OXFORD

Social Cognitive and Affective Neuroscience, 2016, 1638–1649

doi: 10.1093/scan/nsw066
Advance Access Publication Date: 14 May 2016
Original article

Neural decoding of discriminative auditory object features depends on their socio-affective valence

Sascha Frühholz,[1,2] Wietske van der Zwaag,[3] Melissa Saenz,[4,5] Pascal Belin,[6] Anne-Kathrin Schobert,[7] Patrik Vuilleumier,[2,7] and Didier Grandjean[2,8]

[1]Department of Psychology, University of Zurich, 8050 Zurich, Switzerland, [2]Swiss Center for Affective Sciences, University of Geneva, 1202 Geneva, Switzerland, [3]Center for Biomedical Imaging, Ecole Polytechnique Fédérale de Lausanne 1015 Lausanne, Switzerland, [4]Laboratoire de Recherche en Neuroimagerie, Department of Clinical Neurosciences, CHUV, 1011 Lausanne, Switzerland, [5]Institute of Bioengineering, Ecole Polytechnique Fédérale de Lausanne, 1015 Lausanne, Switzerland, [6]Department of Psychology, University of Glasgow, Glasgow G12 8QQ, UK, [7]Laboratory for Neurology and Imaging of Cognition, Department of Neurology and Department Neuroscience, Medical School, University of Geneva, 1211 Geneva, Switzerland and [8]Neuroscience of Emotion and Affective Dynamics Laboratory, Department of Psychology, University of Geneva, Geneva 1205, Switzerland

Correspondence should be addressed to Sascha Frühholz, University of Zürich, Department of Psychology, Binzmühlestrasse 14, Box 18, 8050 Zürich, Switzerland. E-mail: sascha.fruehholz@psychologie.uzh.ch

Abstract

Human voices consist of specific patterns of acoustic features that are considerably enhanced during affective vocalizations. These acoustic features are presumably used by listeners to accurately discriminate between acoustically or emotionally similar vocalizations. Here we used high-field 7T functional magnetic resonance imaging in human listeners together with a so-called experimental 'feature elimination approach' to investigate neural decoding of three important voice features of two affective valence categories (i.e. aggressive and joyful vocalizations). We found a valence-dependent sensitivity to vocal pitch (f0) dynamics and to spectral high-frequency cues already at the level of the auditory thalamus. Furthermore, pitch dynamics and harmonics-to-noise ratio (HNR) showed overlapping, but again valence-dependent sensitivity in tonotopic cortical fields during the neural decoding of aggressive and joyful vocalizations, respectively. For joyful vocalizations we also revealed sensitivity in the inferior frontal cortex (IFC) to the HNR and pitch dynamics. The data thus indicate that several auditory regions were sensitive to multiple, rather than single, discriminative voice features. Furthermore, some regions partly showed a valence-dependent hypersensitivity to certain features, such as pitch dynamic sensitivity in core auditory regions and in the IFC for aggressive vocalizations, and sensitivity to high-frequency cues in auditory belt and parabelt regions for joyful vocalizations.

Key words: voice; auditory features; emotion; multivariate analysis; fMRI

Introduction

The fast and accurate decoding of the affective valence of conspecific vocalizations is of high importance for social interactions and for individual survival in terms of an adaptive behavioral response. Vocalizations consist of specific patterns of acoustic features that support inter- (e.g. voice *vs* car) and intra-class categorizations (e.g. female *vs* male voices). These voice features are considerably enhanced and modulated during emotional states that result in affective vocalizations (Banse and Scherer, 1996; Patel *et al.*, 2011). Some of these features,

Received: 30 September 2015; **Revised:** 10 December 2015; **Accepted:** 11 May 2016

such as vocal pitch, allow to differentiate vocalizations that are either close in their acoustic profile or in their affective meaning. Thus, the fast and accurate affective classification of voices might be accomplished by a processing focused on relevant and discriminative sensory voice features that most likely indicate the valence of the voice.

In terms of sensory voice feature processing, affective vocalizations consistently elicit activity in cortical auditory regions to receive acoustical decoding (Witteman *et al.*, 2012; Frühholz and Grandjean, 2013b; Belyk and Brown, 2014). These regions are represented by bilateral primary and secondary auditory cortex (AC), referred to as 'core' and 'belt' auditory regions (Kaas and Hackett, 2000), and in regions of higher-level auditory cortex, referred to the 'parabelt' auditory region (Kaas and Hackett, 2000). The latter is located in the lateral superior temporal cortex (STC), which is composed of the superior temporal gyrus (STG) and the superior temporal sulcus (STS) (Frühholz and Grandjean, 2013b). These activations in AC/STC in response to affective vocalizations are often accompanied by activity in the amygdala (Irwin *et al.*, 2011; Frühholz and Grandjean, 2013a; Frühholz *et al.*, 2014b), which is thought to decode the affective valence of sounds that were acoustically decoded by the AC/STC according to central voice features (Kumar *et al.*, 2012).

Regarding these relevant acoustic voice features, the pitch of vocalizations or spectral high-frequency voice cues, for example, are important acoustic cues for specific affective vocalizations (Patel *et al.*, 2011). The decoding of these acoustic voice features is associated with left and especially right STC activity, located in mid STC (mSTC) and posterior STC (pSTC) (Ethofer *et al.*, 2006; Wiethoff *et al.*, 2008; Leitman *et al.*, 2010b; Frühholz *et al.*, 2012). Other important vocal features determine the voice quality, such as the harmonics-to-noise ratio (HNR) (Lewis *et al.*, 2009) or the ratio of energy between high- and low-frequency bands (Banse and Scherer, 1996; Leitman *et al.*, 2010b; Patel *et al.*, 2011). The latter ratio is called the 'alpha ratio' (Sundberg and Nordenberg, 2006) indicating the relative rather than the absolute amount of energy in high- compared to low-frequency bands (Patel *et al.*, 2011). By contrast, the HNR refers to the ratio of spectral energy of frequency bands at the level of the voice harmonics relative to all other frequency bands. The HNR level of vocalizations seems to be neurally decoded between the low- and higher level bilateral auditory regions at the border between the lateral Heschl's gyrus (HG) and the mSTC (Lewis *et al.*, 2009), whereas the neural effects of the variations in the alpha ratio are so far unknown. Thus, affective voices consist of several acoustic cues that are relevant for their classification, and the neural decoding of these acoustic cues seems to predominantly involve regions in the AC and pSTC (Wiethoff *et al.*, 2008; Lewis *et al.*, 2009).

Until now, only a few studies have used functional imaging to investigate the sensitivity of the AC/STC to acoustic features of affective vocalizations that are relevant to discriminate between different vocal valences (Ethofer *et al.*, 2006; Wiethoff *et al.*, 2008; Leitman *et al.*, 2010b). Furthermore, these studies used only univariate data modeling approaches, and the results point to sensitivity of a *single* cortical area in the pSTC to a multitude of affective voice features (Ethofer *et al.*, 2006; Wiethoff *et al.*, 2008) close to the planum temporale (PTe) (Leitman *et al.*, 2010b). An adjacent area in pSTC lateral to the HG also seems to be sensitive to a variety of acoustic features (e.g., sound pitch and pitch strength, spectral variability, etc.) across several object classes, such as animal and human vocalizations as well as musical instruments (Leaver and Rauschecker, 2010).

These previous data point to a circumscribed area in pSTC that seems sensitive to variety of important voice features. However, affective vocalizations consist of several first- and second-order acoustic features. First-order features mainly concern the overall *mean* level of such features, whereas the *variability* of these features represents a second-order feature (McGillivray *et al.*, 2012). Furthermore, they are able to elicit distributed neuronal activity across several low- and higher level auditory regions (Frühholz *et al.*, 2012; Frühholz and Grandjean, 2013b). We thus assumed that *several* areas rather than a *single* cortical brain area should be differentially sensitive to acoustic features of affective vocalizations. Furthermore, given that reduced clarity and distinctiveness of acoustic cues can lead to enhanced top-down decoding in the inferior frontal cortex (IFC) (Leitman *et al.*, 2010b), we may assume that the acoustical decoding of affective vocalizations not only involves a local decoding in auditory cortical areas, but also in an extended neural network involving frontal cortices (Frühholz and Grandjean, 2012).

To determine this extended neural network we here accordingly used high-field functional magnetic resonance imaging (fMRI) including a high-spatial resolution partial volume acquisition of the AC, the STC, the IFC and subcortical auditory brainstem structures in response to affective vocalizations. Unlike recently used univariate linear modeling approaches of fMRI data, we used a systematic experimental approach, which was termed 'feature elimination approach' (Rauschecker, 1998), in combination with a multivariate pattern classification analysis (Kriegeskorte *et al.*, 2006). The feature elimination approach is based on the rationale that if the elimination of one feature leads to a change in neural categorical responses, this feature is deemed essential for the recognition of the stimulus it belongs to. Multivoxel pattern classification was performed here using searchlight analysis (Kriegeskorte *et al.*, 2006; Kriegeskorte and Bandettini, 2007), which assesses the ability of local regions to discriminate between experimental conditions.

In contrast to presenting acoustic features in isolation, we presented these features embedded in human affective vocalizations because of two reasons. First, we hypothesized that an initial coarse processing or prediction of a voice valence might be a meaningful and behaviorally important context, which might weight certain relevant features that allow to properly discriminating between acoustically or emotionally similar vocalizations. Second, instead of a single cortical brain region that would be sensitive to a variety of acoustic features (Ethofer *et al.*, 2006; Wiethoff *et al.*, 2008; Leitman *et al.*, 2010b), we hypothesized to find an extended neural network of auditory cortical and frontal brain regions that weights and integrates the valence-dependent context and the discriminative voice features. Concerning the latter we also expected to find activity in subcortical brain structures of the ascending auditory pathway, such as the medial geniculate nucleus (MGN), which are generally sensitive to the spectro-temporal patterns of complex sounds (Wenstrup, 1999; De Martino *et al.*, 2013) that support their classification.

