OXFORD

Social Cognitive and Affective Neuroscience, 2017, 902–909

doi: 10.1093/scan/nsx020
Advance Access Publication Date: 17 February 2017
Original article

Temporal signatures of processing voiceness and emotion in sound

Annett Schirmer[1,2] and Thomas C. Gunter[1]

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

Correspondence should be addressed to Annett Schirmer, Department of Neuropsychology, Max Planck Institute for Human Cognitive and Brain Sciences, Stephanstrasse 1A, 04103 Leipzig, Germany. E-mail: aschirmer@cbs.mpg.de.

Abstract

This study explored the temporal course of vocal and emotional sound processing. Participants detected rare repetitions in a stimulus stream comprising neutral and surprised non-verbal exclamations and spectrally rotated control sounds. Spectral rotation preserved some acoustic and emotional properties of the vocal originals. Event-related potentials elicited to unrepeated sounds revealed effects of voiceness and emotion. Relative to non-vocal sounds, vocal sounds elicited a larger centro-parietally distributed N1. This effect was followed by greater positivity to vocal relative to non-vocal sounds beginning with the P2 and extending throughout the recording epoch (N4, late positive potential) with larger amplitudes in female than in male listeners. Emotion effects overlapped with the voiceness effects but were smaller and differed topographically. Voiceness and emotion interacted only for the late positive potential, which was greater for vocal-emotional as compared with all other sounds. Taken together, these results point to a multi-stage process in which voiceness and emotionality are represented independently before being integrated in a manner that biases responses to stimuli with socio-emotional relevance.

Key words: gender; sex differences; prosody; auditory cortex; implicit perception

Introduction

Of the multitude of sounds reaching our ears, the voices of other humans—especially if they are emotional—stand out. Attempting to explain this phenomenon, neuroscience has compared the processing of vocal with non-vocal and that of emotional with neutral stimuli. Resulting insights point to specialized brain mechanisms and networks underpinning representations of voiceness and emotion, respectively (for recent reviews see Schirmer *et al.*, 2016b; Schirmer and Adolphs, in press). Yet, whether and how these representations are integrated is still unexplored. Here we sought to address this issue by manipulating both voiceness and emotion in the context of an event-related potential (ERP) study.

Evidence for the special processing of voiceness comes from functional magnetic resonance imaging (fMRI) and ERPs. FMRI research has helped characterize the brain's auditory system and identified regions that are more excited by human vocalizations as compared with non-human vocalizations (Fecteau *et al.*, 2004), inanimate nature sounds, or man-made environmental noises (Belin *et al.*, 2000). These regions are located in the middle aspect of the superior temporal gyrus (STG) and sulcus (STS) and are referred to as temporal voice areas (Yovel and Belin, 2013).

ERP evidence has come from the passive oddball paradigm in which participants perform a foreground activity on the backdrop of a task-irrelevant sound sequence comprising frequent standards and rare deviants. Relative to standards, deviants elicit a mismatch negativity around 200 ms following sound onset (Näätänen *et al.*, 2007) and this negativity is larger when deviants are voiced as compared with synthesized (Schirmer *et al.*, 2007). Subsequent ERP studies presenting vocal and non-vocal sounds equiprobably identified temporally overlapping effects. They found that voiceness enhances a positive deflection around

Received: 1 October 2016; **Revised:** 17 January 2017; **Accepted:** 7 February 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

200 ms referred to as fronto-temporal positivity to voice (FTPV) (Charest *et al.*, 2009; Bruneau *et al.*, 2013) with potential sources along the STG/STS in temporal voice areas (Capilla *et al.*, 2013).

Yet unlike fMRI evidence, ERP evidence has delineated early markers of voiceness preference inconsistently (Levy *et al.*, 2001; De Lucia *et al.*, 2010; Rigoulot *et al.*, 2015). For example, a recent study comparing human voices against a range of other sounds including animal vocalizations, music and sounds from man-made objects failed to observe an overall human effect. Instead, there was an early differentiation between living and non-living sources (70 and 119 ms), followed by an enhancement for human voices relative to animal sounds (169 and 219 ms), and for music relative to other man-made objects (251 and 357 ms) (De Lucia *et al.*, 2010). As such the authors questioned the special status of the human voice.

Apart from signaling the presence of another person, social stimuli inform about that person's identity (e.g. age, sex) and mental state. Of particular interest here is the emotion content of social stimuli and how that content is processed. Both fMRI and ERP research suggest that emotionality enhances auditory representations in general and vocal representations in particular. Looking at sounds in general, there is evidence that auditory cortex, amygdala and medial prefrontal cortex are more active for positive and negative as compared with neutral conditions (Viinikainen *et al.*, 2012; Escoffier *et al.*, 2013). In the ERP, a late positive potential (LPP) is modulated by emotion. Task-relevant deviants presented in an active oddball paradigm elicit a larger LPP than standards and this effect is greater when deviants differ from standards in affect as compared with intensity (Thierry and Roberts, 2007).

Looking at voices more specifically, emotionality excites the temporal voice areas especially in the right hemisphere as well as in left inferior frontal gyrus (Kotz *et al.*, 2003; Warren *et al.*, 2006; Leitman *et al.*, 2011; Frühholz *et al.*, 2012). Perhaps surprisingly, the amygdala is rarely implicated (Ethofer *et al.*, 2007; Brück *et al.*, 2011; Mothes-Lasch *et al.*, 2011) unless more lenient statistical thresholds are used (Beaucousin *et al.*, 2007; Fecteau *et al.*, 2007). In the ERP, the LPP shows larger amplitudes for emotional as compared with neutral expressions (Pell *et al.*, 2015; Pinheiro *et al.*, 2016). Additionally, there are earlier emotion effects temporally overlapping with the FTPV. In a passive oddball paradigm, emotional relative to neutral voices enhance the mismatch negativity around 200 ms following stimulus onset (Schirmer *et al.*, 2005, 2016a). For equiprobable stimulation, a P200 modulation shows fairly consistently and may be related to the FTPV (see "Discussion" section). The P200, a centrally distributed component, differentiates between different kinds of emotional expression or is larger when voices are emotional as compared with neutral (Paulmann and Kotz, 2008; Sauter and Eimer, 2010; Schirmer *et al.*, 2013a).

