainesville, FL.

Darwin, C.J. (2008). Listening to speech in the presence of other sounds. *Philosophical Transactions of the Royal Society of London Series B Biological Sciences*, **363**, 1011–21.

Deichmann, R., Gottfried, J.A., Hutton, C., Turner, R. (2003). Optimized EPI for fMRI studies of the orbitofrontal cortex. *Neuroimage*, **19**, 430–41.

Ethofer, T., Bretscher, J., Gschwind, M., Kreifelts, B., Wildgruber, D., Vuilleumier, P. (2012). Emotional voice areas: anatomic location, functional properties, and structural connections revealed by combined fMRI/DTI. *Cerebral Cortex*, **22**, 191–200.

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**, 1255–68.

Forman, S.D., Cohen, J.D., Fitzgerald, M., Eddy, W.F., Mintun, M.A., Noll, D.C. (1995). Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magnetic Resonance in Medicine*, **33**, 636–47.

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

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

Gamer, M., Büchel, C. (2009). Amygdala activation predicts gaze toward fearful eyes. *Journal of Neuroscience*, **29**, 9123–26.

Goebel, R., Esposito, F., Formisano, E. (2006). Analysis of functional image analysis contest (FIAC) data with brainvoyager QX: from single-subject to cortically aligned group general linear model analysis and self-organizing group independent component analysis. *Human Brain Mapping*, **27**, 392–401.

Grandjean, D., Sander, D., Pourtois, G., *et al.* (2005). The voices of wrath: brain responses to angry prosody in meaningless speech. *Nature Neuroscience*, **8**, 145–6.

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

Lancaster, J.L., Woldorff, M.G., Parsons, L., *et al.* (2000). Automated talairach atlas labels for functional brain mapping. *Human Brain Mapping*, **10**, 120–31.

Latinus, M., McAleer, P., Bestelmeyer, P.E., Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, **23**, 1075–80.

LeDoux, J.E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience*, **23**, 155–84.

Leonard, M.K., Chang, E.F. (2014). Dynamic speech representations in the human temporal lobe. *Trends Cognitive Science* **18**, 472–9.

Maldjian, J.A., Laurienti, P.J., Kraft, R.A., Burdette, J.H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage*, **19**, 1233–9.

Mesgarani, N., Chang, E.F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, **485**, 233–6.

Mesgarani, N., Cheung, C., Johnson, K., Chang, E.F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, **343**, 1006–10.

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

Mothes-Lasch, M., Miltner, W.H.R., Straube, T. (2012). Processing of angry voices is modulated by visual load. *Neuroimage*, **63**, 485–90.

Nelken, I., Fishbach, A., Las, L., Ulanovsky, N., Farkas, D. (2003). Primary auditory cortex of cats: feature detection or something else? *Biological Cybernetics*, **89**, 397–406.

Ohman, A. (2005). The role of the amygdala in human fear: automatic detection of threat. *Psychoneuroendocrinology*, **30**, 953–8.

Ohman, A., Mineka, S. (2001). Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychological Review*, **108**, 483–522.

Oldfield, R.C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, **9**, 97–113.

Pessoa, Luiz (2015). Précis on *The Cognitive-Emotional Brain*. *Behavioral and Brain Sciences*, **38**, e71. doi:10.1017/S0140525X14000120

Pessoa, L., Adolphs, R. (2010). Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nature Reviews Neuroscience*, **11**, 773–83.

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**, 129–37.

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**, 848–58.

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

Talairach, J., Tournoux, P. (1988). Co-planar stereotaxic atlas of the human brain. Stuttgart: Thieme.

Vuilleumier, P. (2005). How brains beware: neural mechanisms of emotional attention. *Trends in Cognitive Science*, **9**, 585–94.

Wiethoff, S., Wildgruber, D., Kreifelts, B., *et al.* (2008). Cerebral processing of emotional prosody—influence of acoustic parameters and arousal. *Neuroimage*, **39**, 885–93.

Zündorf, I.C., Lewald, J., Karnath, H.-O. (2013) Neural correlates of sound localization in complex acoustic environments. *PLoS One*, **8**(5), e64259.