We investigated neural decoding, first, of negative (i.e., aggressive expressions of 'hot' anger; Patel *et al.*, 2011) and positive vocalizations (i.e., joyful expressions), which imply different behavioral adaptive responses, but which are acoustically very similar (Patel *et al.*, 2011). Second, we hypothesized that this efficient processing is accomplished by decoding relevant and discriminative acoustic features of vocalizations. We accordingly tested the processing of vocalizations that were manipulated according to three features. These three features were

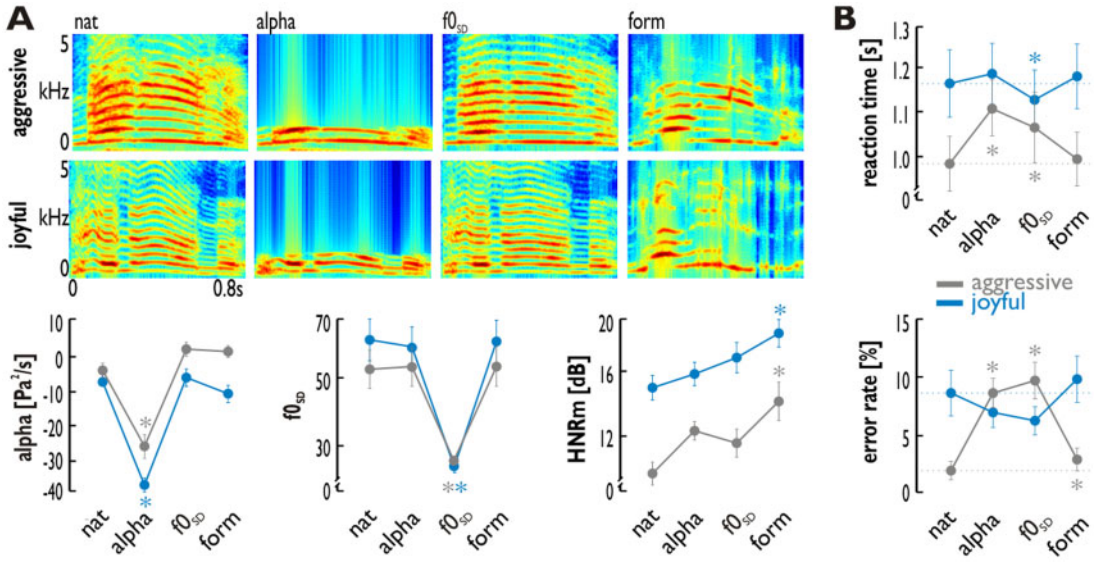

Fig. 1. Stimulus spectrograms, acoustic features and behavioral data. (**A**) The experiment included 16 aggressive and 16 joyful voices presented as native voices ('nat') or as manipulated voices. Shown are spectrograms (time on the *x*-axis, frequency on the *y*-axis) of example stimuli. The latter were created by eliminating one critical acoustic feature, resulting in 1 kHz low-pass filtered stimuli ('alpha', defined by the ratio of energy in high, >1 kHz, and low frequency bands, <1 kHz), in voices with reduced variation of the f0 ('$f0_{SD}$'), or in formant-filtered voices ('HNR_M') to increase the HNR level. Low-pass filtering resulted in a significant decrease in the alpha level of the stimuli (left panel); the variation of the f0 was significantly decreased in $f0_{SD}$ manipulated voices (middle panel); and formant-filtered voices had a significantly increased level of the mean HNR (right panel). Asterisks indicate a significant change ($P < 0.05$; ANOVA post-hoc pairwise comparison) of a specific acoustic feature elimination of manipulated compared to the native vocalizations. (**B**) Reaction times (upper panel) and percentage error rates (lower panel) for the discrimination task computed on the behavioral data of the 13 participants. Asterisks indicate a significant ($P < 0.05$; post-hoc paired *t*-tests) difference for manipulated compared with native voices. Error bars indicate the standard error of the mean (SEM).

high-frequency spectral cues (i.e. the 'alpha' ratio), the variation or fluctuation of the vocal pitch (i.e. the standard deviation of the fundamental frequency f0, '$f0_{SD}$', which mainly determines the vocal pitch), and the level of the HNR (i.e. as changed by formant-filtering, 'form') (Figure 1A).

The rationale for using these three types of acoustic manipulations was based on the observation that these features contribute specifically to the acoustic pattern of affective vocalizations (Banse and Scherer, 1996; Patel *et al.*, 2011) and that listeners usually strongly rely on these features for decoding and discriminating the affective valence of these vocalizations (Banse and Scherer, 1996). Joyful and aggressive expressions show increased spectral energy in the high-frequency band (above 1 kHz) as an important cue for listeners (Juslin and Laukka, 2003; Leitman *et al.*, 2010a; Patel *et al.*, 2011). Similarly, joyful and aggressive expressions usually show an increased pitch variation indicated by an increased temporal standard deviation of the f0 (Banse and Scherer, 1996; Juslin and Laukka, 2003; Leitman *et al.*, 2010a). Finally, the HNR is a strong acoustic cue for recognizing vocalizations of aggression and joy (Patel *et al.*, 2011). Though these acoustic cues are often common to aggressive and joyful vocalizations, they can also discriminate aggressive from joyful vocalizations, such as for the variation of the vocal pitch (Banse and Scherer, 1996) or the HNR level (Hammerschmidt and Jurgens, 2007).

Methods

Participants

Thirteen healthy participants recruited from the École Polytechnique Fédérale de Lausanne (EPFL) and the University of Geneva took part in the experiment (seven male; mean age 24.00 years, SD = 5.65, range 19–39 years). All participants were right-handed, had normal or corrected-to-normal vision, and had normal hearing abilities. No participant presented a neurologic or psychiatric history. All participants gave written informed consent for their participation in accordance with ethical and data security guidelines of the University of Geneva and the EPFL. The study was approved by the Cantonal Ethics Committee of the Canton Vaud (Switzerland).

Stimulus material and trial sequence

Amongst the different types of auditory objects, human vocalizations are the predominant and socially most important auditory objects, which are processed in a circumscribed area of the auditory cortex (Belin *et al.*, 2000). The stimulus material used in our experiments thus consisted of four speech-like but semantically meaningless two-syllable words – '*loman*', '*belam*', '*minad*', '*namil*' – spoken in either a neutral, positive (i.e., joyful vocalizations signaling 'affiliation'), or negative tone (i.e., hot anger that displays an aggressive vocal tone indicating 'social confrontation' or withdrawal) by two male and two female speakers (Frühholz *et al.*, 2014a), resulting in 16 neutral and 32 affective voice stimuli (Figure 1A). Auditory stimuli had a mean duration of 627 ms (SD = 177 ms) with a similar duration for aggressive and joyful voices (*t*-test with degrees of freedom in lower case; $t_{30} = 0.561$, $P = 5.79 \times 10^{-1}$; two-tailed independent samples *t*-test) and were equated for mean energy ($M_{erg} = 4.93 \times 10^{-3}$ Pa²/s).

All stimuli were rated on five different dimensional scales ('neutrality', 'anger', 'fear', 'happiness' and 'sadness'; ratings indicated from '0' = low to '100' = high the perceived amount of each dimension) by an independent sample of 21 participants (10 male; mean age 25.57 years, SD = 3.69, age range 22–34

years). Using separate one-way ANOVAs including the five dimensional scales as within-subject factor, the ratings revealed that neutral, joyful and aggressive stimuli were significantly rated as neutral (*F* test with between- and within-group degrees of freedom in lower case; $F_{4,80} = 207.002$, $P = 2.36 \times 10^{-41}$), joyful ($F_{4,80} = 27.551$, $P = 2.17 \times 10^{-14}$) and aggressive ($F_{4,80} = 216.442$, $P = 4.65 \times 10^{-42}$) on the relevant target dimension, respectively. Based on a two-tailed independent samples *t*-test we ensured that joyful and aggressive voices did not differ in arousal ratings ($t_{30} = 1.436$, $P = 1.66 \times 10^{-1}$). During the experiment, we presented only joyful and aggressive voices, whereas neutral voices served as a reference for creating fundamental frequency (f0) modulated affective voices (see below).

To reveal the influence of central acoustic features of affective vocalizations on subcortical and cortical activations, we adopted a feature elimination approach, which has been previously described to determine the acoustic feature sensitivity of neural populations (Rauschecker, 1998). The experiment included four different conditions for each of the joyful and aggressive vocalizations (Figure 1A): (i) native vocal expressions ('nat'), (ii) low-pass filtered vocalizations using a cutoff frequency of 1 kHz ('alpha'), (iii) affective vocalizations with a reduced variation of the f0 such that it matched the level of f0 variation in the corresponding neutral expressions on the same word by the same speaker ('$f0_{SD}$') and (iv) formant-filtered vocal expressions with increased HNR ('form'). The first voice manipulation, creating low-pass filtered vocalizations ('alpha ratio'), was based on the above-mentioned observation that increased spectral energy in high-frequency bands above 1 kHz or the relative higher energy in high-compared to low frequency bands is an important acoustic feature of joyful and aggressive voices, which makes them distinct from other affective vocalizations, such as fearful or sad vocalizations (Juslin and Laukka, 2003; Leitman *et al.*, 2010a; Patel *et al.*, 2011). For this manipulation, we used a low-pass filtering approach in the Praat software (Boersma, 2001). Low-pass filtering was done by using a Hann pass band filter of 0–1 kHz, including a roll-off of 100 Hz. This filtering nearly eliminated all energy of frequencies above 1 kHz, and thus led to a significant drop of the alpha ratio for these modified vocalizations.