Although the brain structures and mechanisms supporting social perception seem fairly universal, they differ somewhat between the sexes (Proverbio *et al.*, 2008; Schirmer *et al.*, 2013b; Proverbio and Galli, 2016). For example, sex differences have been reported for the temporal voice areas, which are larger and more voice-sensitive in women as compared with men (Ahrens *et al.*, 2014). Women are also more sensitive than men to acoustic change in vocal sounds (Schirmer *et al.*, 2007) as well as to vocal emotions. For example, when emotions are task-irrelevant, the P200 amplitude difference between emotional and neutral voices is greater in women than in men (Schirmer *et al.*, 2013a).

In sum, a substantial number of both fMRI and ERP studies have tackled the perception of voiceness and emotionality suggesting that more processing resources are dedicated towards vocal as compared with non-vocal and emotional as compared with neutral sounds. Additionally, compared with men, women seem more sensitive to voices and the emotions encoded in them. Notably, however, past research pursued voiceness and emotionality effects separately. To the best of our knowledge, both features have been manipulated within the same study only in the visual modality. In this study, participants saw positive and negative scenes that did or did not contain people (Proverbio *et al.*, 2009). Scenes with but not without people elicited a greater positivity around 100 ms when scenes were positive as compared with negative. Humanness and emotionality also interacted for a negativity peaking around 200 ms and the following LPP and these latter interactions were more prominent in female than in male participants. Thus, it seems that humanness may be processed in combination with rather than separate from emotionality and that their later and perhaps more top-down integration is stronger in women than in men.

In the present study, we sought to explore the temporal course of voiceness effects in the ERP and to determine their relation with emotionality. We presented neutral and surprised exclamations and their spectrally rotated counterparts in random order and with equal probability. The rotated stimuli, although distinctly non-human, were acoustically similar to their originals and preserved some emotionality (Scott *et al.*, 2000; Warren *et al.*, 2006; Obleser *et al.*, 2006; Sauter and Eimer, 2010). Participants detected rare sound repetitions. Our expectation was that, in line with some reports, voiceness and emotion enhance a positive component peaking around 200 ms following stimulus onset. Additionally, based on evidence from the visual modality (Proverbio *et al.*, 2009), we speculated that voiceness and emotionality interact in this component and, perhaps, subsequently. Last, we anticipated that, compared with men, women are more sensitive to voiceness and emotion.

Methods

Participants

Thirty-five participants were recruited for the experiment. Three participants were excluded from data analysis because of too many movement artifacts in the EEG. Half of the remaining participants were female with an average age of 24.8 years (s.d. = 2.9). Male participants had an average age of 25.2 years (s.d. = 2.8). All participants were right-handed and reported an absence of hearing or neurological impairments.

Stimulus materials

Twenty-seven individuals expressed 'Ah' with surprise and neutrality. Non-vocal controls were created by spectral rotation (http://www.phon.ucl.ac.uk/resource/software-other.php) as to retain basic similarity with the vocal originals (Obleser *et al.*, 2006; Warren *et al.*, 2006). Nevertheless, sound acoustics necessarily changed as described in the Supplementary Materials. Although sounding distinctly non-human, rotated surprised sounds were perceived as more emotional than rotated neutral sounds (see Supplementary Materials).

Procedure

Participants were tested individually. To prepare them for the EEG recording, a 64-channel cap with empty electrode holders was placed on their head. The electrode holders, which were organized according to the modified 10–20 system, were filled with electrolyte gel and electrodes placed into them. Individual

electrodes were attached above and below the right eye and at the outer canthus of each eye to measure eye movements. One electrode was attached to the nose for data referencing as to enable the exploration of dipoles situated in auditory cortex (Näätänen *et al.*, 2007). The data were recorded at 500 Hz with a BrainAmp EEG system. Only an anti-aliasing filter was applied during data acquisition (i.e. sinc filter with a half-power cutoff at half the sampling rate).

Following the EEG set-up, participants were seated in front of a computer screen that was framed by two speakers. On-screen instructions informed participants that they would hear a sequence of sounds and that their task was to press a button using their right hand any time a sound was immediately repeated. The task comprised three blocks in which sounds (i.e. 27 neutral/vocal, 27 surprised/vocal, 27 neutral/non-vocal and 27 surprised/non-vocal) were played in random order. Thus, stimuli were played thrice across separate blocks. The two stimuli forming a given vocal/non-vocal sound pair occurred in separate block halves as to minimize the emergence of potential acoustic associations. In addition to the trials described thus far, 24 sounds were randomly selected for repetition within each block as to engage participants with the auditory material without highlighting the nature of the sounds and without necessitating a confounding motor response on unrepeated, experimental trials.

Each trial started with a white fixation cross centered on a black background. After 500 ms, a sound played (average duration = 506 ms; s.d. = 25 ms) and the fixation cross remained for 1000 ms. An empty inter-trial interval had a random duration ranging between 2000 and 4000 ms.

Data analysis

EEG data were processed with EEGLAB (Delorme and Makeig, 2004). The recordings were subjected to low- and high-pass filtering with a half-power cut-off at 30 and 0.1 Hz, respectively. The transition band was 7.5 Hz for the low pass filter (−6 dB/octave; 221 pts) and 0.1 Hz for the high pass filter (−6 dB/octave; 16 501 pts). The continuous data were epoched and baseline-corrected using a 200 ms pre-stimulus baseline and a 1000 ms post-stimulus window. The resulting epochs were visually scanned for non-typical artifacts caused by drifts or muscle movements. Epochs containing such artifacts were removed. Infomax, an independent component analysis algorithm, was applied to the remaining data, and components reflecting typical artifacts (i.e. horizontal and vertical eye movements and eye blinks) were removed. Back-projected single trials were again screened visually for residual artifacts and ERPs were derived by averaging individual epochs for each condition and participant including only non-repeated trials on which participants correctly withheld a response. A minimum of 62 and an average of 75 trials per condition entered statistical analysis.