The second manipulation ('$f0_{SD}$') was based on the observation that increased pitch variations and pitch dynamics is a central feature of many affective vocalizations, especially of joyful and aggressive vocalizations (Banse and Scherer, 1996; Juslin and Laukka, 2003; Leitman *et al.*, 2010a). For the second manipulation of vocalizations with reduced f0 variation, we determined the f0 standard deviation for each joyful and aggressive voice, as well as in neutral expressions spoken by the same actors. We used the STRAIGHT toolbox for Matlab (www.wakayama-u.ac.jp/~kawahara/STRAIGHTadv/index_e.html) to determine the f0 in each vocalization, as well as for the resynthesis of the vocalizations with a reduced f0 standard deviation. We first extracted the f0 standard deviation of all neutral vocalizations of the same words by the same actor. Subsequently, we changed the variation of the f0 along the mean f0 of the corresponding joyful and aggressive vocalizations of the same word spoken by the same actors, such that the f0 standard variation matched that of the neutral expression. This manipulation changed only the temporal variation of the f0, but the general pattern of the f0 in terms of rises and falls remained unchanged. We resynthesized the joyful and aggressive expressions with the manipulated f0 by using the acoustic resynthesis algorithms implemented in the STRAIGHT toolbox. During the resynthesis, spectral information is sampled at each vocal excitation pulse time and converted to the minimum phase impulse response. An overlap-and-add process was used to generate the synthetic vocalizations. For unvoiced components a convolution with shaped noise was added. We have to note here that native and manipulated neutral voices could not be included in the experimental design as additional conditions, because the f0 variation of neutral voices served as reference for the manipulation of aggressive and joyful voices.

Finally, the third manipulation of formant-filtered vocalizations ('form') was based on the observation that many high-arousing vocalizations usually show a reduced HNR and thus more 'noisiness' or 'roughness' because of an increased spectral energy between the harmonic frequency bands, especially of vocal expressions of aggression or fear (Patel *et al.*, 2011). To increase the HNR of affective vocalizations, we therefore used formant filtering as implemented in the Praat software to increase the 'harmonicity' of the vocal stimuli. Vocal harmonicity is represented by voice formants that result when sound waves with a certain f0 are emitted from vocal fold vibrations and passed through the vocal tract. Sound waves are reflected in the vocal tract leading to resonances that represent the voice harmonics. The shape of the vocal tract determines its specific resonance frequencies, which emphasizes certain vocal harmonics while it dampens others. For each vocalization, we therefore determined the temporal pattern of the mean and the bandwidth (BW) of the first five formants (F1–F5) by using a sliding temporal time window of 25 ms, a pre-emphasis of 50 Hz to create a flatter spectrum and a maximum formant frequency of 5.5 kHz. While keeping the temporal pattern of the mean F1–F5 unchanged, we reduced the BW of each formant at each time point to 66% of the original size ($0.66 \times$ BW), thus narrowing the region around the center frequency of the formant at each time point. We used this manipulated temporal formant pattern as a filter template and superimposed this template on the original sound. This resulted in manipulated voices with an emphasis on high-energy areas in formant bands and with a filtering of spectral energy between formant frequency bands that fell outside the +/-0.66 BW range, thus decreasing spectral 'noise' between the formants that leads to increased 'harmonicity' of the voices.

The $f0_{SD}$ manipulation resulted in vocalizations that sounded marginally artificial as confirmed by two of the authors (SF and DG). This artificiality was introduced by the resynthesis procedure. We therefore subjected all final stimuli to the resynthesis procedure as used for the $f0_{SD}$ manipulated voices to equate all stimuli in terms of artificiality. Finally, all vocalizations were low-pass filtered with a cutoff frequency of 5 kHz to account for the fact that the formant-filtered voices had a maximum formant frequency of 5.5 kHz, and all acoustic stimuli were scaled to a mean sound pressure level (SPL) of 70 dB. This scaling was done with the Praat software.

We scored the main acoustic features of the final stimuli to ensure that the manipulation resulted in the elimination of the intended feature while other features were unaffected. These scores were subjected to a 2×4 repeated measures analysis of variance (ANOVA) with the within-subject factors *affective* valence (aggressive, joyful) and *manipulation condition* (nat, alpha, $f0_{SD}$, form). The low-pass filtering of the voices resulted in a significant decrease of the alpha level for these stimuli indicated by a significant main effect of the factor *manipulation condition* ($F_{3,93} = 112.003$, $P = 9.17 \times 10^{-31}$). Planned post-hoc comparisons revealed a significant difference of the alpha level of low-pass filtered voices compared with the alpha level of native voices, as indicated by a Bonferroni-corrected post-hoc comparison

($P = 1.90 \times 10^{-11}$), while the alpha level of other voices did not differ significantly from native voices (all P's $> 6.11 \times 10^{-2}$). The f0 standard deviation was significantly lower in $f0_{SD}$ manipulated voices (main factor *manipulation condition*: $F_{3,93} = 68.306$, $P = 1.99 \times 10^{-23}$) compared with native voices ($P = 4.40 \times 10^{-9}$), while other voices did not differ from native voices (all P's > 0.98), as indicated by a Bonferroni-corrected post hoc comparison. Finally, the mean HNR (HNR_M) was significantly increased in formant-filtered voices (main factor *manipulation condition*: $F_{3,90} = 9.867$, $P = 1.01 \times 10^{-5}$), whereby the HNR_M was different from native voices only in formant-filtered voices ($P = 4.99 \times 10^{-7}$), but not for all other voices (all P's $> 8.06 \times 10^{-2}$), as indicated by a Bonferroni-corrected post hoc comparison. Other important acoustic features of affective vocalizations did not differ across the manipulations conditions, such as the intensity standard variation ($F_{3,90} = 2.352$, $P = 7.81 \times 10^{-2}$), the f0 mean ($F_{3,90} = 2.035$, $P = 1.15 \times 10^{-1}$), or the HNR standard deviation ($F_{3,90} = 1.347$, $P = 2.64 \times 10^{-1}$). Altogether, these acoustical analyses indicated that the amount of acoustic *change* for all three types of acoustic manipulations was similar for aggressive and joyful voices (i.e. the acoustic manipulation did not influence the affective voices differentially). The neural processing of these changes of acoustic features embedded in the context of affective vocalizations were the main focus of the current study.

We performed a separate evaluation of these manipulated stimuli by another independent sample of 16 participants (seven male; mean age 22.88 years, SD = 2.82, age range 19–29 years), which were asked rate the arousal level and to classify the vocalizations as aggressive or joyful. This results of this evaluation indicated that, according to a 2 (aggressive, joyful) × 4 (nat, alpha, $f0_{SD}$, form) repeated measures ANOVA, the different *manipulation procedures* did not change the affective valence of the stimuli across the different manipulation conditions ($F_{3,45} = 0.502$, $P = 6.83 \times 10^{-1}$) and between the *affective valences* ($F_{1,15} = 0.165$, $P = 6.90 \times 10^{-1}$). There was no *valence* × *manipulation* interaction ($F_{3,45} = 0.099$, $P = 9.60 \times 10^{-1}$). The arousal level did not differ for aggressive compared with joyful voices ($F_{1,15} = 1.621$, $P = 2.22 \times 10^{-1}$), but did differ between the different *manipulation levels* ($F_{3,45} = 19.499$, $P = 3.02 \times 10^{-8}$). Bonferroni-corrected comparisons indicated that all manipulated voices were rated as significantly less arousing compared with native voices ($M = 48.64$, SEM = 3.16), that is, alpha manipulated voices ($M = 33.30$, SEM = 4.09; $P = 1.57 \times 10^{-4}$), $f0_{SD}$ voices ($M = 40.67$, SEM = 3.71; $P = 1.01 \times 10^{-3}$) and formant-filtered voices ($M = 37.35$, SEM = 3.76; $P = 1.13 \times 10^{-2}$) revealed lower arousal ratings.

During scanning, auditory stimuli were presented binaurally with magnetic resonance imaging-compatible in-ear headphones (Sensimetrics®) at a sound pressure level of 70 dB. Auditory stimuli were presented in the silent gap between image acquisitions (see below). They were preceded by a visual fixation cross (1° × 1°) for 665 ± 66 ms, which cued the onset of the auditory stimulus and remained on the screen until the offset of the auditory stimulus. A blank screen appeared after the offset of the auditory stimulus. After the presentation of the voices, participants had to indicate the affective valence of the vocal expressions in a two-alternative forced-choice decision task (*aggressive* or *joyful*). Participants responded with their right index and middle finger, and response buttons were counterbalanced across participants. We presented a total of 304 trials, including 48 silent events with no auditory stimulation.

To localize human voice-sensitive regions in the bilateral STC, we used 8 s sound clips taken from an existing database (Belin *et al.*, 2000). These sound clips contained 20 sequences of human voices and 20 sequences of animal or environmental sounds. Each sound clip was presented once with a fixation cross on the screen and a 4 s gap between each clip. The scanning sequence also contained twenty 8 s silent events and participants had to passively listen to the stimuli.

For the purpose of a tonotopic mapping of the AC (see Da Costa *et al.*, 2011), we used pure sine wave tones (PSTs) presented in an ascending progression of 14 tones from low to high frequency in half-octave steps in the range 0.88–8 kHz (i.e., 0.88, 0.125, 0.177, 0.25, 0.354, 0.5, 0.707, 1, 1.414, 2, 2.828, 4, 5657 and 8 kHz). Pure tone bursts of a random rhythmic pattern of single PSTs (50 or 200 ms) of a specific frequency were presented for every 2 s with a duration of 1.1–1.7 s, starting with the lowest frequency and then immediately stepping to the next higher frequency. Thus, there were fourteen 2 s-PST-bursts in progression, thus resulting in a 28 s block of a low-to-high PST sequence. This 28 s block was repeated 15 times and the blocks were separated by a 4 s silent pause. Sound intensity of the PSTs was adjusted in terms of a standard equal loudness curve (ISO 226, phon 65) to achieve an equal perceived volume of PSTs.

Image acquisition

For the main experiment, we obtained high-resolution imaging data on a 7-T Siemens Magnetom® System (Siemens, Erlangen, Germany) by using a T2*-weighted gradient echo-planar imaging sequence with sinusoidal readout gradient (Speck *et al.*, 2008). The use of this ultra-high magnetic field system allowed us to record with an increased signal-to-noise ratio together with higher spatial resolution in terms of voxel size. Twenty-five axial slices were obtained in an oblique orientation, rotated about 30° in reference to the AC-PC plane (flip angle 90°, slice thickness/ga P = 1.5/0.3 mm, field of view 222 × 222, in-plane 1.5 × 1.5 mm, matrix size 148 × 148, GRAPPA factor 2, bandwidth/voxel was 1876 Hz). This study had a focus on neural processing auditory features in the neural auditory system, and the slices were thus positioned to cover the cortical and subcortical neural auditory system (i.e. auditory cortex, MGN, inferior colliculi), but did not include a full coverage of other subcortical regions, such as the limbic system (i.e. the amygdala). We used a short-gap sparse temporal acquisition protocol with a repetition time (TR) of 3.29 s, which consisted of 1.47 s for volume acquisition (TA) and 1.82 s for a silent gap. For the voice localizer and tonotopy localizer scans, we used an acquisition volume of 25 slices with the same slice orientation as above, but a continuous volume acquisition (TR/TE = 2 s/25 ms, flip angle 90°, slice thickness/ga P = 1.5/0.3 mm, field of view 222 × 222, in-plane 1.5 × 1.5 mm, matrix size 148 × 148, GRAPPA factor 2, bandwidth/voxel was 1876 Hz). Finally, a high-resolution magnetization-prepared rapid acquisition gradient echo (MP2RAGE; Marques *et al.*, 2010) T1-weighted structural brain image, optimized for 7T MRI (sagittal orientation, slice thickness 1 mm, TR/TE = 5.5 s/2.84 ms, field of view 256 × 240 mm, in-plane 1 × 1 mm), was obtained to provide an anatomical reference for each participant.

Image analysis

We used the statistical parametric mapping software SPM (Version 8; Welcome Department of Cognitive Neurology, London, UK) for preprocessing and statistical analysis of all functional images. Fieldmap images were recorded with two echo times (TE 4 and 5.02 ms) and were used to correct EPI

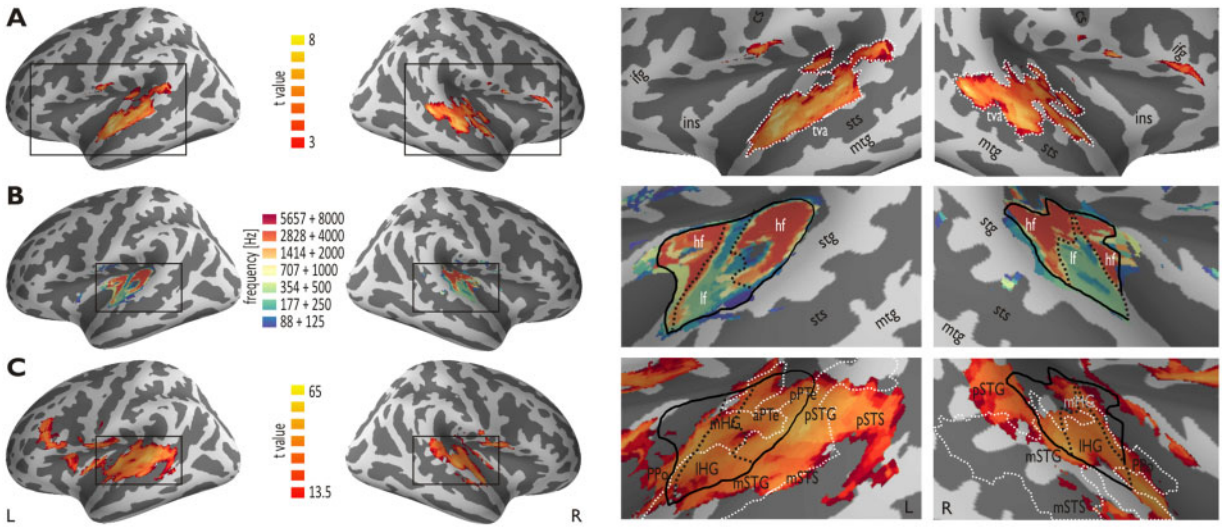

Fig. 2. Functional Localizer Scans. (**A**) The voice localizer scan across all participants ($n = 13$) revealed extended activity (FDR $P < 5.00 \times 10^{-2}$, $k = 50$) in the bilateral STC, which was mainly located in the STG, but partly also in the lateral HG and the PTe and the planum polare (PPo). The right panel shows enlarged views of the black rectangle in the left panel. The temporal voice area (TVA) is outlined by the white dotted line. (**B**) The tonotopic localizer scan across all participants revealed bilateral tonotopic maps located on the HG, the PTe and the PPo (FDR $P < 5.00 \times 10^{-2}$, $k = 50$). The maps consisted of a high-low-high gradient from anterior-to-posterior with the anterior high-low frequency boundary located approximately on the HG. The dotted line marks the high-low frequency boundaries. The color bar refers to the seven combined PST frequency levels, such that two successive frequency levels (e.g. 88 Hz + 125 Hz) were combined in one condition, resulting in seven different frequency conditions out of the fourteen frequency levels. (**C**) Cortical regions where decoding accuracy for the distinction of all eight experimental conditions was above chance level (12.5%) across all participants (FDR $P < 10^{-3}$, $k = 50$). The regions covered large parts of the areas determined by the voice (white outline; TVA) and the tonotopic localizer scan (black outline). The labels in the bottom-right part of the image represent the functional ROIs as defined by the boundaries resulting from the functional localizer scans. Abbreviations: *cs* central sulcus; *hf* high frequency; *ifg* inferior frontal gyrus; *ins* insula; *lf* low frequency; *mtg* middle temporal gyrus; *sts* superior temporal sulcus; *tva* temporal voice area.

distortions in the functional images during preprocessing. Functional images were realigned and unwarped (to correct for EPI distortions), and were subsequently coregistered to the anatomical image. The *New Segment* option in SPM8 was used to perform a unified segmentation approach of individual T1 anatomical images. Individual DARTEL flow fields were estimated on the basis of segmented grey and white matter tissue classes and used for normalizing T1 and EPI images to the MNI space.

We used a general linear model (GLM) for the first-level statistical analyses of the localizer scans, including boxcar functions defined by the onset and duration of the auditory stimuli. These boxcar functions were convolved with a canonical hemodynamic response function. Separate regressors were created for each experimental condition. Six motion correction parameters were finally included as regressors of no interest to minimize false positive activations that were due to task-correlated motion. Linear contrasts for the experimental conditions for each participant were taken to a second-level random effects analysis.

For the functional data of the main experiment, we performed a searchlight decoding analysis as implemented in the PyMVPA package (http://www.pymvpa.org/) for the purpose of an information-based brain mapping according to multivoxel activity patterns (Kriegeskorte *et al.*, 2006; Kriegeskorte and Bandettini, 2007), which differ between the experimental conditions. The searchlight analysis was performed on normalized but unsmoothed functional data. This analysis was performed on single-trial beta images resulting from an iterative GLM analysis as recommended for better sensitivity in event-related design (Mumford *et al.*, 2012). We obtained beta images for each trial by using a GLM with one regressor modeling a single trial and a second regressor modeling all remaining trials. This GLM modeling was repeated for each trial, including movement parameters as a regressor of no interest to account for false positive activity due to head movements. Only trials were included where participants gave a correct behavioral response (93.13% of all trials); thus participants performed at a high level of accuracy. Trials were convolved with a canonical hemodynamic response function.

For each voxel, we defined a local sphere of 4.5 mm radius to investigate the local multivoxel pattern information in the single-trial beta images that was able to differentiate across the eight experimental conditions (two affective valences, four manipulation conditions). We trained a multivoxel support vector machine classifier by using a linear kernel (linear C-SVM), which was trained on 5/6 of the trials and tested on the remaining 1/6 of the trials. Using a cross-validation approach, we repeated this procedure six times according to a leave-one-part-out test. This procedure finally resulted in a brain map of local decoding accuracy across the experimental conditions for each participant. The resulting accuracy maps were spatially smoothed with an isotropic Gaussian kernel of 4.5 mm^3 FWHM and subjected to second-level random effects GLM analysis consistent with the group analysis procedure of the localizer scans. On the second level, we tested the resulting map for statistical significance against the chance level of 12.5% by using a very conservative statistical threshold of $P < 10^{-3}$ (FDR corrected; corresponding to an initial voxel threshold of $P < 10^{-5}$) and a cluster extent of $k = 50$ voxels to considerably minimize the possibility of false positive activity patterns.