We identified the latency ranges of target ERP components based on visual inspection and prior work (see Supplementary Materials for a figure of all electrode traces). Mean voltages from within these ranges were subjected to an ANOVA with 'Voiceness' (vocal, non-vocal), 'Emotion' (surprised, neutral), 'Hemisphere' (left, right) and 'Region' (anterior, central, posterior) as repeated measures factors and 'Sex' as the between subjects factor. The factors 'Hemisphere' and 'Region' comprised average voltages computed across the following subgroups of electrodes: anterior left, Fp1, AF7, AF3, F5, F3, F1; anterior right, Fp2, AF8, AF4, F6, F4, F2; central left, FC3, FC1, C3, C1, CP3, CP1; central right, FC4, FC2, C4, C2, CP4, CP2; posterior left, P5, P3, P1, PO7, PO3, O1; posterior right, P6, P4, P2, PO8, PO4, O2. This selection of electrodes ensured that the tested subgroups contained equal number of electrodes while providing a broad scalp coverage that allowed the assessment of topographical effects. To facilitate the comparison of the present results with other labs, we included an analysis of data re-referenced to the average of all electrodes and an analysis re-referenced to the average of left and right mastoids into the Supplementary Materials.

We only report effects involving factors of interest (i.e. 'Voiceness', 'Emotion') and interactions for which follow-up analyses reached significance ($P < 0.05$) in the nose-referenced data-set or in any of the other two data sets.

Results

Behavioral results

We computed d-prime sensitivity scores by subtracting the normalized probability of hits (i.e. button presses to repeated sounds) from the normalized probability of false alarms (i.e. button presses to non-repeated sounds). The resulting scores were subjected to an ANOVA with 'Voiceness' (vocal, non-vocal) and 'Emotion' (neutral, surprised) as repeated measures factors and 'Sex' as a between subjects factor. A significant effect of 'Voiceness' [$F(1,30) = 23.2$, $P < 0.0001$, $\eta^2_G = 0.132$] indicated that participants were more sensitive to vocal than to non-vocal repetitions (Figure 1). Hit reaction times were analyzed using a comparable statistical model. This revealed effects of 'Voiceness' [$F(1,30) = 21.3$, $P < 0.0001$, $\eta^2_G = 0.036$] and 'Emotion' [$F(1,30) = 10.9$, $P < 0.01$, $\eta^2_G = 0.017$]. Participants responded faster to vocal and surprised sounds as compared with non-vocal and neutral sounds. All other effects were non-significant (Ps > 0.1).

Electrophysiological results

N1. The N1 was explored between 80 and 120 ms following stimulus onset. Mean amplitudes derived by averaging data points from within this time range were subjected to an ANOVA with 'Emotion', 'Voiceness', 'Hemisphere' and 'Region' as repeated measures factors and 'Sex' as a between subjects factor. A main effect of 'Voiceness' [$F(1,30) = 17.47$, $P < 0.001$, $\eta^2_G = 0.017$] indicated that N1 amplitudes were larger in the vocal than the non-vocal condition. Interactions of 'Voiceness' and 'Hemisphere' [$F(1,30) = 4.91$, $P < 0.05$, $\eta^2_G = 0.0003$] and 'Voiceness' and 'Region' [$F(2,60) = 5.39$, $P < 0.05$, $\eta^2_G = 0.001$] showed that this effect differed across the scalp. Exploring the 'Voiceness' effect for each level of 'Hemisphere' revealed greater effects at right [$F(1,30) = 22.22$, $P < 0.001$, $\eta^2_G = 0.021$] as compared with left recording sites [$F(1,30) = 12.97$, $P < 0.01$, $\eta^2_G = 0.013$]. Exploring the 'Voiceness' effect for each level of 'Region' pointed to greater effects over central [$F(1,30) = 22.45$, $P < 0.0001$, $\eta^2_G = 0.025$] as opposed to anterior [$F(1,30) = 6.52$, $P < 0.05$, $\eta^2_G = 0.001$] and posterior recording sites [$F(1,30) = 17.02$, $P < 0.001$, $\eta^2_G = 0.016$; Figures 2 and 3]. The factor 'Emotion' was significant in an interaction with 'Region' [$F(2,60) = 6.03$, $P < 0.01$, $\eta^2_G = 0.001$] indicating that the N1 tended to be larger for neutral than for surprised voices over anterior [$F(1,30) = 3.47$, $P = 0.072$, $\eta^2_G = 0.004$] but not central and posterior regions (Ps > 0.1).

P2/P3 complex. The P2 was immediately followed by another positivity and effects for both seemed comparable. Hence, we examined their mean voltages jointly between 150 and 350 ms following stimulus onset. Statistical analysis produced

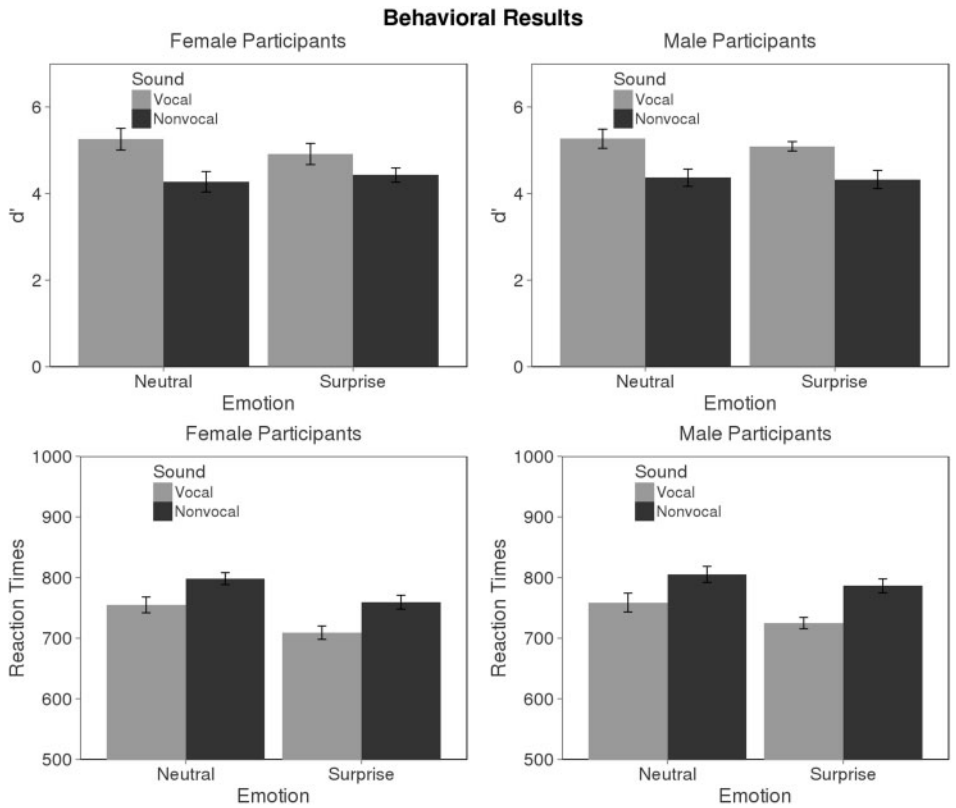

Fig. 1 Experimental task performance. The top row illustrates the sensitivity with which female (left) and male (right) listeners discriminated between repeated and non-repeated sounds. The bottom row illustrates the speed with which female (left) and male (right) listeners pushed the button to sound repetitions.

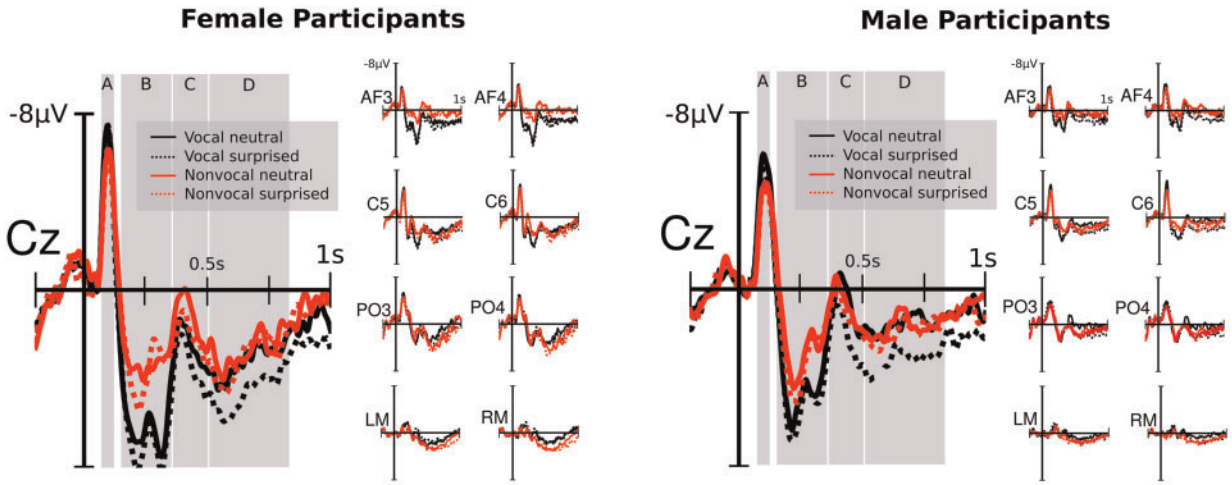

Fig. 2 ERP traces. Illustrated are average voltages recorded to the four sound conditions for female (left) and male (right) participants.

a 'Voiceness' main effect [$F(1,30) = 36.26$, $P < 0.0001$, $\eta^2_G = 0.082$] indicating greater positivity for vocal than for non-vocal sounds. An interaction of 'Voiceness', 'Region' and 'Sex' [$F(2,60) = 3.12$, $P = 0.05$, $\eta^2_G = 0.0016$] was pursued for women and men separately. In both groups, the 'Voiceness' by 'Region' interaction [females, $F(2,30) = 47.08$, $P < 0.0001$, $\eta^2_G = 0.033$; males $F(2,30) = 15.52$, $P < 0.0001$, $\eta^2_G = 0.023$] revealed that the 'Voiceness' effect was maximal over anterior [females, $F(1,15) = 74.33$, $P < 0.0001$, $\eta^2_G = 0.302$; males $F(1,15) = 23.36$, $P < 0.001$, $\eta^2_G = 0.133$], small over central [females, $F(1,15) = 33.75$, $P < 0.0001$, $\eta^2_G = 0.176$; males $F(1,15) = 18.39$, $P < 0.001$, $\eta^2_G = 0.133$) and non-significant over posterior regions ($Ps > 0.1$). Notably, however, the 'Voiceness' effect was considerably greater in women than in men.

The P2/P3 complex was also characterized by an interaction of 'Emotion' and 'Region' [$F(2,60) = 7.41$, $P < 0.01$, $\eta^2_G = 0.0023$]. Follow-up analyses showed that surprised sounds elicited greater positivity than neutral sounds over anterior [$F(1,30) = 7.33$, $P < 0.05$, $\eta^2_G = 0.0115$] and central [$F(1,30) = 5.85$, $P < 0.05$, $\eta^2_G = 0.0072$] but not posterior regions ($P > 0.1$).

N4-like negativity. The P2/P3 complex was followed by an N4-like negativity. An exploration of mean voltages between 350

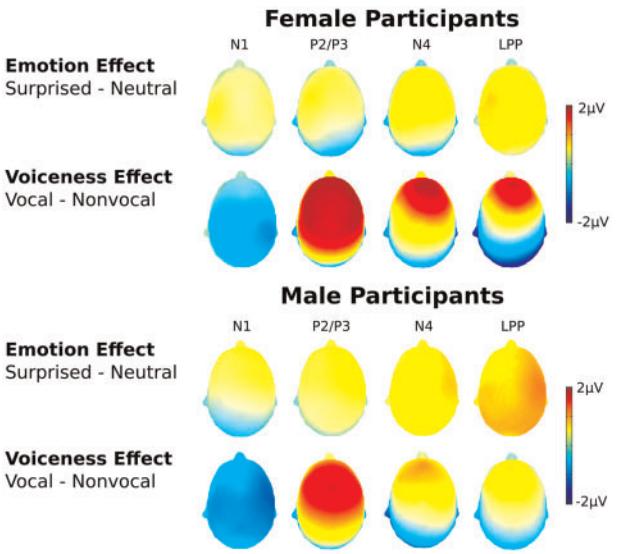

Fig. 3 ERP maps. Topographical maps illustrate the average differences between emotional and neutral as well as vocal and non-vocal sounds for female (top) and male (bottom) participants. Average differences were computed for the four statistical analysis windows capturing the N1, the P2/P3 complex, the N4-like negativity, and the LPP.

and 500 ms revealed effects of 'Voiceness' [$F(1,30) = 5.17$, $P < 0.05$, $\eta^2_G = 0.006$] and 'Emotion' [$F(1,30) = 5.06$, $P < 0.05$, $\eta^2_G = 0.007$] indexing greater amplitudes for non-vocal than for vocal and for neutral than for surprised stimuli.