The resulting thresholded statistical map of informative brain voxels represented brain areas that can discriminate several of the experimental conditions (Figure 2C). Thus, the mask contained informative regions, which were able to distinguish at least two out of the eight conditions, but it was not informative about which specific conditions could be distinguished in a certain brain voxel. We thus performed several follow-up

searchlight decoding analyses, but using a binary decoding procedure (Hebart *et al.*, 2012; Kotz *et al.*, 2013) to perform specific comparisons between the native condition and each of the three manipulation conditions for each of the aggressive and the joyful vocalizations separately. Out of all possible binary comparison, only these specific comparisons out of all possible pairwise comparisons concerned our main experimental hypothesis. These follow-up decoding analyses were performed in voxels that were generally informative as revealed by the first step of the decoding analysis. This resulted in three different statistical maps for each affective category, indicating brain areas where multivoxel classification patterns can distinguish between native and manipulated voices. These maps were again determined for each participant and then subjected to a second-level 2 × 3 factorial design, including the factors *affective valence* (aggressive, joyful) and *manipulation* (alpha, $f0_{SD}$, form). To avoid false positive results, the statistical maps were tested for statistical significance against the chance level of 50% (i.e., binary classification) by using a very conservative statistical threshold of $P < 1.00 \times 10^{-3}$ (FDR corrected) and a cluster extent of $k = 50$ voxels.

Functional images of the voice localizer scan were preprocessed as described for the images of the main experiment, but were additionally smoothed with an isotropic Gaussian kernel of 4.5 mm^3 FWHM. We contrasted vocal against nonvocal animal and environmental stimuli at a threshold of $P < 5.00 \times 10^{-2}$ (FDR corrected) and a cluster extent of $k = 50$ voxels. We determined voice-sensitive regions along the STG and STS in both hemispheres for each participant, as well as for the entire sample.

Functional images of the tonotopic localizer scan were preprocessed as described above and spatially smoothed with an isotropic Gaussian kernel of 4.5 mm^3 FWHM. We set up a 1 × 7 factorial design, where two successive levels of the 14 PST frequency levels were combined in a single condition, thus resulting in seven different conditions. This combination across two frequency levels allowed us to increase the statistical power of the main frequency bands, and resulted in a frequency band resolution, which was still appropriate to determine low- and high-frequency fields in the AC. By using an *F* contrast across these different conditions, we determined areas in the AC that showed a significant difference between these conditions at a combined threshold of $P < 5.00 \times 10^{-2}$ (FDR corrected) and a cluster extent of $k = 50$ voxels. The cluster extend threshold was especially used here to find coherent cortical fields with tonotopic gradients and to exclude small local of sensitivity to only a specific tonotopic frequency. For all voxels in the resulting statistical map, we determined the maximum response across all seven conditions (i.e., winner-take-all). Each voxel was color coded according to its maximum response to one of the seven conditions.

Results

Behavioral data

Participants had to indicate if they heard an aggressive or a joyful voice while listening to native and to feature-manipulated voices, and while their brain response were recorded using high-field magnetic resonance imaging. Reaction times (RTs) and percentage error rates of their classifications according to the affective valence were subjected to a 2 × 4 repeated measures analysis of variance (ANOVA) with the within-subject factors *affective valence* (aggressive, joyful) and *manipulation condition* (nat, alpha, $f0_{SD}$, form; Figure 1B).

RTs did not differ across the affective valence ($F_{1,12} = 3.674$, $P = 8.75 \times 10^{-2}$), but did differ across the different types of manipulation conditions ($F_{3,36} = 7.376$, $P = 9.21 \times 10^{-4}$). There was also an *valence* × *manipulation* interaction ($F_{3,36} = 8.405$, $P = 4.17 \times 10^{-4}$), which was driven by significantly increased RTs for alpha manipulated ($t_{12} = 7.352$, $P = 9.00 \times 10^{-6}$) and $f0_{SD}$ manipulated aggressive voices ($t_{12} = 3.601$, $P = 3.64 \times 10^{-3}$) compared with native aggressive voices, as well as by significantly faster RTs for $f0_{SD}$ manipulated joyful voices ($t_{12} = 5.765$, $P = 2.71 \times 10^{-4}$) compared with joyful native voices. Error rates did not differ across affective valences ($F_{1,12} = 2.087$, $P = 1.74 \times 10^{-1}$), but did differ across the manipulation conditions ($F_{3,36} = 4.773$, $P = 2.38 \times 10^{-2}$). We also found a *valence* by *manipulation* condition interaction ($F_{3,36} = 16.945$, $P = 8.27 \times 10^{-4}$). While all manipulated aggressive voices revealed increased error rates compared with native voices (all $t_{12} > 2.309$, $P < 3.95 \times 10^{-2}$), only $f0_{SD}$ manipulated joyful voices revealed a trend towards a lower error rate compared with native joyful voices ($t_{12} = 2.081$, $P = 5.94 \times 10^{-2}$).

These data suggest that especially pitch variations and the high-frequency spectral cues are important for the discrimination of aggressive vocalizations, while lowering the pitch variation might have rendered joyful vocalizations more distinct from aggressive vocalizations leading to an improved discrimination.

Imaging data

To determine neural voice-sensitive areas, we first performed a voice localizer scan (Belin *et al.*, 2000). Vocal compared with nonvocal sounds elicited extended activity in the bilateral STC, covering regions in the primary and secondary AC and especially in the anterior STC to pSTC (Figure 2A).

Second, to determine the frequency sensitivity of regions in the AC we performed a tonotopic localizer scan. Using 0.88–8 kHz sine wave tones, we revealed a typical pattern of frequency maps with a mirror-symmetric frequency progression (high-low-high) centered on the HG (Da Costa *et al.*, 2011), with the low-frequency area approximately located along Heschl's sulcus (Figure 2B). These two maps are thought to correspond to regions A1 and R (rostral area) of the primate primary auditory core (Saenz and Langers, 2014).

Third, we performed a searchlight decoding analysis for the purpose of an information-based brain mapping. In a first step, we identified informative brain regions that were generally able to discriminate between several of the eight experimental conditions, resulting in an extended brain network consisting of the bilateral AC and STC (Figure 2C). They covered most of the cortical brain areas of interest, as determined by the voice localizer scan and the tonotopic localizer scan. Additional regions, which were able to discriminate between conditions, were found in the bilateral inferior frontal gyrus (IFG), in the frontoparietal operculum, in the bilateral occipital areas, and in the left insula. Finally, we also found that the left MGN was able to discriminate between all eight conditions.

On the basis of the spatial distribution of this discrimination accuracy in the MGN, the AC and the STC, we defined several regions of interest (ROIs). These ROIs were created based on the functional boundaries as defined by the results of different localizer scans on the group level. Within these boundaries (see below) we determined local peak maxima across participants. ROIs were created as 4.5 mm radius spheres around these peak maxima. The ROIs were used to score the local decoding accuracy resulting from the searchlight analysis to specifically perform *interaction analyses* between the experimental factors that

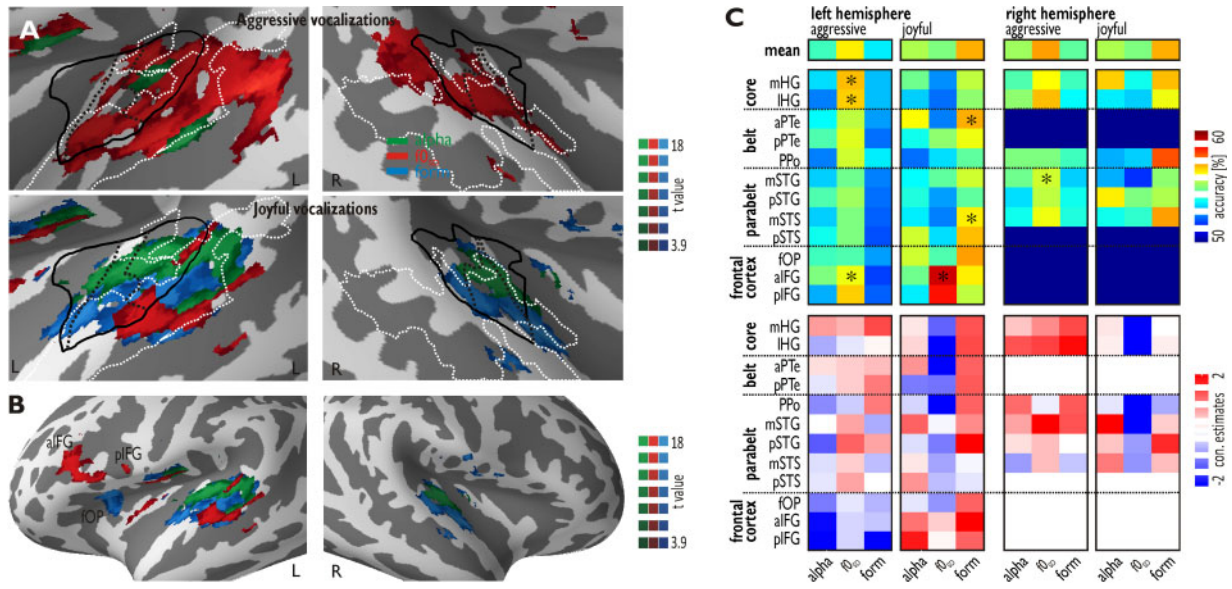

Fig. 3. Cortical Regions for the Decoding of Vocal Acoustic Features. (**A**) Regions in the bilateral STC where decoding accuracy across all participants (**n** = 13) was above chancel level (50%) for the $f0_{SD}$ manipulated (red) and formant-filtered voices (green) compared with native aggressive voices (upper panel) (FDR $P < 10^{-3}$, $k = 50$). The lower panel shows regions with significant decoding accuracy for manipulated compared with native joyful voices. (**B**) For joyful voices, we also found significant activity across all participants in three subregions of the left IFG, located in the anterior IFG (aIFG) and posterior IFG (pIFG), as well as in the frontal operculum (fOP). (**C**) The upper panel shows decoding accuracies above chance level across all participants in all regions, which showed significant decoding accuracy for aggressive or joyful voices. The top row ('mean') shows the mean decoding accuracy across all left and right ROIs. Asterisks indicate a specific significant interaction between *manipulation condition* (alpha, $f0_{SD}$, form) and *affective valence* ($P < 5.00 \times 10^{-2}$, two-way ANOVA). The lower panel shows contrast estimates for the same regions computed by the difference of beta estimates for native compared with manipulated voices. Red indicates higher activity for native voices, while blue indicates higher activity for manipulated voices.

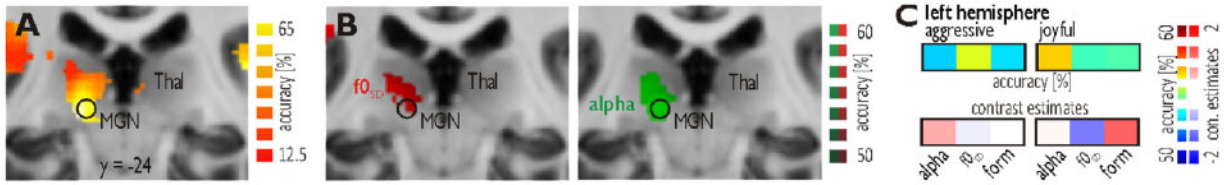

Fig. 4. Decoding of vocal acoustic features in the MGN. (**A**) In the left MGN (encircled in black), decoding accuracy across participants ($n = 13$) for the distinction of all eight experimental conditions was above chance level (12.5%) (FDR $P < 10^{-3}$, $k = 50$). (**B**) Increased decoding accuracy in the left MGN for the $f0_{SD}$ manipulated compared with native aggressive vocalizations (left panel, red), and for the alpha ratio manipulated compared with native joyful voices (right panel, green) (FDR $P < 10^{-3}$, $k = 50$). (**C**) Decoding accuracies and contrast estimates for the left MGN.

were not available with common searchlight analyses. No other statistical analyses were performed to prevent circularity of the analyses. These ROIs were also used to score mean contrast estimates derived from a GLM analysis (see above) across conditions and for each trial separately (Figures 3C and 4A).

One ROI was defined at a local peak maximum in the left MGN (MNI *xyz* [-14 -24 -3]; Figure 4A). This peak was located in the MGN as confirmed with a recent stereotactic description of thalamic nuclei (Tourdias *et al.*, 2014). ROIs in the AC were defined according to the tonotopy localizer scan (Frühholz and Grandjean, 2013b). The ROIs of the medial HG (mHG, left [-41 -26 9] and right [48 -20 7]) and the lateral HG (lHG, left [-54 -15 8] and right [56 -20 10]) were located in the low-frequency fields of the tonotopy localizer scan and largely overlapped with the voice-sensitive cortex. The left anterior PTe (aPTe, [-47 -28 8]) and posterior PTe (pPTe [-49 -37 20]) were located in the posterior high-frequency field. The planum polare (PPo, left [-49 -10 7] and right [52 -9 6]) was located anterior to the anterior high-frequency field. The middle STG (mSTG, left [-62 -22 0] and right [64 -18 11]) and the middle STS (mSTS, left [-52 -20 -2] and right [52 -14 -3]) were located at the same y-level lateral to the border of the low-frequency field, while the posterior STG (pSTG, left [-59 -36 17] and right [66 -30 4]) and the left posterior STS (pSTS [-48 -33 6]) were located approximately at the same y-level and lateral to the posterior low-frequency field. All of these ROIs were located inside or close to the voice-sensitive cortex, except for the right pSTG and the left pSTS. Additional ROIs were created for the left anterior IFG (aIFG [-48 16 12]), the left posterior IFG (pIFG [-54 9 10]) and the left frontal operculum (fOP [-31 18 12]; Figure 3B) (Frühholz and Grandjean, 2013c).

In the second step of the searchlight analysis, we used the pattern of significant decoding accuracies across regions resulting from the first step (i.e. all informative voxels, which could discriminate at least two of the experimental conditions) as a brain mask to perform binary classifications. This was done to determine, which voxels can specifically discriminate between native voices and each of the manipulation conditions. It is important to note that this specific information was not available

in the first step of the searchlight analysis. This information was determined separately for each affective category, resulting in six different binary searchlight analyses. For aggressive voices, we found that $f0_{SD}$ manipulated voices are discriminated from native voices in the left MGN (Figure 4B) and several subregions of the AC and the STC, including all left temporal ROIs except for the PPo and the right HG, mSTG and pSTG (Figure 3A, upper panel). Alpha manipulated aggressive voices were discriminated from native aggressive voices in the left aPTe and mSTS. No discriminant pattern of activity was found for formant-filtered aggressive voices. For joyful voices, we found the most widespread pattern for the discrimination of formant-filtered from native voices. This pattern included all left temporal ROIs except for the pSTS as well as the left fOP. A right pattern of activity was found only in the HG, PPo and mSTS (Figure 3A, lower panel). Alpha manipulated joyful voices were discriminated from native voices in the left MGN (Figure 4B) and in the left mHG, aPTe, pPTe, pSTG and mSTS, as well as in the right mHG. Finally, f0 manipulated voices were discriminated from native voices in the left mSTG, pPTe, mSTS and pSTS, as well as in the left aIFG and pIFG.

For all six binary classification analyses, we extracted the local decoding accuracy above the change level (50%) in the ROIs. First, no specific effects were found in subcortical regions, especially in the MGN. The accuracy values for the MGN were subjected to a 2 × 3 repeated measures ANOVA that included the within-subject factors *affective valence* and the *manipulation condition* and revealed no significant effects (all F's < 2.138, P's $> 1.40 \times 10^{-1}$). Second, cortical auditory regions revealed a general and differential pattern of decoding accuracy. We summed the decoding accuracy separately for each of the six conditions across all cortical ROIs (Figure 3C, topmost row). As indicated by an *affective valence* × *manipulation* condition interaction ($F_{2,22} = 53.099$, $P = 3.47 \times 10^{-11}$), we found that generally across all cortical ROIs, for aggressive voices, the $f0_{SD}$ manipulated voices revealed the highest decoding accuracies ($t_{11} = 3.554$, $P = 2.42 \times 10^{-3}$), while for joyful voices, the formant-filtered voices revealed the highest decoding accuracies in many subregions ($t_{12} = 11.806$, $P = 1.29 \times 10^{-9}$; Figure 3C, topmost row). Exceptions from this general tendency were found in subregions of the left IFG, which showed higher decoding accuracy for $f0_{SD}$ manipulated joyful voices (Figure 3C, upper panel).

Third, we explored whether decoding accuracy in the ROIs is sensitive to a specific feature embedded in a specific vocalization in terms of interaction effects between the factors affective *valence* (aggressive, joyful) and *manipulation condition* (nat, alpha, $f0_{SD}$, form). The accuracy values for each cortical ROI were therefore subjected to the same 2 × 3 repeated measures ANOVA as described above. We found main effects of the factor *manipulation condition* in the left aIFG ($F_{2,24} = 3.741$, $P = 3.85 \times 10^{-2}$), indicative of a higher decoding accuracy for $f0_{SD}$ manipulated voices compared with all other manipulated voices irrespective of the affective valence. More specifically, for aggressive voices, we found a *valence* × *manipulation* condition interactions in the left mHG ($F_{2,24} = 4.203$, $P = 2.72 \times 10^{-2}$) and the lHG ($F_{2,24} = 4.903$, $P = 1.63 \times 10^{-2}$), as well as in the right mSTG ($F_{2,24} = 3.416$, $P = 4.94 \times 10^{-2}$), as indicated by higher decoding accuracy for $f0_{SD}$ manipulated aggressive compared with $f0_{SD}$ manipulated joyful voices (post hoc t-tests: all t's > 2.266, all P's $< 4.34 \times 10^{-2}$). For joyful voices, we found interaction effects in the left aPTe ($F_{2,24} = 4.392$, $P = 2.36 \times 10^{-2}$) and the mSTS ($F_{2,24} = 7.604$, $P = 2.76 \times 10^{-3}$), which were driven by higher decoding accuracies for formant-filtered joyful compared with formant-filtered aggressive voices (post hoc t-tests: all t's > 2.387, all P's $< 3.46 \times 10^{-2}$). The latter interaction effects point to a specific sensitivity of certain acoustic features of vocalizations, but only when embedded in a specific vocalization.

Finally, we computed contrast estimates (i.e., difference scores of beta estimates between native voices and each of the manipulated conditions) across all ROIs (Figure 3C, lower panel), which confirmed that for most of the ROIs with increased decoding accuracy, the underlying brain signal was usually accompanied by a higher signal for native voices (i.e., including a specific feature) compared with manipulated voices (i.e., voices with an eliminated feature). This indicates that most of the regions were sensitive to the presence of a specific feature in the native voices compared with the elimination of these features in manipulated voices. For several ROIs, however, we did not find any strong difference between conditions as scored by the contrast estimates, indicating that the multivariate analysis approach showed a higher sensitivity to differences between conditions.

Discussion

We tested the neural decoding of relevant acoustic object features especially for the neural processing of human voices, as socially relevant auditory objects. The acoustic pattern of human voices is considerably modulated in affective vocalizations, and these acoustic voice features seem relevant for affective classification. We accordingly found a sensitivity to relevant voice features both in cortical and subcortical auditory regions. In terms of subcortical regions we already found a sensitivity at the level of the MGN, pointing to the possibility of an early sensitivity to relevant voice features in the auditory system. The MGN was sensitive to the pitch dynamics (i.e. f0 variation) in aggressive vocalizations and to high-frequency cues in joyful vocalizations. The MGN is generally sensitive to temporal (Bartlett and Wang, 2007) and spectral cues of complex sounds (Wenstrup, 1999) and thus seems involved in auditory affective processing (Weinberger, 2011). The present data thus suggest that this early affective processing in the MGN might be rooted in the sensitivity of the MGN to temporal and spectral features of affective vocalizations. The MGN showed sensitivity to two distinct auditory features, but this sensitivity differentially depended on the affective valence of the vocalizations. This affective valence is a complex acoustical and perceptual background, which might drive the decoding of certain voice features that are relevant for affective classifications and behavioral adaptation. This valence-dependent sensitivity of the MGN to certain vocal features was demonstrated by the second step of our searchlight analysis, but the post-hoc analysis in the MGN ROI data did not reveal consistent results in this direction. Thus, the results of the MGN valence-dependent feature sensitivity has to be taken with some caution, but future studies might help to determine these MGN results more clearly.