The 'Voiceness' effect was further qualified by interactions involving 'Voiceness' and 'Region' [$F(2,60) = 19.72$, $P < 0.0001$, $\eta^2_G = 0.014$] and 'Voiceness', 'Region', 'Hemisphere' and 'Sex' [$F(2,60) = 6.5$, $P < 0.01$, $\eta^2_G = 0.0002$]. In women, the interaction of 'Voiceness', 'Region' and 'Hemisphere' [$F(2,30) = 3.27$, $P = 0.052$, $\eta^2_G = 0.0002$] was pursued for each level of 'Region'. Over anterior and central recording sites, the 'Voiceness' by 'Hemisphere' interaction [anterior, $F(1,15) = 2.24$, $P = 0.15$, $\eta^2_G = 0.001$; central, $F(1,15) = 3.33$, $P = 0.087$, $\eta^2_G = 0.0007$] was non-significant or marginal but the 'Voiceness' effect was significant [anterior, $F(1,15) = 41.37$, $P < 0.0001$, $\eta^2_G = 0.151$; central, $F(1,15) = 7.45$, $P < 0.05$, $\eta^2_G = 0.022$]. There were no effects over posterior recording sites (Ps > 0.1). In men, the interaction of 'Voiceness', 'Region' and 'Hemisphere' [$F(2,30) = 6.15$, $P < 0.01$, $\eta^2_G = 0.0002$] was significant also. However, follow-up analyses revealed only a 'Voiceness' main effect over anterior regions [$F(1,15) = 6.99$, $P < 0.05$, $\eta^2_G = 0.026$]. All other effects were non-significant (Ps > 0.1).

The 'Emotion' effect was qualified by an interaction of 'Emotion' and 'Region' [$F(2,60) = 3.68$, $P = 0.05$, $\eta^2_G = 0.001$] and an interaction of 'Emotion', 'Hemisphere' and 'Sex' [$F(1,30) = 5.65$, $P < 0.05$, $\eta^2_G = 0.0002$]. Across participants, the 'Emotion' effect was present over anterior [$F(1,30) = 5.39$, $P < 0.05$, $\eta^2_G = 0.015$], central [$F(1,30) = 7.88$, $P < 0.01$, $\eta^2_G = 0.014$] but not posterior regions ($P > 0.1$). In female participants, the 'Emotion' effect was independent of 'Hemisphere' ($P > 0.1$). In male participants, the 'Emotion' effect interacted with 'Hemisphere' [$F(1,15) = 5.43$, $P < 0.05$, $\eta^2_G = 0.0004$] in that it was significant over right [$F(1,15) = 4.9$, $P < 0.05$, $\eta^2_G = 0.019$] but not left ($P > 0.1$) recording sites.

Late positive potential. The LPP peaked between 500 and 800 ms following stimulus onset. Analysis of mean voltages within this time window revealed an 'Emotion' effect [$F(1,30) = 9.45$, $P < 0.01$, $\eta^2_G = 0.014$] with greater amplitudes for surprised than neutral stimuli. The 'Emotion' effect was qualified by interactions of 'Emotion', 'Hemisphere' and 'Sex' [$F(1,30) = 8.43$, $P < 0.01$, $\eta^2_G = 0.0003$] and 'Emotion', 'Voiceness' and 'Hemisphere' [$F(1,30) = 5.38$, $P < 0.01$, $\eta^2_G = 0.0002$]. Exploring the first interaction revealed again that the 'Emotion' effect differed by 'Hemisphere' in male [$F(1,15) = 11.15$, $P < 0.01$, $\eta^2_G = 0.0005$] but not female participants ($P > 0.1$). In men, but not in women, the effect was greater over right [$F(1,15) = 7.39$, $P < 0.05$, $\eta^2_G = 0.04$] as compared with left hemisphere leads [$F(1,15) = 4.92$, $P < 0.05$, $\eta^2_G = 0.023$]. Exploring the second interaction revealed an 'Emotion' by 'Voiceness' interaction in the left [$F(1,30) = 4.27$, $P = 0.05$, $\eta^2_G = 0.004$] but not the right hemisphere ($P > 0.1$). Over the left hemisphere, vocal [$F(1,30) = 8.73$, $P < 0.01$, $\eta^2_G = 0.028$] but not non-vocal sounds ($P > 0.1$) elicited a greater positivity for surprised as compared with neutral expressions.

Although the 'Voiceness' effect was non-significant, there was an interaction of 'Voiceness', 'Region' and 'Sex' [$F(2,60) = 8.15$, $P < 0.001$, $\eta^2_G = 0.003$] for which follow-up comparisons were significant in both women [$F(2,30) = 28.74$, $P < 0.0001$, $\eta^2_G = 0.023$] and men [$F(2,30) = 5.48$, $P < 0.01$, $\eta^2_G = 0.005$]. In women, vocal sounds elicited a greater amplitude than non-vocal sounds over anterior electrodes [$F(2,30) = 19.56$, $P < 0.001$, $\eta^2_G = 0.07$]. The effect was non-significant over central electrodes ($P > 0.1$) and reversed polarity over posterior electrodes [$F(2,30) = 8.83$, $P < 0.01$, $\eta^2_G = 0.019$]. In men, the 'Voiceness' effect approached significance over anterior regions only [$F(1,15) = 4.03$, $P = 0.063$, $\eta^2_G = 0.008$].

Mastoid effects. Statistical analysis demonstrated that fronto-central 'Voiceness' but not 'Emotion' effects for P2/P3, N4-like, and LPP components reversed polarity over the mastoids (see Supplementary Materials).

Discussion

In this study, we presented participants with vocal and non-vocal sounds of emotional and neutral quality in order to shed light on the temporal course underpinning vocal-emotional processing.

We found that voiceness modulated the amplitude of a negativity peaking 100 ms following stimulus onset. This negativity belongs to the N1 family, which is modulated by attention in a top-down or bottom-up manner. Attended stimuli or stimuli that capture attention endogenously elicit a greater N1 than less or unattended stimuli (Woldorff and Hillyard, 1991; Escoffier *et al.*, 2015). The present N1 modulation was largest over right and centro-parietal electrodes. It was hence compatible with sources in higher-order auditory regions like the temporal voice areas. Moreover, it agrees with other evidence suggesting that the right hemisphere is more relevant than the left for social processing (for a review see Brauer *et al.*, 2016).

The early voiceness effect identified here aligns with visual work showing P1 differences between images with and without people (Proverbio *et al.*, 2009) and implicating the N170, which, although later than the present N1, belongs to the N1 family (Bentin *et al.*, 1996). In contrast, the present results diverge from prior auditory work reporting voiceness modulations at 200 ms following stimulus onset. Different factors may be responsible for this discrepancy. For example, we presented 162 vocal and 162 non-vocal sounds to 32 participants and was thus better powered than most previous work (Levy *et al.*, 2001; De Lucia *et al.*, 2010; Capilla *et al.*, 2013). Additionally, our high-pass filter settings were lower (Levy *et al.*, 2001; Charest *et al.*, 2009) and thus more appropriate for the examination of early/fast signal aspects. Last, we compared voices with their spectral rotations

whereas other studies implemented other comparisons (e.g. non-human animal vocalizations, music, non-living objects) with other strengths and weaknesses as concerns the control of acoustic and conceptual confounds. Hence, we cannot rule out that the present N1 effects were caused by acoustic and/or conceptual confounds inherent in the present design.

The N1 was succeeded by two positivities named P2 and P3 reflecting stimulus perception and categorization (Johnson and Donchin, 1978; Woldorff and Hillyard, 1991; Schirmer *et al.*, 2011b; Schirmer *et al.*, 2013a). As predicted, their voltages were more positive for vocal as compared with non-vocal sounds and this difference extended into the remainder of the ERP epoch affecting subsequent components. Notably, there was not only a change in polarity but also in topography from the N1 effect. Specifically, the P2/P3 effect showed bilaterally with a maximum at fronto-central electrodes and reversed polarity over the mastoids. Although the ERP inverse problem means that a given scalp topography can be explained by more than one underlying source pattern, the scalp topography observed here is typically linked to a contribution of auditory cortex (Näätänen *et al.*, 2007).

One may speculate that the present P2/P3 effect relates to the FTPV (Charest *et al.*, 2009; Capilla *et al.*, 2013). Both occur within a similar time range over fronto-central electrodes. Moreover, differences over the posterior scalp where the FTPV but not the present P2/P3 reverses polarity are easily explained by differences in the reference electrode. Unlike a single channel reference, the average reference used previously forces the ERP into a dipolar pattern (i.e. potentials across channels sum to 0). This is evident from an analysis of the present data with average reference which revealed a patter akin to that of the FTPV (see Supplementary Materials). Nevertheless, we refrain from using FTPV terminology because we have no clear evidence that the underlying mechanisms are indeed voice-specific. Given their similarity to other kinds of sound processing (De Lucia *et al.*, 2010; Schirmer *et al.*, 2011a), they likely have a more general nature.

Emotion effects emerged simultaneously with voiceness effects. However, in the main (nose-referenced) analysis they were only marginal for the N1 and significant in the P2/P3 complex beginning 150 ms following sound onset. Although initially, emotion effects were similar to those of voiceness, they differed in that they failed to reverse polarity over the mastoids. Moreover, they occurred independently of voiceness processing. The P2/P3 complex was not greater for emotional sounds in the vocal than the non-vocal condition or for vocal sounds in the emotional than the neutral condition. Such super-additivity appeared only later in the epoch, for the LPP, and after the emotion effect gained significance over central electrodes differentiating more clearly from the voiceness effect.

The late interaction of emotion and voiceness diverges from visual evidence (Proverbio *et al.*, 2009). Additionally, it seems at odds with prior auditory evidence for vocal content interacting with verbal content in the N4 (Schirmer and Kotz, 2003). However, ours is the first study to tackle the confluence of vocal and emotion processing and provides a strong test of interactive effects. Specifically, the very nature of our non-vocal sounds should have promoted rather than hampered the interaction of voiceness and emotion during stimulus processing. Although, non-vocal sounds retained emotion aspects of their originals, their emotion was recognized more poorly and arousal differences between surprise and neutral stimuli were perceived as weaker (see Supplementary Materials). This should have hampered emotional processing for non-vocal relative to vocal sounds, which in turn should have a by emotion interaction. That this interaction was absent before the LPP provides convincing support that voiceness and emotion are treated largely independently before being integrated at a later and more controlled processing stage.

The LPP appears to reflect this integration. Its amplitude over the left hemisphere was greater for emotional as compared with neutral vocal but not non-vocal sounds. As apparent in Figure 2, emotional voices differed from all other sounds suggesting that they won the competition for resources. This effect compares to previous reports of the LPP being larger for emotional as compared with neutral stimuli and may reflect emotional strength or simply attention allocation as a function of stimulus significance (Moser *et al.*, 2006; Foti and Hajcak, 2008; Schirmer *et al.*, 2011b). That the voice–emotion interaction was significant in the left hemisphere only accords with fMRI evidence for an involvement of left inferior frontal gyrus for vocal emotions (Kotz *et al.*, 2003; Warren *et al.*, 2006; Leitman *et al.*, 2011; Frühholz *et al.*, 2012) and may be voice-specific (Schirmer and Adolphs, in press).

The evidence discussed thus far highlights temporally and morphologically distinct effects that could map onto different processing stages. In line with proposals made recently (De Lucia *et al.*, 2010; Perrodin *et al.*, 2015), a first stage may involve basic level processing that discriminates living from non-living sources and that may occur around 100 ms following stimulus onset. Although our N1 results concord with this, a direct mapping can only be tentative as acoustic (e.g. HNR) and conceptual factors (e.g. sound familiarity) offer alternative explanations. A second stage could entail subordinate level processing further specifying a sound's source (e.g. human *vs* non-human animal). Presumably this begins around 150 ms (De Lucia *et al.*, 2010) and thus overlaps with the P2/P3 results obtained here. Because emotions or affect are inherent to humans and animals alike, their representations may emerge early in the course of basic level processing. Nevertheless, they are not immediately integrated as is evident from the fact that interaction effects were non-significant for both N1 and P2/P3 complex. Emotion modulated voiceness effects only later for the LPP pointing to a possible third stage during which the different sound properties merge into a holistic sound object and processing prioritizes some objects over others (Figure 4).

Extant research suggests that women engage in preferential social processing more readily than men do (Schirmer *et al.*, 2013a,b; Proverbio *et al.*, 2009; Proverbio and Galli, 2016). The present results corroborate this. The voiceness positivity effect at 200 ms following stimulus onset was significantly greater in women than in men. Moreover, it lasted well into the LPP where it changed into a different pattern over posterior electrodes.