This potential context- and valence-dependent sensitivity of the MGN to important vocal features points to modulatory influences already at the subcortical level of the MGN. This effect might have been driven by a combination of bottom-up sensory processing and by top-down influences of cortical regions, and both mechanisms are represented by the ventral and dorsal division of the MGN, respectively (Sherman and Guillery, 2002). Bottom-up acoustic processing and affective decoding could be supported by the direct anatomical connections of the MGN to the basolateral amygdala (Frühholz *et al.*, 2014b), which projects back to the inferior colliculi as one of the major afferent

connections to the MGN. Top-down influences on the MGN might be represented by the context- and valence-dependent effects that could enhance the decoding of the relevant voice features differentially.

Concerning the voice feature sensitivity at the *cortical* level, we found that high-frequency cues were similarly decoded in the secondary and higher level AC for aggressive and joyful vocalizations, with additional activity in the bilateral primary AC for joyful vocalizations. The sensitivity in the primary and especially the secondary AC could be directly related to the decoding of high-frequency cues of native compared with low-pass filtered voices, since most of this activity overlapped with the high-frequency fields as determined by the tonotopy localizer scan. However, the additional sensitivity in the left pSTG and mSTS seems not directly related to spectral frequency decoding, but might represent high-level feature integration decoding for object representation (Griffiths *et al.*, 2007). The latter regions also consistently overlapped with the voice-sensitive cortex, indicating high-level decoding in these areas to form an auditory percept. High-frequency cues thus seem to be an important ingredient in this feature integration process to form a percept. This percept could be a source for the context- and valence-dependent top-down influences on the MGN, especially for the increased decoding of high-frequency cues in joyful activations, which might be based on anatomical STC-MGN connections (Frühholz *et al.*, 2014b). The slightly more extended activity pattern for high-frequency cues for joyful than for aggressive voices might be due to their increased spectral complexity (Sauter *et al.*, 2010), which seems to specifically influence the connectivity between the HG and the PTe (Griffiths *et al.*, 2007).

This cortical network was not exclusively sensitive to high-frequency cues in vocalizations, but was also involved in the decoding of the pitch dynamics and the HNR. However, unlike for the high-frequency cues, first, the pitch dynamics and the HNR are decoded in a more extended network of AC, STC and IFC, and, second, the decoding of the pitch dynamics and the HNR of vocalizations was differentially driven by the valence of vocalizations. These regions generally were more sensitive to the pitch dynamics of aggressive vocalizations and to the HNR of joyful vocalizations. This indicates that the pitch dynamic is an important feature for the neural decoding of aggressive vocalizations, whereas the HNR seems relevant for joyful vocalizations. We have to note that these differential effects resulted from a similar amount of acoustic change for aggressive and joyful vocalizations across all manipulation conditions, and this differential sensitivity occurred at different levels of auditory processing in cortical regions. This indicates an affective valence-dependent decoding of relevant acoustic voice features rather than a simple decoding of acoustic differences across many cortical auditory regions.

The cortical regions sensitive to the HNR of joyful vocalizations largely resemble activity in the HG, PTe/PPo, mSTG and pSTG previously reported for the decoding of the HNR in primate vocalizations (Lewis *et al.*, 2009) and in human speech (Leaver and Rauschecker, 2010). Lewis *et al.* (2009) reported that the HG and the PTe/PPo were sensitive to the HNR in stimuli with a *simple* harmonic structure, whereas the mSTG/pSTG as reported here seem sensitive to the HNR in auditory stimuli with a more *complex* harmonic structure, such as human voices. For joyful vocalizations, we found that the left PTe as well as the left mSTC showed specifically increased decoding accuracy. This might indicate an auditory cortical sensitivity to both a simple and a complex harmonic structure of joyful vocalizations. Joyful vocalizations show a simple harmonic structure in local temporal segments, but due to their high temporal spectral variability (Sauter *et al.*, 2010) their harmonic structure is temporally more complex. Furthermore, joyful vocalizations commonly show a higher HNR than aggressive vocalizations (Patel *et al.*, 2011), especially in female speakers (Hammerschmidt and Jurgens, 2007). The formant-filtering procedure in the present study might have modulated joyful more than aggressive vocalizations beyond the level of naturally occurring HNR levels, thus probably supporting the neural discrimination of native and formant-filtered joyful voices more than aggressive voices.

The cortical areas that we found for the decoding of the pitch dynamics in aggressive vocalizations resemble the activity in the HG, PTe/PPo and mSTG, which was reported previously for the processing of temporal pitch and melody variations (Patterson *et al.*, 2002; Warren and Griffiths, 2003). However, the present data extend these previous findings by pointing out that the mSTC and pSTC also show sensitivity to the pitch dynamics in affective vocalizations (Frühholz *et al.*, 2012). The latter was found for all vocalizations, indicating a general sensitivity of the mSTC and pSTC to the pitch dynamics in vocalizations (Leitman *et al.*, 2010b; Norman-Haignere *et al.*, 2013). Additionally, we found that the left HG and the right mSTG were specifically sensitive to pitch dynamics in aggressive vocalizations. The lHG is sensitive to (temporal) pitch (Patterson *et al.*, 2002), to spectro-temporal patterns (Schonwiesner and Zatorre, 2009), and also to the temporal structure of auditory objects (Giraud *et al.*, 2000), with a stronger response in the left compared with the right HG (Gourevitch *et al.*, 2008). The latter was mainly determined with intensity variations in sounds, but given the overlap of pitch and temporal structure sensitivity in the bilateral HG, this area might also be sensitive to the temporal structure of the pitch as found in the present study. The sensitivity of the right mSTG corroborates the recent observation of special sensitivity of this area to the pitch-based melody of sounds (Patterson *et al.*, 2002). Although increased pitch variation is an acoustic cue for both joyful and aggressive vocalizations (Juslin and Laukka, 2003; Leitman *et al.*, 2010b), expressions of aggressive anger, as used in the present study, especially show a considerable amount of pitch variation, which listeners use as an important acoustic cue for their decoding (Banse and Scherer, 1996).

In the previous paragraphs we discussed the voice feature sensitivity of subcortical and cortical auditory regions. Beyond the auditory system, we also found some feature sensitivity in the left IFC, especially for joyful vocalizations. The IFC shows sensitivity to the general structure of conspecific vocalizations (Romanski *et al.*, 2005), and the aIFG and the fOP are involved in categorizing and discriminating vocalizations according to their general spectro-temporal morphological characteristics (Frühholz and Grandjean, 2013c). Furthermore, the pIFG decodes suprasegmental prosodic variations (Frühholz and Grandjean, 2013c). The sensitivity of the pIFG for the pitch dynamics closely resembles the latter notion, while the fOP and the aIFG might be involved in categorizing joyful and partly also aggressive vocalizations according to their spectro-temporal morphological characteristics, which seems to be considerably determined by the HNR and the pitch dynamics, respectively. Thus, instead of performing categorizations on acoustic information, which is represented in the AC and the STC, the IFC might to some degree also represent higher-order acoustic information (Cohen *et al.*, 2007).

We finally have to mention two potential limitations of our study. First, native affective voices naturally differ in their

quantity of certain acoustic features. Based on these basic differences, the acoustic change that we introduced during the feature elimination approach might have affected the affective voices differentially depending on the level of certain features in native affective voices. However, our study was mainly interested in the acoustic *change* that is introduced by the feature elimination approach. This acoustic change was similar (i.e. not different) for aggressive and joyful voices. Future studies might select native voices that are similar in their basic level of certain acoustic features, a task that however becomes difficult when controlling for several acoustic features at the same time. Second, we attributed the context-dependent effects to the affective valence of the voices. Although these effects can be largely attributed to the affective valence of voices (i.e. the acoustic change did not change the affective valence of the voices), a clear demonstration of this affective effect might need the inclusion of an additional 'neutral' experimental condition. As outlined in the method section, neutral voices already served as a reference for manipulating aggressive and joyful voices, and thus could not be included as an additional condition (i.e. the manipulation of neutral voices would need another acoustic reference stimulus). Future studies thus might include an additional auditory stimulus with a similar complexity as for human voices, but which has a neutral affective valence, to clearly isolate the effects that are introduced by the affective valence of voices.

Taken together, the data might indicate that the neural decoding of acoustic features embedded in naturalistic and complex stimuli is not a rigorous process based on bottom-up sensory processing. This neural decoding seems rather a context-dependent process, such that not every significant acoustic change leads to neural effects. For example, the significant change of the HNR of aggressive vocalizations was not accompanied by any neural effects. Second, our data indicate that several regions are sensitive to multiple acoustic features, including a hypersensitivity to certain features. Third, the data indicate that the brain location of sensitivity to a specific feature can shift according to the affective valence of vocalizations. All of these effects depended on, and varied with, the socio-affective valence of vocalizations, which provided the acoustic and perceptual context for the features, and which might prioritize the decoding of the most relevant and discriminative voice features. Besides the cortical and subcortical description of the brain regions involved in a context-dependent decoding of discriminative object features, future studies might determine the neural connectivity between these regions, which are not yet available with an information-based analysis of brain data.