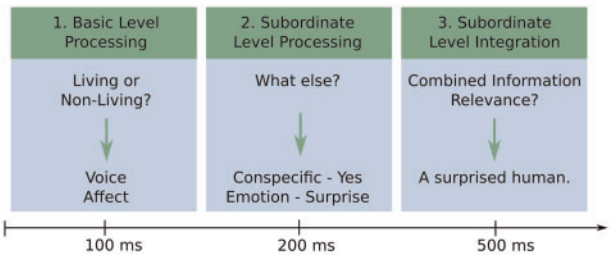

Fig. 4 Interpretative framework. Voice processing is illustrated by example. When presented with a vocal sound we may first represent its animacy and basic affect, before accessing subordinate level sound categories. Subsequently, these separate representations may be integrated into a holistic percept.

This posterior LPP effect was characterized by a greater positivity for non-vocal relative to vocal sounds. Within the interpretational framework outlined above, these findings suggest that initial basic level processing is comparable in men and women. Differences appear only for subsequent and putative subordinate level processing. Perhaps, after having identified a sound source as living, women direct more resources than men at further specifying the sound. Moreover, the posterior LPP effect might reflect additional top-down processes directed at inferring some sort of animacy from the non-vocal sounds—a process known to differ between the sexes (Proverbio and Galli, 2016).

Although this study provides novel insights into vocal-emotional processing, it also raises questions for future research. For example, we have compared vocal with spectrally rotated sounds thus controlling for some but certainly not all acoustic stimulus differences. Moreover, whereas the vocal sounds were highly familiar and natural, their rotated counterparts were not. Thus, it is important for future research to compare human voices with other controls such as animal vocalizations (Fecteau *et al*., 2004), music (Escoffier *et al*., 2013), or environmental noises (Belin *et al*., 2000; De Lucia *et al*., 2010). Consistent results across these conditions would help rule out the confounds that necessarily arise for each control individually.

Another question that should be tackled is why, in this study, women were not more emotionally sensitive than men. Instead, the sexes differed simply in the laterality of emotion effects. In men but not women, the N4-like negativity and the LPP effects were larger at right than left electrodes. Possibly female voiceness sensitivity was so strong as to override any preferential attention to emotion. Future research could test this possibility by presenting vocal and non-vocal sounds in separate blocks. Moreover, new studies could employ different neuroimaging techniques as to characterize underlying spatial sources. A candidate here is functional near-infrared spectroscopy, which yields both high spatial and high temporal resolution in the cortex (Tse and Penney, 2008; Tse *et al*., 2013).

Despite these open questions, however, this work allows for some conclusions to be made. Specifically, our findings show that listeners, especially women, direct more processing resources to vocal than to non-vocal sounds if voiceness is task-irrelevant. This effect unfolds 100 ms following sound onset and is characterized by different processing stages potentially reflecting basic level categorization, subordinate level categorization, and the integration of subordinate-level information, respectively. Voices when compared against their spectral rotations produce effects that are larger but temporally overlapping with emotion effects. Moreover, both are independent before interacting in a manner that enhances the processing of vocal-emotional over vocal-neutral and non-vocal sounds. Taken together, our findings underline the human bias towards conspecifics and the social nature of the human brain.

Supplementary data

Supplementary data are available at *SCAN* online.

Conflict of interest. None declared.

References

Ahrens, M.M., Awwad Shiekh Hasan, B., Giordano, B.L., Belin, P. (2014). Gender differences in the temporal voice areas. *Frontiers in Neuroscience*, **8**, 228.

Beaucousin, V., Lacheret, A., Turbelin, M.R., Morel, M., Mazoyer, B., Tzourio-Mazoyer, N. (2007). FMRI study of emotional speech comprehension. *Cerebral Cortex*, **17**, 339–52.

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

Bentin, S., Allison, T., Puce, A., Perez, E., McCarthy, G. (1996). Electrophysiological studies of face perception in humans. *Journal of Cognitive Neuroscience*, **8**, 551–65.

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**, 3544–52.

Brück, C., Kreifelts, B., Kaza, E., Lotze, M., Wildgruber, D. (2011). Impact of personality on the cerebral processing of emotional prosody. *NeuroImage*, **58**, 259–68.

Bruneau, N., Roux, S., Cléry, H., Rogier, O., Bidet-Caulet, A., Barthélémy, C. (2013). Early neurophysiological correlates of vocal versus non-vocal sound processing in adults. *Brain Research*, **1528**, 20–7.

Capilla, A., Belin, P., Gross, J. (2013). The early spatio-temporal correlates and task independence of cerebral voice processing studied with MEG. *Cerebral Cortex (New York, N.Y.: 1991)*, **23**, 1388–95.

Charest, I., Pernet, C.R., Rousselet, G.A., *et al*. (2009). Electrophysiological evidence for an early processing of human voices. *BMC Neuroscience*, **10**, 127.

De Lucia, M., Clarke, S., Murray, M.M. (2010). A temporal hierarchy for conspecific vocalization discrimination in humans. *The Journal of Neuroscience*, **30**, 11210–21.

Delorme, A., Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, **134**, 9–21.

Escoffier, N., Herrmann, C.S., Schirmer, A. (2015). Auditory rhythms entrain visual processes in the human brain: evidence from evoked oscillations and event-related potentials. *NeuroImage*, **111**, 267–76.

Escoffier, N., Zhong, J., Schirmer, A., Qiu, A. (2013). Emotional expressions in voice and music: Same code, same effect?. *Human Brain Mapping*, **34**, 1796–810.

Ethofer, T., Wiethoff, S., Anders, S., Kreifelts, B., Grodd, W., Wildgruber, D. (2007). The voices of seduction: cross-gender effects in processing of erotic prosody. *Social Cognitive and Affective Neuroscience*, **2**, 334–7.

Fecteau, S., Armony, J.L., Joanette, Y., Belin, P. (2004). Is voice processing species-specific in human auditory cortex? An fMRI study. *NeuroImage*, **23**, 840–8.

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

Foti, D., Hajcak, G. (2008). Deconstructing reappraisal: descriptions preceding arousing pictures modulate the subsequent neural response. *Journal of Cognitive Neuroscience*, **20**, 977–88.