Funding

SF and DG were supported by the Swiss National Science Foundation (SNSF 105314_124572/1 and 105314_146559/1) and by the NCCR Affective Sciences at the University of Geneva (51NF40-104897).

Conflict of interest. None declared.

References

Banse, R., Scherer, K.R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, **70**(3), 614–36.

Bartlett, E.L., Wang, X. (2007). Neural representations of temporally modulated signals in the auditory thalamus of awake primates. *Journal of Neurophysiology*, **97**(2), 1005–17.

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

Belyk, M., Brown, S. (2014). Perception of affective and linguistic prosody: an ALE meta-analysis of neuroimaging studies. *Social Cognitive and Affective Neuroscience*, **9**(9), 1395–403.

Boersma, P. (2001). 'Praat, a system for doing phonetics by computer. *Glot International*, **5**, 341–5.

Cohen, Y.E., Theunissen, F., Russ, B.E., Gill, P. (2007). Acoustic features of rhesus vocalizations and their representation in the ventrolateral prefrontal cortex. *Journal of Neurophysiology*, **97**(2), 1470–84.

Da Costa, S., van der Zwaag, W., Marques, J.P., Frackowiak, R.S., Clarke, S., Saenz, M. (2011). Human primary auditory cortex follows the shape of Heschl's gyrus. *Journal of Neuroscience*, **31**(40), 14067–75.

De Martino, F., Moerel, M., van de Moortele, P.F., *et al.* (2013). Spatial organization of frequency preference and selectivity in the human inferior colliculus. *Nature Communications*, **4**, 1386.

Ethofer, T., Anders, S., Wiethoff, S., *et al.* (2006). Effects of prosodic emotional intensity on activation of associative auditory cortex. *Neuroreport*, **17**(3), 249–53.

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

Frühholz, S., Grandjean, D. (2012). Towards a fronto-temporal neural network for the decoding of angry vocal expressions. *Neuroimage*, **62**(3), 1658–66.

Frühholz, S., Grandjean, D. (2013a). Amygdala subregions differentially respond and rapidly adapt to threatening voices. *Cortex*, **49**(5), 1394–403.

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

Frühholz, S., Grandjean, D. (2013c). Processing of emotional vocalizations in bilateral inferior frontal cortex. *Neuroscience and Biobehavioral Reviews*, **37**(10 Pt 2), 2847–55.

Frühholz, S., Klaas, H.S., Patel, S., Grandjean, D. (2014a). Talking in fury: the cortico-subcortical network underlying angry vocalizations. *Cerebral Cortex*, **25**(9), 2752–62.

Frühholz, S., Trost, W., Grandjean, D. (2014b). The role of the medial temporal limbic system in processing emotions in voice and music. *Progress in Neurobiology*, **123**, 1–17.

Giraud, A.L., Lorenzi, C., Ashburner, J., *et al.* (2000). Representation of the temporal envelope of sounds in the human brain. *Journal of Neurophysiology*, **84**(3), 1588–98.

Gourevitch, B., Le Bouquin Jeannes, R., Faucon, G., Liegeois-Chauvel, C. (2008). Temporal envelope processing in the human auditory cortex: response and interconnections of auditory cortical areas. *Hearing Research*, **237**(1-2), 1–18.

Griffiths, T.D., Kumar, S., Warren, J.D., Stewart, L., Stephan, K.E., Friston, K.J. (2007). Approaches to the cortical analysis of auditory objects. *Hearing Research*, **229**(1–2), 46–53.

Hammerschmidt, K., Jurgens, U. (2007). Acoustical correlates of affective prosody. *Journal of Voice*, **21**(5), 531–40.

Hebart, M.N., Donner, T.H., Haynes, J.D. (2012). Human visual and parietal cortex encode visual choices independent of motor plans. *Neuroimage*, **63**(3), 1393–403.

Irwin, A., Hall, D.A., Peters, A., Plack, C.J. (2011). Listening to urban soundscapes: Physiological validity of perceptual dimensions. *Psychophysiology*, **48**(2), 258–68.

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

Kaas, J.H., Hackett, T.A. (2000). Subdivisions of auditory cortex and processing streams in primates. *Proceedings of the National Academy of Sciences of the United States of America*, **97**(22), 11793–9.

Kotz, S.A., Kalberlah, C., Bahlmann, J., Friederici, A.D., Haynes, J.D. (2013). Predicting vocal emotion expressions from the human brain. *Human Brain Mapping*, **34**(8), 1971–81.

Kriegeskorte, N., Bandettini, P. (2007). Combining the tools: activation- and information-based fMRI analysis. *Neuroimage*, **38**(4), 666–8.

Kriegeskorte, N., Goebel, R., Bandettini, P. (2006). 'Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, **103**(10), 3863–8.

Kumar, S., von Kriegstein, K., Friston, K., Griffiths, T.D. (2012). Features versus feelings: dissociable representations of the acoustic features and valence of aversive sounds. *Journal of Neuroscience*, **32**(41), 14184–92.

Leaver, A.M., Rauschecker, J.P. (2010). Cortical representation of natural complex sounds: effects of acoustic features and auditory object category. *Journal of Neuroscience*, **30**(22), 7604–12.

Leitman, D.I., Laukka, P., Juslin, P.N., Saccente, E., Butler, P., Javitt, D.C. (2010a). Getting the cue: sensory contributions to auditory emotion recognition impairments in schizophrenia. *Schizophrenia Bulletin*, **36**(3), 545–56.

Leitman, D.I., Wolf, D.H., Ragland, J.D., *et al.* (2010b). 'It's Not What You Say, But How You Say it': a Reciprocal Temporo-frontal Network for Affective Prosody. *Frontiers in Human Neuroscience*, **4**, 1–13.

Lewis, J.W., Talkington, W.J., Walker, N.A., *et al.* (2009). Human cortical organization for processing vocalizations indicates representation of harmonic structure as a signal attribute. *Journal of Neuroscience*, **29**(7), 2283–96.

Marques, J.P., Kober, T., Krueger, G., van der Zwaag, W., Van de Moortele, P.F., Gruetter, R. (2010). 'MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. *Neuroimage*, **49**(2), 1271–81.

McGillivray, P., Vonderschen, K., Fortune, E.S., Chacron, M.J. (2012). Parallel coding of first- and second-order stimulus attributes by midbrain electrosensory neurons. *Journal of Neuroscience*, **32**(16), 5510–24.

Mumford, J.A., Turner, B.O., Ashby, F.G., Poldrack, R.A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *Neuroimage*, **59**(3), 2636–43.

Norman-Haignere, S., Kanwisher, N., McDermott, J.H. (2013). Cortical pitch regions in humans respond primarily to resolved harmonics and are located in specific tonotopic regions of anterior auditory cortex. *Journal of Neuroscience*, **33**(50), 19451–69.

Patel, S., Scherer, K.R., Bjorkner, E., Sundberg, J. (2011). Mapping emotions into acoustic space: the role of voice production. *Biological Psychology*, **87**(1), 93–8.

Patterson, R.D., Uppenkamp, S., Johnsrude, I.S., Griffiths, T.D. (2002). The processing of temporal pitch and melody information in auditory cortex. *Neuron*, **36**(4), 767–76.

Rauschecker, J.P. (1998). Cortical processing of complex sounds. *Current Opinion in Neurobiology*, **8**(4), 516–21.

Romanski, L.M., Averbeck, B.B., Diltz, M. (2005). 'Neural representation of vocalizations in the primate ventrolateral prefrontal cortex. *Journal of Neurophysiology*, **93**(2), 734–47.

Saenz, M., Langers, D.R. (2014). Tonotopic mapping of human auditory cortex. *Hearing Research*, **307**, 42–52.

Sauter, D.A., Eisner, F., Calder, A.J., Scott, S.K. (2010). Perceptual cues in nonverbal vocal expressions of emotion. *Quarterly Journal of Experimental Psychology*, **63**(11), 2251–72.

Schonwiesner, M., Zatorre, R.J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences of the United States of America*, **106**(34), 14611–6.

Sherman, S.M., Guillery, R.W. (2002). The role of the thalamus in the flow of information to the cortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, **357**(1428), 1695–708.

Speck, O., Stadler, J., Zaitsev, M. (2008). High resolution single-shot EPI at 7T. *Magma*, **21**(1–2), 73–86.

Sundberg, J., Nordenberg, M. (2006). Effects of vocal loudness variation on spectrum balance as reflected by the alpha measure of long-term-average spectra of speech. *Journal of the Acoustical Society of America*, **120**(1), 453–7.

Tourdias, T., Saranathan, M., Levesque, I.R., Su, J., Rutt, B.K. (2014). 'Visualization of intra-thalamic nuclei with optimized white-matter-nulled MPRAGE at 7T. *Neuroimage*, **84**, 534–45.

Warren, J.D., Griffiths, T.D. (2003). 'Distinct mechanisms for processing spatial sequences and pitch sequences in the human auditory brain. *Journal of Neuroscience*, **23**(13), 5799–804.

Weinberger, N.M. (2011). The medial geniculate, not the amygdala, as the root of auditory fear conditioning. *Hearing Research*, **274**(1–2), 61–74.

Wenstrup, J.J. (1999). Frequency organization and responses to complex sounds in the medial geniculate body of the mustached bat. *Journal of Neurophysiology*, **82**(5), 2528–44.

Wiethoff, S., Wildgruber, D., Kreifelts, B., *et al.* (2008). Cerebral processing of emotional prosody – influence of acoustic parameters and arousal. *Neuroimage*, **39**(2), 885–93.

Witteman, J., Van Heuven, V.J., Schiller, N.O. (2012). Hearing feelings: a quantitative meta-analysis on the neuroimaging literature of emotional prosody perception. *Neuropsychologia*, **50**(12), 2752–63.