Frühholz, S., Ceravolo, L., Grandjean, D. (2012). Specific brain networks during explicit and implicit decoding of emotional prosody. *Cerebral Cortex (New York, N.Y.: 1991)*, **22**, 1107–17.

Johnson, R., Jr., Donchin, E. (1978). On how P300 amplitude varies with the utility of the eliciting stimuli. *Electroencephalography and Clinical Neurophysiology*, **44**, 424–37.

Kotz, S.A., Meyer, M., Alter, K., Besson, M., von Cramon, D.Y., Friederici, A.D. (2003). On the lateralization of emotional prosody: An event-related functional MR investigation. *Brain and Language*, **86**, 366–76.

Leitman, D.I., Wolf, D.H., Laukka, P., *et al.* (2011). Not Pitch Perfect: Sensory Contributions to Affective Communication Impairment in Schizophrenia. *Biological Psychiatry*, **70**, 611–8.

Levy, D.A., Granot, R., Bentin, S. (2001). Processing specificity for human voice stimuli: electrophysiological evidence. *Neuroreport*, **12**, 2653–7.

Moser, J.S., Hajcak, G., Bukay, E., Simons, R.F. (2006). Intentional modulation of emotional responding to unpleasant pictures: An ERP study. *Psychophysiology*, **43**, 292–6.

Mothes-Lasch, M., Mentzel, H.J., Miltner, W.H.R., Straube, T. (2011). Visual Attention Modulates Brain Activation to Angry Voices. *The Journal of Neuroscience*, **31**, 9594–8.

Näätänen, R., Paavilainen, P., Rinne, T., Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: a review. *Clinical Neurophysiology: Official Journal of the International Federation of Clinical Neurophysiology*, **118**, 2544–90.

Obleser, J., Scott, S.K., Eulitz, C. (2006). Now you hear it, now you don't: transient traces of consonants and their nonspeech analogues in the human brain. *Cerebral Cortex*, **16**, 1069–76.

Paulmann, S., Kotz, S.A. (2008). Early emotional prosody perception based on different speaker voices. *Neuroreport*, **19**, 209–13.

Pell, M.D., Rothermich, K., Liu, P., Paulmann, S., Sethi, S., Rigoulot, S. (2015). Preferential decoding of emotion from human non-linguistic vocalizations versus speech prosody. *Biological Psychology*, **111**, 14–25.

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**, 783–96.

Pinheiro, A.P., Barros, C., Pedrosa, J. (2016). Salience in a social landscape: electrophysiological effects of task-irrelevant and infrequent vocal change. *Social Cognitive and Affective Neuroscience*, **11**, 127–39.

Proverbio, A.M., Adorni, R., Zani, A., Trestianu, L. (2009). Sex differences in the brain response to affective scenes with or without humans. *Neuropsychologia*, **47**, 2374–88.

Proverbio, A.M., Galli, J. (2016). Women are better at seeing faces where there are none: an ERP study of face pareidolia. *Social Cognitive and Affective Neuroscience*, **11**, 1501–12.

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

Rigoulot, S., Pell, M.D., Armony, J.L. (2015). Time course of the influence of musical expertise on the processing of vocal and musical sounds. *Neuroscience*, **290**, 175–84.

Sauter, D.A., Eimer, M. (2010). Rapid detection of emotion from human vocalizations. *Journal of Cognitive Neuroscience*, **22**, 474–81.

Schirmer, A., Adolphs, R. (in press). Emotion perception from face, voice, and touch: comparisons and convergence. *Trends in Cognitive Sciences*, doi:http://dx.doi.org/10.1016/j.tics.2017.01.001.

Schirmer, A., Chen, C.B., Ching, A., Tan, L., Hong, R.Y. (2013a). Vocal emotions influence verbal memory: neural correlates and interindividual differences. *Cognitive, Affective, and Behavioral Neuroscience*, **13**, 80–93.

Schirmer, A., Escoffier, N., Cheng, X., Feng, Y., Penney, T.B. (2016a). Detecting temporal change in dynamic sounds: on the role of stimulus duration, speed, and emotion. *Frontiers in Psychology*, doi:10.3389/fpsyg.2015.02055.

Schirmer, A., Kotz, S.A. (2003). ERP evidence for a sex-specific Stroop effect in emotional speech. *Journal of Cognitive Neuroscience*, **15**, 1135–48.

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

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

Schirmer, A., Simpson, E., Escoffier, N. (2007). Listen up! Processing of intensity change differs for vocal and nonvocal sounds. *Brain Research*, **1176**, 103–12.

Schirmer, A., Soh, Y.H., Penney, T.B., Wyse, L. (2011a). Perceptual and conceptual priming of environmental sounds. *Journal of Cognitive Neuroscience*, **23**, 3241–53.

Schirmer, A., Striano, T., Friederici, A.D. (2005). Sex differences in the preattentive processing of vocal emotional expressions. *Neuroreport*, **16**, 635–9.

Schirmer, A., Teh, K.S., Wang, S., *et al.* (2011b). Squeeze me, but don't tease me: Human and mechanical touch enhance visual attention and emotion discrimination. *Social Neuroscience*, **6**, 219–30.

Scott, S.K., Blank, C.C., Rosen, S., Wise, R.J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain: A Journal of Neurology*, **123** (Pt 12), 2400–6.

Thierry, G., Roberts, M.V. (2007). Event-related potential study of attention capture by affective sounds. *Neuroreport*, **18**, 245–8.

Tse, C.Y., Penney, T.B. (2008). On the functional role of temporal and frontal cortex activation in passive detection of auditory deviance. *NeuroImage*, **41**, 1462–70.

Tse, C.Y., Rinne, T., Ng, K.K., Penney, T.B. (2013). The functional role of the frontal cortex in pre-attentive auditory change detection. *NeuroImage*, **83**, 870–9.

Viinikainen, M., Kätsyri, J., Sams, M. (2012). Representation of perceived sound valence in the human brain. *Human Brain Mapping*, **33**, 2295–305.

Warren, J.E., Sauter, D.A., Eisner, F., *et al.* (2006). Positive emotions preferentially engage an auditory–motor "mirror" system. *The Journal of Neuroscience*, **26**, 13067–75.

Woldorff, M.G., Hillyard, S.A. (1991). Modulation of early auditory processing during selective listening to rapidly presented tones. *Electroencephalography and Clinical Neurophysiology*, **79**, 170–91.

Yovel, G., Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, **17**, 263–71.